From Classical to Quantum Shannon Theory

Mark M. Wilde
Hearne Institute for Theoretical Physics
Department of Physics and Astronomy
Center for Computation and Technology
Louisiana State University
Baton Rouge, Louisiana 70803, USA

March 24, 2016

Copyright Notice:

Contents

I Introduction 19

1 Concepts in Quantum Shannon Theory 21
- 1.1 Overview of the Quantum Theory . . . 24
- 1.2 The Emergence of Quantum Shannon Theory . . . 28

2 Classical Shannon Theory 43
- 2.1 Data Compression . . . 43
- 2.2 Channel Capacity . . . 52
- 2.3 Summary . . . 66

II The Quantum Theory 69

3 The Noiseless Quantum Theory 71
- 3.1 Overview . . . 72
- 3.2 Quantum Bits . . . 72
- 3.3 Reversible Evolution . . . 78
- 3.4 Measurement . . . 85
- 3.5 Composite Quantum Systems . . . 90
- 3.6 Entanglement . . . 98
- 3.7 Summary and Extensions to Qudit States . . . 109
- 3.8 Schmidt Decomposition . . . 115
- 3.9 History and Further Reading . . . 117

4 The Noisy Quantum Theory 119
- 4.1 Noisy Quantum States . . . 120
- 4.2 Measurement in the Noisy Quantum Theory . . . 132
- 4.3 Composite Noisy Quantum Systems . . . 136
- 4.4 Quantum Evolutions . . . 148
- 4.5 Interpretations of Quantum Channels . . . 159
- 4.6 Quantum Channels are All Encompassing . . . 161

4.7 Examples of Quantum Channels . . . 172
4.8 Summary . . . 180
4.9 History and Further Reading . . . 181

5 The Purified Quantum Theory 183
5.1 Purification . . . 184
5.2 Isometric Evolution . . . 187
5.3 Coherent Quantum Instrument . . . 198
5.4 Coherent Measurement . . . 199
5.5 History and Further Reading . . . 200

III Unit Quantum Protocols 203

6 Three Unit Quantum Protocols 205
6.1 Non-Local Unit Resources . . . 206
6.2 Protocols . . . 208
6.3 Optimality of the Three Unit Protocols . . . 217
6.4 Extensions for Quantum Shannon Theory . . . 219
6.5 Three Unit Qudit Protocols . . . 220
6.6 History and Further Reading . . . 226

7 Coherent Protocols 227
7.1 Definition of Coherent Communication . . . 229
7.2 Implementations of a Coherent Bit Channel . . . 230
7.3 Coherent Dense Coding . . . 232
7.4 Coherent Teleportation . . . 233
7.5 Coherent Communication Identity . . . 235
7.6 History and Further Reading . . . 236

8 Unit Resource Capacity Region 237
8.1 The Unit Resource Achievable Region . . . 237
8.2 The Direct Coding Theorem . . . 242
8.3 The Converse Theorem . . . 242
8.4 History and Further Reading . . . 246

IV Tools of Quantum Shannon Theory 247

9 Distance Measures 249
9.1 Trace Distance . . . 250
9.2 Fidelity . . . 263
9.3 Relations between Trace Distance and Fidelity . . . 273
9.4 Gentle Measurement . . . 277

9.5 Fidelity of a Quantum Channel 280
9.6 The Hilbert–Schmidt Distance Measure 284
9.7 History and Further Reading 285

10 Classical Information and Entropy 287
10.1 Entropy of a Random Variable 288
10.2 Conditional Entropy 293
10.3 Joint Entropy 294
10.4 Mutual Information 295
10.5 Relative Entropy 296
10.6 Conditional Mutual Information 297
10.7 Entropy Inequalities 298
10.8 Near Saturation of Entropy Inequalities 310
10.9 Classical Information from Quantum Systems 316
10.10History and Further Reading 318

11 Quantum Information and Entropy 319
11.1 Quantum Entropy 320
11.2 Joint Quantum Entropy 324
11.3 Potential yet Unsatisfactory Definitions of Conditional Quantum Entropy . . 328
11.4 Conditional Quantum Entropy 329
11.5 Coherent Information 332
11.6 Quantum Mutual Information 334
11.7 Conditional Quantum Mutual Information 337
11.8 Quantum Relative Entropy 339
11.9 Quantum Entropy Inequalities 345
11.10Continuity of Quantum Entropy 357
11.11History and Further Reading 363

12 Quantum Entropy Inequalities and Recoverability 365
12.1 Recoverability Theorem 365
12.2 Schatten Norms and Complex Interpolation 366
12.3 Petz Recovery Map 371
12.4 Rényi Information Measure 373
12.5 Proof of the Recoverability Theorem 378
12.6 Refinements of Quantum Entropy Inequalities 380
12.7 History and Further Reading 384

13 The Information of Quantum Channels 387
13.1 Mutual Information of a Classical Channel 388
13.2 Private Information of a Wiretap Channel 396
13.3 Holevo Information of a Quantum Channel 400
13.4 Mutual Information of a Quantum Channel 406

13.5 Coherent Information of a Quantum Channel . . . 410
13.6 Private Information of a Quantum Channel . . . 415
13.7 Summary . . . 421
13.8 History and Further Reading . . . 422

14 Classical Typicality 423
14.1 An Example of Typicality . . . 424
14.2 Weak Typicality . . . 425
14.3 Properties of the Typical Set . . . 427
14.4 Application: Data Compression . . . 429
14.5 Weak Joint Typicality . . . 431
14.6 Weak Conditional Typicality . . . 434
14.7 Strong Typicality . . . 437
14.8 Strong Joint Typicality . . . 446
14.9 Strong Conditional Typicality . . . 448
14.10 Application: Channel Capacity Theorem . . . 454
14.11 Concluding Remarks . . . 458
14.12 History and Further Reading . . . 459

15 Quantum Typicality 461
15.1 The Typical Subspace . . . 462
15.2 Conditional Quantum Typicality . . . 471
15.3 The Method of Types for Quantum Systems . . . 481
15.4 Concluding Remarks . . . 483
15.5 History and Further Reading . . . 483

16 The Packing Lemma 485
16.1 Introductory Example . . . 486
16.2 The Setting of the Packing Lemma . . . 486
16.3 Statement of the Packing Lemma . . . 488
16.4 Proof of the Packing Lemma . . . 490
16.5 Derandomization and Expurgation . . . 496
16.6 Sequential Decoding . . . 498
16.7 History and Further Reading . . . 502

17 The Covering Lemma 503
17.1 Introductory Example . . . 504
17.2 Setting and Statement of the Covering Lemma . . . 506
17.3 Operator Chernoff Bound . . . 507
17.4 Proof of the Covering Lemma . . . 512
17.5 History and Further Reading . . . 518

V Noiseless Quantum Shannon Theory 519

18 Schumacher Compression 521
18.1 The Information-Processing Task 522
18.2 The Quantum Data-Compression Theorem 524
18.3 Quantum Compression Example 528
18.4 Variations on the Schumacher Theme 529
18.5 Concluding Remarks 531
18.6 History and Further Reading 531

19 Entanglement Manipulation 533
19.1 Sketch of Entanglement Manipulation 534
19.2 LOCC and Relative Entropy of Entanglement 538
19.3 Entanglement Manipulation Task 540
19.4 The Entanglement Manipulation Theorem 541
19.5 Concluding Remarks 550
19.6 History and Further Reading 551

VI Noisy Quantum Shannon Theory 553

20 Classical Communication 557
20.1 Naive Approach: Product Measurements 559
20.2 The Information-Processing Task 561
20.3 The Classical Capacity Theorem 564
20.4 Examples of Channels 571
20.5 Superadditivity of the Holevo Information 580
20.6 Concluding Remarks 583
20.7 History and Further Reading 584

21 Entanglement-Assisted Classical Communication 587
21.1 The Information-Processing Task 589
21.2 A Preliminary Example 590
21.3 Entanglement-Assisted Capacity Theorem 593
21.4 The Direct Coding Theorem 594
21.5 The Converse Theorem 602
21.6 Examples of Channels 608
21.7 Concluding Remarks 612
21.8 History and Further Reading 613

22 Coherent Communication with Noisy Resources 615
22.1 Entanglement-Assisted Quantum Communication 616
22.2 Quantum Communication 621
22.3 Noisy Super-Dense Coding 622

22.4 State Transfer 625
22.5 Trade-off Coding 628
22.6 Concluding Remarks 636
22.7 History and Further Reading 637

23 Private Classical Communication 639
23.1 The Information-Processing Task 640
23.2 The Private Classical Capacity Theorem 642
23.3 The Direct Coding Theorem 643
23.4 The Converse Theorem 651
23.5 Discussion of Private Classical Capacity 652
23.6 History and Further Reading 655

24 Quantum Communication 657
24.1 The Information-Processing Task 658
24.2 No-Cloning and Quantum Communication 660
24.3 The Quantum Capacity Theorem 661
24.4 The Direct Coding Theorem 661
24.5 Converse Theorem 668
24.6 An Interlude with Quantum Stabilizer Codes 671
24.7 Example Channels 677
24.8 Discussion of Quantum Capacity 680
24.9 Entanglement Distillation 686
24.10History and Further Reading 689

25 Trading Resources for Communication 693
25.1 The Information-Processing Task 694
25.2 The Quantum Dynamic Capacity Theorem 696
25.3 The Direct Coding Theorem 701
25.4 The Converse Theorem 703
25.5 Examples of Channels 713
25.6 History and Further Reading 729

26 Summary and Outlook 731
26.1 Unit Protocols 732
26.2 Noiseless Quantum Shannon Theory 733
26.3 Noisy Quantum Shannon Theory 733
26.4 Protocols Not Covered In This Book 736
26.5 Network Quantum Shannon Theory 737
26.6 Future Directions 738

A Supplementary Results 739

B Unique Linear Extension of a Quantum Physical Evolution 743

How To Use This Book

For Students

Prerequisites for understanding the content in this book are a solid background in probability theory and linear algebra. If you are new to information theory, then there should be enough background in this book to get you up to speed (Chapters 2, 10, 13, and 14). However, classics on information theory such as Cover and Thomas (2006) and MacKay (2003) could be helpful as a reference. If you are new to quantum mechanics, then there should be enough material in this book (Part II) to give you the background necessary for understanding quantum Shannon theory. The book of Nielsen and Chuang (2000) (sometimes affectionately known as "Mike and Ike") has become the standard starting point for students in quantum information science and might be helpful as well. Some of the content in that book is available in the dissertation of Nielsen (1998). If you are familiar with Shannon's information theory (at the level of Cover and Thomas (2006), for example), then the present book should be a helpful entry point into the field of quantum Shannon theory. We build on intuition developed classically to help in establishing schemes for communication over quantum channels. If you are familiar with quantum mechanics, it might still be worthwhile to review Part II because some content there might not be part of a standard course on quantum mechanics.

The aim of this book is to develop "from the ground up" many of the major, exciting, pre- and post-millenium developments in the general area of study known as quantum Shannon theory. As such, we spend a significant amount of time on quantum mechanics for quantum information theory (Part II), we give a careful study of the important unit protocols of teleportation, super-dense coding, and entanglement distribution (Part III), and we develop many of the tools necessary for understanding information transmission or compression (Part IV). Parts V and VI are the culmination of this book, where all of the tools developed come into play for understanding many of the important results in quantum Shannon theory.

For Instructors

This book could be useful for self-learning or as a reference, but one of the main goals is for it to be employed as an instructional aid for the classroom. To aid instructors in designing a course to suit their own needs, a draft, pre-publication copy of this book is available under a Creative Commons Attribution-NonCommercial-ShareAlike license. This means that you can modify and redistribute this draft, pre-publication copy as you wish, as long as you attribute the author, you do not use it for commercial purposes, and you share a modification or derivative work under the same license (see

`http://creativecommons.org/licenses/by-nc-sa/3.0/`

for a readable summary of the terms of the license). These requirements can be waived if you obtain permission from the present author. By releasing the draft, pre-publication copy of the book under this license, I expect and encourage instructors to modify it for their own needs. This will allow for the addition of new exercises, new developments in the theory, and the latest open problems. It might also be a helpful starting point for a book on a related topic, such as network quantum Shannon theory.

I used an earlier version of this book in a one-semester course on quantum Shannon theory at McGill University during Winter semester 2011 (in many parts of the USA, this semester is typically called "Spring semester"). We almost went through the entire book, but it might also be possible to spread the content over two semesters instead. Here is the order in which we proceeded:

1. Introduction in Part I.
2. Quantum mechanics in Part II.
3. Unit protocols in Part III.
4. Chapter 9 on distance measures, Chapter 10 on classical information and entropy, and Chapter 11 on quantum information and entropy.
5. The first part of Chapter 14 on classical typicality and Shannon compression.
6. The first part of Chapter 15 on quantum typicality.
7. Chapter 18 on Schumacher compression.
8. Back to Chapters 14 and 15 for the method of types.
9. Chapter 19 on entanglement concentration.
10. Chapter 20 on classical communication.
11. Chapter 21 on entanglement-assisted classical communication.
12. The final explosion of results in Chapter 22 (one of which is a route to proving the achievability part of the quantum capacity theorem).

The above order is just a particular order that suited the needs for the class at McGill, but other orders are of course possible. One could sacrifice the last part of Part III on the unit resource capacity region if there is no desire to cover the quantum dynamic capacity theorem. One could also focus on going from classical communication to private classical communication to quantum communication in order to develop some more intuition behind the quantum capacity theorem. I later did this when teaching the course at LSU in Fall 2013. But just recently in Fall 2015, I went back to the ordering above while including lectures devoted to the CHSH game and the new results in Chapter 12.

Other Sources

There are many other sources to obtain a background in quantum Shannon theory. The standard reference has become the book of Nielsen and Chuang (2000), but it does not feature any of the post-millenium results in quantum Shannon theory. Other excellent books that cover some aspects of quantum Shannon theory are (Hayashi, 2006; Holevo, 2002a, 2012; Watrous, 2015). Patrick Hayden has had a significant hand as a collaborative guide for many PhD and Masters' theses in quantum Shannon theory, during his time as a postdoctoral fellow at the California Institute of Technology and as a professor at McGill University. These include the theses of Yard (2005), Abeyesinghe (2006), Savov (2008, 2012), Dupuis (2010), and Dutil (2011). All of these theses are excellent references. Hayden also had a strong influence over the present author during the development of the first edition of this book.

Preface to the Second Edition

It has now been some years since I completed the first draft of the first edition of this book. In this time, I have learned much from many collaborators and I am grateful to them. During the past few years, Mario Berta, Nilanjana Datta, Saikat Guha, Marco Tomamichel, and Andreas Winter have strongly shaped my thinking about quantum information theory, and Mario, Nilanjana, and Marco in particular have influenced my technical writing style, which is reflected in the new edition of the book. Also, the chance to work with them and others has led me to new research directions in quantum information theory that I never would have imagined on my own.

I am also thankful to Todd Brun, Paul Cuff, Ludovico Lami, Ciara Morgan, and Giannicola Scarpa for using the book as the main text in their graduate courses on quantum information theory and for feedback. One can try as much as possible to avoid typos in a book, but inevitably, they seem to show up in unexpected places. I am grateful to many people for pointing out typos or errors and for suggesting how to fix them, including Todd Brun, Giulio Chiribella, Paul Cuff, Dawei (David) Ding, Will Matthews, Milan Mosonyi, David Reeb, and Marco Tomamichel. I also thank Corsin Pfister for helpful discussions about unique linear extensions of quantum physical evolutions. I am grateful to David Tranah and the editorial staff at Cambridge University Press for their help with publishing the second edition.

So what's new in the second edition? Suffice it to say that every page of the book has been rewritten and there are over 100 pages of new material! I formulated many thoughts about revising during Fall 2013 while teaching a graduate course on quantum information at LSU, and I then formulated many more thoughts and made the actual changes during Fall 2015 (when teaching it again). In that regard, I am thankful to both the Department of Physics and Astronomy and the Center for Computation and Technology at LSU for providing a great environment and support. I also thank the graduate students at LSU who gave feedback during and after lectures. There are many little changes throughout that will probably go unnoticed. For example, I have come to prefer writing a quantum state shared between Alice and Bob as ρ_{AB} rather than ρ^{AB} (i.e., with system labels as subscripts rather than superscripts). Admittedly, several collaborators influenced me here, but there are a few good reasons for this convention: the phrase "state of a quantum system" suggests that the state ρ should be "resting on" the systems AB, the often used partial trace $\mathrm{Tr}_A\{\rho_{AB}\}$ looks better than $\mathrm{Tr}_A\{\rho^{AB}\}$, and the notation ρ_{AB} is more consistent with the standard notation p_X for a probability distribution corresponding to a random variable X.

OK, that's perhaps minor. Major changes include the addition of many new exercises, a detailed discussion of Bell's theorem, the CHSH game, and Tsirelson's theorem, the axiomatic approach to quantum channels, a proof of the Choi–Kraus theorem, a definition of unital and adjoint maps, a discussion of states, channels, and measurements all as quantum channels, the equivalence of purifications, the adjoint map in terms of isometric extension, the definition of the diamond norm and its interpretation, how a measurement achieves the fidelity, how the Hilbert–Schmidt distance is not monotone with respect to channels, more detailed definitions of classical and quantum relative entropies, new continuity bounds for classical and quantum entropies, refinements of classical entropy inequalities, streamlined proofs of data processing inequalities using relative entropy, the equivalence of quantum entropy inequalities like strong subadditivity and monotonicity of relative entropy, Chapter 12 on recoverability, modified proofs of additivity of channel information quantities, sequential decoding for classical communication, simpler proofs of the Schumacher compression theorem, a complete rewrite of Chapter 19, alternate proofs for the achievability part of the HSW theorem, a proof for the classical capacity of the erasure channel, simpler converse proofs for the entanglement-assisted capacity theorem, a revised proof of the trade-off coding resource inequality, a revised proof of the hashing bound, a simplified converse proof of the quantum dynamic capacity theorem, a completely revised discussion of the importance of the quantum dynamic capacity formula, and the addition of many new references that have been influential in recent years. Minor changes include improved presentations of many theorems and definitions throughout.

I am most grateful to my family for all of their support and encouragement throughout my life, including my mother, father, sister, and brother and all of my surrounding family members. I am still indebted to my wife Christabelle and her family for warmth and love. Christabelle has been an unending source of support and love for me. I dedicate this second edition to my nephews David and Matthew.

Mark M. Wilde
Baton Rouge, Louisiana, USA
December 2015

Preface to the First Edition

I began working on this book in the summer of 2008 in Los Angeles, with much time to spare in the final months of dissertation writing. I had a strong determination to review quantum Shannon theory, a beautiful area of quantum information science that Igor Devetak had taught me three years earlier at USC in fall 2005. I was carefully studying a manuscript entitled "Principles of Quantum Information Theory," a text that Igor had initiated in collaboration with Patrick Hayden and Andreas Winter. I read this manuscript many times, and many parts of it I understood well, though other parts I did not.

After a few weeks of reading and rereading, I decided "if I can write it out myself from scratch, perhaps I would then understand it!", and thus began the writing of the chapters on the packing lemma, the covering lemma, and quantum typicality. I knew that Igor's (now former) students Min-Hsiu Hsieh and Zhicheng Luo knew the topic well because they had already written several quality research papers with him, so I requested if they could meet with me weekly for an hour to review the fundamentals. They kindly agreed and helped me quite a bit in understanding the packing and covering techniques.

Not much later, after graduating, I began collaborating with Min-Hsiu on a research project that Igor had suggested to the both of us: "find the triple trade-off capacity formulas of a quantum channel." This was perhaps the best starting point for me to learn quantum Shannon theory because proving this theorem required an understanding of most everything that had already been accomplished in the area. After a month of effort, I continued to work with Min-Hsiu on this project while joining Andreas Winter's Singapore group for a two-month visit. As I learned more, I added more to the notes, and they continued to grow.

After landing a job in the DC area for January 2009, I realized that I had almost enough material for teaching a course, and so I contacted local universities in the area to see if they would be interested. Can Korman, formerly chair of the Electrical Engineering Department at George Washington University, was excited about the possibility. His enthusiasm was enough to keep me going on the notes, and so I continued to refine and add to them in my spare time in preparing for teaching. Unfortunately (or perhaps fortunately?), the course ended up being canceled. This was disheartening to me, but in the mean time, I had contacted Patrick Hayden to see if he would be interested in having me join his group at McGill University for postdoctoral studies. Patrick Hayden and David Avis then offered me a postdoctoral fellowship, and I moved to Montréal in October 2009.

After joining, I learned a lot by collaborating and discussing with Patrick and his group members. Patrick offered me the opportunity to teach his graduate class on quantum Shannon theory while he was away on sabbatical, and this encouraged me further to persist with the notes.

I am grateful to everyone mentioned above for encouraging and supporting me during this project, and I am also grateful to everyone who provided feedback during the course of writing up. In this regard, I am especially grateful to Dave Touchette for detailed feedback on all of the chapters in the book. Dave's careful reading and spotting of errors has immensely improved the quality of the book. I am grateful to my father, Gregory E. Wilde, Sr., for feedback on earlier chapters and for advice and love throughout. I thank Ivan Savov for encouraging me, for feedback, and for believing that this is an important scholarly work. I also thank Constance Caramanolis, Raza-Ali Kazmi, John M. Schanck, Bilal Shaw, and Anna Vershynina for valuable feedback. I am grateful to Min-Hsiu Hsieh for the many research topics we have worked on together that have enhanced my knowledge of quantum Shannon theory. I thank Michael Nielsen and Victor Shoup for advice on Creative Commons licensing and Kurt Jacobs for advice on book publishing. I am grateful to Sarah Payne and David Tranah of Cambridge University Press for their extensive feedback on the manuscript and their outstanding support throughout the publication process. I acknowledge funding from the MDEIE (Quebec) PSR-SIIRI international collaboration grant.

I am indebted to my mentors who took me on as a student during my career. Todd Brun was a wonderful PhD supervisor—helpful, friendly, and encouraging of creativity and original pursuit. Igor Devetak taught me quantum Shannon theory in fall 2005 and helped me once per week during his office hours. He also invited me to join Todd's and his group, and more recently, Igor provided much encouragement and "big-picture" feedback during the writing of this book. Bart Kosko shaped me as a scholar during my early years at USC and provided helpful advice regarding the book project. Patrick Hayden has been an immense bedrock of support at McGill. His knowledge of quantum information and many other areas is unsurpassed, and he has been kind, inviting, and helpful during my time at McGill. I am also grateful to Patrick for giving me the opportunity to teach at McGill and for advice throughout the development of this book.

I thank my mother, father, sister, and brother and all of my surrounding family members for being a source of love and support. Finally, I am indebted to my wife Christabelle and her family for warmth and love. I dedicate this book to the memory of my grandparents Joseph and Rose McMahon, and Norbert Jay and Mary Wilde. *Lux aeterna luceat eis, Domine.*

Mark M. Wilde
Montreal, Quebec, Canada
June 2011

Part I

Introduction

CHAPTER 1

Concepts in Quantum Shannon Theory

In these first few chapters, our aim is to establish a firm grounding so that we can address some fundamental questions regarding information transmission over quantum channels. This area of study has become known as "quantum Shannon theory" in the broader quantum information community, in order to distinguish this topic from other areas of study in quantum information science. In this text, we will use the terms "quantum Shannon theory" and "quantum information theory" somewhat interchangeably. We will begin by briefly overviewing several fundamental aspects of the quantum theory. Our study of the quantum theory, in this chapter and future ones, will be at an abstract level, without giving preference to any particular physical system such as a spin-1/2 particle or a photon. This approach will be more beneficial for the purposes of our study, but, here and there, we will make some reference to actual physical systems to ground us in reality.

You may be wondering, what is *quantum Shannon theory* and why do we name this area of study as such? In short, quantum Shannon theory is the study of the ultimate capability of noisy physical systems, governed by the laws of quantum mechanics, to preserve information and correlations. Quantum information theorists have chosen the name *quantum Shannon theory* to honor Claude Shannon, who single-handedly founded the field of classical information theory, with a groundbreaking paper (Shannon, 1948). In particular, the name refers to the asymptotic theory of quantum information, which is the main topic of study in this book. Information theorists since Shannon have dubbed him the "Einstein of the information age."[1] The name *quantum Shannon theory* is fit to capture this area of study because we often use quantum versions of Shannon's ideas to prove some of the main theorems in quantum Shannon theory.

We prefer the name "quantum Shannon theory" over such names as "quantum information science" or just "quantum information." These other names are too broad, encompassing subjects as diverse as quantum computation, quantum algorithms, quantum complexity the-

[1] It is worthwhile to look up "Claude Shannon—Father of the Information Age" on YouTube and watch several reknowned information theorists speak with awe about "the founding father" of information theory.

ory, quantum communication complexity, entanglement theory, quantum key distribution, quantum error correction, and even the experimental implementation of quantum protocols. Quantum Shannon theory does overlap with some of the aforementioned subjects, such as quantum computation, entanglement theory, quantum key distribution, and quantum error correction, but the name "quantum Shannon theory" should evoke a certain paradigm for quantum communication with which the reader will become intimately familiar after some exposure to the topics in this book. For example, it is necessary for us to discuss *quantum gates* (a topic in quantum computing) because quantum Shannon-theoretic protocols exploit them to achieve certain information-processing tasks. Also, in Chapter 23, we are interested in the ultimate limitation on the ability of a noisy quantum communication channel to transmit private information (information that is secret from any third party besides the intended receiver). This topic connects quantum Shannon theory with quantum key distribution because the private information capacity of a noisy quantum channel is strongly related to the task of using the quantum channel to distribute a secret key. As a final connection, one of the most important theorems of quantum Shannon theory is the *quantum capacity theorem*. This theorem determines the ultimate rate at which a sender can reliably transmit quantum information over a quantum channel to a receiver. The result provided by the quantum capacity theorem is closely related to the theory of quantum error correction, but the mathematical techniques used in quantum Shannon theory and in quantum error correction are so different that these subjects merit different courses of study.

Quantum Shannon theory intersects two of the great sciences of the twentieth century: the quantum theory and information theory. It was really only a matter of time before physicists, mathematicians, computer scientists, and engineers began to consider the convergence of the two subjects because the quantum theory was essentially established by 1926 and information theory by 1948. This convergence has sparked what we may call the "quantum information revolution" or what some refer to as the "second quantum revolution" (Dowling and Milburn, 2003) (with the first one being the discovery of the quantum theory).

The fundamental components of the quantum theory are a set of postulates that govern phenomena on the scale of atoms. Uncertainty is at the heart of the quantum theory—"quantum uncertainty" or "Heisenberg uncertainty" is not due to our lack or loss of information or due to imprecise measurement capability, but rather, it is a fundamental uncertainty inherent in nature itself. The discovery of the quantum theory came about as a total shock to the physics community, shaking the foundations of scientific knowledge. Perhaps it is for this reason that every introductory quantum mechanics course delves into its history in detail and celebrates the founding fathers of the quantum theory. In this book, we do not discuss the history of the quantum theory in much detail and instead refer to several great introductory books for these details (Bohm, 1989; Sakurai, 1994; Griffiths, 1995; Feynman, 1998). Physicists such as Planck, Einstein, Bohr, de Broglie, Born, Heisenberg, Schrödinger, Pauli, Dirac, and von Neumann contributed to the foundations of the quantum theory in the 1920s and 1930s. We introduce the quantum theory by *briefly* commenting on its history and major underlying concepts.

Information theory is the second great foundational science for quantum Shannon theory.

In some sense, it could be viewed as merely an application of probability theory. Its aim is to quantify the ultimate compressibility of information and the ultimate ability for a sender to transmit information reliably to a receiver. It relies upon probability theory because "classical" uncertainty, arising from our lack of total information about any given scenario, is ubiquitous throughout all information-processing tasks. The uncertainty in classical information theory is the kind that is present in the flipping of a coin or the shuffle of a deck of cards, the uncertainty due to imprecise knowledge. "Quantum" uncertainty is inherent in nature itself and is perhaps not as intuitive as the uncertainty that classical information theory measures. We later expand further on these differing kinds of uncertainty, and Chapter 4 shows how a theory of quantum information captures both kinds of uncertainty within one formalism.[2]

The history of classical information theory began with Claude Shannon. Shannon's contribution is heralded as one of the single greatest contributions to modern science because he established the field in his seminal paper (Shannon, 1948). In this paper, he coined the essential terminology, and he stated and justified the main mathematical definitions and the two fundamental theorems of information theory. Many successors have contributed to information theory, but most, if not all, of the follow-up contributions employ Shannon's line of thinking in some form. In quantum Shannon theory, we will notice that many of Shannon's original ideas are present, though they take a particular "quantum" form.

One of the major assumptions in both classical information theory and quantum Shannon theory is that local computation is free but communication is expensive. In particular, for the classical case, we assume that each party has unbounded computation available. For the quantum case, we assume that each party has a fault-tolerant quantum computer available at his or her local station and the power of each quantum computer is unbounded. We also assume that both communication and a shared resource are expensive, and for this reason, we keep track of these resources in a *resource count*. Sometimes however, we might say that classical communication is free in order to simplify a scenario. A simplification like this one can lead to greater insights that might not be possible without making such an assumption.

We should first study and understand the postulates of the quantum theory in order to study quantum Shannon theory properly. Your heart may sink when you learn that the Nobel Prize-winning physicist Richard Feynman is famously quoted as saying, "I think I can safely say that nobody understands quantum mechanics." We should take the liberty of clarifying Feynman's statement. Of course, Feynman does not intend to suggest that no one knows how to work with the quantum theory. Many well-abled physicists are employed to spend their days exploiting the laws of the quantum theory to do fantastic things, such as the trapping of ions in a vacuum or applying the quantum tunneling effect in a transistor to process a single electron. I am hoping that you will give me the license to interpret Feynman's statement. I think he means that it is very difficult for us to understand the quantum theory intuitively because we do not experience the phenomena that it predicts. If we were the size of atoms and we experienced the laws of quantum theory on a daily basis, then perhaps the

[2] Von Neumann established the density operator formalism in his 1932 book on the quantum theory. This mathematical framework captures both kinds of uncertainty (von Neumann, 1996).

quantum theory would be as intuitive to us as Newton's law of universal gravitation.[3] Thus, in this sense, I would agree with Feynman—nobody can really understand the quantum theory because it is not part of our everyday experiences. Nevertheless, our aim in this book is to work with the laws of quantum theory so that we may begin to gather insights about what the theory predicts. Only by exposure to and practice with its postulates can we really gain an intuition for its predictions. It is best to imagine that the world in our everyday life does incorporate the postulates of quantum mechanics, because, indeed, as many, many experiments have confirmed, it does!

We delve into the history of the convergence of the quantum theory and information theory in some detail in this introductory chapter because this convergence does have an interesting history and is relevant to the topic of this book. The purpose of this historical review is not only to become familiar with the field itself but also to glimpse into the minds of the founders of the field so that we may see the types of questions that are important to think about when tackling new, unsolved problems.[4] Many of the most important results come about from asking simple, yet profound, questions and exploring the possibilities.

We first briefly review the history and the fundamental concepts of the quantum theory before delving into the convergence of the quantum theory and information theory. We build on these discussions by introducing some of the initial fundamental contributions to quantum Shannon theory. The final part of this chapter ends by posing some of the questions to which quantum Shannon theory provides answers.

1.1 Overview of the Quantum Theory

1.1.1 Brief History of the Quantum Theory

A physicist living around 1890 would have been well pleased with the progress of physics, but perhaps frustrated at the seeming lack of open research problems. It seemed as though the Newtonian laws of mechanics, Maxwell's theory of electromagnetism, and Boltzmann's theory of statistical mechanics explained most natural phenomena. In fact, Max Planck, one of the founding fathers of the quantum theory, was searching for an area of study in 1874 and his advisor gave him the following guidance:

> "In this field [of physics], almost everything is already discovered, and all that remains is to fill a few holes."

[3] Of course, Newton's law of universal gravitation was a revolutionary breakthrough because the phenomenon of gravity is not entirely intuitive when a student first learns it. But, we do experience the gravitational law in our daily lives and I would argue that this phenomenon is much more intuitive than, say, the phenomenon of quantum entanglement.

[4] Another way to discover good questions is to attend parties that well-established professors hold. The story goes that Oxford physicist David Deutsch attended a 1981 party at the Austin, Texas house of reknowned physicist John Archibald Wheeler, in which many attendees discussed the foundations of computing (Mullins, 2001). Deutsch claims that he could immediately see that the quantum theory would give an improvement for computation. A few years later, he published an algorithm in 1985 that was the first instance of a quantum speed-up over the fastest classical algorithm (Deutsch, 1985).

Two Clouds

Fortunately, Planck did not heed this advice and instead began his physics studies. Not everyone agreed with Planck's former advisor. Lord Kelvin stated in his famous April 1900 lecture that "two clouds" surrounded the "beauty and clearness of theory" (Kelvin, 1901). The first cloud was the failure of Michelson and Morley to detect a change in the speed of light as predicted by an "ether theory," and the second cloud was the ultraviolet catastrophe, the classical prediction that a blackbody emits radiation with an infinite intensity at high ultraviolet frequencies. Also in 1900, Planck started the quantum revolution that began to clear the second cloud. He assumed that light comes in discrete bundles of energy and used this idea to produce a formula that correctly predicts the spectrum of blackbody radiation (Planck, 1901). A great cartoon lampoon of the ultraviolet catastrophe shows Planck calmly sitting fireside with a classical physicist whose face is burning to bits because of the intense ultraviolet radiation that his classical theory predicts the fire is emitting (McEvoy and Zarate, 2004). A few years later, Einstein (1905) contributed a paper that helped to further clear the second cloud (he also cleared the first cloud with his other 1905 paper on special relativity). He assumed that Planck was right and showed that the postulate that light arrives in "quanta" (now known as the photon theory) provides a simple explanation for the photoelectric effect, the phenomenon in which electromagnetic radiation beyond a certain threshold frequency impinging on a metallic surface induces a current in that metal.

These two explanations of Planck and Einstein fueled a theoretical revolution in physics that some now call the first quantum revolution (Dowling and Milburn, 2003). Some years later, de Broglie (1924) postulated that every element of matter, whether an atom, electron, or photon, has both particle-like behavior and wave-like behavior. Just two years later, Schrödinger (1926) used the de Broglie idea to formulate a wave equation, now known as Schrödinger's equation, that governs the evolution of a closed quantum-mechanical system. His formalism later became known as wave mechanics and was popular among physicists because it appealed to notions with which they were already familiar. Meanwhile, Heisenberg (1925) formulated an "alternate" quantum theory called matrix mechanics. His theory used matrices and linear algebra, mathematics with which many physicists at the time were not readily familiar. For this reason, Schrödinger's wave mechanics was more popular than Heisenberg's matrix mechanics. In 1930, Paul Dirac published a textbook (now in its fourth edition and reprinted 16 times) that unified the formalisms of Schrödinger and Heisenberg, showing that they were actually equivalent (Dirac, 1982). In a later edition, he introduced the now ubiquitous "Dirac notation" for quantum theory that we will employ in this book.

After the publication of Dirac's textbook, the quantum theory then stood on firm mathematical grounding and the basic theory had been established. We thus end our historical overview at this point and move on to the fundamental concepts of the quantum theory.

1.1.2 Fundamental Concepts of the Quantum Theory

Quantum theory, as applied in quantum information theory, really has only a few important concepts. We review each of these aspects of quantum theory briefly in this section. Some of these phenomena are uniquely "quantum" but others do occur in the classical theory. In short, these concepts are as follows:[5]

1. indeterminism,
2. interference,
3. uncertainty,
4. superposition,
5. entanglement.

The quantum theory is *indeterministic* because the theory makes predictions about probabilities of events only. This aspect of quantum theory is in contrast with a deterministic classical theory such as that predicted by the Newtonian laws. In the Newtonian system, it is possible to predict, with certainty, the trajectories of all objects involved in an interaction if one knows only the initial positions and velocities of all the objects. This deterministic view of reality even led some to believe in determinism from a philosophical point of view. For instance, the mathematician Pierre-Simon Laplace once stated that a supreme intellect, colloquially known as "Laplace's demon," could predict all future events from present and past events:

> "We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes."

The application of Laplace's statement to atoms is fundamentally incorrect, but we can forgive him because the quantum theory had not yet been established in his time. Many have extrapolated from Laplace's statement to argue the invalidity of human free will. We leave such debates to philosophers.[6]

In reality, we never can possess full information about the positions and velocities of every object in any given physical system. Incorporating probability theory then allows us to make predictions about the probabilities of events and, with some modifications, the classical theory becomes an indeterministic theory. Thus, indeterminism is not a unique aspect of the quantum theory but merely a feature of it. But this feature is so crucial to the quantum theory that we list it among the fundamental concepts.

[5] I have used Todd A. Brun's list from his lecture notes (Brun).

[6] John Archibald Wheeler may disagree with this approach. He once said, "Philosophy is too important to be left to the philosophers" (Misner et al., 2009).

Interference is another feature of the quantum theory. It is also present in any classical wave theory—constructive interference occurs when the crest of one wave meets the crest of another, producing a stronger wave, while destructive interference occurs when the crest of one wave meets the trough of another, canceling out each other. In any classical wave theory, a wave occurs as a result of many particles in a particular medium coherently displacing one another, as in an ocean surface wave or a sound pressure wave, or as a result of coherent oscillating electric and magnetic fields, as in an electromagnetic wave. The strange aspect of interference in the quantum theory is that even a single "particle" such as an electron can exhibit wave-like features, as in the famous double slit experiment (see, e.g., Greene (1999) for a history of these experiments). This quantum interference is what contributes wave–particle duality to every fundamental component of matter.

Uncertainty is at the heart of the quantum theory. Uncertainty in the quantum theory is fundamentally different from uncertainty in the classical theory (discussed in the former paragraph about an indeterministic classical theory). The archetypal example of uncertainty in the quantum theory occurs for a single particle. This particle has two complementary variables: its position and its momentum. The uncertainty principle states that it is impossible to know both the particle's position and momentum to arbitrary accuracy. This principle even calls into question the meaning of the word "know" in the previous sentence in the context of quantum theory. We might say that we can only know that which we measure, and thus, we can only know the position of a particle after performing a precise measurement that determines it. If we follow with a precise measurement of its momentum, we lose all information about the position of the particle after learning its momentum. In quantum information science, the BB84 protocol for quantum key distribution exploits the uncertainty principle and statistical analysis to determine the presence of an eavesdropper on a quantum communication channel by encoding information into two complementary variables (Bennett and Brassard, 1984).

The *superposition* principle states that a quantum particle can be in a linear combination state, or *superposed state*, of any two other allowable states. This principle is a result of the linearity of quantum theory. Schrodinger's wave equation is a linear differential equation, meaning that the linear combination $\alpha\psi + \beta\phi$ is a solution of the equation if ψ and ϕ are both solutions of the equation. We say that the solution $\alpha\psi + \beta\phi$ is a coherent superposition of the two solutions. The superposition principle has dramatic consequences for the interpretation of the quantum theory—it gives rise to the notion that a particle can somehow "be in one location and another" at the same time. There are different interpretations of the meaning of the superposition principle, but we do not highlight them here. We merely choose to use the technical language that the particle is in a superposition of both locations. The loss of a superposition can occur through the interaction of a particle with its environment. Maintaining an arbitrary superposition of quantum states is one of the central goals of a quantum communication protocol.

The last, and perhaps most striking, quantum feature that we highlight here is *entanglement*. There is no true classical analog of entanglement. The closest analog of entanglement might be a secret key that two parties possess, but even this analogy does not come close.

Entanglement refers to the strong quantum correlations that two or more quantum particles can possess. The correlations in quantum entanglement are stronger than any classical correlations in a precise, technical sense. Schrödinger (1935) first coined the term “entanglement” after observing some of its strange properties and consequences. Einstein, Podolsky, and Rosen then presented an apparent paradox involving entanglement that raised concerns over the completeness of the quantum theory (Einstein et al., 1935). That is, they suggested that the seemingly strange properties of entanglement called the uncertainty principle into question (and thus the completeness of the quantum theory) and furthermore suggested that there might be some “local hidden-variable” theory that could explain the results of experiments. It took about 30 years to resolve this paradox, but John Bell did so by presenting a simple inequality, now known as a Bell inequality (Bell, 1964). He showed that any two-particle classical correlations that satisfy the assumptions of the “local hidden-variable theory” of Einstein, Podolsky, and Rosen must be less than a certain amount. He then showed how the correlations of two entangled quantum particles can violate this inequality, and thus, entanglement has no explanation in terms of classical correlations but is instead a uniquely quantum phenomenon. Experimentalists later verified that two entangled quantum particles can violate Bell’s inequality (Aspect et al., 1981).

In quantum information science, the non-classical correlations in entanglement play a fundamental role in many protocols. For example, entanglement is the enabling resource in teleportation, a protocol that disembodies a quantum state in one location and reproduces it in another. We will see many other examples of entanglement throughout this book.

Entanglement theory concerns methods for quantifying the amount of entanglement present not only in a two-particle state but also in a multiparticle state. A large body of literature exists that investigates entanglement theory (Horodecki et al., 2009), but we only address aspects of it that are relevant in our study of quantum Shannon theory.

The above five features capture the essence of the quantum theory, but we will see more aspects of it as we progress through our overview in Chapters 3, 4, and 5.

1.2 The Emergence of Quantum Shannon Theory

In the previous section, we discussed several unique quantum phenomena such as superposition and entanglement, but it is not clear what kind of information these unique quantum phenomena represent. Is it possible to find a convergence of the quantum theory and Shannon’s information theory, and if so, what is the convergence?

1.2.1 The Shannon Information Bit

A fundamental contribution of Shannon is the notion of a *bit* as a measure of information. Typically, when we think of a bit, we think of a two-valued quantity that can be in the state “off” or the state “on.” We represent this bit with a binary number that can be “0” or “1.” We also associate a physical representation with a bit—this physical representation can be whether a light switch is off or on, whether a transistor allows current to flow or does not, whether a large number of magnetic spins point in one direction or another, the list going on and on. These are all physical notions of a bit.

Shannon's notion of a bit is quite different from these physical notions, and we motivate his notion with the example of a fair coin. Without flipping the coin, we have no idea what the result of a coin flip will be—our best guess at the result is to guess randomly. If someone else learns the result of a random coin flip, we can ask this person the question: What was the result? We then learn *one bit of information*.

Though it may seem obvious, it is important to stress that we do not learn any or not as much information if we do not ask the right question. This point becomes even more important in the quantum case. Suppose that the coin is not fair—without loss of generality, suppose the probability of "heads" is greater than the probability of "tails." In this case, we would not be as surprised to learn that the result of a coin flip is "heads." We may say in this case that we learn less than one bit of information if we were to ask someone the result of the coin flip.

The Shannon binary entropy is a measure of information. Given a probability distribution $(p, 1-p)$ for a binary random variable, its Shannon binary entropy is

$$h_2(p) \equiv -p \log p - (1-p) \log(1-p), \tag{1.1}$$

where here and throughout the book (unless stated explicitly otherwise), the logarithm is taken base two. The Shannon binary entropy measures information in units of bits. We will discuss it in more detail in the next chapter and in Chapter 10.

The Shannon bit, or Shannon binary entropy, is a measure of the surprise upon learning the outcome of a random binary experiment. Thus, the Shannon bit has a completely different interpretation from that of the physical bit. The outcome of the coin flip resides in a physical bit, but it is the information associated with the random nature of the physical bit that we would like to measure. It is this notion of bit that is important in information theory.

1.2.2 A Measure of Quantum Information

The above section discusses Shannon's notion of a bit as a measure of information. A natural question is whether there is an analogous measure of quantum information, but before we can even ask that question, we might first wonder: What is *quantum information*? As in the classical case, there is a *physical* notion of quantum information. A quantum state always resides "in" a physical system. Perhaps another way of stating this idea is that every physical system is in some quantum state. The physical notion of a quantum bit, or qubit for short (pronounced "cue · bit"), is a two-level quantum system. Examples of two-level quantum systems are the spin of the electron, the polarization of a photon, or an atom with a ground state and an excited state. The physical notion of a qubit is straightforward to understand once we have a grasp of the quantum theory.

A more pressing question for us in this book is to understand an *informational* notion of a qubit, as in the Shannon sense. In the classical case, we quantify information by the amount of knowledge we gain after learning the answer to a probabilistic question. In the quantum world, what knowledge can we have of a quantum state?

Sometimes we may know the exact quantum state of a physical system because we prepared the quantum system in a certain way. For example, we may prepare an electron in its “spin-up in the z direction” state, where $|\uparrow_z\rangle$ denotes this state. If we prepare the state in this way, we know for certain that the state is indeed $|\uparrow_z\rangle$ and no other state. Thus, we do not gain any information, or equivalently, there is no removal of uncertainty if someone else tells us that the state is $|\uparrow_z\rangle$. We may say that this state has zero qubits of quantum information, where the term “qubit” now refers to a measure of the quantum information of a state.

In the quantum world, we also have the option of measuring this state in the x direction. The postulates of quantum theory, given in Chapter 3, predict that the state will then be $|\uparrow_x\rangle$ or $|\downarrow_x\rangle$ with equal probability after measuring in the x direction. One interpretation of this aspect of quantum theory is that the system does not have any definite state in the x direction: in fact there is maximal uncertainty about its x direction, if we know that the physical system has a definite z direction. This behavior is one manifestation of the Heisenberg uncertainty principle. So before performing the measurement, we have no knowledge of the resulting state and we gain one Shannon bit of information after learning the result of the measurement. If we use Shannon’s notion of entropy and perform an x measurement, this classical measure loses some of its capability here to capture our knowledge of the state of the system. It is inadequate to capture our knowledge of the state because we actually prepared it ourselves and know with certainty that it is in the state $|\uparrow_z\rangle$. With these different notions of information gain, which one is the most appropriate for the quantum case?

It turns out that the first way of thinking is the one that is most useful for quantifying quantum information. If someone tells us the definite quantum state of a particular physical system and this state is indeed the true state, then we have complete knowledge of the state and thus do not learn more “qubits” of quantum information from this point onward. This line of thinking is perhaps similar in one sense to the classical world, but different from the classical world, in the sense of the case presented in the previous paragraph.

Now suppose that a friend (let us call him “Bob”) randomly prepares quantum states as a probabilistic ensemble. Suppose Bob prepares $|\uparrow_z\rangle$ or $|\downarrow_z\rangle$ with equal probability. With only this probabilistic knowledge, we acquire one bit of information if Bob reveals which state he prepared. We could also perform a quantum measurement on the system to determine what state Bob prepared (we discuss quantum measurements in detail in Chapter 3). One reasonable measurement to perform is a measurement in the z direction. The result of the measurement determines which state Bob actually prepared because both states in the ensembles are states with definite z direction. The result of this measurement thus gives us one bit of information—the same amount that we would learn if Bob informed us which state he prepared. It seems that most of this logic is similar to the classical case—i.e., the result of the measurement only gave us one Shannon bit of information.

Another measurement to perform is a measurement in the x direction. If the actual state prepared is $|{\uparrow_z}\rangle$, then the quantum theory predicts that the state becomes $|{\uparrow_x}\rangle$ or $|{\downarrow_x}\rangle$ with equal probability. Similarly, if the actual state prepared is $|{\downarrow_z}\rangle$, then the quantum theory predicts that the state again becomes $|{\uparrow_x}\rangle$ or $|{\downarrow_x}\rangle$ with equal probability. Calculating probabilities, the resulting state is $|{\uparrow_x}\rangle$ with probability $1/2$ and $|{\downarrow_x}\rangle$ with probability $1/2$. So the Shannon bit content of learning the result is again one bit, but we arrived at this conclusion in a much different fashion from the scenario in which we measured in the z direction. How can we quantify the *quantum information* of this ensemble? We claim for now that this ensemble contains one *qubit* of quantum information and this result derives from either the measurement in the z direction or the measurement in the x direction for this particular ensemble.

Let us consider one final example that perhaps gives more insight into how we might quantify quantum information. Suppose Bob prepares $|{\uparrow_z}\rangle$ or $|{\uparrow_x}\rangle$ with equal probability. The first state is spin-up in the z direction and the second is spin-up in the x direction. If Bob reveals which state he prepared, then we learn one Shannon bit of information. But suppose now that we would like to learn the prepared state on our own, without the help of our friend Bob. One possibility is to perform a measurement in the z direction. If the state prepared is $|{\uparrow_z}\rangle$, then we learn this result with probability $1/2$. But if the state prepared is $|{\uparrow_x}\rangle$, then the quantum theory predicts that the state becomes $|{\uparrow_z}\rangle$ or $|{\downarrow_z}\rangle$ with equal probability (while we learn what the new state is). Thus, quantum theory predicts that the act of measuring this ensemble inevitably disturbs the state some of the time. Also, there is no way that we can learn with certainty whether the prepared state is $|{\uparrow_z}\rangle$ or $|{\uparrow_x}\rangle$. Using a measurement in the z direction, the resulting state is $|{\uparrow_z}\rangle$ with probability $3/4$ and $|{\downarrow_z}\rangle$ with probability $1/4$. We learn less than one Shannon bit of information from this ensemble because the probability distribution becomes skewed when we perform this particular measurement.

The probabilities resulting from the measurement in the z direction are the same that would result from an ensemble where Bob prepares $|{\uparrow_z}\rangle$ with probability $3/4$ and $|{\downarrow_z}\rangle$ with probability $1/4$ and we perform a measurement in the z direction. The actual Shannon entropy of the distribution $(3/4, 1/4)$ is about 0.81 bits, confirming our intuition that we learn approximately less than one bit. A similar, symmetric analysis holds to show that we gain 0.81 bits of information when we perform a measurement in the x direction.

We have more knowledge of the system in question if we gain less information from performing measurements on it. In the quantum theory, we learn less about a system if we perform a measurement on it that does not disturb it too much. Is there a measurement that we can perform in which we learn the least amount of information? Recall that learning the least amount of information is ideal because it has the interpretation that we require fewer questions on average to learn the result of a random experiment. Indeed, it turns out that a measurement in the $x+z$ direction reveals the least amount of information. Avoiding details for now, this measurement returns a state that we label $|{\uparrow_{x+z}}\rangle$ with probability $\cos^2(\pi/8)$ and a state $|{\downarrow_{x+z}}\rangle$ with probability $\sin^2(\pi/8)$. This measurement has the desirable effect that it causes the least amount of disturbance to the original states in the ensemble. The entropy of the distribution resulting from the measurement is about 0.6 bits and is less than the one bit that we learn if Bob reveals the state. The entropy ≈ 0.6 is also the least amount of information among all possible sharp measurements that we may perform on the ensemble. We claim that this ensemble contains ≈ 0.6 *qubits* of quantum information.

We can determine the ultimate compressibility of classical data with Shannon's source coding theorem (we overview this technique in the next chapter). Is there a similar way that we can determine the ultimate compressibility of quantum information? This question was one of the early and profitable ones for quantum Shannon theory and the answer is affirmative. The technique for quantum compression is called Schumacher compression, named after Benjamin Schumacher. Schumacher used ideas similar to that of Shannon—he created the notion of a quantum information source that emits random physical qubits, and he invoked the law of large numbers to show that there is a so-called *typical subspace* where most of the quantum information really resides. This line of thought is similar to that which we will discuss in the overview of data compression in the next chapter. The size of the typical subspace for most quantum information sources is exponentially smaller than the size of the space in which the emitted physical qubits resides. Thus, one can "quantum compress" the quantum information to this subspace without losing much. Schumacher's quantum source coding theorem then quantifies, in an operational sense, the amount of actual quantum information that the ensemble contains. The amount of actual quantum information corresponds to the number of qubits, in the informational sense, that the ensemble contains. It is this measure that is equivalent to the "optimal measurement" one that we suggested in the previous paragraph. We will study this idea in more detail later when we introduce the quantum theory and a rigorous notion of a quantum information source.

Some of the techniques of quantum Shannon theory are the direct *quantum* analog of the techniques from classical information theory. We use the law of large numbers and the notion of the typical subspace, but we require generalizations of measures from the classical world to determine how "close" two different quantum states are. One measure, the *fidelity*, has the operational interpretation that it gives the probability that one quantum state would pass a test for being another. The *trace distance* is another distance measure that is perhaps more similar to a classical distance measure—its classical analog is a measure of the closeness of two probability distributions. The techniques in quantum Shannon theory also reside firmly in the quantum theory and have no true classical analog for some cases. Some of the techniques will seem similar to those in the classical world, but the answer to some of the fundamental questions in quantum Shannon theory are rather different from some of the answers in the classical world. It is the purpose of this book to explore the answers to the fundamental questions of quantum Shannon theory, and we now begin to ask what kinds of tasks we can perform.

1.2.3 Operational Tasks in Quantum Shannon Theory

Quantum Shannon theory has several resources that two parties can exploit in a quantum information-processing task. Perhaps the most natural quantum resource is a *noiseless qubit channel*. We can think of this resource as some medium through which a physical qubit can travel without being affected by any noise. One example of a noiseless qubit channel could be the free space through which a photon travels, where it ideally does not interact with any other particles along the way to its destination.[7]

A *noiseless classical bit channel* is a special case of a noiseless qubit channel because we can always encode classical information into quantum states. For the example of a photon, we can say that horizontal polarization corresponds to a "0" and vertical polarization corresponds to a "1." We refer to the dynamic resource of a noiseless classical bit channel as a *cbit*, in order to distinguish it from the noiseless qubit channel.

Perhaps the most intriguing resource that two parties can share is noiseless entanglement. Any entanglement resource is a *static resource* because it is one that they share. Examples of static resources in the classical world are an information source that we would like to compress or a common secret key that two parties may possess. We actually have a way of measuring entanglement that we discuss later on, and for this reason, we can say that a sender and receiver have bits of entanglement or *ebits*.

Entanglement turns out to be a useful resource in many quantum communication tasks. One example where it is useful is in the teleportation protocol, where a sender and receiver use one ebit and two classical bits to transmit one qubit faithfully. This protocol is an example of the extraordinary power of noiseless entanglement. The name "teleportation" is really appropriate for this protocol because the physical qubit vanishes from the sender's station and appears at the receiver's station after the receiver obtains the two transmitted classical bits. We will see later on that a noiseless qubit channel can generate the other two noiseless resources, but it is impossible for each of the other two noiseless resources to generate the noiseless qubit channel. In this sense, the noiseless qubit channel is the strongest of the three unit resources.

The first quantum information-processing task that we have discussed is Schumacher compression. The goal of this task is to use as few noiseless qubit channels as possible in order to transmit the output of a quantum information source reliably. After we understand Schumacher compression in a technical sense, the main focus of this book is to determine what quantum information-processing tasks a sender and receiver can accomplish with the use of a noisy quantum channel. The first and perhaps simplest task is to determine how much classical information a sender can transmit reliably to a receiver, by using a noisy quantum channel a large number of times. This task is known as HSW coding, named after its discoverers Holevo, Schumacher, and Westmoreland. The HSW coding theorem is one quantum generalization of Shannon's channel coding theorem (the latter overviewed in the next chapter). We can also assume that a sender and receiver share some amount of noiseless entanglement prior to communication. They can then use this noiseless entanglement in addition to a large number of uses of a noisy quantum channel. This task is known as *entanglement-assisted classical communication* over a noisy quantum channel. The capacity theorem corresponding to this task again highlights one of the marvelous features of entanglement. It shows that entanglement gives a boost to the amount of noiseless classical communication we can generate using a noisy quantum channel—the classical capacity is generally higher with entanglement assistance than without it.

[7] We should be careful to note here that this is not actually a perfect channel because even empty space can be noisy in quantum mechanics, but nevertheless, it is a simple physical example to imagine.

One of the most important theorems for quantum Shannon theory is the *quantum channel capacity theorem*. Any proof of a capacity theorem consists of two parts: one part establishes a lower bound on the capacity and the other part establishes an upper bound. If the two bounds coincide, then we have a characterization of the capacity in terms of these bounds. The lower bound on the quantum capacity is colloquially known as the LSD coding theorem,[8] and it gives a characterization of the highest rate at which a sender can transmit quantum information reliably over a noisy quantum channel so that a receiver can recover it perfectly. The rate is generally lower than the classical capacity because it is more difficult to keep quantum information intact. As we have said before, it is possible to encode classical information into quantum states, but this classical encoding is only a special case of a quantum state. In order to preserve quantum information, we have to be able to preserve arbitrary quantum states, not merely a classical encoding within a quantum state.

The pinnacle of this book is in Chapter 24 where we finally reach our study of the quantum capacity theorem. All efforts and technical developments in preceding chapters have this goal in mind.[9] Our first coding theorem in the dynamic setting is the HSW coding theorem. A rigorous study of this coding theorem lays an important foundation—an understanding of the structure of a code for reliable communication over a noisy quantum channel. The method for the HSW coding theorem applies to the "entanglement-assisted classical capacity theorem," which is one building block for other protocols in quantum Shannon theory. We then develop a more complex coding structure for sending private classical information over a noisy quantum channel. In *private coding*, we are concerned with coding in such a way that the intended receiver can learn the transmitted message perfectly, but a third-party eavesdropper cannot learn anything about what the sender transmits to the intended receiver. This study of the private classical capacity may seem like a detour at first, but it is closely linked with our ultimate aim. The coding structure developed for sending private information proves to be indispensable for understanding the structure of a quantum code. There are strong connections between the goals of keeping classical information private and keeping quantum information coherent. In the private coding scenario, the goal is to avoid leaking any information to an eavesdropper so that she cannot learn anything about the transmission. In the quantum coding scenario, we can think of quantum noise as resulting from the environment learning about the transmitted quantum information and this act of learning disturbs the quantum information. This effect is related to the information–disturbance trade-off that is fundamental in quantum information theory. If the environment learns something about the state being transmitted, there is inevitably some sort of noisy disturbance that affects the quantum state. Thus, we can see a correspondence between private coding and quantum coding. In quantum coding, the goal is to avoid leaking any information to the environment because the avoidance of such a leak implies that there is no

[8] The LSD coding theorem does not refer to the synthetic crystalline compound, lysergic acid diethylamide (which one may potentially use as a hallucinogenic drug), but refers rather to Lloyd (1997), Shor (2002b), and Devetak (2005), all of whom gave separate proofs of the lower bound on the quantum capacity with increasing standards of rigor.

[9] One goal of this book is to unravel the mathematical machinery behind Devetak's proof of the quantum channel coding theorem (Devetak, 2005).

disturbance to the transmitted state. So the role of the environment in quantum coding is similar to the role of the eavesdropper in private coding, and the goal in both scenarios is to decouple either the environment or eavesdropper from the picture. It is then no coincidence that private codes and quantum codes have a similar structure. In fact, we can say that the quantum code inherits its structure from that of the private code.[10]

We also consider "trade-off" problems in addition to discussing the quantum capacity theorem. Chapter 22 is another high point of the book, featuring a whole host of results that emerge by combining several of the ideas from previous chapters. The most appealing aspect of this chapter is that we can construct virtually all of the protocols in quantum Shannon theory from just one idea in Chapter 21. Also, Chapter 22 provides partial answers to many practical questions concerning information transmission over noisy quantum channels. Some example questions are as follows:

- How much quantum and classical information can a noisy quantum channel transmit?
- An entanglement-assisted noisy quantum channel can transmit more classical information than an unassisted one, but how much entanglement is really necessary?
- Does noiseless classical communication help in transmitting quantum information reliably over a noisy quantum channel?
- How much entanglement can a noisy quantum channel generate when aided by classical communication?
- How much quantum information can a noisy quantum channel communicate when aided by entanglement?

These are examples of trade-off problems because they involve a noisy quantum channel and either the consumption or generation of a noiseless resource. For every combination of the generation or consumption of a noiseless resource, there is a corresponding coding theorem that states what rates are achievable (and in some cases optimal). Some of these trade-off questions admit interesting answers, but some of them do not. Our final aim in these trade-off questions is to determine the full triple trade-off solution where we study the optimal ways of combining all three unit resources (classical communication, quantum communication, and entanglement) with a noisy quantum channel.

The coding theorems for a noisy quantum channel are just as important (if not more important) as Shannon's classical coding theorems because they determine the ultimate capabilities of information processing in a world where the postulates of quantum theory apply. It is thought that quantum theory is the ultimate theory underpinning all physical phenomena and any theory of gravity will have to incorporate the quantum theory in some fashion.

[10] There are other methods of formulating quantum codes using random subspaces (Shor, 2002b; Hayden et al., 2008a,b; Klesse, 2008), but we prefer the approach of Devetak because we learn about other aspects of quantum Shannon theory, such as the private capacity, along the way to proving the quantum capacity theorem.

Thus, it is reasonable that we should be focusing our efforts now on a full Shannon theory of quantum information processing in order to determine the tasks that these systems can accomplish. In many physical situations, some of the assumptions of quantum Shannon theory may not be justified (such as an independent and identically distributed quantum channel), but nevertheless, it provides an ideal setting in which we can determine the capabilities of these physical systems.

1.2.4 History of Quantum Shannon Theory

We conclude this introductory chapter by giving a brief overview of the problems that researchers were thinking about that ultimately led to the development of quantum Shannon theory.

The 1970s—The first researchers in quantum information theory were concerned with transmitting classical data by optical means. They were ultimately led to a quantum formulation because they wanted to transmit classical information by means of a coherent laser. *Coherent states* are special quantum states that a coherent laser ideally emits. Glauber provided a full quantum-mechanical theory of coherent states in two seminal papers (Glauber, 1963a,b), for which he shared the Nobel Prize in 2005 (Glauber, 2005). The first researchers of quantum information theory were Helstrom, Gordon, Stratonovich, and Holevo. Gordon (1964) first conjectured an important bound for our ability to access classical information from a quantum system and Levitin (1969) stated it without proof. Holevo (1973a,b) later provided a proof that the bound holds. This important bound is now known as the Holevo bound, and it is useful in proving converse theorems (theorems concerning optimality) in quantum Shannon theory. The simplest (yet rough) statement of the Holevo bound states that it is not possible to transmit more than one classical bit of information using a noiseless qubit channel, while at the same time being able to decode it reliably—i.e., we get *one cbit per qubit.* Helstrom (1976) developed a full theory of quantum detection and quantum estimation and published a textbook that discusses this theory. Fannes (1973) contributed a useful continuity property of the entropy that is also useful in proving converse theorems in quantum Shannon theory. Wiesner also used the uncertainty principle to devise a notion of "quantum money" in 1970, but unfortunately, his work was not accepted upon its initial submission. This work was *way* ahead of its time, and it was only until much later that it was accepted (Wiesner, 1983). Wiesner's ideas paved the way for the BB84 protocol for quantum key distribution. Fundamental entropy inequalities, such as the strong subadditivity of quantum entropy (Lieb and Ruskai, 1973a,b) and the monotonicity of quantum relative entropy (Lindblad, 1975), were proved during this time as well. These entropy inequalities generalize the Holevo bound and are foundational for establishing optimality theorems in quantum Shannon theory.

The 1980s—The 1980s witnessed only a few advances in quantum information theory because just a handful of researchers thought about the possibilities of linking quantum theory with information-theoretic ideas. The Nobel Prize-winning physicist Richard Feynman published an interesting 1982 article that was one of the first to discuss computing with quantum-mechanical systems (Feynman, 1982). His interest was in using a quantum computer to simulate quantum-mechanical systems—he figured there should be a speed-up over a classical simulation if we instead use one quantum system to simulate another. This work is less quantum Shannon theory than it is quantum computing, but it is still a landmark because Feynman began to think about exploiting the actual quantum information in a physical system, rather than just using quantum systems to process classical information as the researchers in the 1970s suggested.

Wootters and Zurek (1982) produced one of the simplest, yet most profound, results that is crucial to quantum information science (Dieks (1982) also proved this result in the same year). They proved the *no-cloning theorem*, showing that the postulates of the quantum theory imply the impossibility of universally cloning quantum states. Given an arbitrary unknown quantum state, it is impossible to build a device that can copy this state. This result has deep implications for the processing of quantum information and shows a strong divide between information processing in the quantum world and that in the classical world. We will prove this theorem in Chapter 3 and use it time and again in our reasoning. The history of the no-cloning theorem is one of the more interesting "sociology of science" stories that you may come across. The story goes that Nick Herbert submitted a paper to *Foundations of Physics* with a proposal for faster-than-light communication using entanglement. Asher Peres was the referee (Peres, 2002), and he knew that something had to be wrong with the proposal because it allowed for superluminal communication, yet he could not put his finger on what the problem might be (he also figured that Herbert knew his proposal was flawed). Nevertheless, Peres recommended the paper for publication (Herbert, 1982) because he figured it would stimulate wide interest in the topic. Not much later, Wootters and Zurek published their paper, and since then, there have been thousands of follow-up results on the no-cloning theorem (Scarani et al., 2005).

The work of Wiesner on conjugate coding inspired an IBM physicist named Charles Bennett. Bennett and Brassard (1984) published a groundbreaking paper that detailed the first quantum communication protocol: the BB84 protocol. This protocol shows how a sender and a receiver can exploit a quantum channel to establish a secret key. The security of this protocol, roughly speaking, relies on the uncertainty principle. If any eavesdropper tries to learn about the random quantum data that they use to establish the secret key, this act of learning inevitably disturbs the transmitted quantum data and the two parties can discover this disturbance by noticing the change in the statistics of random sample data. The secret key generation capacity of a noisy quantum channel is inextricably linked to the BB84 protocol, and we study this capacity problem in detail when we study the ability of quantum channels to communicate private information. Interestingly, the physics community largely ignored the BB84 paper when Bennett and Brassard first published it, likely because they presented it at an engineering conference and the merging of physics and information had not yet taken effect.

The 1990s—The 1990s were a time of much increased activity in quantum information science, perhaps some of the most exciting years with many seminal results. One of the first major results was from Ekert. He published a different way for performing quantum key distribution, this time relying on the strong correlations of entanglement (Ekert, 1991). He was unaware of the BB84 protocol when he was working on his entanglement-based quantum key distribution. The physics community embraced this result and shortly later, Ekert and Bennett and Brassard became aware of each other's respective works (Bennett et al., 1992a). Bennett, Brassard, and Mermin later showed a sense in which these two seemingly different schemes are equivalent (Bennett et al., 1992b). Bennett later developed the B92 protocol for quantum key distribution using any two non-orthogonal quantum states (Bennett, 1992).

Two of the most profound results that later impacted quantum Shannon theory appeared in the early 1990s. First, Bennett and Wiesner (1992) devised the super-dense coding protocol. This protocol consumes one noiseless ebit of entanglement and one noiseless qubit channel to simulate two noiseless classical bit channels. Let us compare this result to that of Holevo. Holevo's bound states that we can reliably send only one classical bit per qubit, but the super-dense coding protocol states that we can double this rate if we consume entanglement as well. Thus, entanglement is the enabler in this protocol that boosts the classical rate beyond that possible with a noiseless qubit channel alone. The next year, Bennett and some other coauthors reversed the operations in the super-dense coding protocol to devise a protocol that has more profound implications. They devised the *teleportation protocol* (Bennett et al., 1993)—this protocol consumes two classical bit channels and one ebit to transmit a qubit from a sender to receiver. Right now, without any technical development yet, it may be unclear how the qubit gets from the sender to receiver. The original authors described it as the "disembodied transport of a quantum state." Suffice it for now to say that it is the unique properties of entanglement (in particular, the ebit) that enable this disembodied transport to occur. Yet again, it is entanglement that is the resource that enables this protocol, but let us be careful not to overstate the role of entanglement. Entanglement alone does not suffice for implementing quantum teleportation. These protocols show that it is the unique combination of entanglement and quantum communication or entanglement and classical communication that yields these results. These two noiseless protocols are cornerstones of quantum Shannon theory, originally suggesting that there are interesting ways of combining the resources of classical communication, quantum communication, and entanglement to formulate uniquely quantum protocols and leading the way to more exotic protocols that combine the different noiseless resources with noisy resources. Simple questions concerning these protocols lead to quantum Shannon-theoretic protocols. In super-dense coding, how much classical information can Alice send if the quantum channel becomes noisy? What if the entanglement is noisy? In teleportation, how much quantum information can Alice send if the classical channel is noisy? What if the entanglement is noisy? Researchers addressed these questions quite a bit after the original super-dense coding and teleportation protocols were available, and we discuss these important questions in this book.

The year 1994 was a landmark for quantum information science. Shor (1994) published his algorithm that factors a number in polynomial time—this algorithm gives an exponential speed-up over the best known classical algorithm. We cannot overstate the importance of this algorithm for the field. Its major application is to break RSA encryption (Rivest et al., 1978) because the security of that encryption algorithm relies on the computational difficulty of factoring a large number. This breakthrough generated wide interest in the idea of a quantum computer and started the quest to build one and study its capabilities.

Initially, much skepticism met the idea of building a practical quantum computer (Landauer, 1995; Unruh, 1995). Some experts thought that it would be impossible to overcome errors that inevitably occur during quantum interactions, due to the coupling of a quantum system with its environment. Shor met this challenge by devising the first quantum error-correcting code (Shor, 1995) and a scheme for fault-tolerant quantum computation (Shor, 1996). His paper on quantum error correction is the one most relevant for quantum Shannon theory. At the end of this paper, he posed the idea of the quantum capacity of a noisy quantum channel as the highest rate at which a sender and receiver can maintain the fidelity of a quantum state when it is sent over a large number of uses of the noisy channel. This open problem set the main task for researchers interested in quantum Shannon theory. A flurry of theoretical activity then ensued in quantum error correction (Calderbank and Shor, 1996; Steane, 1996; Laflamme et al., 1996; Gottesman, 1996, 1997; Calderbank et al., 1997, 1998) and fault-tolerant quantum computation (Aharonov and Ben-Or, 1997; Kitaev, 1997; Preskill, 1998; Knill et al., 1998). These two areas are now important subfields within quantum information science, but we do not focus on them in any detail in this book.

Schumacher published a critical paper in 1995 as well (Schumacher, 1995) (we discussed some of his contributions in the previous section). This paper gave the first informational notion of a qubit, and it even established the now ubiquitous term "qubit." He proved the quantum analog of Shannon's source coding theorem, giving the ultimate compressibility of quantum information. He used the notion of a typical subspace as an analogy of Shannon's typical set. This notion of a typical subspace proves to be one of the most crucial ideas for constructing codes in quantum Shannon theory, just as the notion of a typical set is so crucial for Shannon's information theory.

Not much later, several researchers began investigating the capacity of a noisy quantum channel for sending classical information (Hausladen et al., 1996). Holevo (1998) and Schumacher and Westmoreland (1997) independently proved that the Holevo information of a quantum channel is an achievable rate for classical communication over it. They appealed to Schumacher's notion of a typical subspace and constructed channel codes for sending classical information. The proof looks somewhat similar to the proof of Shannon's channel coding theorem (discussed in the next chapter) after taking a few steps away from it. The proof of the converse theorem proceeds somewhat analogously to that of Shannon's theorem, with the exception that one of the steps uses Holevo's bound from 1973. In hindsight, it is perhaps somewhat surprising that it took over 20 years between the appearance of the proof of Holevo's bound (the main step in the converse proof) and the appearance of a direct coding theorem for sending classical information.

The quantum capacity theorem is perhaps one of the most fundamental theorems of quantum Shannon theory. Initial work by several researchers provided some insight into the quantum capacity theorem (Bennett et al., 1996b,c, 1997; Schumacher and Westmoreland, 1998), and a series of papers established an upper bound on the quantum capacity (Schumacher, 1996; Schumacher and Nielsen, 1996; Barnum et al., 1998, 2000). For the lower bound, Lloyd (1997) was the first to construct an idea for a proof, but it turns out that his proof was more of a heuristic argument. Shor (2002b) then followed with another proof of the lower bound, and some of Shor's ideas appeared much later in a full publication (Hayden et al., 2008b). Devetak (2005) and Cai et al. (2004) independently solved the private capacity theorem at approximately the same time (with the publication of the CWY paper appearing a year after Devetak's arXiv post). Devetak took the proof of the private capacity theorem a step further and showed how to apply its techniques to construct a quantum code that achieves a good lower bound on the quantum capacity, while also providing an alternate, cleaner proof of the converse theorem (Devetak, 2005). It is Devetak's technique that we mainly explore in this book because it provides some insight into the coding structure (however, we also explore a different technique via the entanglement-assisted classical capacity theorem).

The 2000s—In recent years, we have had many advancements in quantum Shannon theory (technically some of the above contributions were in the 2000s, but we did not want to break the continuity of the history of the quantum capacity theorem). One major result was the proof of the entanglement-assisted classical capacity theorem—it is the noisy version of the super-dense coding protocol where the quantum channel is noisy (Bennett et al., 1999, 2002; Holevo, 2002b). This theorem assumes that Alice and Bob share unlimited entanglement and they exploit the entanglement and the noisy quantum channel to send classical information.

A few fantastic results have arisen in recent years. Horodecki, Oppenheim, and Winter showed the existence of a state-merging protocol (Horodecki et al., 2005, 2007). This protocol gives the minimum rate at which Alice and Bob consume noiseless qubit channels in order for Alice to send her share of a quantum state to Bob. This rate is the conditional quantum entropy—the protocol thus gives an operational interpretation to this entropic quantity. What was most fascinating about this result is that the conditional quantum entropy can be negative in quantum Shannon theory. Prior to their work, no one really understood what it meant for the conditional quantum entropy to become negative (Wehrl, 1978; Horodecki and Horodecki, 1994; Cerf and Adami, 1997), but this state-merging result gave a compelling operational interpretation. A negative rate implies that Alice and Bob gain the ability for future quantum communication, instead of consuming quantum communication as when the rate is positive.

Another fantastic result came from (Smith and Yard, 2008). Suppose we have two noisy quantum channels and each of them individually has zero capacity to transmit quantum information. One would expect intuitively that the "joint quantum capacity" (when using them together) would also have zero ability to transmit quantum information. But this result is not generally the case in the quantum world. It is possible for some particular noisy quantum channels with no individual quantum capacity to have a non-zero joint quantum capacity. It is not clear yet how we might practically take advantage of such a "superactivation" effect, but the result is nonetheless fascinating, counterintuitive, and not yet fully understood.

The latter part of the 2000s saw the unification of quantum Shannon theory. The resource inequality framework was the first step because it unified many previously known results into one formalism (Devetak et al., 2004, 2008). Devetak, Harrow, and Winter provided a family tree for quantum Shannon theory and showed how to relate the different protocols in the tree to one another. We will go into the theory of resource inequalities in some detail throughout this book because it provides a tremendous conceptual simplification when considering coding theorems in quantum Shannon theory. In fact, the last chapter of this book contains a concise summary of many of the major quantum Shannon-theoretic protocols in the language of resource inequalities. Abeyesinghe et al. (2009) published a work showing a sense in which the mother protocol of the family tree can generate the father protocol. We have seen unification efforts in the form of triple trade-off coding theorems (Abeyesinghe and Hayden, 2003; Hsieh and Wilde, 2010a,b). These theorems give the optimal combination of classical communication, quantum communication, entanglement, and an asymptotic noisy resource for achieving a variety of quantum information-processing tasks.

We have also witnessed the emergence of a study of network quantum Shannon theory. Some authors have tackled the quantum broadcasting paradigm (Guha and Shapiro, 2007; Guha et al., 2007; Dupuis et al., 2010; Yard et al., 2011), where one sender transmits to multiple receivers. A multiple-access quantum channel has many senders and one receiver. Some of the same authors (and others) have tackled multiple-access communication (Winter, 2001; Yard, 2005; Yen and Shapiro, 2005; Yard et al., 2005, 2008; Hsieh et al., 2008a; Czekaj and Horodecki, 2009). This network quantum Shannon theory should become increasingly important as we get closer to the ultimate goal of a quantum Internet.

Quantum Shannon theory has now established itself as an important and distinct field of study. The next few chapters discuss the concepts that will prepare us for tackling some of the major results in quantum Shannon theory.

CHAPTER 2

Classical Shannon Theory

We cannot overstate the importance of Shannon's contribution to modern science. His introduction of the field of information theory and his solutions to its two main theorems demonstrate that his ideas on communication were far beyond the other prevailing ideas in this domain around 1948.

In this chapter, our aim is to discuss Shannon's two main contributions in a descriptive fashion. The goal of this high-level discussion is to build up the intuition for the problem domain of information theory and to understand the main concepts before we delve into the analogous quantum information-theoretic ideas. We avoid going into deep technical detail in this chapter, leaving such details for later chapters where we formally prove both classical and quantum Shannon-theoretic coding theorems. We do use some mathematics from probability theory (namely, the law of large numbers).

We will be delving into the technical details of this chapter's material in later chapters (specifically, Chapters 10, 13, and 14). Once you have reached later chapters that develop some more technical details, it might be helpful to turn back to this chapter to get an overall flavor for the motivation of the development.

2.1 Data Compression

We first discuss the problem of data compression. Those who are familiar with the Internet have used several popular data formats such as JPEG, MPEG, ZIP, GIF, etc. All of these file formats have corresponding algorithms for compressing the output of an information source. A first glance at the compression problem might lead one to believe that it is not possible to compress the output of the information source to an arbitrarily small size, and Shannon proved that this is the case. This result is the content of Shannon's first noiseless coding theorem.

2.1.1 An Example of Data Compression

We begin with a simple example that illustrates the concept of an information source. We then develop a scheme for coding this source so that it requires fewer bits to represent its output faithfully.

Suppose that Alice is a sender and Bob is a receiver. Suppose further that a noiseless bit channel connects Alice to Bob—a noiseless bit channel is one that transmits information perfectly from sender to receiver, e.g., Bob receives "0" if Alice transmits "0" and Bob receives "1" if Alice transmits "1." Alice and Bob would like to minimize the number of times that they use this noiseless channel because it is expensive to use it.

Alice would like to use the noiseless channel to communicate information to Bob. Suppose that an information source randomly chooses from four symbols $\{a, b, c, d\}$ and selects them with a skewed probability distribution:

$$\Pr\{a\} = 1/2, \tag{2.1}$$

$$\Pr\{b\} = 1/8, \tag{2.2}$$

$$\Pr\{c\} = 1/4, \tag{2.3}$$

$$\Pr\{d\} = 1/8. \tag{2.4}$$

So it is clear that the symbol a is the most likely one, c the next likely, and both b and d are least likely. We make the additional assumption that the information source chooses each symbol independently of all previous ones and chooses each with the same probability distribution above. After the information source makes a selection, it gives the symbol to Alice for coding.

A noiseless bit channel accepts only bits as input—it does not accept the symbols a, b, c, d as input. So, Alice has to encode her information into bits. Alice could use the following coding scheme:

$$a \to 00, \quad b \to 01, \quad c \to 10, \quad d \to 11, \tag{2.5}$$

where each binary representation of a letter is a *codeword*. How do we measure the performance of a particular coding scheme? The expected length of a codeword is one way to measure performance. For the above example, the expected length is equal to two bits. This measure reveals a problem with the above scheme—the scheme does not take advantage of the skewed nature of the distribution of the information source because each codeword is the same length.

One might instead consider a scheme that uses shorter codewords for symbols that are more likely and longer codewords for symbols that are less likely.[1] Then the expected length

[1] Such coding schemes are common. Samuel F. B. Morse employed this idea in his popular Morse code. Also, in the movie *The Diving Bell and the Butterfly*, a writer becomes paralyzed with "locked-in" syndrome so that he can only blink his left eye. An assistant then develops a "blinking code" where she reads a list of letters in French, beginning with the most commonly used letter and ending with the least commonly used letter. The writer blinks when she says the letter he wishes and they finish an entire book with this coding scheme.

of a codeword with such a scheme should be shorter than that in the former scheme. The following coding scheme gives an improvement in the expected length of a codeword:

$$a \to 0, \quad b \to 110, \quad c \to 10, \quad d \to 111. \tag{2.6}$$

This scheme has the advantage that any coded sequence is uniquely decodable. For example, suppose that Bob obtains the following sequence:

$$0011010111010100010. \tag{2.7}$$

Bob can parse the above sequence as

$$0\ 0\ 110\ 10\ 111\ 0\ 10\ 10\ 0\ 0\ 10, \tag{2.8}$$

and determine that Alice transmitted the message

$$aabcdaccaac. \tag{2.9}$$

We can calculate the expected length of this coding scheme as follows:

$$\frac{1}{2}(1) + \frac{1}{8}(3) + \frac{1}{4}(2) + \frac{1}{8}(3) = \frac{7}{4}. \tag{2.10}$$

This scheme is thus more efficient because its expected length is 7/4 bits as opposed to two bits. It is a *variable-length code* because the number of bits in each codeword depends on the source symbol.

2.1.2 A Measure of Information

The above scheme suggests a way to measure information. Consider the probability distribution in (2.1)–(2.4). Would we be more surprised to learn that the information source produced the symbol a or to learn that it produced the symbol d? The answer is d because the source is less likely to produce it. Let X denote a random variable with distribution given in (2.1)–(2.4). One measure of the surprise of symbol $x \in \{a, b, c, d\}$ is

$$i(x) \equiv \log\left(\frac{1}{p_X(x)}\right) = -\log\left(p_X(x)\right), \tag{2.11}$$

where the logarithm is base two—this convention implies the units of this measure are bits. This measure of surprise has the desirable property that it is higher for lower probability events and lower for higher probability events. Here, we take after Shannon, and we name $i(x)$ the *information content* or *surprisal* of the symbol x. Observe that the length of each codeword in the coding scheme in (2.6) is equal to the information content of its corresponding symbol.

The information content has another desirable property called *additivity*. Suppose that the information source produces two symbols x_1 and x_2, with corresponding random variables

X_1 and X_2. The probability for this event is $p_{X_1X_2}(x_1, x_2)$ and the joint distribution factors as $p_{X_1}(x_1)p_{X_2}(x_2)$ if we assume the source is *memoryless*—that it produces each symbol independently. The information content of the two symbols x_1 and x_2 is additive because

$$\begin{aligned} i(x_1, x_2) &= -\log\left(p_{X_1X_2}(x_1, x_2)\right) && (2.12)\\ &= -\log\left(p_{X_1}(x_1)p_{X_2}(x_2)\right) && (2.13)\\ &= -\log\left(p_{X_1}(x_1)\right) - \log\left(p_{X_2}(x_2)\right) && (2.14)\\ &= i(x_1) + i(x_2). && (2.15)\end{aligned}$$

In general, additivity is a desirable property for any information measure. We will return to the issue of additivity in many different contexts in this book (especially in Chapter 13).

The expected information content of the information source is

$$\sum_x p_X(x)i(x) = -\sum_x p_X(x)\log\left(p_X(x)\right). \tag{2.16}$$

The above quantity is so important in information theory that we give it a name: the *entropy* of the information source. The reason for its importance is that the entropy and variations of it appear as the answer to many questions in information theory. For example, in the above coding scheme, the expected length of a codeword is the entropy of the information source because

$$\begin{aligned} &-\frac{1}{2}\log\frac{1}{2} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8}\\ &\qquad = \frac{1}{2}(1) + \frac{1}{8}(3) + \frac{1}{4}(2) + \frac{1}{8}(3) && (2.17)\\ &\qquad = \frac{7}{4}. && (2.18)\end{aligned}$$

It is no coincidence that we chose the particular coding scheme in (2.6). The effectiveness of the scheme in this example is related to the structure of the information source—the number of symbols is a power of two and the probability of each symbol is the reciprocal of a power of two.

2.1.3 Shannon's Source Coding Theorem

The next question to ask is whether there is any other scheme that can achieve a better compression rate than the scheme in (2.6). This question is the one that Shannon asked in his first coding theorem. To answer this question, we consider a more general information source and introduce a notion of Shannon, the idea of the *set of typical sequences*.

We can represent a more general information source with a random variable X whose realizations x are *letters* in an *alphabet* $\mathcal{X}$. Let $p_X(x)$ be the probability mass function associated with random variable X, so that the probability of realization x is $p_X(x)$. Let $H(X)$ denote the entropy of the information source:

$$H(X) \equiv -\sum_{x\in\mathcal{X}} p_X(x)\log\left(p_X(x)\right). \tag{2.19}$$

The entropy $H(X)$ is also the entropy of the random variable X. Another way of writing it is $H(p)$, but we use the more common notation $H(X)$ throughout this book.

The information content $i(X)$ of random variable X is

$$i(X) \equiv -\log\left(p_X(X)\right), \tag{2.20}$$

and is itself a random variable. There is nothing wrong mathematically here with having random variable X as the argument to the density function p_X, though this expression may seem self-referential at a first glance. This way of thinking turns out to be useful later. Again, the expected information content of X is equal to the entropy:

$$\mathbb{E}_X\left\{-\log\left(p_X(X)\right)\right\} = H(X). \tag{2.21}$$

Exercise 2.1.1 *Show that the entropy of a uniform random variable is equal to* $\log|\mathcal{X}|$*, where* $|\mathcal{X}|$ *is the size of the variable's alphabet.*

We now turn to source coding the above information source. We *could* associate a binary codeword for each symbol x as we did in the scheme in (2.6). But this scheme may lose some efficiency if the size of our alphabet is not a power of two or if the probabilities are not a reciprocal of a power of two as they are in our nice example. Shannon's breakthrough idea was to let the source emit a large number of realizations and then code the emitted data as a large block, instead of coding each symbol as the above example does. This technique is called *block coding.* Shannon's other insight was to allow for a slight error in the compression scheme, but to show that this error vanishes as the block size becomes arbitrarily large. To make the block coding scheme more clear, Shannon suggests to let the source emit the following sequence:

$$x^n \equiv x_1 x_2 \cdots x_n, \tag{2.22}$$

where n is a large number that denotes the size of the block of emitted data and x_i, for all $i = 1, \ldots, n$, denotes the ith emitted symbol. Let X^n denote the random variable associated with the sequence x^n, and let X_i be the random variable for the ith symbol x_i. Figure 2.1 depicts Shannon's idea for a classical source code.

An important assumption regarding this information source is that it is independent and identically distributed (i.i.d.). The i.i.d. assumption means that each random variable X_i has the same distribution as random variable X, and we use the index i merely to track to which symbol x_i the random variable X_i corresponds. Under the i.i.d. assumption, the probability of any given emitted sequence x^n factors as

$$\begin{aligned} p_{X^n}(x^n) &= p_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) && (2.23)\\ &= p_{X_1}(x_1) p_{X_2}(x_2) \cdots p_{X_n}(x_n) && (2.24)\\ &= p_X(x_1) p_X(x_2) \cdots p_X(x_n) && (2.25)\\ &= \prod_{i=1}^{n} p_X(x_i). && (2.26) \end{aligned}$$

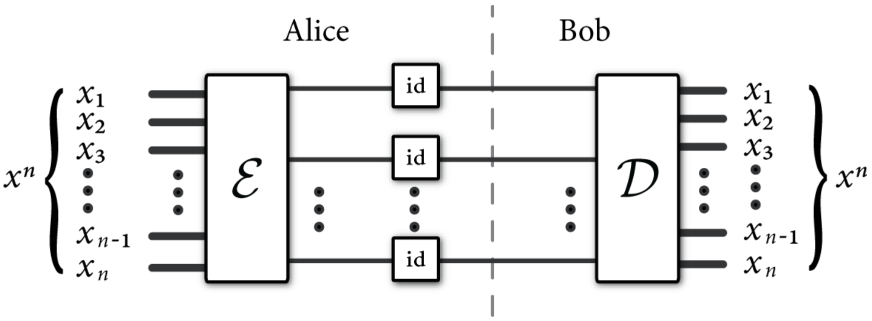

Figure 2.1: This figure depicts Shannon's idea for a classical source code. The information source emits a long sequence x^n to Alice. She encodes this sequence as a block with an encoder $\mathcal{E}$ and produces a codeword whose length is less than that of the original sequence x^n (indicated by fewer lines coming out of the encoder $\mathcal{E}$). She transmits the codeword over noiseless bit channels (each indicated by "id" which stands for the identity bit channel) and Bob receives it. Bob decodes the transmitted codeword with a decoder $\mathcal{D}$ and produces the original sequence that Alice transmitted, only if their chosen code is good, in the sense that the code has a small probability of error.

The above rule from probability theory results in a remarkable simplification of the mathematics. Suppose that we now label the letters in the alphabet $\mathcal{X}$ as $a_1, \ldots, a_{|\mathcal{X}|}$ in order to distinguish the letters from the realizations. Let $N(a_i|x^n)$ denote the number of occurrences of the letter a_i in the sequence x^n (where $i = 1, \ldots, |\mathcal{X}|$). As an example, consider the sequence in (2.9). The quantities $N(a_i|x^n)$ for this example are

$$N(a|x^n) = 5, \tag{2.27}$$

$$N(b|x^n) = 1, \tag{2.28}$$

$$N(c|x^n) = 4, \tag{2.29}$$

$$N(d|x^n) = 1. \tag{2.30}$$

We can rewrite the result in (2.26) as

$$p_{X^n}(x^n) = \prod_{i=1}^{n} p_X(x_i) = \prod_{i=1}^{|\mathcal{X}|} p_X(a_i)^{N(a_i|x^n)}. \tag{2.31}$$

Keep in mind that we are allowing the length n of the emitted sequence to be extremely large, so that it is much larger than the alphabet size $|\mathcal{X}|$:

$$n \gg |\mathcal{X}|\,. \tag{2.32}$$

The formula on the right in (2.31) is much simpler than the formula in (2.26) because it has fewer iterations of multiplications. There is a sense in which the i.i.d. assumption allows us to permute the sequence x^n as

$$x^n \to \underbrace{a_1 \cdots a_1}_{N(a_1|x^n)} \underbrace{a_2 \cdots a_2}_{N(a_2|x^n)} \cdots \underbrace{a_{|\mathcal{X}|} \cdots a_{|\mathcal{X}|}}_{N(a_{|\mathcal{X}|}|x^n)}, \tag{2.33}$$

because the probability calculation is invariant under this permutation. We introduce the above way of thinking right now because it turns out to be useful later when we develop some ideas in quantum Shannon theory (specifically in Section 14.9). Thus, the formula on the right in (2.31) characterizes the probability of any given sequence x^n.

The above discussion applies to a particular sequence x^n that the information source emits. Now, we would like to analyze the behavior of a *random sequence* X^n that the source emits, and this distinction between the realization x^n and the random variable X^n is important. In particular, let us consider the sample average of the information content of the random sequence X^n (divide the information content of X^n by n to get the sample average):

$$-\frac{1}{n}\log\left(p_{X^n}(X^n)\right). \tag{2.34}$$

It may seem strange at first glance that X^n, the argument of the probability mass function p_{X^n} is itself a random variable, but this type of expression is perfectly well defined mathematically. (This self-referencing type of expression is similar to (2.20), which we used to calculate the entropy.) For reasons that will become clear shortly, we call the above quantity the *sample entropy* of the random sequence X^n.

Suppose now that we use the function $N(a_i|\bullet)$ to calculate the number of appearances of the letter a_i in the random sequence X^n. We write the desired quantity as $N(a_i|X^n)$ and note that it is also a random variable, whose random nature derives from that of X^n. We can reduce the expression in (2.34) to the following one with some algebra and the result in (2.31):

$$-\frac{1}{n}\log\left(p_{X^n}(X^n)\right) = -\frac{1}{n}\log\left(\prod_{i=1}^{|\mathcal{X}|} p_X(a_i)^{N(a_i|X^n)}\right) \tag{2.35}$$

$$= -\frac{1}{n}\sum_{i=1}^{|\mathcal{X}|}\log\left(p_X(a_i)^{N(a_i|X^n)}\right) \tag{2.36}$$

$$= -\sum_{i=1}^{|\mathcal{X}|}\frac{N(a_i|X^n)}{n}\log\left(p_X(a_i)\right). \tag{2.37}$$

We stress again that the above quantity is random.

Is there any way that we can determine the behavior of the above sample entropy when n becomes large? Probability theory gives us a way. The expression $N(a_i|X^n)/n$ represents an empirical distribution for the letters a_i in the alphabet $\mathcal{X}$. As n becomes large, one form of the law of large numbers states that it is overwhelmingly likely that a random sequence has its empirical distribution $N(a_i|X^n)/n$ close to the true distribution $p_X(a_i)$, and conversely, it is highly unlikely that a random sequence does not satisfy this property. Thus, a random emitted sequence X^n is highly likely to satisfy the following condition for all $\delta > 0$ as n

becomes large:

$$\lim_{n\to\infty}\Pr\left\{\left|-\frac{1}{n}\log\left(p_{X^n}(X^n)\right)-\sum_{i=1}^{|\mathcal{X}|}p_X(a_i)\log\left(\frac{1}{p_X(a_i)}\right)\right|\leq\delta\right\}=1. \tag{2.38}$$

The quantity $-\sum_{i=1}^{|\mathcal{X}|}p_X(a_i)\log\left(p_X(a_i)\right)$ is none other than the entropy $H(X)$ so that the above expression is equivalent to the following one for all $\delta>0$:

$$\lim_{n\to\infty}\Pr\left\{\left|-\frac{1}{n}\log\left(p_{X^n}(X^n)\right)-H(X)\right|\leq\delta\right\}=1. \tag{2.39}$$

Another way of stating this property is as follows:

> It is highly likely that the information source emits a sequence whose sample entropy is close to the true entropy, and conversely, it is highly unlikely that the information source emits a sequence that does not satisfy this property.[2]

Now we consider a particular realization x^n of the random sequence X^n. We name a particular sequence x^n a *typical sequence* if its sample entropy is close to the true entropy $H(X)$ and the set of all typical sequences is the *typical set*. Fortunately for data compression, the set of typical sequences is not too large. In Chapter 14 on typical sequences, we prove that the size of this set is much smaller than the set of all sequences. We accept it for now (and prove later) that the size of the typical set is $\approx 2^{nH(X)}$, whereas the size of the set of all sequences is equal to $|\mathcal{X}|^n$. We can rewrite the size of the set of all sequences as

$$|\mathcal{X}|^n=2^{n\log|\mathcal{X}|}. \tag{2.40}$$

Comparing the size of the typical set to the size of the set of all sequences, the typical set is exponentially smaller than the set of all sequences whenever the random variable is not equal to the uniform random variable. Figure 2.2 illustrates this concept.We summarize these two crucial properties of the typical set and give another that we prove later:

Property 2.1.1 (Unit Probability) *The probability that an emitted sequence is typical approaches one as n becomes large. Another way of stating this property is that the typical set has almost all of the probability.*

Property 2.1.2 (Exponentially Smaller Cardinality) *The size of the typical set is* $\approx 2^{nH(X)}$ *and is exponentially smaller than the size* $2^{n\log|\mathcal{X}|}$ *of the set of all sequences whenever random variable* X *is not uniform.*

[2] Do not fall into the trap of thinking “The possible sequences that the source emits are typical sequences.” That line of reasoning is quantitatively far from the truth. In fact, what we can show is much different because the set of typical sequences is much smaller than the set of all possible sequences.

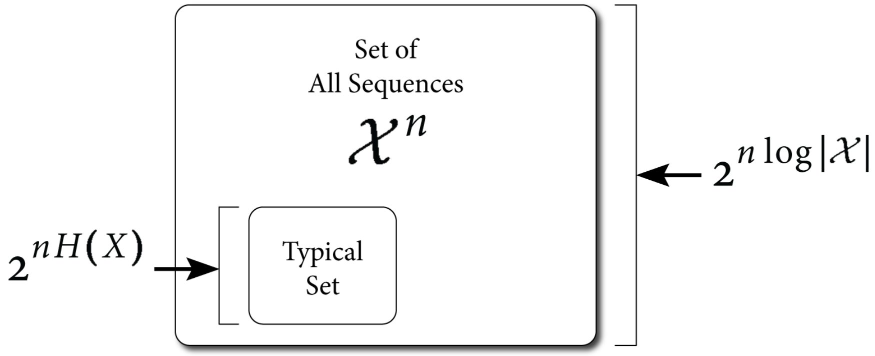

Figure 2.2: This figure indicates that the typical set is much smaller (exponentially smaller) than the set of all sequences. The typical set is roughly the same size as the set of all sequences only when the entropy $H(X)$ of the random variable X is equal to $\log|\mathcal{X}|$—implying that the distribution of random variable X is uniform.

Property 2.1.3 (Equipartition) *The probability of a particular typical sequence is roughly uniform* $\approx 2^{-nH(X)}$. *(The probability* $2^{-nH(X)}$ *is easy to calculate if we accept that the typical set has all of the probability, its size is* $2^{nH(X)}$*, and the distribution over typical sequences is uniform.)*

These three properties together are collectively known as the *asymptotic equipartition theorem*. The word "asymptotic" applies because the theorem exploits the asymptotic limit when n is large and the word "equipartition" refers to the third property above.

With the above notions of a typical set under our belt, a strategy for compressing information should now be clear. The strategy is to compress only the typical sequences that the source emits. We simply need to establish a one-to-one encoding function that maps from the set of typical sequences (size $2^{nH(X)}$) to the set of all binary strings of length $nH(X)$ (this set also has size $2^{nH(X)}$). If the source emits an atypical sequence, we declare an error. This coding scheme is reliable in the asymptotic limit because the probability of an error event vanishes as n becomes large, due to the unit probability property in the asymptotic equipartition theorem. We measure the rate of this block coding scheme as follows:

$$\text{compression rate} \equiv \frac{\#\text{ of noiseless channel bits}}{\#\text{ of source symbols}}. \tag{2.41}$$

For the case of Shannon compression, the number of noiseless channel bits is equal to $nH(X)$ and the number of source symbols is equal to n. Thus, the compression rate is equal to the entropy $H(X)$.

One may then wonder whether this rate of data compression is the best that we can do—whether this rate is optimal (we could achieve a lower rate of compression if it were not optimal). In fact, the above rate is the optimal rate at which we can compress information, and this is the content of Shannon's data compression theorem. We hold off on a formal proof of optimality for now and delay it until we reach Chapter 18. We just mention for now that this data compression protocol gives an *operational interpretation* to the Shannon entropy $H(X)$ because it appears as the optimal rate of data compression.

The above discussion highlights the common approach in information theory for establishing a coding theorem. Proving a coding theorem has two parts—traditionally called the *direct coding theorem* and the *converse theorem*. First, we give a coding scheme that can achieve a given rate for an information-processing task. This first part includes a direct construction of a coding scheme, hence the name *direct coding theorem*. The statement of the direct coding theorem for the above task is as follows:

> If the rate of compression is greater than the entropy of the source, then there exists a coding scheme that can achieve lossless data compression in the sense that it is possible to make the probability of error for incorrectly decoding arbitrarily small.

The second task is to prove that the rate from the direct coding theorem is optimal—that we cannot do any better than the suggested rate. We traditionally call this part the converse theorem because it corresponds to the converse of the above statement:

> If there exists a coding scheme that can achieve lossless data compression with arbitrarily small probability of decoding error, then the rate of compression is greater than the entropy of the source.

The techniques used in proving each part of the coding theorem are completely different. For most coding theorems in information theory, we can prove the direct coding theorem by appealing to the ideas of typical sequences and large block sizes. That this technique gives a good coding scheme is directly related to the asymptotic equipartition properties that govern the behavior of random sequences of data as the length of the sequence becomes large. The proof of a converse theorem relies on information inequalities that give tight bounds on the entropic quantities appearing in the coding constructions. We spend some time with information inequalities in Chapter 10 to build up our ability to prove converse theorems.

Sometimes, in the course of proving a direct coding theorem, one may think to have found the optimal rate for a given information-processing task. Without a matching converse theorem, it is not generally clear that the suggested rate is optimal. So, always prove converse theorems!

2.2 Channel Capacity

The next issue that we overview is the transmission of information over a noisy classical channel. We begin with a standard example—transmitting a single bit of information over a noisy bit-flip channel.

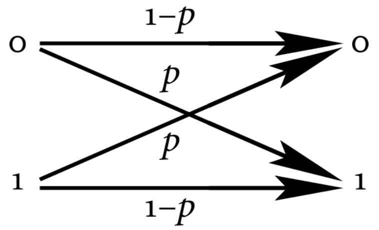

Figure 2.3: This figure depicts the action of the bit-flip channel. It preserves the input bit with probability $1-p$ and flips it with probability p.

2.2.1 An Example of an Error Correction Code

We again have our protagonists, Alice and Bob, as respective sender and receiver. This time, however, we assume that a noisy classical channel connects them, so that information transfer is not reliable. Alice and Bob realize that a noisy channel is not as expensive as a noiseless one, but it still is expensive for them to use. For this reason, they would like to maximize the amount of information that Alice can communicate reliably to Bob, where reliable communication implies that there is a negligible probability of error when transmitting this information.

The simplest example of a noisy classical channel is a bit-flip channel, with the technical name *binary symmetric channel*. This channel flips the input bit with probability p and leaves it unchanged with probability $1-p$. Figure 2.3 depicts the action of the bit-flip channel. Alice and Bob are allowed to use the channel multiple times, and in so doing, we assume that the channel behaves independently from one use to the next and behaves in the same random way as described above. For this reason, we describe the multiple uses of the channel as i.i.d. channels. This assumption will be helpful when we go to the asymptotic regime of a large number of uses of the channel.

Suppose that Alice and Bob just use the channel as is—Alice just sends plain bits to Bob. This scheme works reliably only if the probability of bit-flip error vanishes. So, Alice and Bob could invest their best efforts into engineering the physical channel to make it reliable. But, generally, it is not possible to engineer a classical channel this way for physical or logistical reasons. For example, Alice and Bob may only have local computers at their ends and may not have access to the physical channel because the telephone company may control the channel.

Alice and Bob can employ a "systems engineering" solution to this problem rather than an engineering of the physical channel. They can redundantly encode information in a way such that Bob can have a higher probability of determining what Alice is sending, effectively reducing the level of noise on the channel. A simple example of this systems engineering solution is the three-bit majority vote code. Alice and Bob employ the following encoding:

$$0 \to 000, \qquad 1 \to 111, \tag{2.42}$$

where both "000" and "111" are *codewords*. Alice transmits the codeword "000" with three

Channel Output	**Probability**
000	$(1-p)^3$
001, 010, 100	$p(1-p)^2$
011, 110, 101	$p^2(1-p)$
111	p^3

Table 2.1: The first column gives the eight possible outputs of the noisy bit-flip channel when Alice encodes a "0" with the majority vote code. The second column gives the corresponding probability of Bob receiving the particular outputs.

independent uses of the noisy channel if she really wants to communicate a "0" to Bob and she transmits the codeword "111" if she wants to send a "1" to him. The *physical* or *channel* bits are the actual bits that she transmits over the noisy channel, and the *logical* or *information* bits are those that she intends for Bob to receive. In our example, "0" is a logical bit and "000" corresponds to the physical bits.

The rate of this scheme is 1/3 because it encodes one information bit. The term "rate" is perhaps a misnomer for coding scenarios that do not involve sending bits in a time sequence over a channel. We may just as well use the majority vote code to store one bit in a memory device that may be unreliable. Perhaps a more universal term is *efficiency*. Nevertheless, we follow convention and use the term *rate* throughout this book.

Of course, the noisy bit-flip channel does not always transmit these codewords without error. So how does Bob decode in the case of error? He simply takes a *majority vote* to determine the transmitted message—he decodes as "0" if the number of zeros in the codeword he receives is greater than the number of ones.

We now analyze the performance of this simple "systems engineering" solution. Table 2.1 enumerates the probability of receiving every possible sequence of three bits, assuming that Alice transmits a "0" by encoding it as "000." The probability of no error is $(1-p)^3$, the probability of a single-bit error is $3p(1-p)^2$, the probability of a double-bit error is $3p^2(1-p)$, and the probability of a total failure is p^3. The majority vote solution can "correct" for no error and it corrects for all single-bit errors, but it has no ability to correct for double-bit and triple-bit errors. In fact, it actually incorrectly decodes these latter two scenarios by "correcting" "011", "110", or "101" to "111" and decoding "111" as a "1." Thus, these latter two outcomes are errors because the code has no ability to correct them. We can employ similar arguments as above to the case where Alice transmits a "1" to Bob with the majority vote code.

When does this majority vote scheme perform better than no coding at all? It is exactly when the probability of error with the majority vote code is less than p, the probability of error with no coding. Letting e denote the event that an error occurs, the probability of error is equal to the following quantity:

$$\Pr\{e\} = \Pr\{e|0\}\Pr\{0\} + \Pr\{e|1\}\Pr\{1\}. \tag{2.43}$$

Our analysis above suggests that the conditional probabilities $\Pr\{e|0\}$ and $\Pr\{e|1\}$ are equal for the majority vote code because of the symmetry in the noisy bit-flip channel. This result implies that the probability of error is

$$\Pr\{e\} = 3p^2(1-p) + p^3 \tag{2.44}$$

$$= 3p^2 - 2p^3, \tag{2.45}$$

because $\Pr\{0\}+\Pr\{1\} = 1$. We consider the following inequality to determine if the majority vote code reduces the probability of error:

$$3p^2 - 2p^3 < p. \tag{2.46}$$

This inequality simplifies as

$$0 < 2p^3 - 3p^2 + p \tag{2.47}$$

$$\therefore 0 < p\left(2p-1\right)\left(p-1\right). \tag{2.48}$$

The only values of p that satisfy the above inequality are $0 < p < 1/2$. Thus, the majority vote code reduces the probability of error only when $0 < p < 1/2$, i.e., when the noise on the channel is not too much. Too much noise has the effect of causing the codewords to flip too often, throwing off Bob's decoder.

The majority vote code gives a way for Alice and Bob to reduce the probability of error during their communication, but unfortunately, there is still a non-zero probability for the noisy channel to disrupt their communication. Is there any way that they can achieve reliable communication by reducing the probability of error to zero?

One simple approach to achieve this goal is to exploit the majority vote idea a second time. They can *concatenate* two instances of the majority vote code to produce a code with a larger number of physical bits. Concatenation consists of using one code as an "inner" code and another as an "outer" code. There is no real need for us to distinguish between the inner and outer code in this case because we use the same code for both the inner and outer code. The concatenation scheme for our case first encodes the message i, where $i \in \{0,1\}$, using the majority vote code. Let us label the codewords as follows:

$$\bar{0} \equiv 000, \qquad \bar{1} \equiv 111. \tag{2.49}$$

For the second layer of the concatenation, we encode $\bar{0}$ and $\bar{1}$ with the majority vote code again:

$$\bar{0} \to \bar{0}\bar{0}\bar{0}, \qquad \bar{1} \to \bar{1}\bar{1}\bar{1}. \tag{2.50}$$

Thus, the overall encoding of the concatenated scheme is as follows:

$$0 \to 000\ 000\ 000, \qquad 1 \to 111\ 111\ 111. \tag{2.51}$$

The rate of the concatenated code is $1/9$ and smaller than the original rate of $1/3$. A simple application of the above performance analysis for the majority vote code shows that this concatenation scheme reduces the probability of error as follows:

$$3[\Pr\{e\}]^2 - 2[\Pr\{e\}]^3 = O(p^4). \tag{2.52}$$

The error probability $\Pr\{e\}$ is in (2.45) and $O(p^4)$ indicates that the leading order term of the left-hand side is the fourth power in p.

The concatenated scheme achieves a lower probability of error at the cost of using more physical bits in the code. Recall that our goal is to achieve reliable communication, where there is no probability of error. A first guess for achieving reliable communication is to continue concatenating. If we concatenate again, the probability of error reduces to $O(p^6)$, and the rate drops to $1/27$. We can continue indefinitely with concatenating to make the probability of error arbitrarily small and achieve reliable communication, but the problem is that the rate approaches zero as the probability of error becomes arbitrarily small.

The above example seems to show that there is a trade-off between the rate of the encoding scheme and the desired order of error probability. Is there a way that we can code information for a noisy channel while maintaining a good rate of communication?

2.2.2 Shannon's Channel Coding Theorem

Shannon's second breakthrough coding theorem provides an affirmative answer to the above question. This answer came as a complete shock to communication researchers in 1948. Furthermore, the techniques that Shannon used in demonstrating this fact were rarely used by engineers at the time. We give a broad overview of Shannon's main idea and techniques that he used to prove his second important theorem—the noisy channel coding theorem.

2.2.3 General Model for a Channel Code

We first generalize some of the ideas in the above example. We still have Alice trying to communicate with Bob, but this time, she wants to be able to transmit a larger set of messages with asymptotically perfect reliability, rather than merely sending "0" or "1." Suppose that she selects messages from a message set $[M]$ that consists of M messages:

$$[M] \equiv \{1, \ldots, M\}. \tag{2.53}$$

Suppose furthermore that Alice chooses a particular message m with uniform probability from the set $[M]$. This assumption of a uniform distribution for Alice's messages indicates that we do not really care much about the content of the actual message that she is transmitting. We just assume total ignorance of her message because we only really care about her ability to send any message reliably. The message set $[M]$ requires $\log(M)$ bits to represent it, where the logarithm is again base two. This number becomes important when we calculate the rate of a channel code.

The next aspect of the model that we need to generalize is the noisy channel that connects Alice to Bob. We used the bit-flip channel before, but this channel is not general enough for our purposes. A simple way to extend the channel model is to represent it as a conditional probability distribution involving an input random variable X and an output random variable Y:

$$\mathcal{N}: \qquad p_{Y|X}(y|x). \tag{2.54}$$

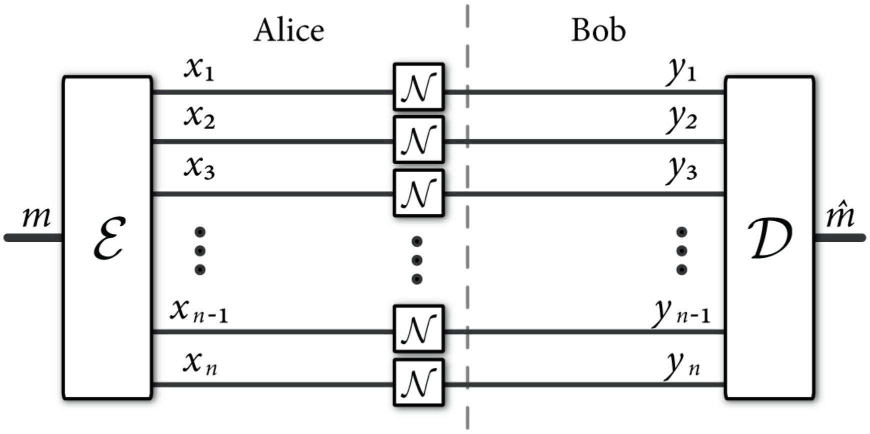

Figure 2.4: This figure depicts Shannon's idea for a classical channel code. Alice chooses a message m from a message set $[M] \equiv \{1, \ldots, M\}$. She encodes the message m with an encoding operation $\mathcal{E}$. This encoding operation assigns a codeword x^n to the message m and inputs the codeword x^n to a large number of i.i.d. uses of a noisy channel $\mathcal{N}$. The noisy channel randomly corrupts the codeword x^n to a sequence y^n. Bob receives the corrupted sequence y^n and performs a decoding operation $\mathcal{D}$ to estimate the codeword x^n. This estimate of the codeword x^n then produces an estimate $\hat{m}$ of the message that Alice transmitted. A reliable code has the property that Bob can decode each message $m \in [M]$ with a vanishing probability of error when the block length n becomes large.

We use the symbol $\mathcal{N}$ to represent this more general channel model. One assumption that we make about random variables X and Y is that they are discrete, but the respective sizes of their outcome sets do not have to match. The other assumption that we make concerning the noisy channel is that it is i.i.d. Let $X^n \equiv X_1X_2 \cdots X_n$ and $Y^n \equiv Y_1Y_2 \cdots Y_n$ be the random variables associated with respective sequences $x^n \equiv x_1x_2 \cdots x_n$ and $y^n \equiv y_1y_2 \cdots y_n$. If Alice inputs the sequence x^n to the n inputs of n respective uses of the noisy channel, a possible output sequence may be y^n. The i.i.d. assumption allows us to factor the conditional probability of the output sequence y^n:

$$p_{Y^n|X^n}(y^n|x^n) = p_{Y_1|X_1}(y_1|x_1)p_{Y_2|X_2}(y_2|x_2) \cdots p_{Y_n|X_n}(y_n|x_n) \tag{2.55}$$

$$= p_{Y|X}(y_1|x_1)p_{Y|X}(y_2|x_2) \cdots p_{Y|X}(y_n|x_n) \tag{2.56}$$

$$= \prod_{i=1}^{n} p_{Y|X}(y_i|x_i). \tag{2.57}$$

The technical name of this more general channel model is a *discrete memoryless channel.*

A coding scheme or *code* translates all of Alice's messages into codewords that can be input to n i.i.d. uses of the noisy channel. For example, suppose that Alice selects a message m to encode. We can write the codeword corresponding to message m as $x^n(m)$ because the input to the channel is some codeword that depends on m.

The last part of the model involves Bob receiving the corrupted codeword y^n over the channel and determining a potential codeword x^n with which it should be associated. We do not get into any details just yet for this last decoding part—imagine for now that it operates similarly to the majority vote code example. Figure 2.4 displays Shannon's model of communication that we have described.

We calculate the *rate* of a given coding scheme as follows:

$$\text{rate} \equiv \frac{\#\text{ of message bits}}{\#\text{ of channel uses}}. \tag{2.58}$$

In our model, the rate of a given coding scheme is

$$R = \frac{1}{n}\log(M), \tag{2.59}$$

where $\log(M)$ is the number of bits needed to represent any message in the message set $[M]$ and n is the number of channel uses. The *capacity* of a noisy channel is the highest rate at which it can communicate information reliably.

We also need a way to determine the performance of any given code. Here, we list several measures of performance. Let $\mathcal{C} \equiv \{x^n(m)\}_{m\in[M]}$ represent a code that Alice and Bob choose, where $x^n(m)$ denotes each codeword corresponding to the message m. Let $p_e(m,\mathcal{C})$ denote the probability of error when Alice transmits a message $m \in [M]$ using the code $\mathcal{C}$. We denote the average probability of error as

$$\bar{p}_e(\mathcal{C}) \equiv \frac{1}{M}\sum_{m=1}^{M} p_e(m,\mathcal{C}). \tag{2.60}$$

The maximal probability of error is

$$p_e^*(\mathcal{C}) \equiv \max_{m\in[M]} p_e(m,\mathcal{C}). \tag{2.61}$$

Our ultimate aim is to make the maximal probability of error $p_e^*(\mathcal{C})$ arbitrarily small, but the average probability of error $\bar{p}_e(\mathcal{C})$ is important in the analysis. These two performance measures are related—the average probability of error is small if the maximal probability of error is. Perhaps surprisingly, the maximal probability is small for at least half of the messages if the average probability of error is. We make this statement more quantitative in the following exercise.

Exercise 2.2.1 *Let $\varepsilon \in [0,1/2]$ and let $p_e(m,\mathcal{C})$ denote the probability of error when Alice transmits a message $m \in [M]$ using the code $\mathcal{C}$. Use Markov's inequality to prove that the following upper bound on the average probability of error:*

$$\frac{1}{M}\sum_{m} p_e(m,\mathcal{C}) \leq \varepsilon \tag{2.62}$$

implies the following upper bound for at least half of the messages m:

$$p_e(m,\mathcal{C}) \leq 2\varepsilon. \tag{2.63}$$

You may have wondered why we use the random sequence X^n to model the inputs to the channel. We have already stated that Alice's message is a uniform random variable, and the codewords in any coding scheme directly depend on the message to be sent. For example, in the majority vote code, the channel inputs are always "000" whenever the intended message is "0" and similarly for the channel inputs "111" and the message "1". So why is there a need to overcomplicate things by modeling the channel inputs as the random variable X^n when it seems like each codeword is a deterministic function of the intended message? We are not yet ready to answer this question but will return to it shortly.

We should also stress an important point before proceeding with Shannon's ingenious scheme for proving the existence of reliable codes for a noisy channel. In the above model, we described essentially two "layers of randomness":

1. The first layer of randomness is the uniform random variable associated with Alice's choice of a message.
2. The second layer of randomness is the noisy channel. The output of the channel is a random variable because we cannot always predict the output of the channel with certainty.

It is not possible to "play around" with these two layers of randomness. The random variable associated with Alice's message is fixed as a uniform random variable because we assume ignorance of Alice's message. The conditional probability distribution of the noisy channel is also fixed. We are assuming that Alice and Bob can learn the conditional probability distribution associated with the noisy channel by estimating it. Alternatively, we may assume that a third party has knowledge of the conditional probability distribution and informs Alice and Bob of it in some way. Regardless of how they obtain the knowledge of the distribution, we assume that they both know it and that it is fixed.

2.2.4 Proof Sketch of Shannon's Channel Coding Theorem

We are now ready to present an overview of Shannon's technique for proving the existence of a code that can achieve the capacity of a given noisy channel. Some of the methods that Shannon uses in his outline of a proof are similar to those in the first coding theorem. We again use the channel a large number of times so that the law of large numbers from probability theory comes into play and allow for a small probability of error that vanishes as the number of channel uses becomes large. If the notion of typical sequences is so important in the first coding theorem, we might suspect that it should be important in the noisy channel coding theorem as well. The typical set captures a certain notion of efficiency because it is a small set when compared to the set of all sequences, but it is the set that has almost all of the probability. Thus, we should expect this efficiency to come into play somehow in the channel coding theorem.

The aspect of Shannon's technique for proving the noisy channel coding theorem that is different from the other ideas in the first theorem is the idea of *random coding.* Shannon's technique adds a *third* layer of randomness to the model given above (recall that the first two are Alice's random message and the random nature of the noisy channel).

The third layer of randomness is to choose the codewords themselves in a random fashion according to a random variable X, where we choose each letter x_i of a given codeword x^n independently according to the distribution $p_X(x_i)$. It is for this reason that we model the channel inputs as a random variable. We can then write each codeword as a random variable $X^n(m)$. The probability distribution for choosing a particular codeword $x^n(m)$ is

$$\begin{aligned}\Pr\left\{X^n(m) = x^n(m)\right\} &= p_{X_1,X_2,\ldots,X_n}(x_1(m), x_2(m), \ldots, x_n(m)) && (2.64)\\ &= p_X(x_1(m))p_X(x_2(m))\cdots p_X(x_n(m)) && (2.65)\\ &= \prod_{i=1}^{n} p_X(x_i(m)). && (2.66)\end{aligned}$$

The important result to notice is that the probability for a given codeword factors because we choose the code in an i.i.d. fashion, and perhaps more importantly, the distribution of each codeword has no explicit dependence on the message m with which it is associated. That is, the probability distribution of the first codeword is exactly the same as the probability distribution of all of the other codewords. The code $\mathcal{C}$ itself becomes a random variable in this scheme for choosing a code randomly. We now let $\mathcal{C}$ refer to the random variable that represents a random code, and we let $\mathcal{C}_0$ represent any particular deterministic code. The probability of choosing a particular code $\mathcal{C}_0 = \left\{x^n(m)\right\}_{m\in[M]}$ is

$$p_{\mathcal{C}}(\mathcal{C}_0) = \prod_{m=1}^{M}\prod_{i=1}^{n} p_X(x_i(m)), \tag{2.67}$$

and this probability distribution again has no explicit dependence on each message m in the code $\mathcal{C}_0$.

Choosing the codewords in a random way allows for a dramatic simplification in the mathematical analysis of the probability of error. One of Shannon's breakthrough ideas was to analyze the *expectation* of the average probability of error, where the expectation is with respect to the random code $\mathcal{C}$, rather than analyzing the average probability of error itself. The expectation of the average probability of error is

$$\mathbb{E}_{\mathcal{C}}\left\{\bar{p}_e(\mathcal{C})\right\}. \tag{2.68}$$

This expectation is much simpler to analyze because of the random way that we choose the code. Consider that

$$\mathbb{E}_{\mathcal{C}}\left\{\bar{p}_e(\mathcal{C})\right\} = \mathbb{E}_{\mathcal{C}}\left\{\frac{1}{M}\sum_{m=1}^{M} p_e(m,\mathcal{C})\right\}. \tag{2.69}$$

Using linearity of the expectation, we can exchange the expectation with the sum so that

$$\mathbb{E}_{\mathcal{C}}\left\{\bar{p}_e(\mathcal{C})\right\} = \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\mathcal{C}}\left\{p_e(m,\mathcal{C})\right\}. \tag{2.70}$$

Now, the expectation of the probability of error for a particular message m does not actually depend on the message m because the distribution of each random codeword $X^n(m)$ does not explicitly depend on m. This line of reasoning leads to the dramatic simplification because $\mathbb{E}_{\mathcal{C}}\left\{p_e(m,\mathcal{C})\right\}$ is then the same for all messages. So we can then say that

$$\mathbb{E}_{\mathcal{C}}\left\{p_e(m,\mathcal{C})\right\} = \mathbb{E}_{\mathcal{C}}\left\{p_e(1,\mathcal{C})\right\}. \tag{2.71}$$

(We could have equivalently chosen any message instead of the first.) We then have that

$$\mathbb{E}_{\mathcal{C}}\left\{\bar{p}_e(\mathcal{C})\right\} = \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\mathcal{C}}\left\{p_e(1,\mathcal{C})\right\} \tag{2.72}$$

$$= \mathbb{E}_{\mathcal{C}}\left\{p_e(1,\mathcal{C})\right\}, \tag{2.73}$$

where the last step follows because the quantity $\mathbb{E}_{\mathcal{C}}\left\{p_e(1,\mathcal{C})\right\}$ has no dependence on m. We now only have to determine the expectation of the probability of error for *one message* instead of determining the expectation of the average error probability of the whole set. This simplification follows because random coding results in the equality of these two quantities.

Shannon then determined a way to obtain a bound on the expectation of the average probability of error (we soon discuss this technique briefly) so that

$$\mathbb{E}_{\mathcal{C}}\left\{\bar{p}_e(\mathcal{C})\right\} \leq \varepsilon, \tag{2.74}$$

where ε is some number $\in (0,1)$ that we can make arbitrarily small by letting the block size n become arbitrarily large. If it is possible to obtain a bound on the expectation of the average probability of error, then surely there exists some deterministic code $\mathcal{C}_0$ whose average probability of error meets this same bound:

$$\bar{p}_e(\mathcal{C}_0) \leq \varepsilon. \tag{2.75}$$

If it were not so, then the original bound on the expectation would not be possible. This step is the *derandomization* step of Shannon's proof. Ultimately, we require a deterministic code with a high rate and arbitrarily small probability of error and this step shows the *existence* of such a code. The random coding technique is only useful for simplifying the mathematics of the proof.

The last step of the proof is the *expurgation* step. It is an application of the result of Exercise 2.2.1. Recall that our goal is to show the existence of a high-rate code that has low maximal probability of error. But so far we only have a bound on the average probability of error. In the expurgation step, we simply throw out the half of the codewords with the worst probability of error. Throwing out the worse half of the codewords reduces the number of messages by a factor of two, but only has a negligible impact on the rate of the code. Consider that the number of messages is 2^{nR} where R is the rate of the code. Thus, the number of messages is $2^{n\left(R-\frac{1}{n}\right)}$ after throwing out the worse half of the codewords, and the rate $R-\frac{1}{n}$ is asymptotically equal to the rate R. After throwing out the worse half of the

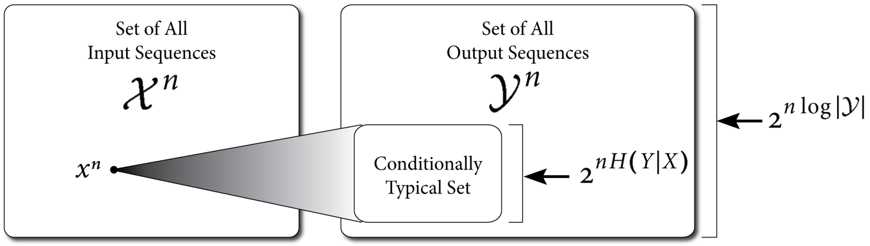

Figure 2.5: This figure depicts the notion of a conditionally typical set. Associated to every input sequence x^n is a conditionally typical set consisting of the likely output sequences. The size of this conditionally typical set is $\approx 2^{nH(Y|X)}$. It is exponentially smaller than the set of all output sequences whenever the conditional random variable is not uniform.

codewords, the result of Exercise 2.2.1 shows that the following bound then applies to the maximal probability of error:

$$p_e^*(\mathcal{C}_0) \leq 2\varepsilon. \tag{2.76}$$

This last expurgation step ends the analysis of the probability of error.

We now discuss the size of the code that Alice and Bob employ. Recall that the rate of the code is $R = \log(M)/n$. It is convenient to define the size M of the message set $[M]$ in terms of the rate R. When we do so, the size of the message set is

$$M = 2^{nR}. \tag{2.77}$$

What is peculiar about the message set size when defined this way is that it grows exponentially with the number of channel uses. But recall that any given code exploits n channel uses to send M messages. So when we take the limit as the number of channel uses tends to infinity, we are implying that there exists a sequence of codes whose messages set size is $M = 2^{nR}$ and number of channel uses is n. We are focused on keeping the rate of the code constant and use the limit $n \to \infty$ to make the probability of error vanish for a certain fixed rate R.

What is the maximal rate at which Alice can communicate to Bob reliably? We need to determine the number of distinguishable messages that Alice can reliably send to Bob, and we require the notion of *conditional typicality* to do so. Consider that Alice chooses codewords randomly according to random variable X with probability distribution $p_X(x)$. By the asymptotic equipartition theorem, it is highly likely that each of the codewords that Alice chooses is a typical sequence with sample entropy close to $H(X)$. In the coding scheme, Alice transmits a particular codeword x^n over the noisy channel and Bob receives a random sequence Y^n. The random sequence Y^n is a random variable that depends on x^n through the conditional probability distribution $p_{Y|X}(y|x)$. We would like a way to determine the number of possible output sequences that are likely to correspond to a particular input sequence x^n. A useful entropic quantity for this situation is the conditional entropy $H(Y|X)$, the technical details of which we leave for Chapter 10. For now, just think of this conditional entropy as measuring the uncertainty of a random variable Y when one already knows the

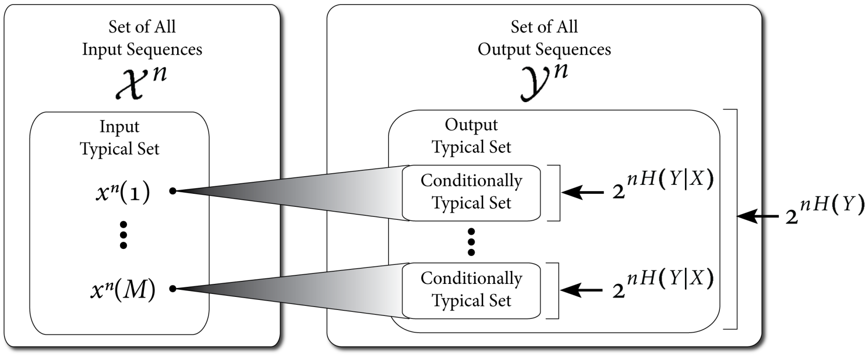

Figure 2.6: This figure depicts the packing argument that Shannon used. The channel induces a conditionally typical set corresponding to each codeword $x^n(i)$ where $i \in \{1, \ldots, M\}$. The size of each conditionally typical output set is $2^{nH(Y|X)}$. The size of the typical set of all output sequences is $2^{nH(Y)}$. These sizes suggest that we can divide the output typical set into M conditionally typical sets and be able to distinguish $M \approx 2^{nH(Y)}/2^{nH(Y|X)}$ messages without error.

value of the random variable X. The conditional entropy $H(Y|X)$ is always less than the entropy $H(Y)$ unless X and Y are independent. This inequality holds because knowledge of a correlated random variable X does not increase the uncertainty about Y. It turns out that there is a notion of conditional typicality (depicted in Figure 2.5), similar to the notion of typicality, and a similar asymptotic equipartition theorem holds for conditionally typical sequences (more details in Section 14.9). This theorem also has three important properties. For each input sequence x^n, there is a corresponding conditionally typical set with the following properties:

1. It has almost all of the probability—it is highly likely that a random channel output sequence is conditionally typical given a particular input sequence.

2. Its size is $\approx 2^{nH(Y|X)}$.

3. The probability of each conditionally typical sequence y^n, given knowledge of the input sequence x^n, is $\approx 2^{-nH(Y|X)}$.

If we disregard knowledge of the input sequence used to generate an output sequence, the probability distribution that generates the output sequences is

$$p_Y(y) = \sum_x p_{Y|X}(y|x) p_X(x). \tag{2.78}$$

We can think that this probability distribution is the one that generates all the possible output sequences. The likely output sequences are in an output typical set of size $2^{nH(Y)}$.

We are now in a position to describe the structure of a random code and the size of the message set. Alice generates 2^{nR} codewords according to the distribution $p_X(x)$ and suppose for now that Bob has knowledge of the code after Alice generates it. Suppose Alice sends one of the codewords over the channel. Bob is ignorant of the transmitted codeword, so from his point of view, the output sequences are generated according to the distribution $p_Y(y)$. Bob then employs typical sequence decoding. He first determines if the output sequence y^n is in the typical output set of size $2^{nH(Y)}$. If not, he declares an error. The probability of this type of error is small by the asymptotic equipartition theorem. If the output sequence y^n is in the output typical set, he uses his knowledge of the code to determine a conditionally typical set of size $2^{nH(Y|X)}$ to which the output sequence belongs. If he decodes an output sequence y^n to the wrong conditionally typical set, then an error occurs. This last type of error suggests how they might structure the code in order to prevent this type of error from happening. If they structure the code so that the output conditionally typical sets do not overlap too much, then Bob should be able to decode each output sequence y^n to a unique input sequence x^n with high probability. This line of reasoning suggests that they should divide the set of output typical sequences into M sets of conditionally typical output sets, each of size $2^{nH(Y|X)}$. Thus, if they set the number of messages $M = 2^{nR}$ as follows:

$$2^{nR} \approx \frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{n(H(Y)-H(Y|X))}, \tag{2.79}$$

then our intuition is that Bob should be able to decode correctly with high probability. Such an argument is a “packing” argument because it shows how to pack information into the space of all output sequences. Figure 2.6 gives a visual depiction of the packing argument. It turns out that this intuition is correct—Alice can reliably send information to Bob if the quantity $H(Y) - H(Y|X)$ bounds the rate R:

$$R < H(Y) - H(Y|X). \tag{2.80}$$

A rate less than $H(Y) - H(Y|X)$ ensures that we can make the expectation of the average probability of error as small as we would like. We then employ the derandomization and expurgation steps, discussed before, in order to show that there exists a code whose maximal probability of error vanishes as the number n of channel uses tends to infinity.

The entropic quantity $H(Y) - H(Y|X)$ deserves special attention because it is another important entropic quantity in information theory. It is the *mutual information* between random variables X and Y and we denote it as

$$I(X;Y) \equiv H(Y) - H(Y|X). \tag{2.81}$$

It is important because it arises as the limiting rate of reliable communication. We will discuss its properties in more detail throughout this book.

There is one final step that we can take to strengthen the above coding scheme. We mentioned before that there are three layers of randomness in the coding construction: Alice's uniform choice of a message, the noisy channel, and Shannon's random coding scheme. The first two layers of randomness we do not have control over. But we actually do have control over the last layer of randomness. Alice chooses the code according to the distribution $p_X(x)$. She can choose the code according to any distribution that she would like. If she chooses it according to $p_X(x)$, the resulting rate of the code is the mutual information $I(X;Y)$. We will prove later on that the mutual information $I(X;Y)$ is a concave function of the distribution $p_X(x)$ when the conditional distribution $p_{Y|X}(y|x)$ is fixed. Concavity implies that there is a distribution $p_X^*(x)$ that maximizes the mutual information. Thus, Alice should choose an optimal distribution $p_X^*(x)$ when she randomly generates the code, and this choice gives the largest possible rate of communication that they could have. This largest possible rate is the *capacity* of the channel and we denote it as

$$C(\mathcal{N}) \equiv \max_{p_X(x)} I(X;Y). \tag{2.82}$$

Our discussion here is just an overview of Shannon's channel capacity theorem. In Section 14.10, we give a full proof of this theorem after having developed some technical tools needed for a formal proof.

We clarify one more point. In the discussion of the operation of the code, we mentioned that Alice and Bob both have knowledge of the code. Well, how can Bob know the code if a noisy channel connects Alice to Bob? One solution to this problem is to assume that Alice and Bob have unbounded computation on their local ends. Thus, for a given code that uses the channel n times, they can both compute the above optimization problem and generate "test" codes randomly until they determine the best possible code to employ for n channel uses. They then both end up with the unique, best possible code for n uses of the given channel. This scheme might be impractical, but nevertheless, it provides a justification for both of them to have knowledge of the code that they use. Another solution to this problem is simply to allow them to meet before going their separate ways in order to coordinate on the choice of code.

We have said before that the capacity $C(\mathcal{N})$ is the maximal rate at which Alice and Bob can communicate. But in our discussion above, we did not prove optimality—we only proved a direct coding theorem for the channel capacity theorem. It took quite some time and effort to develop this elaborate coding procedure—along the way, we repeatedly invoked one of the powerful tools from probability theory, the law of large numbers. It perhaps seems intuitive that typical sequence coding and decoding should lead to optimal code constructions. Typical sequences exhibit some kind of asymptotic efficiency by being the most likely to occur, but in the general case, their cardinality is exponentially smaller than the set of all sequences. But is this intuition about typical sequence coding correct? Is it possible that some other scheme for coding might beat this elaborate scheme that Shannon devised? *Without a converse theorem that proves optimality, we would never know!* If you recall from our previous discussion in Section 2.1.3 about coding theorems, we stressed how important it is to prove a converse theorem that matches the rate that the direct coding theorem suggests is optimal. For now, we delay the proof of the converse theorem because the tools for proving it are much different from the tools we described in this section. For now, accept that the formula in (2.82) is indeed the optimal rate at which two parties can communicate and we will prove this result in a later chapter.

We end the description of Shannon's channel coding theorem by summarizing the statements of the direct coding theorem and the converse theorem. The statement of the direct coding theorem is as follows:

> If the rate of communication is less than the channel capacity, then it is possible for Alice to communicate reliably to Bob, in the sense that a sequence of codes exists whose maximal probability of error vanishes as the number of channel uses tends to infinity.

The statement of the converse theorem is as follows:

> If a reliable sequence of codes exists, then the rate of this sequence of codes is less than the channel capacity.

Another way of stating the converse proves to be useful later on:

> If the rate of a coding scheme is greater than the channel capacity, then a reliable code does not exist, in the sense that the error probability of the coding scheme is bounded away from zero.

2.3 Summary

A general communication scenario involves one sender and one receiver. In the classical setting, we discussed two information-processing tasks that they can perform. The first task was data compression or source coding, and we assumed that the sender and receiver are linked together by a noiseless classical bit channel that they can use a large number of times. We can think of this noiseless classical bit channel as a *noiseless dynamic resource* that the two parties share. The resource is dynamic because we assume that there is some physical medium through which the physical carrier of information travels in order to get from the sender to the receiver. It was our aim to count the number of times they would have to use the noiseless resource in order to send information reliably. The result of Shannon's source coding theorem is that the entropy gives the minimum rate at which they have to use the noiseless resource. The second task we discussed was channel coding and we assumed that the sender and receiver are linked together by a noisy classical channel that they can use a large number of times. This noisy classical channel is a *noisy dynamic resource* that they share. We can think of this information-processing task as a *simulation task*, where the goal is to simulate a noiseless dynamic resource by using a noisy dynamic resource in a redundant way. This redundancy is what allows Alice to communicate reliably to Bob, and reliable communication implies that they have effectively simulated a noiseless resource. We again had a resource count for this case, where we counted n as the number of times they use the noisy resource and nC is the number of noiseless bit channels they simulate (where C is the capacity of the channel). This notion of resource counting may not seem so important for the classical case, but it becomes much more important for the quantum case.

We now conclude our overview of Shannon's information theory. The main points to take home from this overview are the ideas that Shannon employed for constructing source and channel codes. We let the information source emit a large sequence of data, or similarly, we use the channel a large number of times so that we can invoke the law of large numbers from probability theory. The result is that we can show vanishing error for both schemes by taking a limit. In Chapter 14, we develop the theory of typical sequences in detail, proving many of the results taken for granted in this overview.

In hindsight, Shannon's methods for proving the two coding theorems are merely a *tour de force* for one idea from probability theory: the law of large numbers. Perhaps, this viewpoint undermines the contribution of Shannon, until we recall that no one else had even come close to devising these methods for data compression and channel coding. The theoretical development of Shannon is one of the most important contributions to modern science because his theorems determine the ultimate rate at which we can compress and communicate classical information.

Part II

The Quantum Theory

CHAPTER 3

The Noiseless Quantum Theory

The simplest quantum system is the physical quantum bit or *qubit*. The qubit is a two-level quantum system—example qubit systems are the spin of an electron, the polarization of a photon, or a two-level atom with a ground state and an excited state. We do not worry too much about physical implementations in this chapter, but instead focus on the mathematical postulates of the quantum theory and operations that we can perform on qubits. From qubits we progress to a study of physical *qudits*. Qudits are quantum systems that have d levels and are an important generalization of qubits. Again, we do not discuss physical realizations of qudits.

Noise can affect quantum systems, and we must understand methods of modeling noise in the quantum theory because our ultimate aim is to construct schemes for protecting quantum systems against the detrimental effects of noise. In Chapter 1, we remarked on the different types of noise that occur in nature. The first, and perhaps more easily comprehensible type of noise, is that which is due to our lack of information about a given scenario. We observe this type of noise in a casino, with every shuffle of cards or toss of dice. These events are random, and the random variables of probability theory model them because the outcomes are unpredictable. This noise is the same as that in all classical information-processing systems.

On the other hand, the quantum theory features a fundamentally different type of noise. Quantum noise is inherent in nature and is not due to our lack of information, but is due rather to nature itself. An example of this type of noise is the "Heisenberg noise" that results from the uncertainty principle. If we know the momentum of a given particle from performing a precise measurement of it, then we know absolutely nothing about its position—a measurement of its position gives a random result. Similarly, if we know the rectilinear polarization of a photon by precisely measuring it, then a future measurement of its diagonal polarization will give a random result. It is important to keep the distinction clear between these two types of noise.

We explore the postulates of the quantum theory in this chapter, by paying particular attention to qubits. These postulates apply to a closed quantum system that is isolated from everything else in the universe. We label this first chapter "The Noiseless Quantum

Theory" because closed quantum systems do not interact with their surroundings and are thus not subject to corruption and information loss. Interaction with surrounding systems can lead to loss of information in the sense of the classical noise that we described above. Closed quantum systems do undergo a certain type of quantum noise, such as that from the uncertainty principle and the act of measurement, because they are subject to the postulates of the quantum theory. The name "noiseless quantum theory" thus indicates the closed, ideal nature of the quantum systems discussed.

This chapter introduces the four postulates of the quantum theory. The mathematical tools of the quantum theory rely on the fundamentals of linear algebra—vectors and matrices of complex numbers. It may seem strange at first that we need to incorporate the machinery of linear algebra in order to describe a physical system in the quantum theory, but it turns out that this description uses the simplest set of mathematical tools to predict the phenomena that a quantum system exhibits. A hallmark of the quantum theory is that certain operations do not commute with one another, and matrices are the simplest mathematical objects that capture this idea of non-commutativity.

3.1 Overview

We first briefly overview how information is processed with quantum systems. This usually consists of three steps: state preparation, quantum operations, and measurement. State preparation is the initialization of a quantum system to some beginning state, depending on what operation we would like a quantum system to execute. There could be some classical control device that initializes the state of the quantum system. Observe that the input system for this step is a classical system, and the output system is quantum. After initializing the state of the quantum system, we perform some quantum operations that evolve its state. This stage is where we can take advantage of quantum effects for enhanced information-processing abilities. Both the input and output systems of this step are quantum. Finally, we need some way of reading out the result of the computation, and we can do so with a measurement. The input system for this step is quantum, and the output is classical. Figure 3.1 depicts all of these steps. In a quantum communication protocol, spatially separated parties may execute different parts of these steps, and we are interested in keeping track of the non-local resources needed to implement a communication protocol. Section 3.2 describes quantum states (and thus state preparation), Section 3.3 describes the noiseless evolution of quantum states, and Section 3.4 describes "read out" or measurement. For now, we assume that we can perform all of these steps perfectly and later chapters discuss how to incorporate the effects of noise.

3.2 Quantum Bits

The simplest quantum system is a two-state system: a physical qubit. Let $|0\rangle$ denote one possible state of the system. The left vertical bar and the right angle bracket indicate that we

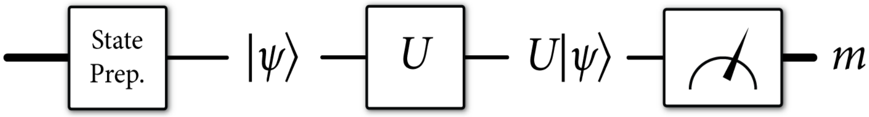

Figure 3.1: All of the steps in a typical noiseless quantum information processing protocol. A classical control (depicted by the thick black line on the left) initializes the state of a quantum system. The quantum system then evolves according to some unitary operation (described in Section 3.3). The final step is a measurement that reads out some classical data m from the quantum system.

are using the Dirac notation to represent this state. The Dirac notation has some advantages for performing calculations in the quantum theory, and we highlight some of these advantages as we progress through our development. Let $|1\rangle$ denote another possible state of the qubit. We can encode a classical bit or *cbit* into a qubit with the following mapping:

$$0 \to |0\rangle, \qquad 1 \to |1\rangle. \tag{3.1}$$

So far, nothing in our description above distinguishes a classical bit from a qubit, except for the funny vertical bar and angle bracket that we place around the bit values. However, the quantum theory predicts that the above states are not the only possible states of a qubit. Arbitrary *superpositions* (linear combinations) of the above two states are possible as well because the quantum theory is a linear theory. Suffice it to say that the linearity of the quantum theory results from the linearity of Schrödinger's equation that governs the evolution of quantum systems.[1] A general noiseless qubit can be in the following state:

$$|\psi\rangle \equiv \alpha|0\rangle + \beta|1\rangle, \tag{3.2}$$

where the coefficients α and β are arbitrary complex numbers with unit norm: $|\alpha|^2+|\beta|^2 = 1$. The coefficients α and β are *probability amplitudes*—they are not probabilities themselves, but they do allow us to calculate probabilities. The unit-norm constraint leads to the *Born rule* (the probabilistic interpretation) of the quantum theory, and we speak more on this constraint and probability amplitudes when we introduce the measurement postulate.

The possibility of superposition states indicates that we cannot represent the states $|0\rangle$ and $|1\rangle$ with the Boolean algebra of the respective classical bits 0 and 1 because Boolean algebra does not allow for superposition states. We instead require the mathematics of *linear algebra* to describe these states. It is beneficial at first to define a vector representation of the states $|0\rangle$ and $|1\rangle$:

$$|0\rangle \equiv \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad |1\rangle \equiv \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{3.3}$$

The $|0\rangle$ and $|1\rangle$ states are called "kets" in the language of the Dirac notation, and it is best at first to think of them merely as column vectors. The superposition state in (3.2) then has

[1] We will not present Schrödinger's equation in this book, but instead focus on a "quantum information" presentation of the quantum theory. Griffith's book on quantum mechanics introduces the quantum theory from the Schrödinger equation if you are interested (Griffiths, 1995).

a representation as the following two-dimensional vector:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}. \tag{3.4}$$

The representation of quantum states with vectors is helpful in understanding some of the mathematics that underpins the theory, but it turns out to be much more useful for our purposes to work directly with the Dirac notation. We give the vector representation for now, but later on, we will exclusively employ the Dirac notation.

The *Bloch sphere*, depicted in Figure 3.2, gives a valuable way to visualize a qubit. Consider any two qubits that are equivalent up to a differing global phase. For example, these two qubits could be

$$|\psi_0\rangle \equiv |\psi\rangle, \qquad |\psi_1\rangle \equiv e^{i\chi}|\psi\rangle, \tag{3.5}$$

where $0 \leq \chi < 2\pi$. These two qubits are physically equivalent because they give the same physical results when we measure them (more on this point when we introduce the measurement postulate in Section 3.4). Suppose that the probability amplitudes α and β have the following representations as complex numbers:

$$\alpha = r_0 e^{i\varphi_0}, \qquad \beta = r_1 e^{i\varphi_1}. \tag{3.6}$$

We can factor out the phase $e^{i\varphi_0}$ from both coefficients α and β, and we still have a state that is physically equivalent to the state in (3.2):

$$|\psi\rangle \equiv r_0|0\rangle + r_1 e^{i(\varphi_1 - \varphi_0)}|1\rangle, \tag{3.7}$$

where we redefine $|\psi\rangle$ to represent the state because of the equivalence mentioned in (3.5). Let $\varphi \equiv \varphi_1 - \varphi_0$, where $0 \leq \varphi < 2\pi$. Recall that the unit-norm constraint requires $|r_0|^2 + |r_1|^2 = 1$. We can thus parametrize the values of r_0 and r_1 in terms of one parameter θ so that

$$r_0 = \cos(\theta/2), \qquad r_1 = \sin(\theta/2). \tag{3.8}$$

The parameter θ varies between 0 and π. This range of θ and the factor of two give a unique representation of the qubit. One may think to have θ vary between 0 and 2π and omit the factor of two, but this parametrization would not uniquely characterize the qubit in terms of the parameters θ and φ. The parametrization in terms of θ and φ gives the Bloch sphere representation of the qubit in (3.2):

$$|\psi\rangle \equiv \cos(\theta/2)|0\rangle + \sin(\theta/2)e^{i\varphi}|1\rangle. \tag{3.9}$$

In linear algebra, column vectors are not the only type of vectors—row vectors are useful as well. Is there an equivalent of a row vector in Dirac notation? The Dirac notation provides an entity called a "bra," that has a representation as a row vector. The bras corresponding to the kets $|0\rangle$ and $|1\rangle$ are as follows:

$$\langle 0| \equiv \begin{bmatrix} 1 & 0 \end{bmatrix}, \qquad \langle 1| \equiv \begin{bmatrix} 0 & 1 \end{bmatrix}, \tag{3.10}$$

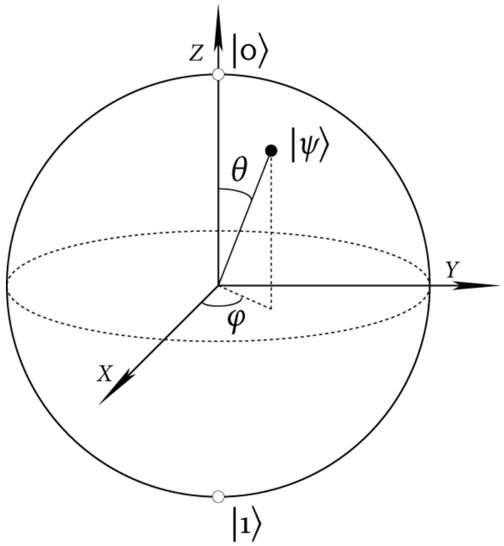

Figure 3.2: The Bloch sphere representation of a qubit. Any qubit $|\psi\rangle$ admits a representation in terms of two angles θ and φ where $0 \leq \theta \leq \pi$ and $0 \leq \varphi < 2\pi$. The state of any qubit in terms of these angles is $|\psi\rangle = \cos(\theta/2)|0\rangle + e^{i\varphi}\sin(\theta/2)|1\rangle$.

and are the matrix conjugate transpose of the kets $|0\rangle$ and $|1\rangle$:

$$\langle 0| = (|0\rangle)^\dagger, \qquad \langle 1| = (|1\rangle)^\dagger. \tag{3.11}$$

We require the conjugate transpose operation (as opposed to just the transpose) because the mathematical representation of a general quantum state can have complex entries.

The bras do not represent quantum states, but are helpful in calculating probability amplitudes. For our example qubit in (3.2), suppose that we would like to determine the probability amplitude that the state is $|0\rangle$. We can combine the state in (3.2) with the bra $\langle 0|$ as follows:

$$\begin{aligned} \langle 0||\psi\rangle &= \langle 0|\left(\alpha|0\rangle + \beta|1\rangle\right) && (3.12)\\ &= \alpha\langle 0||0\rangle + \beta\langle 0||1\rangle && (3.13)\\ &= \alpha \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \beta \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} && (3.14)\\ &= \alpha \cdot 1 + \beta \cdot 0 && (3.15)\\ &= \alpha. && (3.16) \end{aligned}$$

The above calculation may seem as if it is merely an exercise in linear algebra, with a "glorified" Dirac notation, but it is a standard calculation in the quantum theory. A quantity like $\langle 0||\psi\rangle$ occurs so often in the quantum theory that we abbreviate it as

$$\langle 0|\psi\rangle \equiv \langle 0||\psi\rangle, \tag{3.17}$$

and the above notation is known as a "braket."[2] The physical interpretation of the quantity $\langle 0|\psi\rangle$ is that it is the probability amplitude for being in the state $|0\rangle$, and likewise, the quantity $\langle 1|\psi\rangle$ is the probability amplitude for being in the state $|1\rangle$. We can also determine that the amplitude $\langle 1|0\rangle$ (for the state $|0\rangle$ to be in the state $|1\rangle$) and the amplitude $\langle 0|1\rangle$ are both equal to zero. These two states are *orthogonal states* because they have no overlap. The amplitudes $\langle 0|0\rangle$ and $\langle 1|1\rangle$ are both equal to one by following a similar calculation.

Our next task may seem like a frivolous exercise, but we would like to determine the amplitude for any state $|\psi\rangle$ to be in the state $|\psi\rangle$, i.e., to be itself. Following the above method, this amplitude is $\langle\psi|\psi\rangle$ and we calculate it as

$$\begin{aligned}\langle\psi|\psi\rangle &= \left(\langle 0|\alpha^* + \langle 1|\beta^*\right)\left(\alpha|0\rangle + \beta|1\rangle\right) &(3.18)\\ &= \alpha^*\alpha\,\langle 0|0\rangle + \beta^*\alpha\,\langle 1|0\rangle + \alpha^*\beta\,\langle 0|1\rangle + \beta^*\beta\,\langle 1|1\rangle &(3.19)\\ &= |\alpha|^2 + |\beta|^2 &(3.20)\\ &= 1, &(3.21)\end{aligned}$$

where we have used the orthogonality relations of $\langle 0|0\rangle$, $\langle 1|0\rangle$, $\langle 0|1\rangle$, and $\langle 1|1\rangle$, and the unit-norm constraint. We also write this in terms of the Euclidean norm of $|\psi\rangle$ as

$$\left\||\psi\rangle\right\|_2 \equiv \sqrt{\langle\psi|\psi\rangle} = 1. \qquad (3.22)$$

We come back to the unit-norm constraint in our discussion of quantum measurement, but for now, we have shown that any quantum state has a unit amplitude for being itself.

The states $|0\rangle$ and $|1\rangle$ are a particular basis for a qubit that we call the *computational basis*. The computational basis is the standard basis that we employ in quantum computation and communication, but other bases are important as well. Consider that the following two vectors form an orthonormal basis:

$$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}. \qquad (3.23)$$

The above alternate basis is so important in quantum information theory that we define a Dirac notation shorthand for it, and we can also define the basis in terms of the computational basis:

$$|+\rangle \equiv \frac{|0\rangle + |1\rangle}{\sqrt{2}}, \qquad |-\rangle \equiv \frac{|0\rangle - |1\rangle}{\sqrt{2}}. \qquad (3.24)$$

The common names for this alternate basis are the "+/−" basis, the Hadamard basis, or the diagonal basis. It is preferable for us to use the Dirac notation, but we are using the vector representation as an aid for now.

Exercise 3.2.1 *Determine the Bloch sphere angles θ and φ for the states $|+\rangle$ and $|-\rangle$.*

[2] It is for this (somewhat silly) reason that Dirac decided to use the names "bra" and "ket," because putting them together gives a "braket." The names in the notation may be silly, but the notation itself has persisted over time because this way of representing quantum states turns out to be useful. We will avoid the use of the terms "bra" and "ket" as much as we can, only resorting to these terms if necessary.

What is the amplitude that the state in (3.2) is in the state $|+\rangle$? What is the amplitude that it is in the state $|-\rangle$? These are questions to which the quantum theory provides simple answers. We employ the bra $\langle+|$ and calculate the amplitude $\langle+|\psi\rangle$ as

$$\begin{aligned}\langle+|\psi\rangle &= \langle+|\left(\alpha|0\rangle+\beta|1\rangle\right) && (3.25)\\ &= \alpha\,\langle+|0\rangle+\beta\,\langle+|1\rangle && (3.26)\\ &= \frac{\alpha+\beta}{\sqrt{2}}. && (3.27)\end{aligned}$$

The result follows by employing the definition in (3.24) and doing similar linear algebraic calculations as the example in (3.16). We can also calculate the amplitude $\langle-|\psi\rangle$ as

$$\langle-|\psi\rangle = \frac{\alpha-\beta}{\sqrt{2}}. \tag{3.28}$$

The above calculation follows from similar manipulations.

The $+/-$ basis is a *complete* orthonormal basis, meaning that we can represent any qubit state in terms of the two basis states $|+\rangle$ and $|-\rangle$. Indeed, the above probability amplitude calculations and the fact that the $+/-$ basis is complete imply that we can represent the qubit in (3.2) as the following superposition state:

$$|\psi\rangle = \left(\frac{\alpha+\beta}{\sqrt{2}}\right)|+\rangle+\left(\frac{\alpha-\beta}{\sqrt{2}}\right)|-\rangle\,. \tag{3.29}$$

The above representation is an alternate one if we would like to “see” the qubit state represented in the $+/-$ basis. We can substitute the equalities in (3.27) and (3.28) to represent the state $|\psi\rangle$ as

$$|\psi\rangle = \langle+|\psi\rangle\,|+\rangle+\langle-|\psi\rangle\,|-\rangle. \tag{3.30}$$

The amplitudes $\langle+|\psi\rangle$ and $\langle-|\psi\rangle$ are both scalar quantities so that the above quantity is equal to the following one:

$$|\psi\rangle = |+\rangle\langle+|\psi\rangle+|-\rangle\langle-|\psi\rangle. \tag{3.31}$$

The order of the multiplication in the terms $|+\rangle\langle+|\psi\rangle$ and $|-\rangle\langle-|\psi\rangle$ does not matter, i.e., the following equality holds:

$$|+\rangle\left(\langle+|\psi\rangle\right) = \left(|+\rangle\langle+|\right)|\psi\rangle, \tag{3.32}$$

and the same for $|-\rangle\,\langle-|\psi\rangle$. The quantity on the left is a ket multiplied by an amplitude, whereas the quantity on the right is a linear operator multiplying a ket, but linear algebra tells us that these two quantities are equal. The operators $|+\rangle\langle+|$ and $|-\rangle\langle-|$ are special operators—they are rank-one projection operators, meaning that they project onto a one-dimensional subspace. Using linearity, we have the following equality:

$$|\psi\rangle = \left(|+\rangle\langle+|+|-\rangle\langle-|\right)|\psi\rangle. \tag{3.33}$$

The above equation indicates a seemingly trivial, but important point—the operator $|+\rangle\langle+|+|-\rangle\langle-|$ is equal to the identity operator and we can write

$$I = |+\rangle\langle+| + |-\rangle\langle-|, \tag{3.34}$$

where I stands for the identity operator. This relation is known as the *completeness relation* or the *resolution of the identity.* Given any orthonormal basis, we can always construct a resolution of the identity by summing over the rank-one projection operators formed from each of the orthonormal basis states. For example, the computational basis states give another way to form a resolution of the identity operator:

$$I = |0\rangle\langle0| + |1\rangle\langle1|. \tag{3.35}$$

This simple trick provides a way to find the representation of a quantum state in any basis.

3.3 Reversible Evolution

Physical systems evolve as time progresses. The application of a magnetic field to an electron can change its spin and pulsing an atom with a laser can excite one of its electrons from a ground state to an excited state. These are only a couple of ways in which physical systems can change.

The Schrödinger equation governs the evolution of a closed quantum system. In this book, we will not even state the Schrödinger equation, but we will instead focus on an important implication of it. *The evolution of a closed quantum system is reversible if we do not learn anything about the state of the system (that is, if we do not measure it).* Reversibility implies that we can determine the input state of an evolution given the output state and knowledge of the evolution. An example of a single-qubit reversible operation is a NOT gate:

$$|0\rangle \to |1\rangle, \qquad |1\rangle \to |0\rangle. \tag{3.36}$$

In the classical world, we would say that the NOT gate merely flips the value of the input classical bit. In the quantum world, the NOT gate flips the basis states $|0\rangle$ and $|1\rangle$. The NOT gate is reversible because we can simply apply the NOT gate again to recover the original input state—the NOT gate is its own inverse.

In general, a closed quantum system evolves according to a unitary operator U. Unitary evolution implies reversibility because a unitary operator always possesses an inverse—its inverse is merely $U^\dagger$, the conjugate transpose. This property gives the relations:

$$U^\dagger U = UU^\dagger = I. \tag{3.37}$$

The unitary property also ensures that evolution preserves the unit-norm constraint (an important requirement for a physical state that we discuss in Section 3.4). Consider applying the unitary operator U to the example qubit state in (3.2): $U|\psi\rangle$. Figure 3.3 depicts a quantum circuit diagram for unitary evolution.

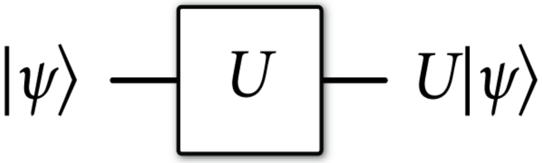

Figure 3.3: A quantum circuit diagram that depicts the evolution of a quantum state $|\psi\rangle$ according to a unitary operator U.

The bra that is dual to the above state is $\langle\psi|U^\dagger$ (we again apply the conjugate transpose operation to get the bra). We showed in (3.18)–(3.21) that every quantum state should have a unit amplitude for being itself. This relation holds for the state $U|\psi\rangle$ because the operator U is unitary:

$$\langle\psi|U^\dagger U|\psi\rangle = \langle\psi|I|\psi\rangle = \langle\psi|\psi\rangle = 1. \tag{3.38}$$

The assumption that a vector always has a unit amplitude for being itself is one of the crucial assumptions of the quantum theory, and the above reasoning demonstrates that unitary evolution complements this assumption.

Exercise 3.3.1 *A linear operator T is norm preserving if $\|T|\psi\rangle\|_2 = \||\psi\rangle\|_2$ holds for all quantum states $|\psi\rangle$ (unit vectors), where the Euclidean norm is defined in* (3.22)*. Prove that an operator T is unitary if and only if it is norm preserving. Hint: For showing the "only-if" part, consider using the polarization identity:*

$$\langle\psi|\phi\rangle = \frac{1}{4}\left(\||\psi\rangle + |\phi\rangle\|_2^2 - \||\psi\rangle - |\phi\rangle\|_2^2 + i\||\psi\rangle + i|\phi\rangle\|_2^2 - i\||\psi\rangle - i|\phi\rangle\|_2^2\right). \tag{3.39}$$

3.3.1 Matrix Representations of Operators

We now explore some properties of the NOT gate. Let X denote the operator corresponding to a NOT gate. The action of X on the computational basis states is as follows:

$$X|i\rangle = |i \oplus 1\rangle, \tag{3.40}$$

where $i = \{0, 1\}$ and $\oplus$ denotes binary addition. Suppose the NOT gate acts on a superposition state:

$$X(\alpha|0\rangle + \beta|1\rangle). \tag{3.41}$$

By linearity of the quantum theory, the X operator distributes so that the above expression is equal to the following one:

$$\alpha X|0\rangle + \beta X|1\rangle = \alpha|1\rangle + \beta|0\rangle. \tag{3.42}$$

Indeed, the NOT gate X merely flips the basis states of any quantum state when represented in the computational basis.

We can determine a *matrix representation* for the operator X by using the bras $\langle 0|$ and $\langle 1|$. Consider the relations in (3.40). Let us combine the relations with the bra $\langle 0|$:

$$\langle 0|X|0\rangle = \langle 0|1\rangle = 0, \qquad \langle 0|X|1\rangle = \langle 0|0\rangle = 1. \tag{3.43}$$

Likewise, we can combine with the bra $\langle 1|$:

$$\langle 1|X|0\rangle = \langle 1|1\rangle = 1, \qquad \langle 1|X|1\rangle = \langle 1|0\rangle = 0. \tag{3.44}$$

We can place these entries in a matrix to give a matrix representation of the operator X:

$$\begin{bmatrix} \langle 0|X|0\rangle & \langle 0|X|1\rangle \\ \langle 1|X|0\rangle & \langle 1|X|1\rangle \end{bmatrix}, \tag{3.45}$$

where we order the rows according to the bras and order the columns according to the kets. We then say that

$$X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \tag{3.46}$$

and adopt the convention that the symbol X refers to both the operator X and its matrix representation (this is an abuse of notation, but it should be clear from the context when X refers to an operator and when it refers to the matrix representation of the operator).

Let us now observe some uniquely quantum behavior. We would like to consider the action of the NOT operator X on the $+/-$ basis. First, let us consider what happens if we operate on the $|+\rangle$ state with the X operator. Recall that the state $|+\rangle = (|0\rangle + |1\rangle)/\sqrt{2}$ so that

$$X|+\rangle = X\left(\frac{|0\rangle + |1\rangle}{\sqrt{2}}\right) = \frac{X|0\rangle + X|1\rangle}{\sqrt{2}} = \frac{|1\rangle + |0\rangle}{\sqrt{2}} = |+\rangle. \tag{3.47}$$

The above development shows that the state $|+\rangle$ is a special state with respect to the NOT operator X—it is an *eigenstate* of X with *eigenvalue* one. An eigenstate of an operator is one that is invariant under the action of the operator. The coefficient in front of the eigenstate is the *eigenvalue* corresponding to the eigenstate. Under a unitary evolution, the coefficient in front of the eigenstate is just a complex phase, but this global phase has no effect on the observations resulting from a measurement of the state because two quantum states are equivalent up to a differing global phase.

Now, let us consider the action of the NOT operator X on the state $|-\rangle$. Recall that $|-\rangle = (|0\rangle - |1\rangle)/\sqrt{2}$. Calculating similarly, we get that

$$X|-\rangle = X\left(\frac{|0\rangle - |1\rangle}{\sqrt{2}}\right) = \frac{X|0\rangle - X|1\rangle}{\sqrt{2}} = \frac{|1\rangle - |0\rangle}{\sqrt{2}} = -|-\rangle. \tag{3.48}$$

So the state $|-\rangle$ is also an eigenstate of the operator X, but its eigenvalue is -1.

We can find a matrix representation of the X operator in the $+/-$ basis as well:

$$\begin{bmatrix} \langle +|X|+\rangle & \langle +|X\,|-\rangle \\ \langle -|X|+\rangle & \langle -|X|-\rangle \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}. \tag{3.49}$$

This representation demonstrates that the X operator is diagonal with respect to the $+/-$ basis, and therefore, the $+/-$ basis is an *eigenbasis* for the X operator. It is always handy to know the eigenbasis of a unitary operator U because this eigenbasis gives the states that are invariant under an evolution according to U.

Let Z denote the operator that flips states in the $+/-$ basis:

$$Z|+\rangle \rightarrow |-\rangle, \qquad Z|-\rangle \rightarrow |+\rangle. \tag{3.50}$$

Using an analysis similar to that which we did for the X operator, we can find a matrix representation of the Z operator in the $+/-$ basis:

$$\begin{bmatrix} \langle+|Z|+\rangle & \langle+|Z\,|-\rangle \\ \langle-|Z|+\rangle & \langle-|Z|-\rangle \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \tag{3.51}$$

Interestingly, the matrix representation for the Z operator in the $+/-$ basis is the same as that for the X operator in the computational basis. For this reason, we call the Z operator the *phase-flip* operator.[3]

We expect the following steps to hold because the quantum theory is a linear theory:

$$Z\left(\frac{|+\rangle+|-\rangle}{\sqrt{2}}\right) = \frac{Z|+\rangle+Z|-\rangle}{\sqrt{2}} = \frac{|-\rangle+|+\rangle}{\sqrt{2}} = \frac{|+\rangle+|-\rangle}{\sqrt{2}}, \tag{3.52}$$

$$Z\left(\frac{|+\rangle-|-\rangle}{\sqrt{2}}\right) = \frac{Z|+\rangle-Z|-\rangle}{\sqrt{2}} = \frac{|-\rangle-|+\rangle}{\sqrt{2}} = -\left(\frac{|+\rangle-|-\rangle}{\sqrt{2}}\right). \tag{3.53}$$

The above steps demonstrate that the states $\left(|+\rangle+|-\rangle\right)/\sqrt{2}$ and $\left(|+\rangle-|-\rangle\right)/\sqrt{2}$ are both eigenstates of the Z operators. These states are none other than the respective computational basis states $|0\rangle$ and $|1\rangle$, by inspecting the definitions in (3.24). Thus, a matrix representation of the Z operator in the computational basis is

$$\begin{bmatrix} \langle0|Z|0\rangle & \langle0|Z|1\rangle \\ \langle1|Z|0\rangle & \langle1|Z|1\rangle \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \tag{3.54}$$

and is a diagonalization of the operator Z. So, the behavior of the Z operator in the computational basis is the same as the behavior of the X operator in the $+/-$ basis.

3.3.2 Commutators and Anticommutators

The *commutator* $[A,B]$ of two operators A and B is as follows:

$$[A,B] \equiv AB - BA. \tag{3.55}$$

Two operators commute if and only if their commutator is equal to zero.

The *anticommutator* $\{A,B\}$ of two operators A and B is as follows:

$$\{A,B\} \equiv AB + BA. \tag{3.56}$$

We say that two operators *anticommute* if their anticommutator is equal to zero.

Exercise 3.3.2 *Find a matrix representation for* $[X,Z]$ *in the basis* $\{|0\rangle,|1\rangle\}$.

[3] A more appropriate name might be the "bit flip in the $+/-$ basis operator," but this name is too long, so we stick with the term "phase flip."

3.3.3 The Pauli Matrices

The convention in quantum theory is to take the computational basis as the *standard basis* for representing physical qubits. The standard matrix representation for the above two operators is as follows when we choose the computational basis as the standard basis:

$$X \equiv \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad Z \equiv \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}. \tag{3.57}$$

The identity operator I has the following representation in any basis:

$$I \equiv \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{3.58}$$

Another operator, the Y operator, is a useful one to consider as well. The Y operator has the following matrix representation in the computational basis:

$$Y \equiv \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}. \tag{3.59}$$

It is easy to check that $Y = iXZ$, and for this reason, we can think of the Y operator as a combined bit and phase flip. The four matrices I, X, Y, and Z are special for the manipulation of physical qubits and are known as the *Pauli matrices.*

Exercise 3.3.3 *Show that the Pauli matrices are all Hermitian, unitary, they square to the identity, and their eigenvalues are* ± 1.

Exercise 3.3.4 *Represent the eigenstates of the* Y *operator in the computational basis.*

Exercise 3.3.5 *Show that the Pauli matrices either commute or anticommute.*

Exercise 3.3.6 *Let us label the Pauli matrices as* $\sigma_0 \equiv I$, $\sigma_1 \equiv X$, $\sigma_2 \equiv Y$, *and* $\sigma_3 \equiv Z$. *Show that* $\operatorname{Tr}\{\sigma_i \sigma_j\} = 2\delta_{ij}$ *for all* $i, j \in \{0, \ldots, 3\}$, *where* Tr *denotes the trace of a matrix, defined as the sum of the entries along the diagonal (see also Definition 4.1.1).*

3.3.4 Hadamard Gate

Another important unitary operator is the transformation that takes the computational basis to the $+/-$ basis. This transformation is the Hadamard transformation:

$$|0\rangle \to |+\rangle, \qquad |1\rangle \to |-\rangle. \tag{3.60}$$

Using the above relations, we can represent the Hadamard transformation as the following operator:

$$H \equiv |+\rangle\langle 0| + |-\rangle\langle 1|. \tag{3.61}$$

It is straightforward to check that the above operator implements the transformation in (3.60).

Now consider a generalization of the above construction. Suppose that one orthonormal basis is $\{|\psi_i\rangle\}_{i\in\{0,1\}}$ and another is $\{|\phi_i\rangle\}_{i\in\{0,1\}}$ where the index i merely indexes the states in each orthonormal basis. Then the unitary operator that takes states in the first basis to states in the second basis is

$$\sum_{i=0,1} |\phi_i\rangle\langle\psi_i|. \tag{3.62}$$

Exercise 3.3.7 *Show that the Hadamard operator H has the following matrix representation in the computational basis:*

$$H = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}. \tag{3.63}$$

Exercise 3.3.8 *Show that the Hadamard operator is its own inverse by employing the above matrix representation and by using its operator form in* (3.61).

Exercise 3.3.9 *If the Hadamard gate is its own inverse, then it takes the states $|+\rangle$ and $|-\rangle$ to the respective states $|0\rangle$ and $|1\rangle$ and we can represent it as the following operator: $H = |0\rangle\langle+| + |1\rangle\langle-|$. Show that $|0\rangle\langle+| + |1\rangle\langle-| = |+\rangle\langle 0| + |-\rangle\langle 1|$.*

Exercise 3.3.10 *Show that $HXH = Z$ and that $HZH = X$.*

3.3.5 Rotation Operators

We end this section on the evolution of quantum states by discussing "rotation evolutions" and by giving a more complete picture of the Bloch sphere. The rotation operators $R_X(\phi)$, $R_Y(\phi)$, $R_Z(\phi)$ are functions of the respective Pauli operators X, Y, Z where

$$R_X(\phi) \equiv \exp\{iX\phi/2\}, \qquad R_Y(\phi) \equiv \exp\{iY\phi/2\}, \qquad R_Z(\phi) \equiv \exp\{iZ\phi/2\}, \tag{3.64}$$

and ϕ is some angle such that $0 \leq \phi < 2\pi$. How do we determine a function of an operator? The standard way is to represent the operator in its diagonal basis and apply the function to the non-zero eigenvalues of the operator. For example, the diagonal representation of the X operator is

$$X = |+\rangle\langle+| - |-\rangle\langle-|. \tag{3.65}$$

Applying the function $\exp\{iX\phi/2\}$ to the non-zero eigenvalues of X gives

$$R_X(\phi) = \exp\{i\phi/2\}\, |+\rangle\langle+| + \exp\{-i\phi/2\}\, |-\rangle\langle-|. \tag{3.66}$$

This is a special case of the following more general convention that we follow throughout this book:

Definition 3.3.1 (Function of a Hermitian operator) *Suppose that a Hermitian operator A has a spectral decomposition $A = \sum_{i:a_i\neq 0} a_i |i\rangle\langle i|$ for some orthonormal basis $\{|i\rangle\}$. Then the operator $f(A)$ for some function f is defined as follows:*

$$f(A) \equiv \sum_{i:a_i\neq 0} f(a_i)|i\rangle\langle i|. \tag{3.67}$$

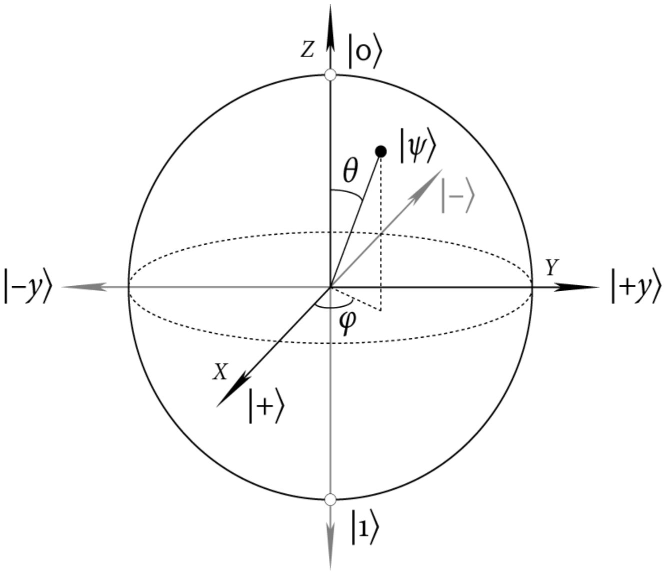

Figure 3.4: This figure provides more labels for states on the Bloch sphere. The Z axis has its points on the sphere as eigenstates of the Pauli Z operator, the X axis has eigenstates of the Pauli X operator, and the Y axis has eigenstates of the Pauli Y operator. The rotation operators $R_X(\phi)$, $R_Y(\phi)$, and $R_Z(\phi)$ rotate a state on the sphere by an angle ϕ about the respective X, Y, and Z axis.

Exercise 3.3.11 *Show that the rotation operators $R_X(\phi)$, $R_Y(\phi)$, $R_Z(\phi)$ are equal to the following expressions:*

$$R_X(\phi) = \cos(\phi/2)I + i\sin(\phi/2)X, \tag{3.68}$$

$$R_Y(\phi) = \cos(\phi/2)I + i\sin(\phi/2)Y, \tag{3.69}$$

$$R_Z(\phi) = \cos(\phi/2)I + i\sin(\phi/2)Z, \tag{3.70}$$

by using the facts that $\cos(\phi/2) = \frac{1}{2}\left(e^{i\phi/2} + e^{-i\phi/2}\right)$ *and* $\sin(\phi/2) = \frac{1}{2i}\left(e^{i\phi/2} - e^{-i\phi/2}\right)$.

Figure 3.4 provides a more detailed picture of the Bloch sphere since we have now established the Pauli operators and their eigenstates. The computational basis states are the eigenstates of the Z operator and are the north and south poles on the Bloch sphere. The $+/-$ basis states are the eigenstates of the X operator and the calculation from Exercise 3.2.1 shows that they are the "east and west poles" of the Bloch sphere. We leave it as another exercise to show that the Y eigenstates are the other poles along the equator of the Bloch sphere.

Exercise 3.3.12 *Determine the Bloch sphere angles θ and φ for the eigenstates of the Pauli Y operator.*

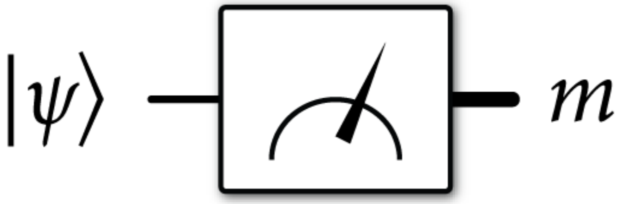

Figure 3.5: This figure depicts our diagram of a quantum measurement. Thin lines denote quantum information and thick lines denote classical information. The result of the measurement is to output a classical variable m according to a probability distribution governed by the Born rule of the quantum theory.

3.4 Measurement

Measurement is another type of evolution that a quantum system can undergo. It is an evolution that allows us to retrieve classical information from a quantum state and thus is the way that we can "read out" information. Suppose that we would like to learn something about the quantum state $|\psi\rangle$ in (3.2). Nature prevents us from learning anything about the probability amplitudes α and β if we have only one quantum measurement that we can perform on one copy of the state. Nature only allows us to measure *observables*. Observables are physical variables such as the position or momentum of a particle. In the quantum theory, we represent observables as Hermitian operators in part because their eigenvalues are real numbers and every measuring device outputs a real number. Examples of qubit observables that we can measure are the Pauli operators X, Y, and Z.

Suppose that we measure the Z operator. This measurement is called a "measurement in the computational basis" or a "measurement of the Z observable" because we are measuring the eigenvalues of the Z operator. The measurement postulate of the quantum theory, also known as the *Born rule*, states that the system reduces to the state $|0\rangle$ with probability $|\alpha|^2$ and reduces to the state $|1\rangle$ with probability $|\beta|^2$. That is, the resulting probabilities are the squares of the probability amplitudes. After the measurement, our measuring apparatus tells us whether the state reduced to $|0\rangle$ or $|1\rangle$—it returns $+1$ if the resulting state is $|0\rangle$ and returns -1 if the resulting state is $|1\rangle$. These returned values are the eigenvalues of the Z operator. The measurement postulate is the aspect of the quantum theory that makes it probabilistic or "jumpy" and is part of the "strangeness" of the quantum theory. Figure 3.5 depicts the notation for a measurement that we will use in diagrams throughout this book.

What is the result if we measure the state $|\psi\rangle$ in the $+/-$ basis? Consider that we can represent $|\psi\rangle$ as a superposition of the $|+\rangle$ and $|-\rangle$ states, as given in (3.29). The measurement postulate then states that a measurement of the X operator gives the state $|+\rangle$ with probability $|\alpha + \beta|^2/2$ and the state $|-\rangle$ with probability $|\alpha - \beta|^2/2$. Quantum interference is now coming into play because the amplitudes α and β interfere with each other. So this effect plays an important role in quantum information theory.

In some cases, the basis states $|0\rangle$ and $|1\rangle$ may not represent the spin states of an electron, but may represent the *location* of an electron. So, a way to interpret this measurement

Quantum State	**Probability of $\vert +\rangle$**	**Probability of $\vert -\rangle$**
Superposition state	$\left\vert \alpha+\beta\right\vert ^{2}/2$	$\left\vert \alpha-\beta\right\vert ^{2}/2$
Probabilistic description	1/2	1/2

Table 3.1: This table summarizes the differences in probabilities for a quantum state in a superposition $\alpha|0\rangle+\beta|1\rangle$ and a classical state that is a probabilistic mixture of $|0\rangle$ and $|1\rangle$.

postulate is that the electron "jumps into" one location or another depending on the outcome of the measurement. But what is the state of the electron before the measurement? We will just say in this book that it is in a superposed, indefinite, or unsharp state, rather than trying to pin down a philosophical interpretation. Some might say that the electron is in "two different locations at the same time."

Also, we should stress that we cannot interpret the measurement postulate as meaning that the state is in $|0\rangle$ or $|1\rangle$ with respective probabilities $|\alpha|^2$ and $|\beta|^2$ before the measurement occurs, because this latter interpretation is physically different from what we described above and is also completely classical. The superposition state $\alpha|0\rangle+\beta|1\rangle$ gives fundamentally different behavior from the probabilistic description of a state that is in $|0\rangle$ or $|1\rangle$ with respective probabilities $|\alpha|^2$ and $|\beta|^2$. Suppose that we have the two different descriptions of a state (superposition and probabilistic) and measure the Z operator. We get the same result for both cases—the resulting state is $|0\rangle$ or $|1\rangle$ with respective probabilities $|\alpha|^2$ and $|\beta|^2$.

But now suppose that we measure the X operator. The superposed state gives the result from before—we get the state $|+\rangle$ with probability $\left|\alpha+\beta\right|^2/2$ and the state $|-\rangle$ with probability $\left|\alpha-\beta\right|^2/2$. The probabilistic description gives a much different result. Suppose that the state is $|0\rangle$. We know that $|0\rangle$ is a uniform superposition of $|+\rangle$ and $|-\rangle$:

$$|0\rangle=\frac{|+\rangle+|-\rangle}{\sqrt{2}}. \tag{3.71}$$

So the state collapses to $|+\rangle$ or $|-\rangle$ with equal probability in this case. If the state is $|1\rangle$, then it collapses again to $|+\rangle$ or $|-\rangle$ with equal probabilities. Summing up these probabilities, it follows that a measurement of the X operator gives the state $|+\rangle$ with probability $\left(|\alpha|^2+|\beta|^2\right)/2=1/2$ and gives the state $|-\rangle$ with the same probability. These results are fundamentally different from those in which the state is the superposition state $|\psi\rangle$, and experiment after experiment has supported the predictions of the quantum theory. Table 3.1 summarizes the results that we just discussed.

Now we consider a "Stern–Gerlach"-like argument to illustrate another example of fundamental quantum behavior (Gerlach and Stern, 1922). The Stern–Gerlach experiment was a crucial one for determining the "strange" behavior of quantum spin states. Suppose that we prepare the state $|0\rangle$. If we measure this state in the Z basis, the result is that we always obtain the state $|0\rangle$ because the prepared state is a definite Z eigenstate. Suppose now that we measure the X operator. The state $|0\rangle$ is equal to a uniform superposition of $|+\rangle$ and $|-\rangle$. The measurement postulate then states that we get the state $|+\rangle$ or $|-\rangle$ with equal probability after performing this measurement. If we then measure the Z operator again, the result is completely random. The Z measurement result is $|0\rangle$ or $|1\rangle$ with equal probability if the result of the X measurement is $|+\rangle$ and the same outcome occurs if the result of the X measurement is $|-\rangle$. This argument demonstrates that the measurement of the X operator “throws off” the measurement of the Z operator. The Stern–Gerlach experiment was one of the earliest to validate the predictions of the quantum theory.

3.4.1 Probability, Expectation, and Variance of an Operator

We have an alternate, more formal way of stating the measurement postulate that turns out to be more useful for a general quantum system. Suppose that we are measuring the Z operator. The diagonal representation of this operator is

$$Z = |0\rangle\langle 0| - |1\rangle\langle 1|. \tag{3.72}$$

Consider the Hermitian operator

$$\Pi_0 \equiv |0\rangle\langle 0|. \tag{3.73}$$

It is a projection operator because applying it twice has the same effect as applying it once: $\Pi_0^2 = \Pi_0$. It projects onto the subspace spanned by the single vector $|0\rangle$. A similar line of analysis applies to the projection operator

$$\Pi_1 \equiv |1\rangle\langle 1|. \tag{3.74}$$

So we can represent the Z operator as $\Pi_0 - \Pi_1$. Performing a measurement of the Z operator is equivalent to asking the question: Is the state $|0\rangle$ or $|1\rangle$? Consider the quantity $\langle\psi|\Pi_0|\psi\rangle$:

$$\langle\psi|\Pi_0|\psi\rangle = \langle\psi|0\rangle\,\langle 0|\psi\rangle = \alpha^*\alpha = |\alpha|^2 . \tag{3.75}$$

A similar analysis demonstrates that

$$\langle\psi|\Pi_1|\psi\rangle = |\beta|^2 . \tag{3.76}$$

These two quantities then give the probability that the state reduces to $|0\rangle$ or $|1\rangle$.

A more general way of expressing a measurement of the Z basis is to say that we have a set $\{\Pi_i\}_{i\in\{0,1\}}$ of measurement operators that determine the outcome probabilities. These measurement operators also determine the state that results after the measurement. If the measurement result is $+1$, then the resulting state is

$$\frac{\Pi_0|\psi\rangle}{\sqrt{\langle\psi|\Pi_0|\psi\rangle}} = |0\rangle, \tag{3.77}$$

where we implicitly ignore the irrelevant global phase factor $\frac{\alpha}{|\alpha|}$. If the measurement result is -1, then the resulting state is

$$\frac{\Pi_1|\psi\rangle}{\sqrt{\langle\psi|\Pi_1|\psi\rangle}} = |1\rangle, \tag{3.78}$$

where we again implicitly ignore the irrelevant global phase factor $\frac{\beta}{|\beta|}$. Dividing by $\sqrt{\langle\psi|\Pi_i|\psi\rangle}$ for $i = 0, 1$ ensures that the state resulting after measurement corresponds to a physical state (a unit vector).

We can also measure any orthonormal basis in this way—this type of projective measurement is called a *von Neumann measurement*. For any orthonormal basis $\{|\phi_i\rangle\}_{i\in\{0,1\}}$, the measurement operators are $\{|\phi_i\rangle\langle\phi_i|\}_{i\in\{0,1\}}$, and the state reduces to $|\phi_i\rangle\langle\phi_i|\psi\rangle / |\langle\phi_i|\psi\rangle|$ with probability $\langle\psi|\phi_i\rangle \langle\phi_i|\psi\rangle = |\langle\phi_i|\psi\rangle|^2$.

Exercise 3.4.1 *Determine the set of measurement operators corresponding to a measurement of the X observable.*

We might want to determine the expectation of the measurement result when measuring the Z operator. The probability of getting the $+1$ value corresponding to the $|0\rangle$ state is $|\alpha|^2$ and the probability of getting the -1 value corresponding to the -1 eigenstate is $|\beta|^2$. Standard probability theory then gives us a way to calculate the expected value of a measurement of the Z operator when the state is $|\psi\rangle$:

$$\mathbb{E}[Z] = |\alpha|^2 (1) + |\beta|^2 (-1) = |\alpha|^2 - |\beta|^2 . \tag{3.79}$$

We can formulate an alternate way to write this expectation, by making use of the Dirac notation:

$$\begin{aligned} \mathbb{E}[Z] &= |\alpha|^2 (1) + |\beta|^2 (-1) && (3.80) \\ &= \langle\psi|\Pi_0|\psi\rangle + \langle\psi|\Pi_1|\psi\rangle (-1) && (3.81) \\ &= \langle\psi|\Pi_0 - \Pi_1|\psi\rangle && (3.82) \\ &= \langle\psi|Z|\psi\rangle. && (3.83) \end{aligned}$$

It is common for physicists to denote the expectation as

$$\langle Z\rangle \equiv \langle\psi|Z|\psi\rangle, \tag{3.84}$$

when it is understood that the expectation is with respect to the state $|\psi\rangle$. This type of expression is a general one and the next exercise asks you to show that it works for the X and Y operators as well.

Exercise 3.4.2 *Show that the expressions $\langle\psi|X|\psi\rangle$ and $\langle\psi|Y|\psi\rangle$ give the respective expectations $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ when measuring the state $|\psi\rangle$ in the respective X and Y basis.*

We also might want to determine the variance of the measurement of the Z operator. Standard probability theory again gives that

$$\mathrm{Var}[Z] = \mathbb{E}\left[Z^2\right] - \mathbb{E}[Z]^2 . \tag{3.85}$$

Physicists denote the standard deviation of the measurement of the Z operator as

$$\Delta Z \equiv \left\langle (Z - \langle Z\rangle)^2 \right\rangle^{1/2}, \tag{3.86}$$

and thus the variance is equal to $(\Delta Z)^2$. Physicists often refer to ΔZ as the uncertainty of the observable Z when the state is $|\psi\rangle$.

In order to calculate the variance $\text{Var}[Z]$, we really just need the second moment $\mathbb{E}\left[Z^2\right]$ because we already have the expectation $\mathbb{E}\left[Z\right]$:

$$\mathbb{E}\left[Z^2\right] = |\alpha|^2 (1)^2 + |\beta|^2 (-1)^2 = |\alpha|^2 + |\beta|^2 . \tag{3.87}$$

We can again calculate this quantity with the Dirac notation. The quantity $\langle\psi|Z^2|\psi\rangle$ is the same as $\mathbb{E}\left[Z^2\right]$ and the next exercise asks you for a proof.

Exercise 3.4.3 *Show that* $\mathbb{E}\left[X^2\right] = \langle\psi|X^2|\psi\rangle$, $\mathbb{E}\left[Y^2\right] = \langle\psi|Y^2|\psi\rangle$, *and* $\mathbb{E}\left[Z^2\right] = \langle\psi|Z^2|\psi\rangle$.

3.4.2 The Uncertainty Principle

The uncertainty principle is a fundamental feature of the quantum theory. In the case of qubits, one instance of the uncertainty principle gives a lower bound on the product of the uncertainty of the Z operator and the uncertainty of the X operator:

$$\Delta Z \Delta X \geq \frac{1}{2} \left|\langle\psi| \left[Z, X\right] |\psi\rangle\right| . \tag{3.88}$$

We can prove this principle using the postulates of the quantum theory. Let us define the operators $Z_0 \equiv Z - \langle Z\rangle$ and $X_0 \equiv X - \langle X\rangle$. First, consider that

$$\Delta Z \Delta X = \langle\psi|Z_0^2|\psi\rangle^{1/2}\langle\psi|X_0^2|\psi\rangle^{1/2} \geq \left|\langle\psi|Z_0 X_0|\psi\rangle\right| . \tag{3.89}$$

The above step follows by applying the Cauchy–Schwarz inequality to the vectors $X_0|\psi\rangle$ and $Z_0|\psi\rangle$. For any operator A, we define its real part $\text{Re}\left\{A\right\}$ as $\text{Re}\left\{A\right\} \equiv (A + A^\dagger)/2$, and its imaginary part $\text{Im}\left\{A\right\}$ as $\text{Im}\left\{A\right\} \equiv (A - A^\dagger)/2i$, so that $A = \text{Re}\left\{A\right\} + i\,\text{Im}\left\{A\right\}$. So the real and imaginary parts of the operator $Z_0 X_0$ are

$$\text{Re}\left\{Z_0 X_0\right\} = \frac{Z_0 X_0 + X_0 Z_0}{2} \equiv \frac{\left\{Z_0, X_0\right\}}{2}, \tag{3.90}$$

$$\text{Im}\left\{Z_0 X_0\right\} = \frac{Z_0 X_0 - X_0 Z_0}{2i} \equiv \frac{\left[Z_0, X_0\right]}{2i}, \tag{3.91}$$

where $\left\{Z_0, X_0\right\}$ is the anticommutator of Z_0 and X_0 and $\left[Z_0, X_0\right]$ is the commutator of the two operators. We can then express the quantity $\left|\langle\psi|Z_0 X_0|\psi\rangle\right|$ in terms of the real and imaginary parts of $Z_0 X_0$:

$$\begin{aligned}
\left|\langle\psi|Z_0 X_0|\psi\rangle\right| &= \left|\langle\psi| \text{Re}\left\{Z_0 X_0\right\} |\psi\rangle + i\langle\psi| \text{Im}\left\{Z_0 X_0\right\} |\psi\rangle\right| &(3.92)\\
&\geq \left|\langle\psi| \text{Im}\left\{Z_0 X_0\right\} |\psi\rangle\right| &(3.93)\\
&= \left|\langle\psi| \left[Z_0, X_0\right] |\psi\rangle\right| /2 &(3.94)\\
&= \left|\langle\psi| \left[Z, X\right] |\psi\rangle\right| /2. &(3.95)
\end{aligned}$$

The first equality follows by substitution. The first inequality follows because the magnitude of any complex number is greater than the magnitude of its imaginary part. The second equality follows by substitution with (3.91). Finally, the third equality follows from the result of Exercise 3.4.4 below. We worked out the above derivation for particular observables acting on qubit states, but note that it holds for general observables and quantum states.

The commutator of the operators Z and X arises in the lower bound, and thus, the non-commutativity of the operators Z and X is the fundamental reason that there is an uncertainty principle for them. Also, there is no uncertainty principle for any two operators that commute with each other.

It is worthwhile to interpret the uncertainty principle in (3.88), which really receives an interpretation after conducting a large number of independent experiments of two different kinds. In the first kind of experiment, one prepares the state $|\psi\rangle$ and measures the Z observable. After repeating this experiment independently many times, one can calculate an estimate of the standard deviation ΔZ, which becomes closer and closer to the true standard deviation ΔZ as the number of independent experiments becomes large. In the second kind of experiment, one prepares the state $|\psi\rangle$ and measures the X observable. After repeating many times, one can calculate an estimate of ΔX. The uncertainty principle then states that the product of the estimates (for a large number of independent experiments) is bounded from below by the expectation of the commutator: $\frac{1}{2}\left|\langle\psi|\left[X,Z\right]|\psi\rangle\right|$.

Exercise 3.4.4 *Show that* $[Z_0, X_0] = [Z, X]$ *and that* $[Z, X] = -2iY$*.*

Exercise 3.4.5 *The uncertainty principle in* (3.88) *has the property that the lower bound has a dependence on the state* $|\psi\rangle$*. Find a state* $|\psi\rangle$ *for which the lower bound on the uncertainty product* $\Delta X \Delta Z$ *vanishes.*[4]

3.5 Composite Quantum Systems

A single physical qubit is an interesting physical system that exhibits uniquely quantum phenomena, but it is not particularly useful on its own (just as a single classical bit is not particularly useful for classical communication or computation). We can only perform interesting quantum information-processing tasks when we combine qubits together. Therefore, we should have a way to describe their behavior when they combine to form a composite quantum system.

Consider two classical bits c_0 and c_1. In order to describe bit operations on the pair of cbits, we write them as an ordered pair (c_1, c_0). The space of all possible bit values is the Cartesian product $\mathbb{Z}_2 \times \mathbb{Z}_2$ of two copies of the set $\mathbb{Z}_2 \equiv \{0, 1\}$:

$$\mathbb{Z}_2 \times \mathbb{Z}_2 \equiv \{(0,0), (0,1), (1,0), (1,1)\}. \tag{3.96}$$

[4] Do not be alarmed by the result of this exercise! The usual formulation of the uncertainty principle only gives a lower bound on the uncertainty product. This lower bound never vanishes for the case of position and momentum observables because the commutator of these two observables is equal to the identity operator multiplied by i, but it can vanish for the operators given in the exercise.

Typically, we make the abbreviation $c_1c_0 \equiv (c_1, c_0)$ when representing cbit states.

We can represent the state of two cbits with particular states of qubits. For example, we can represent the two-cbit state 00 using the following mapping:

$$00 \to |0\rangle|0\rangle. \tag{3.97}$$

Many times, we make the abbreviation $|00\rangle \equiv |0\rangle|0\rangle$ when representing two-cbit states with qubits. Any two-cbit state c_1c_0 has the following representation as a two-qubit state:

$$c_1c_0 \to |c_1c_0\rangle\,. \tag{3.98}$$

The above qubit states are not the only possible states that can occur in the quantum theory. By the superposition principle, any possible linear combination of the set of two-cbit states is a possible two-qubit state:

$$|\xi\rangle \equiv \alpha|00\rangle + \beta|01\rangle + \gamma|10\rangle + \delta|11\rangle. \tag{3.99}$$

The unit-norm condition $|\alpha|^2 + |\beta|^2 + |\gamma|^2 + |\delta|^2 = 1$ again must hold for the two-qubit state to correspond to a physical quantum state. It is now clear that the Cartesian product is not sufficient for representing two-qubit quantum states because it does not allow for linear combinations of states (just as the mathematics of Boolean algebra is not sufficient to represent single-qubit states).

We again turn to linear algebra to determine a representation that suffices. The *tensor product* is a mathematical operation that suffices to give a representation of two-qubit quantum states. Suppose we have two two-dimensional vectors:

$$\begin{bmatrix} a_1 \\ b_1 \end{bmatrix}, \qquad \begin{bmatrix} a_2 \\ b_2 \end{bmatrix}. \tag{3.100}$$

The tensor product of these two vectors is

$$\begin{bmatrix} a_1 \\ b_1 \end{bmatrix} \otimes \begin{bmatrix} a_2 \\ b_2 \end{bmatrix} \equiv \begin{bmatrix} a_1 \begin{bmatrix} a_2 \\ b_2 \end{bmatrix} \\ b_1 \begin{bmatrix} a_2 \\ b_2 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} a_1a_2 \\ a_1b_2 \\ b_1a_2 \\ b_1b_2 \end{bmatrix}. \tag{3.101}$$

One can understand this operation as taking the vector on the right and stacking two copies of it together, while multiplying each copy by the corresponding number in the first vector.

Recall, from (3.3), the vector representation of the single-qubit states $|0\rangle$ and $|1\rangle$. Using these vector representations and the above definition of the tensor product, the two-qubit basis states have the following vector representations:

$$|00\rangle = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad |01\rangle = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad |10\rangle = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad |11\rangle = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}. \tag{3.102}$$

A simple way to remember these representations is that the bits inside the ket index the element equal to one in the vector. For example, the vector representation of $|01\rangle$ has a one as its second element because 01 is the second index for the two-bit strings. The vector representation of the superposition state in (3.99) is

$$\begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{bmatrix}. \tag{3.103}$$

There are actually many different ways that we can write two-qubit states, and we list all of these right now. Physicists have developed many shorthands, and it is important to know each of these because they often appear in the literature (this book even uses different notations depending on the context). We may use any of the following two-qubit notations if the two qubits are local to one party and only one party is involved in a protocol:

$$\alpha|0\rangle \otimes |0\rangle + \beta|0\rangle \otimes |1\rangle + \gamma|1\rangle \otimes |0\rangle + \delta|1\rangle \otimes |1\rangle, \tag{3.104}$$

$$\alpha|0\rangle|0\rangle + \beta|0\rangle|1\rangle + \gamma|1\rangle|0\rangle + \delta|1\rangle|1\rangle, \tag{3.105}$$

$$\alpha|00\rangle + \beta|01\rangle + \gamma|10\rangle + \delta|11\rangle. \tag{3.106}$$

We can put labels on the qubits if two or more parties, such as A and B, are involved

$$\alpha|0\rangle_A \otimes |0\rangle_B + \beta|0\rangle_A \otimes |1\rangle_B + \gamma|1\rangle_A \otimes |0\rangle_B + \delta|1\rangle_A \otimes |1\rangle_B, \tag{3.107}$$

$$\alpha|0\rangle_A|0\rangle_B + \beta|0\rangle_A|1\rangle_B + \gamma|1\rangle_A|0\rangle_B + \delta|1\rangle_A|1\rangle_B, \tag{3.108}$$

$$\alpha|00\rangle_{AB} + \beta|01\rangle_{AB} + \gamma|10\rangle_{AB} + \delta|11\rangle_{AB}. \tag{3.109}$$

This second scenario is different from the first scenario because two spatially separated parties share the two-qubit state. If the state has quantum correlations, then it can be valuable as a communication resource. We go into more detail on this topic in Section 3.6, which discusses *entanglement*.

3.5.1 Evolution of Composite Systems

The postulate on unitary evolution extends to the two-qubit scenario as well. First, let us establish that the tensor product $A \otimes B$ of two operators A and B is

$$A \otimes B \equiv \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \otimes \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \tag{3.110}$$

$$\equiv \begin{bmatrix} a_{11}\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} & a_{12}\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \\ a_{21}\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} & a_{22}\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \end{bmatrix} \tag{3.111}$$

$$= \begin{bmatrix} a_{11}b_{11} & a_{11}b_{12} & a_{12}b_{11} & a_{12}b_{12} \\ a_{11}b_{21} & a_{11}b_{22} & a_{12}b_{21} & a_{12}b_{22} \\ a_{21}b_{11} & a_{21}b_{12} & a_{22}b_{11} & a_{22}b_{12} \\ a_{21}b_{21} & a_{21}b_{22} & a_{22}b_{21} & a_{22}b_{22} \end{bmatrix}. \tag{3.112}$$

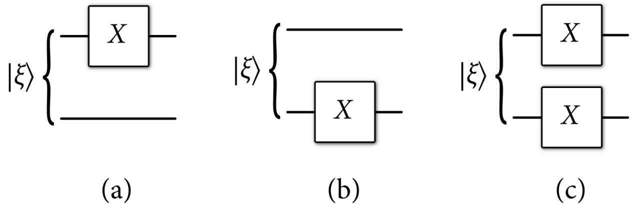

Figure 3.6: This figure depicts circuits for the example two-qubit unitaries $X_1 I_2$, $I_1 X_2$, and $X_1 X_2$.

The tensor-product operation for matrices is similar to what we did for vectors, but now we are stacking copies of the matrix on the right both vertically and horizontally, and multiplying each copy by the corresponding number in the first matrix.

Consider the two-qubit state in (3.99). We can perform a NOT gate on the first qubit so that it changes to $\alpha|10\rangle + \beta|11\rangle + \gamma|00\rangle + \delta|01\rangle$. We can alternatively flip its second qubit: $\alpha|01\rangle + \beta|00\rangle + \gamma|11\rangle + \delta|10\rangle$, or flip both at the same time: $\alpha|11\rangle + \beta|10\rangle + \gamma|01\rangle + \delta|00\rangle$. Figure 3.6 depicts quantum circuit representations of these operations. These are all reversible operations because applying them again gives the original state in (3.99). In the first case, we did nothing to the second qubit, and in the second case, we did nothing to the first qubit. The identity operator acts on the qubits that have nothing happen to them.

Let us label the first qubit as "1" and the second qubit as "2." We can then label the operator for the first operation as $X_1 I_2$ because this operator flips the first qubit and does nothing (applies the identity) to the second qubit. We can also label the operators for the second and third operations respectively as $I_1 X_2$ and $X_1 X_2$. The matrix representation of the operator $X_1 I_2$ is the tensor product of the matrix representation of X with the matrix representation of I—this relation similarly holds for the operators $I_1 X_2$ and $X_1 X_2$. We show that it holds for the operator $X_1 I_2$ and ask you to verify the other two cases. We can use the two-qubit computational basis to get a matrix representation for the two-qubit operator $X_1 I_2$:

$$\begin{bmatrix} \langle 00|X_1 I_2|00\rangle & \langle 00|X_1 I_2|01\rangle & \langle 00|X_1 I_2|10\rangle & \langle 00|X_1 I_2|11\rangle \\ \langle 01|X_1 I_2|00\rangle & \langle 01|X_1 I_2|01\rangle & \langle 01|X_1 I_2|10\rangle & \langle 01|X_1 I_2|11\rangle \\ \langle 10|X_1 I_2|00\rangle & \langle 10|X_1 I_2|01\rangle & \langle 10|X_1 I_2|10\rangle & \langle 10|X_1 I_2|11\rangle \\ \langle 11|X_1 I_2|00\rangle & \langle 11|X_1 I_2|01\rangle & \langle 11|X_1 I_2|10\rangle & \langle 11|X_1 I_2|11\rangle \end{bmatrix}$$
$$= \begin{bmatrix} \langle 00|10\rangle & \langle 00|11\rangle & \langle 00|00\rangle & \langle 00|01\rangle \\ \langle 01|10\rangle & \langle 01|11\rangle & \langle 01|00\rangle & \langle 01|01\rangle \\ \langle 10|10\rangle & \langle 10|11\rangle & \langle 10|00\rangle & \langle 10|01\rangle \\ \langle 11|10\rangle & \langle 11|11\rangle & \langle 11|00\rangle & \langle 11|01\rangle \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}. \qquad (3.113)$$

This last matrix is equal to the tensor product $X \otimes I$ by inspecting the definition of the tensor product for matrices in (3.110).

Exercise 3.5.1 *Show that the matrix representation of the operator $I_1 X_2$ is equal to the tensor product $I \otimes X$. Show the same for $X_1 X_2$ and $X \otimes X$.*

3.5.2 Probability Amplitudes for Composite Systems

We relied on the orthogonality of the two-qubit computational basis states for evaluating amplitudes such as $\langle 00|10\rangle$ or $\langle 00|00\rangle$ in the above matrix representation. It turns out that there is another way to evaluate these amplitudes that relies only on the orthogonality of the single-qubit computational basis states. Suppose that we have four single-qubit states $|\phi_0\rangle$, $|\phi_1\rangle$, $|\psi_0\rangle$, $|\psi_1\rangle$, and we make the following two-qubit states from them:

$$|\phi_0\rangle \otimes |\psi_0\rangle, \qquad |\phi_1\rangle \otimes |\psi_1\rangle. \tag{3.114}$$

We may represent these states equally well as follows:

$$|\phi_0, \psi_0\rangle, \qquad |\phi_1, \psi_1\rangle, \tag{3.115}$$

because the Dirac notation is versatile (virtually anything can go inside a ket as long as its meaning is not ambiguous). The bra $\langle \phi_1, \psi_1|$ is dual to the ket $|\phi_1, \psi_1\rangle$, and we can use it to calculate the following amplitude:

$$\langle \phi_1, \psi_1|\phi_0, \psi_0\rangle . \tag{3.116}$$

This amplitude is equal to the multiplication of the single-qubit amplitudes:

$$\langle \phi_1, \psi_1|\phi_0, \psi_0\rangle = \langle \phi_1|\phi_0\rangle \langle \psi_1|\psi_0\rangle . \tag{3.117}$$

Exercise 3.5.2 *Verify that the amplitudes* $\{\langle ij|kl\rangle\}_{i,j,k,l\in\{0,1\}}$ *are respectively equal to the amplitudes* $\{\langle i|k\rangle \langle j|l\rangle\}_{i,j,k,l\in\{0,1\}}$. *By linearity, this exercise justifies the relation in* (3.117) *(at least for two-qubit states).*

3.5.3 Controlled Gates

An important two-qubit unitary evolution is the controlled-NOT (CNOT) gate. We consider its classical version first. The classical gate acts on two cbits. It does nothing if the first bit is equal to zero, and flips the second bit if the first bit is equal to one:

$$00 \to 00, \qquad 01 \to 01, \qquad 10 \to 11, \qquad 11 \to 10. \tag{3.118}$$

We turn this gate into a quantum gate[5] by demanding that it act in the same way on the two-qubit computational basis states:

$$|00\rangle \to |00\rangle, \qquad |01\rangle \to |01\rangle, \qquad |10\rangle \to |11\rangle, \qquad |11\rangle \to |10\rangle. \tag{3.119}$$

By linearity, this behavior carries over to superposition states as well:

$$\alpha|00\rangle + \beta|01\rangle + \gamma|10\rangle + \delta|11\rangle \quad \underrightarrow{\text{CNOT}} \quad \alpha|00\rangle + \beta|01\rangle + \gamma|11\rangle + \delta|10\rangle. \tag{3.120}$$

[5] There are other terms for the action of turning a classical operation into a quantum one. Some examples are “making it coherent,” “coherifying,” or the quantum gate is a “coherification” of the classical one. The term “coherify” is not a proper English word, but we will use it regardless at certain points.

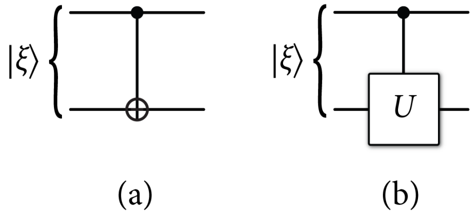

Figure 3.7: Circuit diagrams that we use for (a) a CNOT gate and (b) a controlled-U gate.

A useful operator representation of the CNOT gate is

$$\text{CNOT} \equiv |0\rangle\langle 0| \otimes I + |1\rangle\langle 1| \otimes X. \tag{3.121}$$

The above representation truly captures the coherent quantum nature of the CNOT gate. In the classical CNOT gate, we can say that it is a conditional gate, in the sense that the gate applies to the second bit conditioned on the value of the first bit. In the quantum CNOT gate, the second operation is *controlled* on the basis state of the first qubit (hence the choice of the name "controlled-NOT"). That is, the gate acts on superpositions of quantum states and maintains these superpositions, shuffling the probability amplitudes around while it does so. The one case in which the gate does not act is when the first qubit is prepared in the state $|0\rangle$ and the state of the second qubit is arbitrary.

A controlled-U gate is similar to the CNOT gate in (3.121). It simply applies the unitary U (assumed to be a single-qubit unitary) to the second qubit, controlled on the first qubit:

$$\text{controlled-}U \equiv |0\rangle\langle 0| \otimes I + |1\rangle\langle 1| \otimes U. \tag{3.122}$$

The control qubit could alternatively be controlled with respect to any orthonormal basis $\{|\phi_0\rangle, |\phi_1\rangle\}$:

$$|\phi_0\rangle\langle\phi_0| \otimes I + |\phi_1\rangle\langle\phi_1| \otimes U. \tag{3.123}$$

Figure 3.7 depicts the circuit diagrams for a controlled-NOT and controlled-U operation.

Exercise 3.5.3 *Verify that the matrix representation of the* CNOT *gate in the computational basis is*

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \tag{3.124}$$

Exercise 3.5.4 *Consider applying Hadamards to the first and second qubits before and after a* CNOT *acts on them. Show that this gate is equivalent to a* CNOT *in the $+/-$ basis (recall that the Z operator flips the $+/-$ basis):*

$$H_1 H_2 \;\; \text{CNOT}\, H_1 H_2 = |+\rangle\langle +| \otimes I + |-\rangle\langle -| \otimes Z. \tag{3.125}$$

Exercise 3.5.5 *Show that two* CNOT *gates with the same control qubit commute.*

Exercise 3.5.6 *Show that two* CNOT *gates with the same target qubit commute.*

3.5.4 The No-Cloning Theorem

The no-cloning theorem is one of the simplest results in the quantum theory, yet it has some of the most profound consequences. It states that it is impossible to build a *universal copier* of quantum states. A universal copier would be a device that could copy any arbitrary quantum state that is input to it. It may be surprising at first to hear that copying quantum information is impossible because copying classical information is ubiquitous.

We now give a simple proof of the no-cloning theorem. Suppose for a contradiction that there is a two-qubit unitary operator U acting as a universal copier of quantum information. That is, if we input an arbitrary state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ as the first qubit and input an ancilla qubit $|0\rangle$ as the second qubit, such a device should "write" the first qubit to the second qubit slot as follows:

$$\begin{aligned} U|\psi\rangle|0\rangle &= |\psi\rangle|\psi\rangle && (3.126)\\ &= (\alpha|0\rangle + \beta|1\rangle)(\alpha|0\rangle + \beta|1\rangle) && (3.127)\\ &= \alpha^2|0\rangle|0\rangle + \alpha\beta|0\rangle|1\rangle + \alpha\beta|1\rangle|0\rangle + \beta^2|1\rangle|1\rangle. && (3.128)\end{aligned}$$

The copier is universal, meaning that it copies an arbitrary state. In particular, it also copies the states $|0\rangle$ and $|1\rangle$:

$$U|0\rangle|0\rangle = |0\rangle|0\rangle, \qquad U|1\rangle|0\rangle = |1\rangle|1\rangle. \tag{3.129}$$

Linearity of the quantum theory then implies that the unitary operator acts on a superposition $\alpha|0\rangle + \beta|1\rangle$ as follows:

$$U(\alpha|0\rangle + \beta|1\rangle)|0\rangle = \alpha|0\rangle|0\rangle + \beta|1\rangle|1\rangle. \tag{3.130}$$

However, the consequence in (3.128) contradicts the consequence in (3.130) because these two expressions do not have to be equal for all α and β:

$$\exists\alpha,\beta : \alpha^2|0\rangle|0\rangle + \alpha\beta|0\rangle|1\rangle + \alpha\beta|1\rangle|0\rangle + \beta^2|1\rangle|1\rangle \neq \alpha|0\rangle|0\rangle + \beta|1\rangle|1\rangle. \tag{3.131}$$

Thus, linearity of the quantum theory contradicts the existence of a universal quantum copier.

We would like to stress that this proof does not mean that it is impossible to copy certain quantum states—it only implies the impossibility of a *universal* copier. Observe that (3.131) is satisfied for $\alpha = 1, \beta = 0$ or $\alpha = 0, \beta = 1$, so that we can copy unknown classical states prepared in the basis $|0\rangle, |1\rangle$ (or any other orthonormal basis for that matter).

Another proof of the no-cloning theorem arrives at a contradiction by exploiting unitarity of quantum evolutions. Let us again suppose that a universal copier U exists. Consider two arbitrary states $|\psi\rangle$ and $|\phi\rangle$. If a universal copier U exists, then it performs the following copying operation for both states:

$$U|\psi\rangle|0\rangle = |\psi\rangle|\psi\rangle, \qquad U|\phi\rangle|0\rangle = |\phi\rangle\,|\phi\rangle. \tag{3.132}$$

Consider the probability amplitude $\langle\psi|\langle\psi||\phi\rangle|\phi\rangle$:

$$\langle\psi|\langle\psi||\phi\rangle|\phi\rangle = \langle\psi|\phi\rangle \langle\psi|\phi\rangle = \langle\psi|\phi\rangle^2 . \tag{3.133}$$

The following relation for $\langle\psi|\langle\psi||\phi\rangle|\phi\rangle$ holds as well by using the results in (3.132) and the unitarity property $U^\dagger U = I$:

$$\begin{aligned}\langle\psi|\langle\psi||\phi\rangle|\phi\rangle &= \langle\psi|\langle 0|U^\dagger U|\phi\rangle|0\rangle & (3.134)\\ &= \langle\psi|\langle 0||\phi\rangle|0\rangle & (3.135)\\ &= \langle\psi|\phi\rangle \langle 0|0\rangle & (3.136)\\ &= \langle\psi|\phi\rangle . & (3.137)\end{aligned}$$

As a consequence, we find that

$$\langle\psi|\langle\psi||\phi\rangle|\phi\rangle = \langle\psi|\phi\rangle^2 = \langle\psi|\phi\rangle , \tag{3.138}$$

by employing the above two results. The equality $\langle\psi|\phi\rangle^2 = \langle\psi|\phi\rangle$ holds for exactly two cases, $\langle\psi|\phi\rangle = 1$ and $\langle\psi|\phi\rangle = 0$. The first case holds only when the two states are the same state and the second case holds when the two states are orthogonal to each other. Thus, it is impossible to copy quantum information in any other case because we would contradict unitarity.

The no-cloning theorem has several applications in quantum information processing. First, it underlies the security of the quantum key distribution protocol because it ensures that an attacker cannot copy the quantum states that two parties use to establish a secret key. It finds application in quantum Shannon theory because we can use it to reason about the quantum capacity of a certain quantum channel known as the erasure channel. We will return to this point in Chapter 24.

Exercise 3.5.7 *Suppose that two states $|\psi\rangle$ and $|\psi^\perp\rangle$ are orthogonal: $\langle\psi|\psi^\perp\rangle = 0$. Construct a two-qubit unitary that can copy the states, i.e., find a unitary U that acts as follows: $U|\psi\rangle|0\rangle = |\psi\rangle|\psi\rangle$, $U|\psi^\perp\rangle|0\rangle = |\psi^\perp\rangle|\psi^\perp\rangle$.*

Exercise 3.5.8 (No-Deletion Theorem) *Somewhat related to the no-cloning theorem, there is a no-deletion theorem. Suppose that two copies of a quantum state $|\psi\rangle$ are available, and the goal is to delete one of these states by a unitary interaction. That is, there should exist a universal quantum deleter U that has the following action on the two copies of $|\psi\rangle$ and an ancilla state $|A\rangle$, regardless of the input state $|\psi\rangle$:*

$$U|\psi\rangle|\psi\rangle |A\rangle = |\psi\rangle|0\rangle |A'\rangle , \tag{3.139}$$

where $|A'\rangle$ is another state. Show that this is impossible.

3.5.5 Measurement of Composite Systems

The measurement postulate also extends to composite quantum systems. Suppose again that we have the two-qubit quantum state in (3.99). By a straightforward analogy with the single-qubit case, we can determine the following probability amplitudes:

$$\langle 00|\xi\rangle = \alpha, \quad \langle 01|\xi\rangle = \beta, \quad \langle 10|\xi\rangle = \gamma, \quad \langle 11|\xi\rangle = \delta. \tag{3.140}$$

We can also define the following projection operators:

$$\Pi_{00} \equiv |00\rangle\langle 00|, \quad \Pi_{01} \equiv |01\rangle\langle 01|, \quad \Pi_{10} \equiv |10\rangle\langle 10|, \quad \Pi_{11} \equiv |11\rangle\langle 11|, \tag{3.141}$$

and apply the Born rule to determine the probabilities for each result:

$$\langle\xi|\,\Pi_{00}\,|\xi\rangle = |\alpha|^2, \quad \langle\xi|\,\Pi_{01}\,|\xi\rangle = |\beta|^2, \quad \langle\xi|\,\Pi_{10}\,|\xi\rangle = |\gamma|^2, \quad \langle\xi|\,\Pi_{11}\,|\xi\rangle = |\delta|^2. \tag{3.142}$$

Suppose that we wish to perform a measurement of the Z operator on the first qubit only. What is the set of projection operators that describes this measurement? The answer is similar to what we found for the evolution of a composite system. We apply the identity operator to the second qubit because no measurement occurs on it. Thus, the set of measurement operators is

$$\{\Pi_0 \otimes I, \Pi_1 \otimes I\}, \tag{3.143}$$

where the definition of Π_0 and Π_1 is in (3.73)–(3.74). The state reduces to

$$\frac{(\Pi_0 \otimes I)\,|\xi\rangle}{\sqrt{\langle\xi|\,(\Pi_0 \otimes I)\,|\xi\rangle}} = \frac{\alpha|00\rangle + \beta|01\rangle}{\sqrt{|\alpha|^2 + |\beta|^2}}, \tag{3.144}$$

with probability $\langle\xi|\,(\Pi_0 \otimes I)\,|\xi\rangle = |\alpha|^2 + |\beta|^2$, and reduces to

$$\frac{(\Pi_1 \otimes I)\,|\xi\rangle}{\sqrt{\langle\xi|\,(\Pi_1 \otimes I)\,|\xi\rangle}} = \frac{\gamma|10\rangle + \delta|11\rangle}{\sqrt{|\gamma|^2 + |\delta|^2}}, \tag{3.145}$$

with probability $\langle\xi|\,(\Pi_1 \otimes I)\,|\xi\rangle = |\gamma|^2 + |\delta|^2$. Normalizing by $\sqrt{\langle\xi|\,(\Pi_0 \otimes I)\,|\xi\rangle}$ and $\sqrt{\langle\xi|\,(\Pi_1 \otimes I)\,|\xi\rangle}$ again ensures that the resulting vector corresponds to a physical state.

3.6 Entanglement

Composite quantum systems give rise to a uniquely quantum phenomenon: *entanglement*. Schrödinger first observed that two or more quantum systems can be entangled and coined the term after noticing some of the bizarre consequences of this phenomenon.[6]

[6] Schrödinger actually used the German word “Verschränkung” to describe the phenomenon, which literally translates as “little parts that, though far from one another, always keep the exact same distance from each other.” The one-word English translation is “entanglement.” Einstein described the “Verschränkung” as a “spukhafte Fernwirkung,” most closely translated as “long-distance ghostly effect” or the more commonly stated “spooky action at a distance.”

We first consider a simple, unentangled state that two parties, Alice and Bob, may share, in order to see how an unentangled state contrasts with an entangled state. Suppose that they share the state

$$|0\rangle_A|0\rangle_B, \tag{3.146}$$

where Alice has the qubit in system A and Bob has the qubit in system B. Alice can definitely say that her qubit is in the state $|0\rangle_A$ and Bob can definitely say that his qubit is in the state $|0\rangle_B$. There is nothing really too strange about this scenario.

Now, consider the composite quantum state $|\Phi^+\rangle_{AB}$:

$$\left|\Phi^+\right\rangle_{AB} \equiv \frac{1}{\sqrt{2}}\left(|0\rangle_A|0\rangle_B + |1\rangle_A|1\rangle_B\right). \tag{3.147}$$

Alice again has possession of the first qubit in system A and Bob has possession of the second qubit in system B. But now, it is not clear from the above description how to determine the individual state of Alice or the individual state of Bob. The above state is really a uniform superposition of the joint state $|0\rangle_A|0\rangle_B$ and the joint state $|1\rangle_A|1\rangle_B$, and it is not possible to describe either Alice's or Bob's individual state in the noiseless quantum theory. We also cannot describe the entangled state $|\Phi^+\rangle_{AB}$ as a product state of the form $|\phi\rangle_A|\psi\rangle_B$, for any states $|\phi\rangle_A$ or $|\psi\rangle_B$. This leads to the following general definition:

Definition 3.6.1 (Pure-State Entanglement) *A pure bipartite state $|\psi\rangle_{AB}$ is entangled if it cannot be written as a product state $|\phi\rangle_A \otimes |\varphi\rangle_B$ for any choices of states $|\phi\rangle_A$ and $|\varphi\rangle_B$.*

Exercise 3.6.1 *Show that the entangled state $|\Phi^+\rangle_{AB}$ has the following representation in the $+/-$ basis:*

$$\left|\Phi^+\right\rangle_{AB} = \frac{1}{\sqrt{2}}\left(|+\rangle_A|+\rangle_B + |-\rangle_A|-\rangle_B\right). \tag{3.148}$$

Figure 3.8 gives a graphical depiction of entanglement. We use this depiction often throughout this book. Alice and Bob must receive the entanglement in some way, and the diagram indicates that some source distributes the entangled pair to them. It indicates that Alice and Bob are spatially separated and they possess the entangled state after some time. If they share the entangled state in (3.147), we say that they share one bit of entanglement, or one *ebit*. The term "ebit" implies that there is some way to quantify entanglement and we will make this notion clear in Chapter 19.

3.6.1 Entanglement as a Resource

In this book, we are interested in the use of entanglement as a resource. Much of this book concerns the theory of quantum information-processing resources and we have a standard notation for the theory of resources. Let us represent the resource of a shared ebit as

$$[qq], \tag{3.149}$$

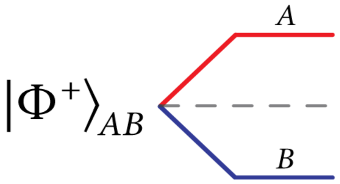

Figure 3.8: We use the above diagram to depict entanglement shared between two parties A and B. The diagram indicates that a source location creates the entanglement and distributes one system to A and the other system to B. The standard unit of entanglement is the ebit $|\Phi^+\rangle_{AB} \equiv (|00\rangle_{AB} + |11\rangle_{AB})/\sqrt{2}$.

meaning that the ebit is a noiseless, quantum resource shared between two parties. Square brackets indicate a noiseless resource, the letter q indicates a quantum resource, and the two copies of the letter q indicate a two-party resource.

Our first example of the use of entanglement is its role in generating *shared randomness.* We define one bit of shared randomness as the following probability distribution for two binary random variables X_A and X_B:

$$p_{X_A,X_B}(x_A, x_B) = \frac{1}{2}\delta(x_A, x_B), \tag{3.150}$$

where δ is the Kronecker delta function. Suppose Alice possesses random variable X_A and Bob possesses random variable X_B. Thus, with probability 1/2, they either both have a zero or they both have a one. We represent the resource of one bit of shared randomness as

$$[cc]\,, \tag{3.151}$$

indicating that a bit of shared randomness is a noiseless, classical resource shared between two parties.

Now suppose that Alice and Bob share an ebit and they decide that they will each measure their qubits in the computational basis. Without loss of generality, suppose that Alice performs a measurement first. Thus, Alice performs a measurement of the Z_A operator, meaning that she measures $Z_A \otimes I_B$ (she cannot perform anything on Bob's qubit because they are spatially separated). The projection operators for this measurement are the same from (3.143), and they project the joint state. Just before Alice looks at her measurement result, she does not know the outcome, and we can describe the system as being in the following ensemble of states:

$$|0\rangle_A|0\rangle_B \text{ with probability } \frac{1}{2}, \tag{3.152}$$

$$|1\rangle_A|1\rangle_B \text{ with probability } \frac{1}{2}. \tag{3.153}$$

The interesting thing about the above ensemble is that Bob's result is already determined even before he measures, just after Alice's measurement occurs. Suppose that Alice knows the result of her measurement is $|0\rangle_A$. When Bob measures his system, he obtains the state