Exploring Practical Methodologies for the Characterization and Control of Small Quantum Systems

by

Ian Hincks

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Applied Mathematics

Waterloo, Ontario, Canada, 2018

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Kenneth Brown Department of Electrical and Computer Engineering, Duke University
Supervisor:	David Cory Department of Chemistry, University of Waterloo Joseph Emerson Department of Applied Math, University of Waterloo
Internal Member:	Achim Kempf Department of Applied Math, University of Waterloo
Internal-External Member:	Crystal Senko Department of Physics, University of Waterloo
Other Member(s):	Vern Paulsen Deptartment of Pure Math, University of Waterloo

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

We explore methodologies for characterizing and controlling small quantum systems. We are interested in starting with a description of a quantum system, designing estimators for parameters of the system, developing robust and high-fidelity gates for the system using knowledge of these parameters, and experimentally verifying the performance of these gates. A strong emphasis is placed on using rigorous statistical methods, especially Bayesian ones, to analyze quantum system data. Throughout this thesis, the Nitrogen Vacancy system is used as an experimental testbed. Characterization of system parameters is done using quantum Hamiltonian learning, where we explore the use of adaptive experiment design to speed up learning rates. Gates for the full three-level system are designed with numerical optimal control methods that take into account imperfections of the control hardware. Gate quality is assessed using randomized benchmarking protocols, including standard randomized benchmarking, unitarity benchmarking, and leakage/loss benchmarking.

Acknowledgements

I am deeply grateful for the support from many people throughout the course of my graduate studies. I am thankful to my supervisor David Cory, for his inspiration, kindness and seemingly endless research ideas, and to my co-supervisor Joseph Emerson, who taught the memorable quantum theory course that first set me off along this path. It has been a privilege to work with the changing members of my research group over the years, with special mention to Pat Gumann, Sarah Sheldon, Troy Borneman, Chris Wood, Josh Combes, Holger Haas, Joachim Nsofini, Dusan Sarenac, Maryam Mirkamali, Mohamad Niknam, Carly Turnbull, the NV group including Madelaine Liddy, Zimeng Wang, Romain Ruhlmann, and my past office-mates Om Patange and Thomas Alexander. I single out Chris Granade, who continually inspires me both in research and in character, and Dan Puzzuoli for his lasting friendship. I am indebted to those who have acted as mentors, in particular Osama Moussa and Chris Ferrie. I have appreciated being a member of the Institute for Quantum Computing, which has provided an excellent venue for learning, and whose high quality courses have been particularly valuable.

Lastly, I thank my friends, my parents and siblings, my wife Miriam and our son Ezra, whose collective support has been my bedrock.

To Amandil

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Two important aspects of developing quantum devices that push beyond the capabilities of their classical counterparts—such as quantum sensors, quantum information processors, and quantum actuators—include gaining high-fidelity coherence control of their dynamics, and efficiently probing their properties through quantum measurement. The goal of this thesis is to explore practical methods for performing these two tasks with a focus on small quantum systems currently available. A central theme will be the relationship between the complexity of systems and their practical utility. For instance, we study the benefit that adaptive experiment design can have on magnetometry, the cost of using a non-linear resonator to control a quantum device, and the quality of qutrit gatesets composed of numerically optimized control pulses.

To test our methodologies, we will use a quantum system called a Nitrogen Vacancy (NV) defect in diamond as a testbed on which to apply ideas, and so several chapters will include detailed information about the dynamics of this system. Throughout, we place emphasis on applied statistical inference, and in particular Bayesian methods, to analyze data from quantum experiments. Therefore, this introductory chapter contains relevant background information on quantum dynamics and some central concepts of statistical inference.

1.1 Thesis Outline

Figure 1.1 shows the natural organization of the chapters in this thesis, which are summarized below. Chapter 1, the current introductory chapter, overviews notation and concepts

used throughout the thesis. It consists of three sections, which respectively introduce quantum theory and quantum dynamics, statistical inference, and bayesian data analysis.

Chapter 2 introduces the Nitrogen Vacancy (NV) defect in diamond as an experimental platform for quantum information processing (QIP). Its ground state Hamiltonian is defined, and numerically-ammenable methods for entering rotating frames and taking average stroboscopic approximations are provided in detail. Optical dynamics of the system are introduced, and used to explore the process of room temperature measurment, including dark counts and finite visibility. These dynamics are also used to decribe state initialization of the system. Finally, the experimental setup is briefly presented. This section, in particular the optical dynamics, measurement and initialization are based partly on Reference [84], work done in collaboration with Chris Granade. This section also derives certain block structures to help with average Hamiltonian theory, which is unpublished work done in collaboration with Romain Ruhlmann.

Chapter 3 takes the NV measurement description and considers it in the framework of statistical inference. The risk of frequentist and Bayesian estimators are compared against each other where we find a slight advantage in the latter. We discuss numerical methods for integrating measurement data into more involved inference tasks like tomography. The section concludes with a detailed example where we perform quantum Hamiltonian learning with Bayesian inference, implemented with a sequential Monte Carlo sampler, using data from an experimental system. Quantum Hamiltonian learning is the task of estimating parameters from a system's Hamiltonian given data from the system, and so Bayesian methods return a posterior distribution over some parameterization of Hamiltonians. This section is based on the remaining part of Reference [84].

Chapter 4 seeks to improve the Hamiltonian learning example from the previous section by employing online experiment design. Estimating parameters of quantum systems is usually done by performing a sequence of predetermined experiments and post-processing the resulting data. It is known that online design, where the choice of the next experiment is based on the most up-to-date knowledge about the system, can offer speedups to parameter estimation. We use risk-based Bayesian experiment design to learn the values of a five-parameter model describing its Hamiltonian and decoherence process. Comparing this to standard pre-determined experiment sweeps, we find that we can achieve median posterior variances on some parameters that are between 10 and 100 times better given the same amount of data. This has applications to NV magnetometry where one of the Hamiltonian coefficients is the parameter of interest. This chapter is based on Reference [83], done in collaboration with Thomas Alexander, Michal Kononenko, and Benjamin Soloway.

Chapter 5 moves away from inference and toward quantum control, where the primary

interest is gaining high-fidelity control in the presence of imperfect classical hardware. This is done by itroducing a framework of distortion operators, and detailing how they can be integrated with existing numerical optimal control routines like gradient-ascent pulse engineering (GRAPE). We provide examples of distortions for many types of common hardware imperfections, including transfer functions and cross talk. The main example presented discusses controlling a qubit by a non-linear resonator, whose inductance depends on the amount of current present within. We show that it is possible to optimize gates in this sequence, and moreover, it is possible to include ringdown compensation to drive the power out of the resonator when the pulse is complete. This chapter is based on Reference [82], done in collaboration with Chris Granade, and Troy Borneman.

Chapter 6 is concerned with the data analysis of randomized benchmarking (RB) and related protocols (together denoted by RB+). RB+ protocols are standard tools for characterizing quantum devices. Prior analyses of RB protocols have not provided a complete method for analyzing realistic data, resulting in a variety of ad-hoc methods. The main confounding factor in rigorously analyzing data from RB protocols is an unknown and noise-dependent distribution of survival probabilities over random sequences. We propose a hierarchical Bayesian method where these survival distributions are modeled as nonparametric Dirichlet process mixtures. Our method infers parameters of interest without additional assumptions about the underlying physical noise process. We show with numerical examples that our method works robustly for both standard and highly pathological error models. Our method also works reliably at low noise levels and with little data because we avoid the asymptotic assumptions of commonly used methods such as least-squares fitting. For example, our method produces a narrow and consistent posterior for the average gate fidelity from ten random sequences per sequence length in the standard RB protocol. This work is based on Reference [85], done in collaboration with Joel Wallman, Chris Ferrie, and Chris Granade.

Chapter 7 combines ideas from the previous chapters to test our ability to design high-fidelity quantum gates for the NV system, while laying out a methodology to do so. We find a unitary 2-design of order 72 to be used with randomized benchmarking protocols on a qutrit. Gates for a this gateset, as well as for a qubit 2-design group, are designed with GRAPE, taking into account circuit distortions and power constraints. These gates are performed experimentally in RB+ protocols, namely, standard randomized benchmarking, interleaved randomized benchmarking, leakage benchmarking, and unitarity benchmarking. We find average gate fidelities of around $98 - 99\%$ for these pulses. Pulse shapes are then tuned-up using classical closed-loop feedback with an oscilliscope to match the desired profile. We find that this process improves average gate fidelities to $> 99.5\%$ for qubits. This work is to be published, and was performed in collaboration with Chris Granade, and

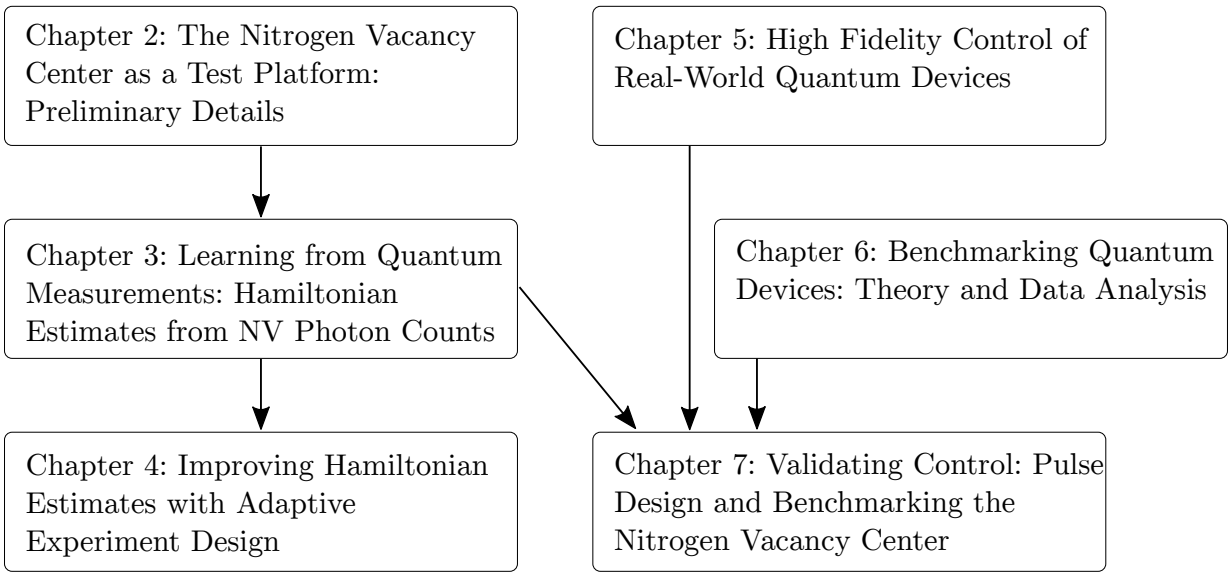

Figure 1.1: The organization of this thesis.

Joel Wallman.

Finally, Chapter 8 makes some concluding remarks.

1.2 Quantum Systems

This section briefly describes the postulates and mechanics of finite-dimensional quantum theory, with the main goal of introducing the notation used throughout this thesis.

This section, and indeed this thesis, presupposes some basic knowledge of linear algebra, such as understanding of vector spaces, norms, inner products, operators acting on vectorspaces, the vector spaces they form in turn, their decompositions into eigenstructures, and so on. Standard linear algebra will introduce these concepts, and we in particular promote the introductory sections of Reference [183] (available free online).

Quantum mechanics takes place, at least mathematically, within a seperable Hilbert space over the complex numbers, labelled $\mathcal{H}$. A Hilbert space is an inner-product space such that the limit of every Cauchy sequence of vectors (under the induced metric of the inner product) itself exists within the space. Separability is a topological constraint which, in Hilbert spaces, is equivalent to the existence of an orthonormal basis that can be

indexed by the natural numbers. However, since we are restricting our attention to finite-dimensional spaces in this thesis, i.e. $\dim \mathcal{H} = d < \infty$, seperability and completeness come for free, so that a Hilbert space for us is simply an inner product space—our convention of keeping Hilbert's name attached in this simpler setting is nearly universal in the field.

1.2.1 Quantum Theory Notation

We will find little advantage in working with abstract vector spaces, and therefore always consider the concrete vector space $\mathcal{H} = \mathbb{C}^d$ over $\mathbb{C}$ for some dimension d, with the standard complex inner product between two column vectors $(x_1, \ldots, x_d)^T, (y_1, \ldots, y_d)^T \in \mathcal{H}$ given by $\sum_{i=1}^d \overline{x}_i y_i$. In an often convenient tradition beginning with Dirac [36], we use the notation $|\cdot\rangle$ (called a *ket*) to denote a member of a Hilbert space $\mathcal{H}$. For instance, $|\psi\rangle, |\phi\rangle \in \mathcal{H}$ might be two different vectors, and canonical basis elements are often labeled by integers, $|0\rangle, |1\rangle, \ldots, |\dim \mathcal{H} - 1\rangle$.

Members of the adjoint space $\mathcal{H}^\dagger$—formally being the set of linear functions from $\mathcal{H}$ into $\mathbb{C}$—are written using the notation $\langle\cdot|$ (called a *bra*), and are best considered as row vectors, whereas members of $\mathcal{H}$ are column vectors. Riesz' one-to-one correspondence between vectors and their duals is therefore made, in our concrete case, by conjugation and transposition from a row vector into a column vector, or vice versa. That is, if $|\psi\rangle \in \mathcal{H}$ is a column vector, then $\langle\psi| = |\psi\rangle^\dagger \in \mathcal{H}^\dagger$ is the correspnoding dual vector, whose row elements are the complex conjugates of the column elements of $|\psi\rangle$.

One of the main advantages of Dirac's notation is that the application of a dual vector $\langle\psi| \in \mathcal{H}^\dagger$ to a vector $|\phi\rangle \in \mathcal{H}$, namely $\langle\psi|(|\psi\rangle)$, looks very similar to the usual notation of an inner-product, $\langle\cdot,\cdot\rangle$, when the parentheses are removed. Indeed, it is easy to see that this application *is* equal to the inner product between $|\psi\rangle$ and $|\phi\rangle$, and is therefore condensed into the notation $\langle\psi|\phi\rangle$.

Example 1.1. Suppose that $|1\rangle, \ldots, |d\rangle \in \mathcal{H}$ is an orthonormal basis and $|\psi\rangle \in \mathcal{H}$. Since $|1\rangle, \ldots, |d\rangle$ is a basis there must exist coefficients $x_i \in \mathbb{C}$ such that $|\psi\rangle = \sum_{i=1}^d x_i |i\rangle$, and since it is orthornormal, we must have the inner-products $\langle i|j\rangle = \delta_{i,j}$. Then for any $1 \leq j \leq d$ we have

$$\langle j|\psi\rangle = \langle j| \left(\sum_{i=1}^d x_i |i\rangle \right) = \sum_{i=1}^d x_i \langle j|i\rangle = x_j \tag{1.1}$$

hence the decomposition $|\psi\rangle = \sum_{i=1}^d \langle i|\psi\rangle |i\rangle$. ⊣

Given a Hilbert space $\mathcal{H} = \mathbb{C}^d$, we denote the set of linear maps from $\mathcal{H}$ into itelf by $\mathcal{L}(\mathcal{H})$, which we conflate with the set of $d \times d$ matrices with complex entries. The set $\mathcal{L}(\mathcal{H})$ itself forms an inner product space with the Hilbert-Schmidt inner product

$$\langle A, B \rangle = \mathrm{Tr}(A^\dagger B) \tag{1.2}$$

where $A, B \in \mathcal{L}(\mathcal{H})$ and $\cdot^\dagger$ is complex conjugation.

Note that if $|\psi\rangle, |\phi\rangle \in \mathcal{H}$, then $|\psi\rangle\langle\phi| \in \mathcal{L}(\mathcal{H})$ is an operator. We call this the *outer product* of $|\psi\rangle$ and $|\phi\rangle$. Indeed, $\langle\phi|$ is a functional that maps vectors in $\mathcal{H}$ to complex numbers, so that when multiplied by $|\psi\rangle$, as in $|\psi\rangle\langle\phi|$, the result is a vector in $\mathcal{H}$; $|\psi\rangle\langle\phi|$ is a map from vectors to vectors. Therefore a third state, $|\alpha\rangle \in \mathcal{H}$, applied to $|\psi\rangle\langle\phi|$ gives $|\psi\rangle\langle\phi|(|\alpha\rangle) = |\psi\rangle\langle\phi|\alpha\rangle = \langle\phi|\alpha\rangle|\psi\rangle$, where we can put $\langle\phi|\alpha\rangle$ on the left to emphasize the inner product, or we can put $|\psi\rangle\langle\phi|$ on the left to emphasize the outer product. Perhaps a simpler explanation is just that $|\psi\rangle\langle\phi|$ is the matrix product of a column vector with a row, which gives a rank-1 matrix, where all matrix mechanics apply as usual.

Example 1.2. Let $|1\rangle, \ldots, |d\rangle \in \mathcal{H}$ be the canonical basis, so that $|i\rangle$ has a 1 in its i^{th} column vector entry, and zeros elsewhere. Then $|i\rangle\langle j|$ is a matrix with a 1 in the entry (i, j), and zeros elsewhere. Thus any matrix $A \in \mathcal{L}(\mathcal{H})$ can be expanded as $A = \sum_{i,j=1}^d A_{ij}|i\rangle\langle j|$. Observe, for example, that

$$A^\dagger = \left(\sum_{i,j=1}^d A_{ij}|i\rangle\langle j|\right)^\dagger = \sum_{i,j=1}^d \overline{A_{ij}}\langle j|^\dagger |i\rangle^\dagger = \sum_{i,j=1}^d \overline{A_{ij}}|j\rangle\langle i| = \sum_{i,j=1}^d \overline{A_{ji}}|i\rangle\langle j|, \tag{1.3}$$

as expected. ⊣

Example 1.3. Let $A \in \mathcal{L}(\mathcal{H})$ be a normal matrix, $AA^\dagger = A^\dagger$. Then it admits eigenvalues $\lambda_1, \ldots, \lambda_d \in \mathbb{C}$ with corresponding eigenvectors $|\psi_1\rangle, \ldots, |\psi_d\rangle$ that form an orthonormal basis. For any $|\psi\rangle \in \mathcal{H}$ we have

$$\begin{aligned} A|\psi\rangle &= A\sum_{i=1}^d \langle\psi_i|\psi\rangle|\psi_i\rangle = \sum_{i=1}^d \langle\psi_i|\psi\rangle A|\psi_i\rangle = \sum_{i=1}^d \lambda_i\langle\psi_i|\psi\rangle|\psi_i\rangle \\ &= \sum_{i=1}^d \lambda_i|\psi_i\rangle\langle\psi_i|\psi\rangle = \left(\sum_{i=1}^d \lambda_i|\psi_i\rangle\langle\psi_i|\right)|\psi\rangle \end{aligned} \tag{1.4}$$

which shows the notation for a spectral decomposition in Dirac's notation:

$$A = \sum_{i=1}^d \lambda_i|\psi_i\rangle\langle\psi_i|$$

. ⊣

The following subsets of $\mathcal{L}(\mathcal{H})$ are commonly used.

Definition 1.1. The set of *unitary operators* $\mathcal{U}(\mathcal{H}) = \{U \in \mathcal{L}(\mathcal{H}) | U^\dagger U = \mathbb{I}\}$, of which the set of *special unitary operators* is a subset, $\mathcal{SU}(\mathcal{H}) = \{U \in \mathcal{U}(\mathcal{H}) | \det U = 1\}$.

Theorem 1. *Suppose $U \in \mathcal{L}(\mathcal{H})$. The following are equivalent:*

1. $U \in \mathcal{U}(\mathcal{H})$
2. $U^\dagger U = U U^\dagger = \mathbb{I}$
3. $\forall |\psi\rangle, |\phi\rangle \mathcal{H}$, $\langle\psi|\phi\rangle = \langle\psi'|\phi'\rangle$ *where* $|\psi'\rangle = U|\psi\rangle$ *and* $|\phi'\rangle = U|\phi\rangle$
4. $\exists \phi_1, \ldots, \phi_d \in \mathbb{R}$, $|\psi_1\rangle, \ldots, |\psi_d\rangle \in \mathcal{H}$ *an orthonormal basis, such that*

$$U = \sum_{i=1}^{d} \mathrm{e}^{\mathrm{i}\phi_i} |\psi_i\rangle\langle\psi_i|$$

Definition 1.2. The set of *hermitian operators* $\mathfrak{u}(\mathcal{H}) = \{H \in \mathcal{L}(\mathcal{H}) | H = H^\dagger\}$, of which the set of *traceless hermition operators* is a subset, $\mathfrak{su}(\mathcal{H}) = \{H \in \mathfrak{u}(\mathcal{H}) | \operatorname{Tr} H = 0\}$. Both of these collections are vector subspaces of $\mathcal{L}(\mathcal{H})$.

Theorem 2. *Suppose $H \in \mathcal{L}(\mathcal{H})$. The following are equivalent:*

1. $H \in \mathfrak{u}(\mathcal{H})$
2. $H = H^\dagger$
3. $\langle\psi| H |\psi\rangle \in \mathbb{R} \forall |\psi\rangle \mathcal{H}$
4. $\langle\psi| H |\phi\rangle = (H|\psi\rangle)^\dagger |\phi\rangle \, \forall |\psi\rangle, |\phi\rangle \in \mathcal{H}$
5. $\exists \lambda_1, \ldots, \lambda_d \in \mathbb{R}$, $|\psi_1\rangle, \ldots, |\psi_d\rangle \in \mathcal{H}$ *an orthonormal basis, such that*

$$H = \sum_{i=1}^{d} \lambda_i |\psi_i\rangle\langle\psi_i|$$

Example 1.4. If $\mathrm{i}H \in \mathfrak{isu}(\mathcal{H})$, then we may obtain a decomposition $\mathrm{i}H = \sum_{i=1}^{d} \mathrm{i}\lambda_i |\psi_i\rangle\langle\psi_i|$ as in Theorem 2. Because $\operatorname{Tr} H = 0$, we see that $\sum_{i=1}^{d} \lambda_i = 0$. The matrix exponential of $\mathrm{i}H$ obtained by applying the scalar exponential to each of the eigenvalues; we have

$$\mathrm{e}^{\mathrm{i}H} = \sum_{i=1}^{d} \mathrm{e}^{\mathrm{i}\lambda_i} |\psi_i\rangle\langle\psi_i| \tag{1.5}$$

where it is seen, by Theorem 1, that $e^{iH} \in \mathcal{UH}$. Since the determinant is the product of the eigenvalues, we have moreover that $\det e^{iH} = e^{i\sum_{i=1}^{d} \lambda_i} = 0$, so that $e^{iH} \in \mathcal{SUH}$. This shows that $e^{i\mathfrak{su}(\mathcal{H})} \subseteq \mathcal{SU}(\mathcal{H})$. In fact, it holds that the reverse containment is true, so that any special unitary operator can be written as the exponential of some traceless hermitian operator times i.

Similarly, $e^{i\mathfrak{u}(\mathcal{H})} = \mathcal{U}(\mathcal{H})$. ⊣

Definition 1.3. The set of *positive semi-definite* operators

$$\mathcal{P}(\mathcal{H}) = \{P \in \mathfrak{u}(\mathcal{H}) | \langle\psi| P |\psi\rangle \geq 0 \forall |\psi\rangle \in \mathcal{H}\}. \tag{1.6}$$

We write $P \geq 0$ to denote $P \in \mathcal{P}(\mathcal{H})$, and $P \geq Q$ to denote $P - Q \geq 0$.

Theorem 3. *Suppose $P \in \mathcal{L}(\mathcal{H})$. The following are equivalent:*

1. $P \in \mathcal{P}(\mathcal{H})$
2. $P = AA^\dagger$ *for some* $A \in \mathcal{L}(\mathcal{H})$
3. $\langle\psi| P |\psi\rangle \geq 0, \forall |\psi\rangle \mathcal{H}$
4. $\exists \vec{p}_1, \ldots, p_d \geq 0$, $|\psi_1\rangle, \ldots, |\psi_d\rangle \in \mathcal{H}$ *an orthonormal basis, such that*

$$P = \sum_{i=1}^{d} p_i |\psi_i\rangle\langle\psi_i| .$$

There are two important ways to join vector spaces together to make larger vector spaces. The first is simply the cartesian product of the two spaces, and linear operations on this space can be thought of in terms of block matrices.

Definition 1.4. Given two Hilbert spaces $\mathcal{H}_1 = \mathbb{C}^{d_1}$ and $\mathcal{H}_2 = \mathbb{C}^{d_2}$, the *direct sum* $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$ is the inner product space $\mathbb{C}^{d_1+d_2}$. If $|\psi_1\rangle \in \mathcal{H}_1$ and $|\psi_2\rangle \in \mathcal{H}_2$, then $|\psi_1\rangle \oplus |\psi_2\rangle$ is notation for the column stacking of the two vectors.

The second, which is of fundamental importance in quantum theory, is the tensor product, which results in a space whose dimension is the product of the constituent spaces, rather than the sum.

Definition 1.5. Given two Hilbert spaces $\mathcal{H}_1 = \mathbb{C}^{d_1}$ and $\mathcal{H}_2 = \mathbb{C}^{d_2}$, the *tensor product space* $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$ is the inner product space $\mathbb{C}^{d_1 d_2}$. Given $|\psi_1\rangle = (x_1, x_2, \ldots, x_{d_1})^T \in \mathcal{H}_1$ and $|\psi_2\rangle = (y_1, y_2, \ldots, y_{d_2})^T \in \mathcal{H}_2$, then $|\psi_1\rangle \otimes |\psi_2\rangle$ is notation for the column vector

$$|\psi_1\rangle \otimes |\psi_2\rangle = (x_1 \cdot (y_1, y_2, \ldots, y_{d_2}), x_2 \cdot (y_1, y_2, \ldots, y_{d_2}), \ldots, x_{d_1} \cdot (y_1, y_2, \ldots, y_{d_2}))^T. \quad (1.7)$$

This form satisfies bilinearity; we have

$$\begin{aligned}&(|\psi_1\rangle + \alpha\,|\phi_1\rangle) \otimes (|\psi_2\rangle + \beta\,|\phi_2\rangle) = \\ &|\psi_1\rangle \otimes |\psi_2\rangle + \alpha\,|\phi_1\rangle \otimes |\psi_2\rangle + \beta\,|\psi_1\rangle \otimes |\phi_2\rangle + \alpha\beta\,|\phi_1\rangle \otimes |\phi_2\rangle\,. \end{aligned} \quad (1.8)$$

The tensor product of linear operator spaces functions similarly. If $A \in \mathcal{L}(\mathcal{H}_1)$ and $B \in \mathcal{L}(\mathcal{H}_2)$, then $A \otimes B \in \mathcal{L}(\mathcal{H}_1) \otimes \mathcal{L}(\mathcal{H}_2)$ is a matrix with entries

$$A \otimes B = \begin{pmatrix} A_{11}B & A_{12}B & \ldots \\ A_{21}B & A_{22}B & \ldots \\ \vdots & \vdots & \ddots \end{pmatrix}. \quad (1.9)$$

It holds that $(A \otimes B)(|\psi_1\rangle \otimes |\psi_2\rangle) = A\,|\psi_1\rangle \otimes B\,|\psi_2\rangle$.

Example 1.5. Suppose $\mathcal{H}_1 = \mathbb{C}^3$ and $\mathcal{H}_2 = \mathbb{C}^2$, with respective canonical orthonormal bases written as $|0\rangle, |1\rangle, |2\rangle$ and $|0\rangle, |1\rangle$. Then $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2 = \mathbb{C}^6$, and we have, for example,

$$\begin{aligned}\frac{|0\rangle + |2\rangle}{\sqrt{2}} \otimes |0\rangle &= \begin{pmatrix} 1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} \end{pmatrix} \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} 1/\sqrt{2} & 0 & 0 & 0 & 1/\sqrt{2} & 0 \end{pmatrix}^T \\ &= (|0\rangle \otimes |0\rangle + |2\rangle \otimes |0\rangle)/\sqrt{2} \end{aligned} \quad (1.10)$$

⊣

1.2.2 Postulates of Quantum Theory

Entities existing in the physical world, a phrase we leave nebulous on-purpose, can be described by vectors in a Hilbert space. However, its operational meaning will hopefully be made clearer once we introduce the time-evolution and measurement of these states, along with a few examples. We begin with the definition of the simplest quantum object.

Definition 1.6. A *pure state* is a vector $|\psi\rangle \in \mathcal{H}$ with unit norm, $\langle\psi|\psi\rangle = 1$.

This definition is generally insufficient because, due to the probabalistic nature of quantum mechanices which we will see, it is very common to want to talk about probabilistic combinations of pure states. For example, one may have an entity that should described by the pure state $|\psi\rangle$ but only with probability $p \in [0,1]$, and should described by another pure state $|\phi\rangle$ with the remaining probability $1-p$. It turns out that such a description of a system, which is a perfectly reasonable demand, cannot be described by any single vector in $\mathcal{H}$, at least not with respect to an arbitrary choice of measurement. This leads us to the following generalization.

Definition 1.7. A *mixed state* (or *density matrix*, or often simply *state*) is an operator $\rho \in \mathcal{L}(\mathcal{H})$ that is positive-semidefinite, $\rho \geq 0$, with unit trace, $\operatorname{Tr}\rho = 1$. We write $\mathcal{D}(\mathcal{H}) = \{X \in \mathcal{P}(\mathcal{H}) | \operatorname{Tr} X = 1\}$ for the set of density matrices. If a mixed state ρ is rank-one, so that there exists some $|\psi\rangle \in \mathcal{H}$ such that $\rho = |\psi\rangle\langle\psi|$, then we may sometimes refer to ρ as a pure state even though it is not a vector.

Example 1.6. If $\rho_1, \ldots, \rho_n \in \mathcal{D}(\mathcal{H})$ are density matrices and $p_1, \ldots, p_n \in [0,1]$ form a probability vector, so that $\sum_{i=1}^n p_i = 1$, then $\rho = \sum_{i=1}^n p_i \rho_i$ is also a density matrix. Indeed, the sum of positive-semidefinite operators is positive-semidefinite, and $\operatorname{Tr}\rho = \sum_{i=1}^n p_i \operatorname{Tr}\rho_i = 1$. ⊣

Example 1.7. Fix a mixed state $\rho \in \mathcal{D}(\mathcal{H})$. Since positivity implies normality, the spectral theorem gives us the eigendecomposition

$$\rho = \sum_{i=1}^n p_i |i\rangle\langle i| \tag{1.11}$$

where $d = \dim \mathcal{H}$, $\sum_{i=1}^d p_i = 1$, and $|1\rangle, \ldots, |d\rangle$ forms an orthonormal basis for $\mathcal{H}$. Thus, we may interpret our system as having the state $|1\rangle$ with probability p_1, the state $|2\rangle$ with probability p_2, and so on. However, there may exist some other decomposition (and usually there will be infinitely many such), say

$$\rho = \sum_{i=1}^n q_i |\psi_i\rangle\langle\psi_i|, \tag{1.12}$$

which is different from the first in a non-trivial way. This is an equally valid intepretation of the state, as no experiment will be able to distinguish it from the first. ⊣

Actions on pure states transform a given pure state into another pure state. Supposing such operations are linear, they must be isometric to preserve the unit length (so that outputs are pure states), and hence unitary, since all isometric operators are unitary in finite dimensions. Hence given a unitary matrix $U \in \mathcal{U}(\mathcal{H})$, we might have the transformation of pure states $|\psi\rangle \mapsto U|\psi\rangle$. Therefore, in terms of dual pure states we must have the mapping $\langle\psi| \mapsto \langle\psi| U^\dagger$ for consistency. This, in turn, implies that the rank-1 density matrix $|\psi\rangle\langle\psi|$ transforms as $|\psi\rangle\langle\psi| \mapsto U|\psi\rangle\langle\psi|U^\dagger$ under the operation U. To be consistent with linearity, we infer that a density matrix $\rho \in \mathcal{D}(\mathcal{H})$ of any rank must transform as $\rho \mapsto U\rho U^\dagger$ inder the operation U.

However, just as we claimed pure states were insufficient to describe reasonably motivated situations above, likewise unitary transformations of density matrices are insufficient to describe transformations of mixed states. We may have, for example, a situation where an operation $U \in \mathcal{U}(\mathcal{H})$ is applied with probability p, and another operation $W \in \mathcal{U}(\mathcal{H})$ is applied with probability $(1-p)$. Such an operation would map a density matrix as $\rho \mapsto pU\rho U^\dagger + (1-p)W\rho W^\dagger$, which is not generally describable by a single unitary. Therefore, we consider more generally the set of linear operations acting on on $\mathcal{L}(\mathcal{H})$ itself.

Definition 1.8. We call a linear operator $\Lambda : \mathcal{L}(\mathcal{H}) \to \mathcal{L}(\mathcal{H})$ a *superoperator*, denoting the space of all superoperators as $\mathcal{T}(\mathcal{H}) = \mathcal{L}(\mathcal{L}(\mathcal{H}))$.

For a superoperator $\Lambda \in \mathcal{T}(\mathcal{H})$ to preserve mixed states, it must be *positivity preserving*, $\rho \in \mathcal{P}(\mathcal{H}) \implies \Lambda(\rho) \in \mathcal{P}(\mathcal{H})$, and it must be *trace preserving*, $A \in \mathcal{L}(\mathcal{H}) \implies \operatorname{Tr}\Lambda(A) = \operatorname{Tr} A$. Preserving positivity of states is not quite enough, it turns out, and we also need to worry about preserving the positivity when joint to some other hypothetical system of arbitrary size. We also want a notion of time evolution of such maps.

Definition 1.9. We call a superoperator $\Lambda \in \mathcal{T}(\mathcal{H})$ *completely positive* if for any $d' \in \mathbb{N}$ and any $P \in \mathcal{P}(\mathcal{H} \otimes \mathcal{H}')$ it holds that $(\Lambda \otimes \mathbb{I}_{d'})(P) \in \mathcal{P}(\mathcal{H} \otimes \mathcal{H}')$, where $\mathbb{I}_{d'}$ is the identity operator in $\mathrm{T}(\mathcal{H}')$ and $\mathcal{H}' = \mathbb{C}^{d'}$. We denote the set of all completely positive trace preserving maps as $\mathcal{C}(\mathcal{H})$, variously calling them *CPTP maps*, *quantum channels*, or just *channels*.

Definition 1.10. Suppose $\Lambda_t \in \mathcal{C}(\mathcal{H})$ is CPTP for each $t \in \mathbb{R}$, and that the following conditions hold:

1. Λ_t is continuous with respect to t
2. $\Lambda_t \circ \Lambda_s = \Lambda_{s+t}, \forall s, t \in \mathbb{R}$

then we call $\{\Lambda_t\}_{t\in\mathbb{R}}$ a *continuous dynamical semigroup*. (Since we are working in finite dimensions, we don't need to specify the type of continuity—they're all the same.)

Finally, we move to measurement of quantum states. Fixing an orthonormal basis $|\psi_1\rangle, \ldots, |\psi_d\rangle$ for a Hilbert space $\mathcal{H}$, a pure state $|\psi\rangle$ can be interpreted as a discrete probability distribution, where the indices $1, 2, \ldots, d$ are the outcomes, and the values $p_i = |\langle\psi_i|\psi\rangle|^2$ are their respective probabilities. This is possible because $|\psi\rangle$, being a pure state, has unit norm. We see that the outcomes correspond to different regions of the state space $\mathcal{H}$, and the probabilities are based on the amplitude of the projection of $|\psi\rangle$ into that subspace. Generalizing this idea to density matrices, and taking into account other desired features such as the ability to bin multiple directions into single events, or feature some outcomes very weakly, we have the following description of a measurement.

Definition 1.11. A *positive operator valued measurement* (POVM) is a set $\{E_i\}_{i=1}^n \subset \mathcal{P}(\mathcal{H})$ such that $\sum_{i=1}^n E_i = \mathbb{I}$. In the special case each E_i is a hermitian projector, $E_i \in \mathfrak{u}(\mathcal{H})$ and $E_i E_i = E_i$, we call $\{E_i\}_{i=1}^n$ a *projection valued measurement* (PVM). Given a density matrix $\rho \in \mathcal{D}(\mathcal{H})$, the discrete probability distribution associated with this POVM with outcome events $1, 2, \ldots, n$ is given by $\Pr(i) = p_i = \operatorname{Tr}(E_i\rho)$. Note that condition $\sum_{i=1}^n E_i = \mathbb{I}$ (along with $\operatorname{Tr}\rho = 1$) ensures that $\sum_{i=1}^n p_i = 1$, and the positivity of each E_i (along with ρ) ensures that each $p_i \geq 0$. Also note that $n \neq d$ in general, and we may have both $n < d$ and $n > d$.

The motivating prose and definitions laid out above are summarized in the following four postulates of quantum theory.

Postulate 1. *States.* The state of a physical system at time t is represented by a density matrix $\rho(t) \in \mathcal{D}(\mathcal{H})$ acting on the Hilbert space $\mathcal{H}$.

Postulate 2. *Composition.* If two distinct physical systems (or two distinct aspects of the same physical system) are described in Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, then states of the joint system are represented in the Hilbert space $\mathcal{H}_1 \otimes \mathcal{H}_2$.

Postulate 3. *Evolution.* States in a static environment evolve in time via continuous dynamical semigroups; if $\{\Lambda_t\}_t$ is a continuous dynamical semigroup describing evolution, and $\rho(t_1) \in \mathcal{D}(\mathcal{H})$ is the state of the system at time t_1, then the state of the system at time t_2 is given by $\rho_2 = \Lambda_{t_2-t_1}(\rho_1) \in \mathcal{D}(\mathcal{H})$.

Postulate 4. *Measurement.* A quantum state $\rho \in \mathcal{D}(\mathcal{H})$ is measured with a POVM $\{E_i\}_{i=1}^n$. If event $1 \leq i \leq n$ is measured (which happens with probability $\operatorname{Tr}(E_i\rho)$), then the post-measurement state is given by $\rho' = \frac{E_i\rho E_i'}{\operatorname{Tr}(E_i\rho E_i')}$.

1.2.3 Time Evolution of Quantum Systems

Postulate 3 asserts that quantum states in a static environment evolve via continuous dynamical semigroups, objects which are characterized by well motivated properties: complete positivity, trace preservation, continuity, and consistency of composition. Lindblad's theorem neatly characterizes all such objects.

Definition 1.12. A superoperator $L \in \mathcal{T}(\mathcal{H})$ is said to be in *Lindblad form* if there exists $H \in \mathfrak{u}(\mathcal{H})$ and $L_k \in \mathcal{L}(\mathcal{H})$ for $k = 1, 2, \cdots, d^2 - 1$ such that

$$L(X) = -\mathrm{i}[H, X] + \sum_{k=1}^{d^2-1} L_k X L_k^\dagger - \{L_k^\dagger L_k, X\}/2 \tag{1.13}$$

where $[A, B] = AB - BA$ is the commutator and $\{A, B\} = AB + BA$ is the anti-commutator. We call H the *Hamiltonian* of the system and $\{L_k\}$ the *Lindblad dissipators* of the system. Given the theorem that follows, we often refer to L as a *supergenerator* to informally distinguish it from a superoperator, as it does not usually act on quantum states before it is exponentiated.

Theorem 4. *(Lindblad [112]) The family $\{\Lambda_t\}_{t\in\mathbb{R}} \subseteq \mathcal{C}(\mathcal{H})$ is a continuous dynamical semigroup if and only if $\Lambda_t = \mathrm{e}^{tL}$ where $L \in \mathcal{T}(\mathcal{H})$ is of Lindblad form.*

The notation e^{tL} represents the solution to the differential equation

$$\frac{\mathrm{d}X(t)}{\mathrm{d}t} = L(X(t)) = -\mathrm{i}[H, X(t)] + \sum L_k X(t) L_k^\dagger - \{L_k^\dagger L_k, X(t)\}/2. \tag{1.14}$$

which we call the *Lindblad master equation*. Therefore if $X(t)$ is the solution to this equation under initial condition $X(0) \in \mathcal{L}(\mathcal{H})$, then $X(t) = \mathrm{e}^{tL}(X(0))$.

The Hamiltonian of the master equation is responsible for all unitary evolution, and the Lindblad dissipators are responsible for all non-unitary dynamics. Indeed, if we set all of the Lindblad dissipators to 0, we end up with purely unitary dynamics—setting $U(t) = \mathrm{e}^{-\mathrm{i}tH}$, which is unitary as per Example 1.4, it holds that $\Lambda_t(\rho) = U(t)\rho U(t)^\dagger$. This is the solution to the *Shrödinger equation*, a special case of the Lindblad master equation,

$$\frac{\mathrm{d}X(t)}{\mathrm{d}t} = -\mathrm{i}[H, X(t)]. \tag{1.15}$$

While a continuous dynamical semigroup can always be expressed as the exponential of a supergenerator, we often wish to deviate from the condition of a static environment.

This might be due to, for example, an external, classical magnetic field whose amplitude and direction are changing as a function of time. In this case, we can imagine that we still have a static environment over infinitesimal time intervals, during which Lindblad's form will apply. However, each of these infinitesimal intervals will receive different Lindblad forms, which motivates the use of time dependent Hamiltonians, $H(t)$, and/or time dependent Lindblad dissipators, $L_k(t)$. We used classical magnetic fields as an example above. It is worth mentioning that, by contrast, quantum environments are best dealt with by expanding the Hilbert space to include them, rather than adding time dependence to the supergenerator. However, the computational cost of doing so can be prohibitive, in which case master equation approximations in the original space are very valuable when back-action and memory effects are not important. Given a time dependent Lindblad form, we have the master equation

$$\frac{\mathrm{d}X(t)}{\mathrm{d}t} = L[t](X(t)) = -\mathrm{i}[H(t), X(t)] + \sum L_k(t)X(t)L_k(t)^\dagger - \{L_k(t)^\dagger L_k(t), X(t)\}/2. \tag{1.16}$$

whose formal solution is written as

$$\Lambda_t = \mathcal{T}\mathrm{e}^{\int_0^t L[t']\mathrm{d}t'}. \tag{1.17}$$

This solution does not necessarily obey $\Lambda_{t+s} = \Lambda_t \circ \Lambda_s$. The T out front is Dyson's time ordering symbol, which serves to remind one that this is not a standard matrix exponential; the non-commutativity of $L[t]$ with itself at different times must be accounted for.

It is worthwhile to consider how to solve equations such as Equation 1.14 and Equation 1.16 in practical terms, given only classical computational resources, as numerical solutions will be needed throughout this thesis. We introduce vectorization for this purpose.

Definition 1.13. Given a linear operator $X \in \mathcal{L}(\mathcal{H})$, its *vectorization* is written as $|X\rangle\!\rangle \in \mathcal{H} \otimes \mathcal{H} = \mathbb{C}^{d^2}$, obtained by stacking all of the columns of X into a single column vector. We can also consider the corresponding dual vectorization $\langle\!\langle X|$, a row vector, so that, for example, $|X\rangle\!\rangle \langle\!\langle X| \in \mathcal{L}(\mathcal{H} \otimes \mathcal{H})$. Note that vectorization is a linear operation.

Lemma 1. *(Roth [90]) Suppose $X, Y, Z \in \mathcal{L}(\mathcal{H})$. Then $|XYZ\rangle\!\rangle = (Z^{\mathrm{T}} \otimes X)\,|Y\rangle\!\rangle$.*

This useful Lemma can be applied to Lindblad's form, giving

$$|L(X)\rangle\!\rangle = \left[-\mathrm{i}(\mathbb{I} \otimes H - H^{\mathrm{T}} \otimes \mathbb{I}) + \sum \overline{L_k} \otimes L_k - (\mathbb{I} \otimes L_k^\dagger L_k + (L_k^\dagger L_k)^{\mathrm{T}} \otimes \mathbb{I})/2\right] |X\rangle\!\rangle \tag{1.18}$$

which conveniently moves all appearances of X to the right of the formula. Setting $\hat{L}$ to the $d^2 \times d^2$ matrix in the square brackets of this formula, we have the vectorized differential equation

$$\frac{\mathrm{d}\,|X(t)\rangle\!\rangle}{\mathrm{d}t} = \hat{L[t]}\,|X(t)\rangle\!\rangle\,. \tag{1.19}$$

In the case that this equation is time independent, the solution is given by $|X(t)\rangle\!\rangle = \mathrm{e}^{t\hat{L}}\,|X(0)\rangle\!\rangle$, where $\mathrm{e}^{t\hat{L}}$ is no longer just formal notation for the solution, but the concrete exponentiation of a matrix. Most numerical computer libraries will contain such a function, which can in principle be computed with the Taylor series

$$\mathrm{e}^{t\hat{L}} = \sum_{k=0}^{\infty} \frac{(t\hat{L})^k}{k!}, \tag{1.20}$$

but is more efficiently computed with Padé approximants [81]. In the case of time dependence, one convenient approach is to slice a time interval $[0, T]$ into N short slices $[(n-1)\delta t, n\delta t]$ for $n = 1, \ldots, N$, over each of which the Lindblad form is well approximated as being constant. This produces the time-sliced approximate solution, using a mid-ponit sampling rule, given by

$$\mathcal{T}\mathrm{e}^{\int_0^T \hat{L[t']}\mathrm{d}t'} \approx \prod_{n=1}^{N} \mathrm{e}^{\delta t \hat{L}[(n-0.5)\delta t]} \tag{1.21}$$

where the product is taken in the left-to-right order $n = N, N-1, \ldots, 2, 1$.

1.3 Statistical Inference

Once data have been collected from a quantum system, they are no longer quantum but classical. To be of use, they must be analyzed, and this thesis places particular importance on performing these relevant statistics in detail, where raw data is used to infer parameters of interest about a quantum system. Indeed, this is a central theme of several chapters, including Chapter 3, which studies statistical inference using the data from a Nitrogen vacancy system, and Chapter 6, which considers a rigorous Bayesian analysis of data from randomized benchmarking experiments. Therefore, in this section, we lay out some notation and background material regarding statistical inference.

1.3.1 Notation

If F is a probability distribution, we denote a random variable X drawn from this distribution as $X \sim F$, and we denote the probability (or probability density for continuous random variables) variously as $\Pr(X = x)$, $\Pr(x)$, or $p(x)$. If f is a measurable function between the sample space of X and some other sample space, then the *expectation of* f is defined as

$$\mathbb{E}[f(X)] = \int f(x)p(x)\mathrm{d}x. \tag{1.22}$$

We will sometimes underscript the expectation to make it clear which distribution is being integrated over when confusion might be possible, e.g. $\mathbb{E}_X[F(X)]$. In the case of discrete random variables, the integral becomes a sum over possible values of x. Expectations are linear; if X and Y are random variables and a, b are scalars, then it holds that

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]. \tag{1.23}$$

The *(raw) moments* of X are defined by taking $f(x) = x^n$, which gives

$$\mu_n = \mathbb{E}[X^n] = \int x^n p(x)\mathrm{d}x \tag{1.24}$$

for $n = 0, 1, 2, 3, \ldots$, where the special case μ_1 is known as the *expectation value of* X. Similarly, the central moments of X are found by taking the raw moments of $X - \mu_1$,

$$\mu'_n = \mathbb{E}[(X - \mathbb{E}[X])^n], \tag{1.25}$$

where in particular we have

$$\mu'_1 = \mathbb{E}[X] = \mu_1 \tag{1.26}$$

$$\mu'_2 = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mu_2 - \mu_1^2 =: \mathrm{Var}[X] \tag{1.27}$$

with the latter quantity named the *variance* of X. The *standard deviation* of a random variable is the square root of its variance,

$$\mathrm{Std}[X] = \sqrt{\mathrm{Var}[X]}. \tag{1.28}$$

The notation $\Pr(x, y)$ denotes the joint density of two random variables X and Y, and the notation $\Pr(x|y)$ denotes the density of the random variable X conditioned on the particular value $Y = y$. We say that two random variables X and Y are *independent* if

their joint density factorizes, $\Pr(x, y) = \Pr(x)\Pr(y)$, and we call them *dependent* otherwise. Their *covariance* is the quantity

$$\mathrm{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \tag{1.29}$$

which is equal to 0 if X and Y are independent.

We have the following fundamental relationships of densities.

Law 1. *(Conditional Density)* $\Pr(y|x)\Pr(x) = \Pr(x, y) = \Pr(x|y)\Pr(y)$

Law 2. *(Marginal Density)* $\Pr(x) = \int \Pr(x, y)\mathrm{d}y$

Law 3. *(Total Probability)* $\Pr(x) = \int \Pr(x|y)\Pr(y)\mathrm{d}y$

Law 4. *(Total Conditional Probability)* $\Pr(x|z) = \int \Pr(x|y, z)\Pr(y|z)\mathrm{d}y$

Law 5. *(Bayes)* $\Pr(x|y) = \frac{\Pr(y|x)\Pr(x)}{\Pr(y)} = \frac{\Pr(y|x)\Pr(x)}{\int \Pr(y|x)\Pr(x)\mathrm{d}x}$

In each case, integration takes place over the entire sample space, and integrals can be replaced by sums in the case of discrete distributions. One sees that these laws are not all independent—the law of total probability is a simple consequence of the definition of marginal density and conditional density, and Bayes law is just a rearangement of conditional density.

If $\Pr(x|y)$ is a conditional density conditioned on the particular value y, then $X|y$ denotes a random variable drawn from this distribution. It holds that X has distribution $\Pr(x) = \int \Pr(x|y)\Pr(y)\mathrm{d}y$ by the law of total probability. We will often find the resulting laws of total expectation and total variance useful:

Law 6. *(Total Expectation)* $\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}[X|y]]$

Law 7. *(Total Variance)* $\mathrm{Var}[X] = \mathbb{E}_Y[\mathrm{Var}[X|y]] + \mathrm{Var}_Y[\mathbb{E}[X|y]]$

Law 8. *(Total Raw Moment)* $\mathrm{Var}[X^m] = \mathbb{E}_Y[\mathbb{E}[X^m|y]]$

Law 9. *(Total Covariance)* $\mathrm{Cov}[X, Y] = \mathbb{E}_Z[\mathrm{Cov}[X, Y|z]] + \mathrm{Cov}[\mathbb{E}[X|z], \mathbb{E}[Y|z]]$

1.3.2 The Likelihood Function

Let $Pr(y|\theta)$ be the marginal density of the random variable $Y|\theta$, conditioned on some particular value θ. In the case that Y is an observable result of some experiment or trial, and θ is an unknown parameter (or vector of unknown paramaters) that we wish to learn, then we call $Pr(y|\theta)$ the *likelihood function*, and denote it as

$$\mathcal{L}(\theta; y) = Pr(y|\theta). \tag{1.30}$$

Notice that the order in which the arguments appear is switched—this standard convention is to emphasize functional dependence on θ; this quantity represents the likelihood that θ is the 'true' value given that some value y is measured.

Often our experiments will contain additional settings which are known exactly, which we label e. In this case we sometimes explicitly add conditional dependence on e to help us keep track of which experiment was performed, writing

$$\mathcal{L}(\theta; y, e) = Pr(y|\theta, e) \tag{1.31}$$

as the likelihood function.

Example 1.8. Suppose a coin with unknown bias is flipped n times. Let $K \in \{0, 1, \ldots, n\}$ be the random variable representing the number of times heads lands. Then, in the notation above, we have $Y = K$, $y = k$, $\theta = p$, and $e = n$. The likelihood function is the distribution of $K|p$, the binomial distribution, given by

$$\mathcal{L}(p; k, n) = \Pr(k|p, n) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{1.32}$$

⊣

1.3.3 Estimators

Definition 1.14. Let $Y_1|\theta, Y_2|\theta, \ldots, Y_n|\theta$ be identical and independent conditional random variables with a likelihood function $\mathcal{L}(\theta; y)$ where the sample space of the unknown parameter θ is $\mathbb{R}^m$ for some $m \geq 1$. Let $\hat{\theta}(Y_1, Y_2, ..., Y_n)|\theta$ be a new random variable that takes output values in $\mathbb{R}^m$. Then we call $\hat{\theta}$ an *estimator* of the unknown quantity θ.

This broad definition "an estimator is just some function of the data" means that estimators don't need to be sensible. For example, $\hat{\theta}(\cdot) = 5$ is a perfectly valid estimator

whenever $m = 1$, but it will usually be a bad one. The estimator itself is a random variable, with its own distribution, and therefore we can study properties such as its expectation value or variance. These properties can be used to compare the quality of several estimators.

Estimator Bias

For instance, we can define the bias of an estimator as follows.

Definition 1.15. Let $\hat{\theta}(Y_1, Y_2, ..., Y_n)|\theta$ be an estimator of θ. We call

$$\text{Bias}[\hat{\theta}] = \mathbb{E}[\hat{\theta} - \theta] \tag{1.33}$$

the *bias* of $\hat{\theta}$, and we say that $\hat{\theta}$ is biased if $\text{Bias}[\hat{\theta}] \neq 0$. If $\lim_{n\to\infty} \text{Bias}[\hat{\theta}] = 0$, then we say that $\hat{\vec{p}}$ is *consistent*.

Example 1.9. As in the previous example, suppose a coin with bias p is flipped n times, where $p \in [0, 1]$ is the unknown quantity. Let $K|p$ be the random variable representing the number of times the coin lands heads. Then $K|p$ has a binomial distribution with central moments $\mathbb{E}[K|p] = np$ and $\text{Var}[K|p] = np(1-p)$. Let $\hat{p}_0 = K/n$ and $\hat{p}_1 = (K+1)/(n+1)$ be two different estimators of p. It holds that

$$\mathbb{E}[\hat{p}_0] = \mathbb{E}[K/n] = p \tag{1.34}$$

$$\mathbb{E}[\hat{p}_1] = \mathbb{E}[(K+1)/(n+2)] = \frac{n}{n+2}p + \frac{1}{n+2} \tag{1.35}$$

and therefore $\text{Bias}[\hat{p}_0] = 0$ so that $\hat{p}$ is unbiased, but $\hat{p}_1$ is biased though consistent.

Note that in the definition n refers to the number of samples collected, whereas in this example it appears as a parameter of the binomial distribution with a single sample K. This difference is artificial—a binomial variate is the sum of n independent trials, and is a *sufficient statistic*, which in this cases, informally, means there is nothing to gain in any situation by keeping trials separate, so one might as well sum them. ⊣

Estimator Risk

Although no one will complain about an unbiased estimator, there are other properties that one might reasonably prefer if a compromise must be made. One such property is low risk.

Definition 1.16. Let $\hat{\theta}(Y_1, Y_2, ..., Y_n)|\theta$ be an estimator of θ, and let $\mathrm{L}(\hat{\theta}, \theta) \in [0, \infty)$ be a loss function which specifies the distance between the estimate and the true value. In this thesis we will only be concerned with the *squared-error loss function*

$$\mathrm{L}(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2. \tag{1.36}$$

We call the quantity

$$\mathrm{R}_{\mathrm{L}}(\hat{\theta}, \theta) = \mathbb{E}[L(\hat{\theta}, \theta)] \tag{1.37}$$

the *risk* of the estimator $\hat{\theta}$ with respect to the loss function L.

Notice that in the case of squared-error, the risk of an estimator is the sum of its variance and squared bias:

$$\begin{aligned}
\mathrm{R}[\hat{\theta}] &= \mathbb{E}[(\hat{\theta} - \theta)^2] \\
&= \mathbb{E}[\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2] \\
&= \mathbb{E}[\hat{\theta}^2 - 2\theta\hat{\theta} + \theta^2 + (\mathbb{E}[\hat{\theta}]^2 + \mathbb{E}[\hat{\theta}]^2 - 2\hat{\theta}\mathbb{E}[\hat{\theta}])] \\
&= \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2] + \mathbb{E}[\hat{\theta} - \theta]^2 \\
&= \mathrm{Var}[\hat{\theta}] + \mathrm{Bias}[\hat{\theta}]^2
\end{aligned} \tag{1.38}$$

This shows that risk is slightly more nuanced than bias—it is not only concerned with the deviation from the true value, but how large the uncertainty of the estimate is.

Example 1.10. We continue Example 1.9, introducing the estimators

$$\hat{p}_\beta = \frac{K + \beta}{n + 2\beta} \tag{1.39}$$

for any $\beta \geq 0$, whereas the previous example looked at $\beta = 0$ and $\beta = 1$ in particular. We can compute that the (squared-error) risk of $\hat{p}_\beta$ is given by

$$\mathrm{R}[\hat{p}_\beta] = \frac{\beta^2 + 4p(1-p)\,(n - \beta^2)}{4(n+\beta)^2}. \tag{1.40}$$

Straight-forward calculus shows that the value of β that minimizes the average risk, $\min_\beta \int_0^1 \mathrm{R}[\hat{p}_\beta]\mathrm{d}p$, is $\beta = 2$. Similarly, the value of β that minimizes the maximum risk, $\min_\beta \max_p \mathrm{R}[\hat{p}_\beta]$, is given by $\sqrt{n}$. Risk is plotted as a function of p in Figure 1.2 for these two cases, as well as for the unbiased estimator, $\beta = 0$,

⊣

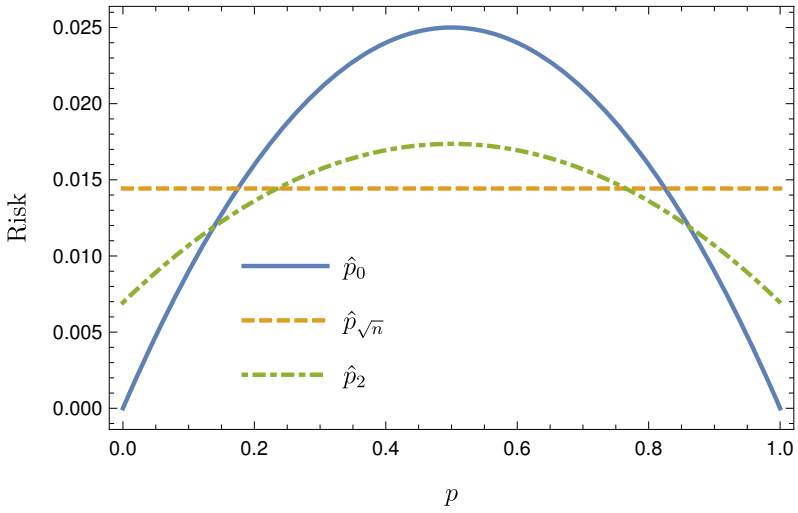

Figure 1.2: The squared-error risk of the estimator $\hat{p}_\beta$ (see Equation 1.39) for $\beta = 0, 2$ and $\sqrt{n}$. Here, n is fixed at the arbitrary choice $n = 10$.

Some Common Estimators

One of the mostly widely considered estimators is the maximum likelihood estimator (MLE). Given a likelihood function $\mathcal{L}(\theta; y)$, the MLE is the estimator denoted θ_{MLE} that maximizes the likelihood as a function of θ. The MLE is always consistent, and asympotically saturates the Cramer-Rao bound (discussed in the next subsection).

As in Example 1.10, we might be interested in estimators with low risk under relevant loss functions. In that example, we saw an estimator based on minimizing the averege squared-error loss, and minimizing the maximum risk over parameter space.

The final type of estimator we mention here are Bayes estimators, which minimize Bayes risk, discussed in the next section. In particular, we will see that for squared-error loss, the Bayes estimator is the mean of the posterior distribution. As it happens, the family of add-beta estimators in Example 1.10 are all Bayes estimators for symmetric Beta-distributed priors.

1.3.4 The Cramer-Rao Bound

Intuitively, if a likelihood function $\mathcal{L}(\theta; y)$ is tightly concentrated about some value of θ given some y, then y is highly informative about θ. The converse is true when the

likelihood is broad given some result y. We can quantify this by considering the negative $m \times m$ Hessian of the log-likelihood function, whose entries are given by

$$\mathcal{J}(\theta, y)_{i,j} = -\frac{\partial^2 \log \mathcal{L}(\theta|y)}{\partial\theta_i \partial\theta_j}. \tag{1.41}$$

This matrix is called the *observed information* of the system. It has a large magnitude in regions of the likelihood function that are tightly peaked. The dependence on the observed data y may be inconvenient if we have not measured anything yet, or if we want to make general statements that don't depend on what we end up measuring. Therefore consider the average observed information over possible data,

$$\mathcal{I}(\theta) = \mathbb{E}[\mathcal{J}(\theta, y)], \tag{1.42}$$

which is a famous matrix known as the *Fisher Information* . Its importance stems from the *Cramer-Rao bound* (CRB) which states that [74]

$$\mathrm{Cov}[\hat{\theta}|\theta] \geq (J\mu(\theta))\mathcal{I}(\theta)^{-1}(J\mu(\theta))^{\mathrm{T}} \tag{1.43}$$

for any estimator $\hat{\theta}$ of θ, where $\mu(\theta) = \mathbb{E}[\hat{\theta}(y)]$ is the expectation of the estimator, and $J\mu(\theta)$ its Jacobian matrix. The matrix inequality is as defined in Definition 1.3. This puts a lower bound on the performance of the specific estimator $\hat{\theta}$ conditioned on the true value θ, as quantified by its second moment.

The linearity of derivatives in combination with having taken the logarithm of the likelihood conspire to give the Fisher information a very useful property. If Y_1 and Y_2 are independent random variables with the same parameter space, then their joint distribution is separable making their joint loglikelihood additive, and so the Fisher information of taking a sample from each is $\mathcal{I}(\theta) = \mathcal{I}_{Y_1}(\theta) + \mathcal{I}_{Y_2}(\theta)$. Or if we take n independent and identically distributed samples like Y, then the total Fisher information is $n\mathcal{I}(\theta)$.

A nicer form of the CRB is available in the special case of unbiased estimators, that is, when $\mu(\theta) = \theta$, which gives

$$\mathrm{Cov}[\hat{\theta}|\theta] \geq \mathcal{I}(\theta)^{-1}. \tag{1.44}$$

Since the right hand side doesn't depend on the estimator at all, it is a generic lower bound for *every* unbiased estimator. Additionally, many unbiased but consistent estimators have a bias that scales as $O(1/n)$ (as in Example 1.9), and therefore this form of the CRB often holds up to $O(1/n^2)$ for such estimators.

In order to compute the inverse of the Fisher information matrix, we need to compute all elements of $\mathcal{I}(\theta)$. If this is inconvenient, or if taking the inverse is inconvenient, and we only want a bound on the variance of the i^{th} entry of $\hat{\theta}$, then we can use the looser inequality

$$\text{Var}[\hat{\theta}_i|\theta] \geq \left(\mathcal{I}(\theta)^{-1}\right)_{1,1} \geq 1/\mathcal{I}(\theta)_{i,i}. \tag{1.45}$$

1.4 Bayesian Data Analysis

Bayesian data analysis can conceptually be divided into three steps [58]:

1. Setting up the full probability model of the system in question. Typically, this involves writing down the likelihood function and deciding on a *prior* distribution to place on unknown parameters. The prior distribution is subjective and encodes prior knowledge about their values, which may be broad (not much is known) or peaked (much is known).

2. Collecting data and computing the distribution of the unknown parameters conditioned on the particular data that was observed, the *posterior* distribution. This step, except for simple examples or cases where strong assumptions can be made, is performed numerically with one of many numerical algorithms designed to sample from posteriors.

3. Evaluate the fit of the model, determine if it is reasonable, and study the predictions that it makes.

The premise of Bayesian analysis is that one should describe one's uncertainty in unknown parameters by probability distributions, and update these distrubutions as data is obtained. This is in contrast to frequentist analysis where the true parameters do not have distributions attached to them—they are fixed and absolute in reality, obtained only in the limit of infinite data. Instead, one analyses the distributions of statistics (functions) of the data, which are often arranged to estimate the true parameters.

One schema is not fundamentally superior to the other, it ultimately comes down to personal preference, and which route is more practical and efficient for the task at hand. Frequentist methods, for example, often excel at producing estimators of parameters and uncertainties that are efficient to compute. However, they come at the cost of needing

specialists to study one problem at a time, deriving appropriate custom tools. Tools such as bootstrapping [35] can help mitigate against this, but they also often require fine-tuning. Bayesian analysis, on the other hand, is more of a one-size-fits-all affair: the probability model of the system is defined, a prior is chosen, and Bayes' theorem provides the result: a posterior distribution[1].

1.4.1 Bayes' Theorem and Inference

Recall Bayes' law, Law 5:

$$\Pr(\theta|y) = \frac{\Pr(y|\theta)\Pr(\theta)}{\Pr(y)} = \frac{\Pr(y|\theta)\Pr(\theta)}{\int \Pr(y|\theta)\Pr(\theta)\mathrm{d}\theta} = \frac{\mathcal{L}(\theta;y)\Pr(\theta)}{\int \mathcal{L}(\theta;y)\Pr(\theta)\mathrm{d}\theta} \tag{1.46}$$

Bayes' law has a striking interpretation. If we take θ to mean a parameter (or vector of parameters) that we wish to learn about, y to be the data (or collection of data $y = (y_1, \ldots, y_n)$) from an experiment that depends on θ, and $\Pr(\theta)$ to be our *prior* knowledge distribution about θ before data, then Bayes law can be interpretted as follows: the distribution of θ conditioned on having observed some particular y, $\Pr(\theta|y)$ (the *posterior* distribution, exactly what a Bayesian wants to know), is proportional to the likelihood of having gotten y conditioned on a hypothetical θ times the *a priori* probability of this hypothetical value θ being correct, $\Pr(\theta)$. The denominator, $\int \mathcal{L}(\theta;y)\Pr(\theta)\mathrm{d}\theta$, does not depend on θ, which is the only functional dependece on the left-hand-side, and therefore serves only to normalize the right-hand-side.

Therefore, Bayes law connects inference to simulation: if we are able to compute the likelihood of data $\mathcal{L}(\theta;y)$ for hypothetical true values of y, which is at heart a simulation of the parameterized physical system, then we are able to make inferences about the parameters of the system, θ.

Bayes Estimators

The *Bayes risk* of an estimator $\hat{\theta}$ for θ is the risk (see Definition 1.16) for some loss function L marginalized over the prior, $\pi = \Pr(\theta)$,

$$r_{\mathrm{L}}(\hat{\theta}|\pi) = \mathbb{E}_\theta[R_L(\hat{\theta}|\theta)]. \tag{1.47}$$

[1]Numerical Bayesian methods often also require fine-tuning, though such fine tuning is becoming less relevant as methods improve [89, 13, 12]. However, this fine-tuning is of a completely different nature; parameters of the Bayesian algorithms are tweaked to make them more efficient, rather than being tuned to make them correct.

An estimator which minimizes (over estimators) the Bayes risk is known as a *Bayes estimator*. For us, a Bayes Estimator is the mean value of the posterior,

$$\hat{\theta}_{\text{Bayes}} = \int \theta \Pr(\theta|y) \mathrm{d}\theta \tag{1.48}$$

because this choice always minimizes the Bayes risk under mean-squared loss [109].

Example 1.11. Suppose a coin with unknown bias p is flipped n times, and k sides land heads. The likelihood function is given by the binomial distribution,

$$\mathcal{L}(p;k) = \Pr(k|p) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{1.49}$$

The prior distribution, $\Pr(p)$, is chosen subjectively. In this example, we choose to assign it as a uniform distribution, $\Pr(p) = 1$ for all $p \in [0,1]$. Suppose some particular value $k \in \{0, 1, ..., n\}$ is measured, then the posterior distribution is computed as

$$\begin{aligned} \Pr(p|k) &= \frac{\Pr(k|p)\Pr(p)}{\int_0^1 \Pr(k|p)\Pr(p)\mathrm{d}p} = \frac{\binom{n}{k} p^k (1-p)^{n-k}}{\binom{n}{k} \int_0^1 p^k (1-p)^{n-k} \mathrm{d}p} \\ &= (n+1)\binom{n}{k} p^k (1-p)^{n-k}. \end{aligned} \tag{1.50}$$

This distribution is plotted in Figure 1.3 for several values of n and k. We see the distribution is converging on $p = 0.8$ as we increase n, keeping the ratio $k/n = 0.8$ fixed. In fact, the posterior has a mean value of $\frac{k+1}{n+2}$ and a variance of $\frac{(k+1)(n-k+1)}{(n+2)^2(n+3)}$. Thus Bayes estimator is given by

$$\hat{\theta}_{\text{Bayes}} = \frac{k+1}{n+2}. \tag{1.51}$$

⊣

Error Bars

The interpretations of error bars in Bayesian analysis and frequentist analysis are analagous, but different. Bayesian error bars are specified as *credible intervals* and frequentist error bars are specified as *confidence intervals*. We replace the word 'intervals' with 'regions' in the case of multi-dimensional parameters.

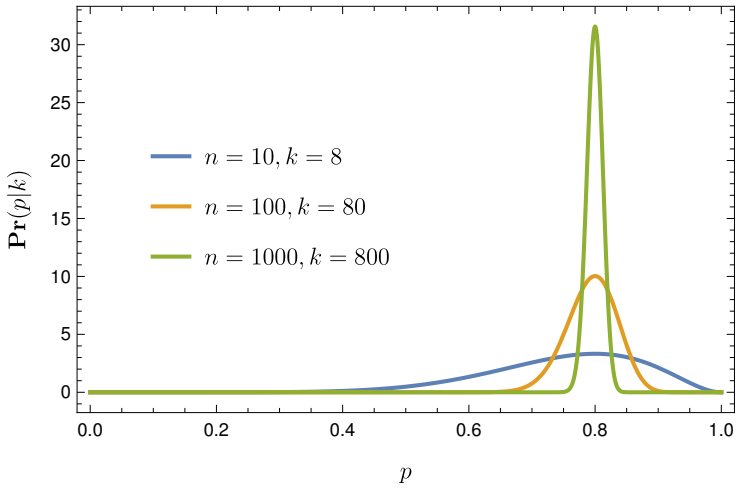

Figure 1.3: The posterior distribution of the bias of a coin, p, after it has been flipped n times, receiving k heads as the outcome.

- (Frequentist) Suppose a dataset $y_1, y_2, ..., y_n$ is collected, from which a 95% confidence interval $[\hat{l}, \hat{u}]$ is constructed for some parameter of interest θ. The data are variates of some random variable, and the limits of the confidence interval, being functions of the data, are therefore also variates of some distribution. The interpretation of confidence intervals is as follows: if, hypothetically, an infinite number of like datasets were collected from the same distribution, and confidence intervals were computed in the same way for each dataset, then the true value of x would lie in these hypothetical intervals with an asymptotic frequency of 0.95.
- (Bayesian) Suppose that $\Pr(\theta|y_1, ..., y_n)$ is the posterior distribution of θ. Then the probability that the true value of θ satisfies $l \leq \theta \leq u$ is equal to $\int_l^u \Pr(\theta|y_1, ..., y_n)\mathrm{d}\theta$. A 95% credible interval is a choice of l and u such that this integral is equal to 0.95.

1.4.2 Posterior Sampling Methods

The coin example in Example 1.11 has an analytic posterior—this turns out to be a special case of a conjugate relationship between the prior and the likelihood, discussed below. In practice, however, exact posteriors are not usually not possible to compute. In this section, after we discuss *conjugate priors*, we overview two popular numerical methods to sample from posterior distributions.

In terms of numerics, being able to sample from a continuous distribution is nearly as good as having a formula for its density function, and sometimes better. Often, quantities that we would most like to know about are integrals over the posterior, such as moment and volume integrals.

Suppose we are interested in the quantity $A = \int f(x) \Pr(x) \mathrm{d}x$ for some function f, where the integral is taken over the whole sample space. Let $x_1, x_2, \ldots, x_n$ be samples from $\Pr(x)$. Then assigning $\hat{A} = \frac{1}{n} \sum_{i=1}^{n} f(x_i)$, we have that

$$\mathbb{E}[\hat{A}] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[f(x_i)] = \int f(x) \Pr(x) \mathrm{d}x = A, \tag{1.52}$$

hence $\hat{A}$ is an unbiased estimator of A. This is called the *Monte Carlo integral* of f. Next,

$$\mathrm{Var}[\hat{A}] = \frac{1}{N^2} \sum_{i=1}^{n} \mathrm{Var}[f] = \frac{\mathrm{Var}[f]}{N} \tag{1.53}$$

hence

$$\hat{A} \approx A + O(1/\sqrt{N}) \tag{1.54}$$

assuming $\mathrm{Var}[f]$ is bounded. Notice this analysis is independent of the dimension of x. Therefore, one can well estimate integral properties of distribution given only samples from the distribution. If one is interested in plotting the distribution, then the dimension of x is likely either 1 or 2, and either histograms or kernel density estimates (KDE) can be used.

Conjugate Priors

A *conjugate prior* is a family of distributions which, relative to a fixed likelihood function, yield posteriors in the same family as the prior. Many of the common named distributions, when considered as likelihood functions, have conjugate priors.

For instance, in Example 1.11 we found an analytic posterior for the binomial likelihood when the prior used was a uniform distibution. The uniform distribution on the unit interval is a special case of the beta distribution, $\mathrm{Beta}\,(a, b)$, where $a = b = 1$. The beta distribution has a density function given by

$$f_{\mathrm{Beta}}(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a, b)} \tag{1.55}$$

where $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the beta function and $\Gamma(\cdot)$ is the usual gamma function. If $\text{Beta}\,(a,b)$ is the prior distibution for the binomial likelihood with n trials, then the posterior distribution is given as $\text{Beta}\,(a+k, b+n-k)$ if $0 \leq k \leq n$ is the outcome of the experiment.

SMC Sampling

Sequential Monte Carlo (SMC) sampling is a method of sampling from a posterior by updating the locations and weights of a swarm of particles in the sample space sequentially for each datum in a dataset. The sequential nature of this sampler makes it particularly appropriate for data that is being streamed, and indeed, SMC samplers are often referred to as *particle filters*, sharing similarities with other filters such as Kalman filters. SMC has receieved attention in the field of quantum information processing for this reason [63], and the main analyses of Chapter 3 and Chapter 4 use this method.

The premise is to approximate the prior, intermediate, and posterior distributions as finite sums of weighted points in sample space. Suppose that $\Pr(x)$ is the prior distribution of the unknown quantity $x \in \mathbb{R}^m$, and the dataset $(y_1, y_2, \cdots, y_n)$ consists of variates drawn identically and independently with the likelihood function $\mathcal{L}(x; y)$. We begin by sampling elements (called *particles*) $x_1, x_2, ..., x_I \in \mathbb{R}^m$ from $\Pr(x)$ and setting uniform weights $w_1 = w_2 = \cdots = w_I = 1/I$ to approximate the distribution $\Pr(x)$ as

$$\Pr(x) \approx \sum_{i=1}^{I} w_i \delta(x - x_i) \tag{1.56}$$

where $\delta(\cdot)$ is the Dirac delta function centered at the origin. Although this is not a good approximation to $\Pr(x)$ pointwise, it is accurate at estimating quantities that are based on integrals of $\Pr(x)$, as discussed in the introduction of this section.

Next, notice that Bayes law can be expanded as follows due to the independence of data

$$\begin{aligned} \Pr(x|y_1, y_2, \ldots, y_n) &= \frac{\Pr(y_1, y_2, \ldots|x)\Pr(x)}{\Pr(y_1, y_2, \ldots, y_n)} \\ &= \frac{\Pr(y_1|x)}{\Pr(y_1)}\frac{\Pr(y_2|x)}{\Pr(y_2)}\cdots\frac{\Pr(y_n|x)}{\Pr(y_n)}\Pr(x). \end{aligned} \tag{1.57}$$

Therefore, the intermediate distrubution $\frac{\Pr(y_k|x)}{\Pr(y_k)}\cdots\frac{\Pr(y_n|x)}{\Pr(y_n)}\Pr(x)$ can be interpreted as the prior distribution of the next data point, y_{k+1}. This allows us to update our particle

approximation sequentially by applying Bayes rule. If $\sum_{i=1}^{I} w_i \delta(x - x_i)$ is our particle approximation at some $1 \leq k < n$, then the particle approximation at step $k+1$ is given by

$$\begin{aligned} \Pr(x|y_{k+1}, y_k, \ldots, y_1) &\approx \frac{\mathcal{L}(x; y_{k+1})}{\Pr(y_{k+1})} \sum_{i=1}^{I} w_i \delta(x - x_i) \\ &= \frac{1}{\Pr(y_{k+1})} \sum_{i=1}^{I} \mathcal{L}(x_i; y_{k+1}) w_i \delta(x - x_i) \end{aligned} \tag{1.58}$$

where the normalization constant is

$$\mathcal{N} := \Pr(y_{k+1}) = \int \mathcal{L}(x_i; y_{k+1}) \sum_{i=1}^{I} w_i \delta(x - x_i) \mathrm{d}x = \sum_{i=1}^{I} \mathcal{L}(x_i; y_{k+1}) w_i. \tag{1.59}$$

Hence updating a particle distribution with the data point y_{k+1} amounts to updating the weights using the rule

$$w_i \mapsto \mathcal{L}(x_i; y_{k+1}) w_i / \mathcal{N} \tag{1.60}$$

while keeping the particles constant.

As this algorithm progresses, the weights of unlikely particles x_i will diminish, and the weights of likely particles x_i will increase. Eventually, this will lead to instability as the number of effective particles

$$n_{\text{eff}} = 1 / \sum_{i=1}^{I} w_i^2, \tag{1.61}$$

will tend to 0, unless the original sampling from $\Pr(x)$ was quite dense, which is usually impractical. Therefore, periodically, a resampling operation is performed, where the weights are reset to uniform, while the particles are moved closer to where they are needed most. In this thesis, we use the Liu-West resampler [113], which samples new particles from the mixture distribution

$$\sum_{i=1}^{n} w_i \,\text{Normal}\left(a x_i + (1-a)\mu, h^2 \sigma^2\right) \tag{1.62}$$

where $\mu = \sum_{i=1}^{I} w_i x_i$ and $\sigma^2 = \sum_{i=1}^{I} w_i (x_i - \mu)^2$. If $a^2 + h^2 = 1$, then this resampler maintains the first two moments of the particle distribution; we use $a = 0.98$.

MCMC Sampling

A Markov chain Monte Carlo (MCMC) method is an algorithm used to sample independent elements from some desired distribution using the following general principle: an instance of a Markov chain is simulated, where the Markov chain has been designed to have a steady-state distribution equal to the distribution of interest.

The Metropolis–Hastings is one of the simplest such algorithms. It is designed for the scenario where one wants to sample from the density function $f(x)$ but one only has access to an unnormalized version $g(x)$, where $f(x) = g(x)/\int g(x)dx$ [76]. This is often useful in the context of Bayesian inference where $f(x) = \Pr(x|y)$ is the posterior of x given the data d. By Bayes' law, $f(x) \propto \Pr(d|x)\pi(x)$ where both the likelihood $\Pr(y|x)$ and prior $\pi(x)$ are known, but the normalization constant $\int \Pr(y|x)\pi(x)dx$ is intractable.

The Metropolis–Hastings trick is to construct a Markov chain whose steady state distribution is given by $f(x)$, but for which simulating a random instance requires only evaluations of ratios of f, which are the same as ratios of g. Then we may start with an arbitrary initial value and evolve until we have reason to believe we are in the steady state, which is determined either empirically or theoretically. The last time sample represents a random sample drawn from $f(x)$. If multiple samples are required, it is common to, say, throw out the first 1000 transient time points of the process (the *burn-in* period), and keep every 100^{th} subsequent time step as a random sample of $f(x)$. A short auto-correlation time post burn-in, known as a fast *mixing rate*, is desired, so that fewer samples need to be thrown out.

The algorithm requires a *proposal density* $h(x'|x)$ whose job is to propose the next value of the process, x', given the previous value, x. The prototypical choice is a normal distribution $h(x'|x) \propto e^{-(x-x')^2/2\sigma^2}$. This choice affects the burn-in time and mixing rate. For example, with a normal proposal density, a small variance will mean it takes many steps to move around the domain of f leading to a slow mixing rate. On the other hand, a large variance may usually propose new locations well outside the likely support of f, leading to high rejection rates and therefore also slow mixing. A well-tuned proposal density will hit the sweet spot. The algorithm is as follows:

1. Somehow pick an initial value, x_0.
2. For $k \geq 1$, draw a proposal and a random number,

$$\begin{aligned} x' &\sim h(x'|x_{k-1}) \\ r &\sim \text{Unif}\left([0,1]\right), \end{aligned}$$

and then set $x_k = \begin{cases} x' & \text{if } r \leq g(x')/g(x_{k-1}) \\ x_{k-1} & \text{else} \end{cases}$.

3. Iterate the previous step until the desired number of samples from the steady-state have been aquired.

Intuitively this makes sense; we move from the previous location x_{k-1} to the proposed location x' with a probability that prefers a higher density of f, characterized by $f(x')/f(x_{k-1}) = g(x')/g(x_{k-1})$. Therefore samples will end up in the densest regions of $f(x)$.

The sampling method used in Chapter 6 is Hamiltonian Monte carlo (HMC) which is a sophisticated, modern MCMC method [133]. In particular we use an library implementation called *stan* [22]. Here, the term Hamiltonian is used in the classical context. The unknown distribution $f(x)$ is treated as being the Boltzmann distribution of some energy function over states in the sample space. The proposal for the next step in the Markov chain simulation is drawn by simulating the dynamics of this Hamiltonian system, using the previous sample as the starting point and a random initial momentum, for some amount of time—the endpoint of the trajectory is the proposal. This results in very large steps and greatly decreases the mixing time; a well tuned HMC sampler has nearly no correlation between adjacent points. A recent improvement to HMC is the No-U-Turns sampler, which introduces an automatic way to determine the number of steps to simulate each Hamiltonian trajectory [89]. A recent conceptual tutorial on Hamiltonian Monte Carlo has been provided by [13].

Chapter 2

The Nitrogen Vacancy Center as a Test Platform: Preliminary Details[1]

2.1 Introduction

This thesis contains several experimental demonstrations of concepts and protocols, all of which are based on a quantum system called the Nitrogen Vacancy (NV^-) center. In this section we describe the properties and dynamics of this system. Later sections discuss methods to design control schemes and assess the quality of this control experimentally.

Pure diamond is an allotrope of carbon arranged in a diamond cubic crystal structure. An NV^- defect is an impurity in the diamond lattice consisting of a nitrogen atom adjacent to an empty lattice site, replacing two carbon atoms that would normally be in those positions. This defect, when negatively charged, has six bound electrons whose spatial wavefunction extends on the order of a dozen lattice sites before coupling to 13-carbon atoms through Fermi-contact becomes negligible compared to coupling through dipolar interactions [161, 57, 41]. These six electrons form an effective spin-1 electron in the

[1]This section, in particular the optical dynamics, measurement and initialization subsections, is based partly on Reference [84], work done in collaboration with Chris Granade, whose main contributions appear in Chapter 3, which is also based on this reference. This section also derives certain block structures to help with average Hamiltonian theory, which is unpublished work done in collaboration with Romain Ruhlmann, who helped to extend the idea to Bloch-Siegert approximations in the first rotating frame, with a corresponding numerical implementation (found in Appendix A 'nvham'). Little of this section is novel, except perhaps the block treatment of Hamiltonian approximations, but serves as one of the most complete (and hopefully beginner-friendly) references for the dynamics of NV systems.

electronic ground state [37]. Each NV defect in a single crystal diamond has slightly different properties, such as mechanical strain, electric fields, and magnetic fields. These differences in the local environment are caused by randomly distributed nearby defects, such as 13-carbon and other lattice defects, or features on the surface if sufficiently close ($\lessapprox 10\,\mathrm{nm}$). It is standard practice to pick two out of these three spin energy levels to define a qubit, though in this thesis we will often be interested in working with all three.

NV^- centers have been studied extensively due to a number of remarkable physical properties [37]. These include long coherence times at room temperature [4], the ability to address and readout a single defect in isolation [70, 95], the ability to initialize to a (nearly) pure state on demand [75], and the ability to selectively interact with nearby nuclei [94, 43, 48, 96, 135, 41]. Moreover, the NV^- center's sensitivity to external macroscopic properties like magnetic fields, electric fields, and temperature, in combination with its nanoscopic spatial profile, have shown it to be highly suitable as a quantum sensor [39, 1, 148].

2.2 Ground State Hamiltonian

There are four possible orientations for an NV^-defect to have, corresponding to the four crystalline directions of the diamond lattice. In low magnetic and electric fields the dominant energy is given by the zero field splitting (ZFS) which lies along this direction, which we call the principal axis (PA). There can additionally be small distortions on the defect's electron wave-function due to stray electric fields and crystal strain. As shown in detail in Reference [37], these effects can be ignored without approximation if we rotate our reference frame so that it is centered along the combined effect of the ZFS, the strain, and the constant electric field. For small strains and electric fields, this simply results in a slightly modified ZFS along a slightly different direction than the PA. Henceforward, all spatial coordinates are with respect to this modified frame.

The total Hamiltonian for an NV^-center in the electronic ground state manifold, con-

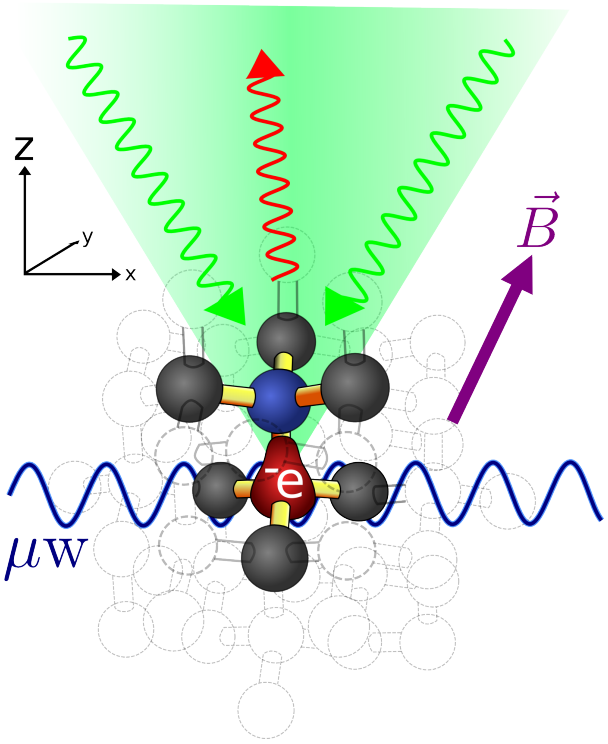

Figure 2.1: Illustration of an NV^- defect in diamond, consisting of a nitrogen atom (blue) and an adjacent vacancy filled with six electrons (red). Green laser light incident on the center is used for initialization and measurement stimulation. Emitted red photons are are detected and used for measurement. An applied microwave field drives spin transitions to control dynamics, and a static external magnetic field splits degenerate energy levels.

sisting of an effective spin-1 electron and nearby nuclei, is given by

$$\begin{aligned} H(t) =& \Delta \mathbf{S}_z^2 && \text{ZFS} \\ &+ \gamma_e \vec{B} \cdot \mathbf{S} + \gamma_N \vec{B} \cdot \mathbf{I} + \sum_i \gamma_C \vec{B} \cdot \mathbf{I}_i && \text{Zeeman} \\ &+ \mathbf{S} \cdot A \cdot \mathbf{I} + \sum_i \mathbf{S} \cdot A_i \cdot \mathbf{I}_i && \text{Hyperfine} \\ &+ + \sum_i \mathbf{I} \cdot D_i \cdot \mathbf{I}_i + \sum_{i<j} \mathbf{I}_i \cdot D_{ij} \cdot \mathbf{I}_j && \text{Dipole-dipole} \\ &+ \mathbf{I} \cdot Q \cdot \mathbf{I} && \text{Quadrapolar} \\ &+ 2\Omega(t) \cos\left(\omega_{\mu w} t + \phi(t)\right) (\mathbf{S}_x + b \mathbf{S}_z). && \text{Microwave drive} \end{aligned} \tag{2.1}$$

All terms of this Hamiltonian are written in the (modified) PA reference frame, and definitions are provided in Table 2.1. We adopt an implicit tensor space ordering $\mathcal{H} = \mathcal{H}_{NV} \otimes \mathcal{H}_N \otimes \mathcal{H}_{13C}$ with a dimension of either $3 \times 3 \times 2^n$ or $3 \times 2 \times 2^n$ (in the case of 14-nitrogen or 15-nitrogen, respectively), where n is the number of relevant 13-carbon atoms in the environment. Tensor products with the identity operator are implicit, for example, $\mathbf{S}_z$, in this context, refers to $\mathbf{S}_z \otimes \mathbb{I}_N \otimes \mathbb{I}_{13C}^{\otimes n}$, and $\mathbf{S}_z \mathbf{I}_{z,2}$ refers to $\mathbf{S}_z \otimes \mathbb{I}_N \otimes \mathbb{I}_{13C} \otimes \mathbf{I}_z \otimes \mathbb{I}_{13C}^{\otimes (n-2)}$. Spin operators are defined in Table 2.3 (spin-1/2) and Table 2.2 (spin-1). Note that we are using standard spin-vector notation is used, so that, for example, we have

$$\gamma_e \vec{B} \cdot \mathbf{S} = \gamma_e \left(B_x \mathbf{S}_x + B_y \mathbf{S}_y + B_z \mathbf{S}_z\right) \tag{2.2}$$

and

$$\mathbf{S} \cdot A \cdot \mathbf{I} = \sum_{i,j \in \{x,y,z\}} A_{ij} \mathbf{S}_i \mathbf{I}_j \tag{2.3}$$

where A is a 3×3 hyperfine tensor indexed by $\{x, y, z\}$.

This Hamiltonian applies to both natural isotopes of nitrogen, 14-nitrogen and 15-nitrogen, which are spin-1 and spin-1/2 respectively. However, 14-nitrogen is by far more common, and therefore appears in all simulations and experiments presented in this thesis. Since the nitrogen lies on the z-axis, its hyperfine tensor is diagonal, with a 14-nitrogen value of roughly $A = \text{diag}(-2.7, -2.7, -2.14)\text{MHz}$ [48].

Dipole-dipole interactions between nuclear spins are determined geometrically. If two spins i and j have respective locations $\vec{r}_i$ and $\vec{r}_j$ in our coordinate system, then letting $\vec{r} = \vec{r}_i - \vec{r}_j$, $r = \|\vec{r}\|$ and $\hat{r} = \vec{r}/r$, we get the dipolar tensor components

$$D_{ij} = -2\pi \frac{\mu_0 \gamma_i \gamma_j \hbar}{4\pi r^3} \left(3 r_i \cdot r_j - \delta_{ij}\right) \tag{2.4}$$

Symbol	Value	Description
$\mathbf{S}$	$\{\mathbf{S}_\mathrm{x}, \mathbf{S}_\mathrm{y}, \mathbf{S}_\mathrm{z}\}$	Electron spin-1 vector
$\mathbf{I}$	$\{\mathbf{I}_\mathrm{x}, \mathbf{I}_\mathrm{y}, \mathbf{I}_\mathrm{z}\}$	Nitrogen spin-1 or spin-1/2 vector
$\mathbf{I}_i$	$\{\mathbf{I}_\mathrm{x}, \mathbf{I}_\mathrm{y}, \mathbf{I}_\mathrm{z}\}$	Carbon spin-1/2 vector
Δ	$2\pi \cdot 2.87\,\mathrm{GHz}$	Nominal ZFS value
$\vec{B}$	$\lesssim 100\,\mathrm{G}$	External magnetic field
γ_e	$2\pi \cdot 2.8025\,\mathrm{MHz/G}$	Electron gyromagnetic ratio
γ_N	$2\pi \cdot 0.3077(-0.4316)\mathrm{kHz/G}$	Nitrogen 14(15) gyromagnetic ratio
γ_C	$2\pi \cdot 1.071\,\mathrm{kHz/G}$	Carbon gyromagnetic ratio
A, A_i	$\mathrm{kHz} - \mathrm{MHz}$	Hyperfine tensors, dipolar + Fermi contact
D_i, D_{ij}	kHz	Nuclear-nuclear dipolar tensors
Q	$-2\pi \cdot 5.01\,\mathrm{MHz}$	14-nitrogen quadrapolar (0 for 15-nitrogen)
$\Omega(t)$	$\lesssim 2\pi \cdot 20\,\mathrm{MHz}$	Microwave amplitude envelope
$\phi(t)$	$[0, 2\pi]$	Microwave phase profile
$\omega_{\mu w}$	$\approx \Delta$	Microwave carrier frequency
b	unitless	Microwave field direction misalignment

Table 2.1: Glossary of Hamiltonian parameters. Spin operators are defined in Table 2.3 (spin-1/2) and Table 2.2 (spin-1). Note that the difference between spin labels is just that: labeling—historically, electron spins get the $\mathbf{S}$ symbol and nuclear spins get the $\mathbf{I}$ symbol.

for $i, j \in \{x, y, z\}$, where δ_{ij} is the kronecker delta, $\mu_0 = 4\pi \cdot 10 \times 10^{-7}\,\mathrm{H/m}\cdot$ is the magnetic constant. Note that Planck's reduced constant $\hbar$ in the numerator is not squared because of our convention to omit $\hbar$ it from Shrödinger's equation. Hyperfine tensors A between the electron and nuclear spins arise as a combination of dipole-dipole interaction (see above), which must be integrated over the spatial extend of the electron wavefunction, and the Fermi contact interaction—this calculation is difficult, and experimental measurements are preferred [57]. The exception is when the nucleus is sufficiently far away from the NV^- so that Fermi contact is negligible, and a mean field approximation is acceptable.

The z axis is aligned with the PA, which leave the x and y axes undefined. Here, our choice is to fix the x axis such that the microwave field lies in the xz plane, hence the lack of $\mathbf{S}_\mathrm{y}$ term in the microwave drive Hamiltonian. If a single 13-carbon is present, it is sometimes preferable to let its spatial location determine the x-axis instead.

We see in Table 2.1 that the ZFS is the dominant energy term when we limit the external magnetic field value. In subsequent sections we explore entering the rotating frame induced by this term in order to simplify, ease simulations, and provide insight.

$$\mathbf{S}_{\mathrm{x}} = \tfrac{1}{\sqrt{2}}\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \qquad \mathbf{S}_{\mathrm{y}} = \tfrac{1}{\sqrt{2}}\begin{pmatrix} 0 & -\mathrm{i} & 0 \\ \mathrm{i} & 0 & -\mathrm{i} \\ 0 & \mathrm{i} & 0 \end{pmatrix} \qquad \mathbf{S}_{\mathrm{z}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

$$\mathbf{S}'_{\mathrm{x}} = \tfrac{1}{\sqrt{2}}\begin{pmatrix} 0 & -1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \qquad \mathbf{S}'_{\mathrm{y}} = \tfrac{1}{\sqrt{2}}\begin{pmatrix} 0 & -\mathrm{i} & 0 \\ \mathrm{i} & 0 & \mathrm{i} \\ 0 & -\mathrm{i} & 0 \end{pmatrix} \qquad P_0 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$\mathbf{S}^{+}_{\mathrm{x}} = \tfrac{1}{\sqrt{2}}\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad \mathbf{S}^{-}_{\mathrm{x}} = \tfrac{1}{\sqrt{2}}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \qquad \mathbf{S}_{\mathrm{xx}} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

$$\mathbf{S}^{+}_{\mathrm{y}} = \tfrac{1}{\sqrt{2}}\begin{pmatrix} 0 & -\mathrm{i} & 0 \\ \mathrm{i} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad \mathbf{S}^{-}_{\mathrm{y}} = \tfrac{1}{\sqrt{2}}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -\mathrm{i} \\ 0 & \mathrm{i} & 0 \end{pmatrix} \qquad \mathbf{S}_{\mathrm{yy}} = \begin{pmatrix} 0 & 0 & -\mathrm{i} \\ 0 & 0 & 0 \\ \mathrm{i} & 0 & 0 \end{pmatrix}$$

Table 2.2: Our extended spin-1 operators. The first row cantains the standard spin-1 operators, and defines the spin vector $\{\mathbf{S}_{\mathrm{x}}, \mathbf{S}_{\mathrm{y}}, \mathbf{S}_{\mathrm{z}}\}$. The second row contains the 0-projector and the 'twisted' spin operators that naturally arise in the NV^{-} rotating frame due to conjugation by $\mathbf{S}_{\mathrm{z}}^2$. The third and fourth rows isolate individual transitions.

$$X = 2\sigma_x = \mathbf{S}_{\mathrm{x}} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad Y = 2\sigma_y = \mathbf{S}_{\mathrm{y}} = \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix}, \quad Z = 2\sigma_z = \mathbf{S}_{\mathrm{z}} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

Table 2.3: Spin-1/2 operators. The spin vector is given by $\{\sigma_x, \sigma_y, \sigma_z\}$. When the 2×2 identity matrix is added, this forms an orthogonal basis.

2.3 Secular and Higher Order Hamiltonian Approximations

The Hamiltonian in Equation 2.1 has time dependence which enters at both large frequencies on the order of $\omega_{\mu w} \approx 3\,\mathrm{GHz}$, and at smaller frequencies on the order of kHz. Therefore, it will be helpful to enter the unitary rotating frame $U_{\mathrm{rot}}(t) = \mathrm{e}^{\mathrm{i}tH_{\mathrm{rot}}}$ where $H_{\mathrm{rot}} = \omega_{\mu w}\mathbf{S}_{\mathrm{z}}^2$. It is easy to verify that for any differentiable pure state function $|\psi(t)\rangle$ satisfying the Shrödinger equation $\frac{\mathrm{d}}{\mathrm{d}t}|\psi(t)\rangle = -\mathrm{i}H(t)|\psi(t)\rangle$, the frame-rotated state $|\tilde{\psi}(t)\rangle = U_{\mathrm{rot}}(t)|\psi(t)\rangle$ satisfies the Shrödinger equation

$$\frac{\mathrm{d}}{\mathrm{d}t}|\tilde{\psi}(t)\rangle = -\mathrm{i}H_{\mathrm{eff}}(t)|\tilde{\psi}(t)\rangle \tag{2.5}$$

using the effective Hamiltonian

$$H_{\mathrm{eff}}(t) := U_{\mathrm{rot}}(t)\left(H - H_{\mathrm{rot}}\right)U_{\mathrm{rot}}^{\dagger}(t). \tag{2.6}$$

The operator solution of Equation 2.5 is denoted as

$$U(t) = \mathcal{T}\exp\left(-i\int_{0,t} H_{\mathrm{eff}}(t')\mathrm{d}t'\right), \tag{2.7}$$

where $\mathcal{T}$ is Dyson's time ordering symbol, so that $|\tilde{\psi}(t)\rangle = U(t)|\psi(0)\rangle$ for any time t. One of the main purposes of entering this rotating frame is to make the action of the microwave drive term appear quasi-static, that is, to eliminate time-dependence at the high frequency $\omega_{\mu w}$, leaving behind only the slower functions $\Omega(t)$ and $\phi(t)$. As currently formulated, this purpose is not achieved, and has in fact been made worse—terms exist in $H_{\mathrm{eff}}(t)$ at frequency $\mathrm{e}^{2\mathrm{i}\omega_{\mu w}t}$ which arise via the scalar conjugation $\mathrm{e}^{\mathrm{i}\omega_{\mu w}t}\cos(\omega_{\mu w}t + \phi(t))\mathrm{e}^{-\mathrm{i}\omega_{\mu w}t}$. As is standard practice, we use average Hamiltonian theory [73] to divide the mechanics into successive orders of importance, while also eliminating high-frequency noise. For a fixed time t, the average Hamiltonian expansion is given by

$$U(t) = \exp(-\mathrm{i}t(\overline{H}_0 + \overline{H}_1 + \overline{H}_2 + ...)) \tag{2.8}$$

where

$$\begin{aligned}
\overline{H}^0 &= \frac{1}{t}\int_0^t dt_1 H_{\mathrm{eff}}(t_1) \\
\overline{H}^1 &= \frac{-\mathrm{i}}{2t}\int_0^t dt_1 \int_0^{t_1} dt_2 [H_{\mathrm{eff}}(t_1), H_{\mathrm{eff}}(t_2)] \\
&\dots
\end{aligned} \tag{2.9}$$

In the regime where time dependence of $\Omega(t)$ and $\phi(t)$ is negligible over the period of rotation $T = 2\pi/\omega_{\mu w}$, as it typically is, it is useful to take this period T as our average Hamiltonian expansion period. Then, because T is so small, any simulation time $t > T$ with constant $\Omega(t)$ and $\phi(t)$ can be achieved as

$$U(t) \approx U(T\lfloor t/T \rfloor) = U(T)^{\lfloor t/T \rfloor}; \tag{2.10}$$

the so-called stroboscopic approximation.

The naive method of computing each of the average Hamiltonian terms in Equation 2.9 at duration T is to perform the integrals, which becomes increasingly cumbersome as the number of nested commutants increases. However, by exploiting the known structure of the rotating frame transformation, and using some formulas emerging from Floquet theory due to Leskes et al. [110], we can avoid integration altogether. To begin, consider operators acting on $\mathcal{H} = \mathcal{H}_{NV} \otimes \mathcal{H}_N \otimes \mathcal{H}_{13C}$ as being 3×3 block matrices with each block having size compatible with the nuclear subspace, $\mathcal{H}_N \otimes \mathcal{H}_{13C}$. Our rotating frame unitary can then be written as the block diagonal matrix $U_{\text{rot}}(t) = \text{diag}(e^{i\omega_{\mu w}t}\mathbb{I}, \mathbb{I}, e^{i\omega_{\mu w}t}\mathbb{I})$. Conjugating any operator $X \in \text{L}(\mathcal{H})$ by this transformation results in

$$\begin{aligned}
U_{\text{rot}}(t) X U_{\text{rot}}(t)^\dagger &= \begin{pmatrix} e^{i\omega_{\mu w}t}\mathbb{I} & 0 & 0 \\ 0 & \mathbb{I} & 0 \\ 0 & 0 & e^{i\omega_{\mu w}t}\mathbb{I} \end{pmatrix} \begin{pmatrix} A & B & C \\ D & E & F \\ G & H & I \end{pmatrix} \begin{pmatrix} e^{-i\omega_{\mu w}t}\mathbb{I} & 0 & 0 \\ 0 & \mathbb{I} & 0 \\ 0 & 0 & e^{-i\omega_{\mu w}t}\mathbb{I} \end{pmatrix} \\
&= e^{-i\omega_{\mu w}t} \begin{pmatrix} 0 & 0 & 0 \\ D & 0 & F \\ 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} A & 0 & C \\ 0 & E & 0 \\ G & 0 & I \end{pmatrix} + e^{+i\omega_{\mu w}t} \begin{pmatrix} 0 & B & 0 \\ 0 & 0 & 0 \\ 0 & H & 0 \end{pmatrix} \\
&\equiv e^{-i\omega_{\mu w}t} P_-[X] + P_\times[X] + e^{+i\omega_{\mu w}t} P_|[X]
\end{aligned} \tag{2.11}$$

where we have implicitly defined three linear operations $P_\times$, $P_|$, and P_- that project onto the shown disjoint blocks.

Next, expanding the cosine term in our original Hamiltonian Equation 2.1 into complex exponentials, we can write it (minus the rotation frame) as

$$H - H_{\text{rot}} = e^{-i\omega_{\mu w}t} H_- + H_s + e^{+i\omega_{\mu w}t} H_+ \tag{2.12}$$

where H_s contains all of the static terms, and H_+ and H_- contain the control terms. Combining Equation 2.6, Equation 2.11, and Equation 2.12 then gives

$$\begin{aligned}
H_{\text{eff}} = {} & e^{-2i\omega_{\mu w}t}(P_-[H_-]) + e^{-i\omega_{\mu w}t}(P_-[H_s] + P_\times[H_-]) \\
& + (P_-[H_+] + P_\times[H_s] + P_|[H_-]) \\
& + e^{+i\omega_{\mu w}t}(P_\times[H_+] + P_|[H_s]) + e^{+2i\omega_{\mu w}t}(P_|[H_+])
\end{aligned} \tag{2.13}$$

which provides simple formulas to expand H_{eff} into the Floquet form

$$H_{\text{eff}} = \sum_{-2 \le k \le 2} \mathrm{e}^{\mathrm{i}k\omega_{\mu w}t} H_k. \tag{2.14}$$

The matrix terms $\{H_{-2}, H_{-1}, H_0, H_{+1}, H_{+2}\}$ of this expansion can be plugged directly into the formulas of Leskes [110] to compute average Hamiltonians:

$$\overline{H}_0 = H_0 \tag{2.15a}$$

$$\overline{H}_1 = -\frac{1}{2}\sum_{k\neq 0} \frac{[H_{-k}, H_k]}{k\omega_{\mu w}} + \sum_{k\neq 0} \frac{[H_0, H_k]}{k\omega_{\mu w}} \tag{2.15b}$$

$$\begin{aligned}\overline{H}_2 = \frac{1}{3}\sum_{k,k'\neq 0, k\neq k'} \frac{[H_{k'}, [H_{k-k'}, H_{-k}]]}{k'k\omega_{\mu w}^2} \\ + \frac{1}{2}\sum_{k\neq 0} \frac{[H_k, [H_0, H_{-k}]]}{k^2\omega_{\mu w}^2} - \frac{1}{2}\sum_{k\neq 0} \frac{[H_0, [H_0, H_k]]}{k^2\omega_{\mu w}^2} \\ + \sum_{k\neq 0, k'\neq 0} \frac{[H_{k'}, [H_{-k'}, H_k]]}{kk'\omega_{\mu w}^2} + \frac{1}{2}\sum_{k\neq 0, k'\neq 0} \frac{[H_{k'}, [H_k, H_0]]}{kk'\omega_{\mu w}^2}\end{aligned} \tag{2.15c}$$

$$\overline{H}_3 = \dots$$

As an example, the zeroth order average Hamiltonian—which includes two approximations often discussed in the literature, namely the secular and rotating wave approximation—can be deduced through visual inspection by considering the blocks of H in Equation 2.1 projected using $P_\times$, $P_|$ and P_-, giving

$$\begin{aligned}\overline{H}_0 = H_0 &= P_\times[H_s] + P_-[H_+] + P_|[H_-] \\ &= (\Delta - \omega_{\mu w})\mathbf{S}_\mathrm{z}^2 + \gamma_e B_z \mathbf{S}_\mathrm{z} + \gamma_N \vec{B}\cdot\mathbf{I} + \sum_i \gamma_C \vec{B}\cdot\mathbf{I}_i \\ &\quad + A_{zz}\mathbf{S}_\mathrm{z}\mathbf{I}_\mathrm{z} + \sum_i \mathbf{S}_\mathrm{z}(A_{zx,i}\mathbf{I}_\mathrm{x} + A_{zy,i}\mathbf{I}_\mathrm{y} + A_{zz,i}\mathbf{I}_\mathrm{z}) \\ &\quad + \sum_i \mathbf{I}\cdot D_i\cdot\mathbf{I}_i + \sum_{i<j}\mathbf{I}_i\cdot D_{ij}\cdot\mathbf{I}_j + \mathbf{I}\cdot Q\cdot\mathbf{I} \\ &\quad + \Omega(t)(\cos\phi(t)\mathbf{S}_\mathrm{x} + \sin\phi(t)\mathbf{S}'_\mathrm{y}).\end{aligned} \tag{2.16}$$

Note the twisted spin operator $\mathbf{S}'_\mathrm{y}$ in the control term, defined in Table 2.2. Pseudocode for a general algorithm to compute average Hamiltonians is outlined in Algorithm 2.

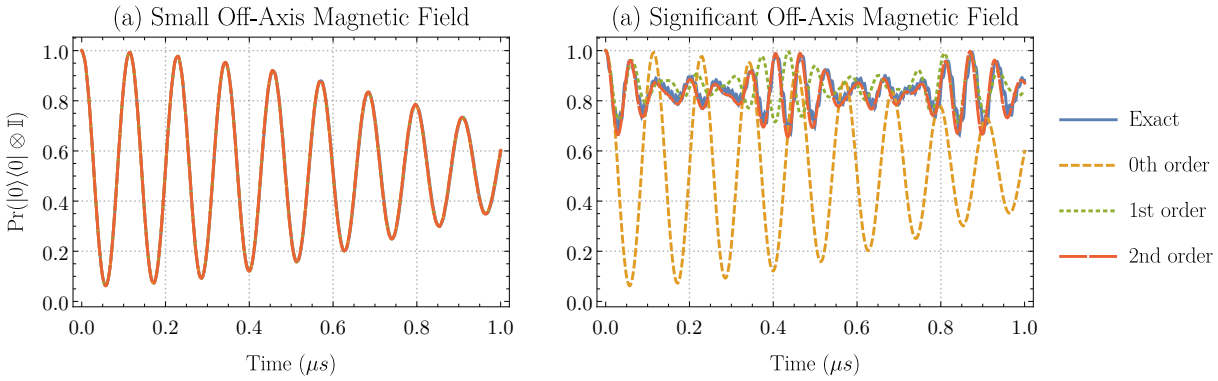

Figure 2.2: Comparison between an 'exact' simulation and various average Hamiltonian simulations for a system with one NV^- and one 14-nitrogen. The simulation consists of a 1 us evolution under the constant parameters $\Delta = 2\pi \cdot 2.87\,\text{GHz}$, $B_z = 20\,\text{G}$, $\omega_{\mu w} = \Delta - \gamma_e B_z$, $Q = -2\pi \cdot 5.01\,\text{MHz}$, $A_{xx} = A_{yy} = 2\pi \cdot 2.14\,\text{MHz}$, $A_{zz} = 2\pi \cdot 2.7\,\text{MHz}$, $\Omega = 2\pi \cdot 6\,\text{MHz}$, $\phi = 0$, and $B_x = 2\,\text{G}$ for (a), $B_x = 60\,\text{G}$ for (b). The initial state is $\rho_0 = |0\rangle\langle 0| \otimes \mathbb{I}_3/3$ and the measurement projection is $P_0 = |0\rangle\langle 0| \otimes \mathbb{I}_3$. Since these both commute with the frame tranformation $U_{\text{rot}}(t)$, we don't need to worry about exiting and entering the rotating frame in this comparison. Exact simulation is done by time-slicing Equation 2.1 at 10 ps intervals. Average Hamiltonian simulation is done by exponentiating sums of expressions in Equation 2.15 at 1 ns intervals.

Simulation under $H_{\text{avg}} = \overline{H}_0$, $H_{\text{avg}} = \overline{H}_0 + \overline{H}_1$, $H_{\text{avg}} = \overline{H}_0 + \overline{H}_1 + \overline{H}_2$, etc. successively improves the approximation, as demonstrated in Figure 2.2. Indeed, in the secular approximation Equation 2.16 we see that static magnetic fields off-axis to z on the NV^- center are completely ignored. However, higher order corrections account for this; the first order correction will add terms of order $B_z B_x / \omega_{\mu w}$ which we see are relevant even 50 ns into the simulation. Given that the formulas in Equation 2.9 only involve a couple dozen matrix multiplications in total (essentially the price of a single matrix exponential), from the perspective of simulation, the cost of using an approximation better than $\overline{H}_0$ is negligible.

2.4 A Second Hamiltonian Rotating Frame

The previous section described a practical method to enter the rotating frame $H_{\text{rot}} = \omega_{\mu w} \mathbf{S}_z^2$ and calculate the average Hamiltonians of the effective Hamiltonian in this frame. One of the primary assumptions of that derivation was that the microwave profile defined through

$\Omega(t)$ and $\phi(t)$ should appear at a much lower frequency than the carrier frequency $\omega_{\mu w}$. Specifically, we mean that the Fourier transform of $\Omega(t)\mathrm{e}^{\mathrm{i}\phi(t)}$ should contain most of its support on an interval $[-\omega_0, \omega_0]$ for some $0 < \omega_0 \ll \Delta$. The electron transition frequencies $|0\rangle \leftrightarrow |+1\rangle$ and $|0\rangle \leftrightarrow |-1\rangle$ happen at roughly the frequencies $\Delta + \gamma_e B_z$ and $\Delta - \gamma_e B_z$, respectively. If one is interested in doing experiments that involve only the states $|0\rangle$ and $|+1\rangle$ (with $|-1\rangle$ being a nuissance level), for example, then the results of the previous section are sufficient, since one may set $\omega_{\mu w}$ near one of the transitions $\Delta \pm \gamma_e B_z$.

If, however, one is interested in involving both transitions in the same experiment—as we will be—then it is sometimes helpful to enter a second rotating frame. In particular, in section Section 7.2, we found entering this frame when designing quantum gates was essential to success, as it dramatically improved the control landscape. We begin by defining $\omega_e = \gamma_e B_z$, and noting that this strategy will only work in the intermediate regime $0 \ll \omega_e \ll \Delta$, for example, $\omega_e = 2\pi \cdot 50\,\mathrm{MHz} \approx \Delta/60$ is a reasonable value.

Next, note that the pulse profile $(\Omega(t), \phi(t))$ can be alternatively written in Cartesian form with the transformation

$$\omega_x(t) = \Omega(t)\cos\phi(t) \quad \text{and} \quad \omega_y(t) = \Omega(t)\sin\phi(t). \tag{2.17}$$

If we wish to concentrate microwave power around both electron transitions, then it is sensible to set $\omega_{\mu w} \approx \Delta$ and enforce the forms

$$\omega_x(t) = \omega_1(t)\cos(\omega_e t + \phi_1(t)) \quad \text{and} \quad \omega_y(t) = \omega_2(t)\cos(\omega_e t + \phi_2(t)) \tag{2.18}$$

on our profile. In this way, in the absence of very nearby 13-carbon atoms, the static part of H_{eff} will be dominated by the term $\omega_e \mathbf{S}_z$, and ω_e will be the primary frequency appearing in our controls in the rotating frame.

Therefore, after computing our average Hamiltonian from the first frame, we enter the second rotating frame $U_{\text{rot},2}(t) = \mathrm{e}^{\mathrm{i}tH_{\text{rot},2}}$ where $H_{\text{rot},2} = \omega_e \mathbf{S}_z$. The strategy is the same as before, except now we have the conjugation rule

$$\begin{aligned} U_{\text{rot},2}(t) X U_{\text{rot},2}(t)^\dagger &= \begin{pmatrix} \mathrm{e}^{\mathrm{i}\omega_e t}\mathbb{I} & 0 & 0 \\ 0 & \mathbb{I} & 0 \\ 0 & 0 & \mathrm{e}^{-\mathrm{i}\omega_e t}\mathbb{I} \end{pmatrix} \begin{pmatrix} A & B & C \\ D & E & F \\ G & H & I \end{pmatrix} \begin{pmatrix} \mathrm{e}^{-\mathrm{i}\omega_e t}\mathbb{I} & 0 & 0 \\ 0 & \mathbb{I} & 0 \\ 0 & 0 & \mathrm{e}^{\mathrm{i}\omega_e t}\mathbb{I} \end{pmatrix} \\ &= \mathrm{e}^{-i\omega_e t}\begin{pmatrix} 0 & 0 & C \\ D & 0 & 0 \\ 0 & H & 0 \end{pmatrix} + \begin{pmatrix} A & 0 & 0 \\ 0 & E & 0 \\ 0 & 0 & I \end{pmatrix} + \mathrm{e}^{+i\omega_e t}\begin{pmatrix} 0 & B & 0 \\ 0 & 0 & F \\ G & 0 & 0 \end{pmatrix} \\ &\equiv \mathrm{e}^{-i\omega_e t} P_{\smile}[X] + P_{\backslash}[X] + \mathrm{e}^{+i\omega_e t} P_{\frown}[X] \end{aligned} \tag{2.19}$$

that can be used in equations similar to Equation 2.13, Equation 2.14, and Equation 2.15, and as outlined in Algorithm 2. This completes the description of the approximation.

We write

$$H_{m,n} \tag{2.20}$$

to denote the average effective Hamiltonian that results from using average Hamiltonian orders $\overline{H}_0 + \cdots + \overline{H}_m$ in the first rotating frame, and average Hamiltonian orders $\overline{H}_0 + \cdots + \overline{H}_n$ in the second rotating frame. For example, rewriting $\overline{H}_0$ from Section 2.3 in terms of Equation 2.18 we get

$$\overline{H}_0 = H_{\text{static}} + \frac{1}{2}\left[\omega_1 \mathrm{e}^{\mathrm{i}\omega_e t}\mathrm{e}^{\mathrm{i}\phi_1}\mathbf{S}_\mathrm{x} + \omega_1 \mathrm{e}^{-\mathrm{i}\omega_e t}\mathrm{e}^{-\mathrm{i}\phi_1}\mathbf{S}_\mathrm{x}\right] - \frac{\mathrm{i}}{2}\left[\omega_2 \mathrm{e}^{\mathrm{i}\omega_e t}\mathrm{e}^{\mathrm{i}\phi_2}\mathbf{S}'_\mathrm{y} - \omega_2 \mathrm{e}^{-\mathrm{i}\omega_e t}\mathrm{e}^{-\mathrm{i}\phi_2}\mathbf{S}'_\mathrm{y}\right] \tag{2.21}$$

so that

$$\begin{aligned} H_{0,0} &= P_{\backslash}[H_{\text{static}} - \omega_e \mathbf{S}_\mathrm{z}] + P_{\smile}[\omega_1 \mathrm{e}^{\mathrm{i}\phi_1}\mathbf{S}_\mathrm{x} - \mathrm{i}\omega_2 \mathrm{e}^{\mathrm{i}\phi_2}\mathbf{S}'_\mathrm{y}]/2 + P_{\frown}[\omega_1 \mathrm{e}^{-\mathrm{i}\phi_1}\mathbf{S}_\mathrm{x} + \mathrm{i}\omega_2 \mathrm{e}^{-\mathrm{i}\phi_2}\mathbf{S}'_\mathrm{y}]/2 \\ &= (\Delta - \omega_{\mu w})\mathbf{S}_\mathrm{z}^2 + \gamma_N \vec{B}\cdot\mathbf{I} + \sum_i \gamma_C \vec{B}\cdot\mathbf{I}_i \\ &\quad + A_{zz}\mathbf{S}_\mathrm{z}\mathbf{I}_\mathrm{z} + \sum_i \mathbf{S}_\mathrm{z}(A_{zx,i}\mathbf{I}_\mathrm{x} + A_{zy,i}\mathbf{I}_\mathrm{y} + A_{zz,i}\mathbf{I}_\mathrm{z}) \\ &\quad + \sum_i \mathbf{I}\cdot D_i \cdot \mathbf{I}_i + \sum_{i<j} \mathbf{I}_i \cdot D_{ij}\cdot \mathbf{I}_j + \mathbf{I}\cdot Q\cdot \mathbf{I} \\ &\quad + \frac{1}{2}\left[\omega_1 \cos\phi_1 \mathbf{S}_\mathrm{x} + \omega_1 \sin\phi_1 \mathbf{S}_\mathrm{y} - \omega_2 \cos\phi_2 \mathbf{S}'_\mathrm{x} + \omega_2 \sin\phi_2 \mathbf{S}'_\mathrm{y}\right]. \end{aligned} \tag{2.22}$$

The control term of $H_{0,0}$ can be separated into simple transitions of the eigenstates by defining

$$\begin{aligned} \omega_{+1,x} &= \omega_1 \cos\phi_1 + \omega_2 \cos\phi_2 & \omega_{-1,x} &= \omega_1 \cos\phi_1 - \omega_2 \cos\phi_2 \\ \omega_{+1,y} &= \omega_1 \sin\phi_1 + \omega_2 \sin\phi_2 & \omega_{-1,y} &= \omega_1 \sin\phi_1 - \omega_2 \sin\phi_2 \end{aligned} \tag{2.23}$$

which gives

$$H_{0,0,\text{control}} = \omega_{+1,x}\mathbf{S}_\mathrm{x}^+ + \omega_{-1,x}\mathbf{S}_\mathrm{x}^- + \omega_{+1,y}\mathbf{S}_\mathrm{y}^+ + \omega_{-1,y}\mathbf{S}_\mathrm{y}^-. \tag{2.24}$$

The reverse mapping is discussed in Section 7.2.3, where control pulses are converted from this second rotating frame back into the first. This shows, for example, that we can drive the $|0\rangle \leftrightarrow |-1\rangle$ transition along the y direction by arranging the phases ϕ_1, ϕ_2 and amplitudes ω_1, ω_2 so that only $\omega_{-1,y} \neq 0$: simply choose $\omega_1 = \omega_2$, $\phi_1 = \pi/2$, and $\phi_2 = -\pi/2$.

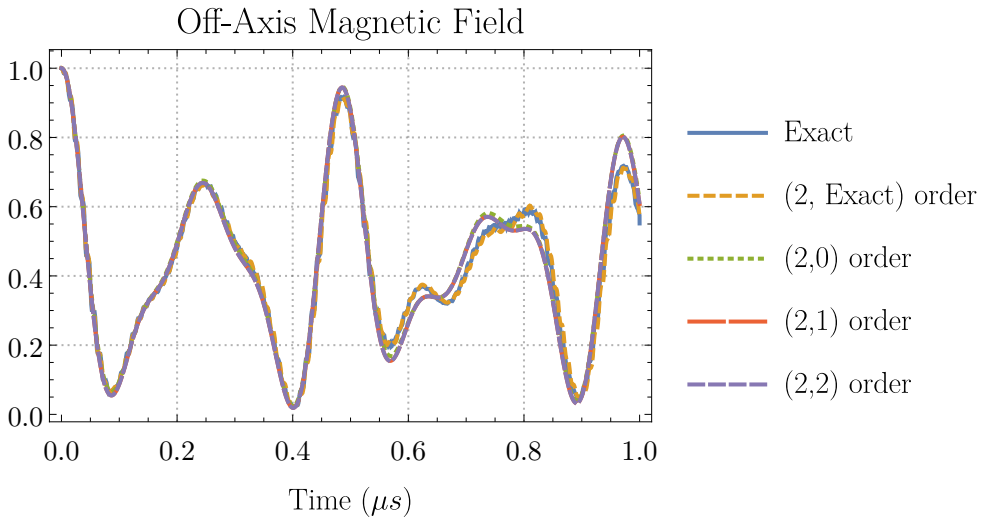

Figure 2.3: Comparison between an 'exact' simulation and various stacked average Hamiltonian simulations for a system with one NV^- and one 14-nitrogen, in which we are modulating the microwave control field at a frequency of $\omega_e = 2\pi \cdot 50\,\text{MHz}$. The simulation consists of a 1 us evolution under the constant parameters $\Delta = 2\pi \cdot 2.87\,\text{GHz}$, $\omega_{\mu w} = \Delta$, $Q = -2\pi \cdot 5.01\,\text{MHz}$, $A_{xx} = A_{yy} = 2\pi \cdot 2.14\,\text{MHz}$, $A_{zz} = 2\pi \cdot 2.7\,\text{MHz}$, $\omega_1 = 2\pi \cdot 6\,\text{MHz}$, $\phi_1 = \omega_2 = \phi_2 = 0$, and $B_x = 20\,\text{G}$. Simulation details are the same as Figure 2.2, where '(m, n) order' refers to the average Hamiltonian $H_{m,n}$ of Equation 2.20.

It is certainly worthwhile to ask whether it was necessary to enter these two rotating frames separately. The issue with entering both frames at once, using $H_{\text{rot}} = \Delta \mathbf{S}_{\text{z}}^2 + \omega_e \mathbf{S}_{\text{z}}$, is that multiple frequencies appear in the effective Hamiltonian. Using a stroboscopic scheme in such a case is not possible unless one can find a common period of all frequencies that is not too long. Therefore, using a joint frame transformation is less general than what we have described above. However, if one is free to precisely set ω_e in an experiment, then it is worthwhile to enforce a common period by design—a good choice would be to make ω_e an integer divisor of Δ, such as $\omega_e = 2\pi \cdot 47.8333\,\text{MHz} = \Delta/N$, where $N = 60$, in which case one would have a Floquet expansion of the form

$$H = \sum_{k=\pm\{0,N-1,N+1\}} \mathrm{e}^{\mathrm{i}k\omega_e} H_k \tag{2.25}$$

for our original Hamiltonian Equation 2.1 assuming the controls from Equation 2.18. The conjugation rule for this transformation has seven terms instead of three, given by

$$\begin{aligned}
&U_{\text{rot}}(t) X U_{\text{rot}}(t)^\dagger \\
&= \begin{pmatrix} \mathrm{e}^{\mathrm{i}(N+1)\omega_e t}\mathbb{I} & 0 & 0 \\ 0 & \mathbb{I} & 0 \\ 0 & 0 & \mathrm{e}^{\mathrm{i}(N-1)\omega_e t}\mathbb{I} \end{pmatrix} \begin{pmatrix} A & B & C \\ D & E & F \\ G & H & I \end{pmatrix} \begin{pmatrix} \mathrm{e}^{-\mathrm{i}(N+1)\omega_e t}\mathbb{I} & 0 & 0 \\ 0 & \mathbb{I} & 0 \\ 0 & 0 & \mathrm{e}^{\mathrm{i}(N-1)\omega_e t}\mathbb{I} \end{pmatrix} \\
&= \mathrm{e}^{-\mathrm{i}(N+1)\omega_e t} \begin{pmatrix} 0 & 0 & 0 \\ D & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \mathrm{e}^{-\mathrm{i}(N-1)\omega_e t} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & F \\ 0 & 0 & 0 \end{pmatrix} \\
&\quad + \mathrm{e}^{-2\mathrm{i}\omega_e t} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ G & 0 & 0 \end{pmatrix} + \begin{pmatrix} A & 0 & 0 \\ 0 & E & 0 \\ 0 & 0 & I \end{pmatrix} + \mathrm{e}^{2\mathrm{i}\omega_e t} \begin{pmatrix} 0 & 0 & C \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \\
&\quad + \mathrm{e}^{\mathrm{i}(N-1)\omega_e t} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & H & 0 \end{pmatrix} + \mathrm{e}^{\mathrm{i}(N+1)\omega_e t} \begin{pmatrix} 0 & B & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}
\end{aligned} \tag{2.26}$$

for any $X \in \mathrm{L}(\mathcal{H})$, where $U_{\text{rot}}(t) = \mathrm{e}^{\mathrm{i}t(N\omega_e \mathbf{S}_{\text{z}}^2 + \omega_e \mathbf{S}_{\text{z}})}$, which again, can be used in Algorithm 1 and Algorithm 2. The time-dependent effective Hamiltonian ends up with Floquet terms at the following orders:

$$\pm\{0, 2, N-3, N-1, N+1, N+3, 2N-2, 2N, 2N+2\}. \tag{2.27}$$

Algorithm 1 The function EffectiveHamiltonian takes as input a Hamiltonian Floquet expansion (see Equation 2.14) in the form of a mapping $\langle k : H_k \rangle$ with integer keys k, and outputs a new Floquet expansion, representing the same Hamiltonian in the rotating frame defined by ω and the block projections of BlockProject.

```
function BlockProject(p, X)                    ▷ See ex. Equation 2.11, Equation 2.19, Equation 2.26
    return <Relevant projection of X for e^{ipω}>
end function
function EffectiveHamiltonian(H_floq = ⟨k : H_k⟩, P)
    H_eff ← ⟨⟩                                 ▷ Empty mapping with default value 0
    for all k, H_k ∈ H_floq do
        for all p ∈ P do                       ▷ Loop over projection integers
            H_eff[p + k] += BlockProject(p, H_k)   ▷ Matrix coefficient of e^{i(p+k)ω}
        end for
    end for
    return H_eff
end function
```

2.5 Using Average Hamiltonian Approximations

In the previous two subsections we have provided detailed methods for computing average Hamiltonians of the NV system. It merits briefly discussing, now, when to use average Hamiltonians, how many terms to include, and other practical matters.

Fundamentally, average Hamiltonian theory provides two resources: (i) it makes simulation more efficient by increasing the time resolution at which one needs to simulate, and (ii) it provides insight into quantum systems by dividing dynamics into successive orders of importance, each of which are likely easier to understand than the original time-dependent Hamiltonian. The latter resource is very powerful, and has been one of the main methods of analyzing and developing pulse sequences in NMR. This thesis is primarily interested in the former resource, owing to our heavy use of numerics and inference. In the remainder of this subsection, advice is given with respect to simulation accuracy.

The number of average Hamiltonian terms one should use depends on the parameter regime one is in, the intended application, and the types of experiments being simulated. Therefore, we offer the following generic, best practice advice (which also serves as a sanity check against implementation errors):

1. Simulate the intended application in the regime of interest under the exact Hamilto-

Algorithm 2 The function AVERAGEHAMILTONIAN computes the average Hamiltonian of the effective Hamiltonian (see Algorithm 1) generated by moving H into the rotating frame H_{rot}, assuming this frame has the correct projection rules defined by BLOCKPROJECT. Floquet expansions as found in Equation 2.14 are written as mappings $\langle k : H_k \rangle$ with integer keys k.

function COM($H_{\text{floq}} = \langle k : H_k \rangle$, m, n)
  return $H_{\text{floq}}[m] \cdot H_{\text{floq}}[n] - H_{\text{floq}}[n] \cdot H_{\text{floq}}[m]$
end function
function AVGHAMTERM(0, $H_{\text{eff}} = \langle k : H_k \rangle$, ω)
  return $H_{\text{eff}}[0]$
end function
function AVGHAMTERM(1, $H_{\text{eff}} = \langle k : H_k \rangle$, ω)
  $\overline{H} \leftarrow 0$
  for all $k \in H_{\text{eff}}$ **do**
    if $k \neq 0$ **then**
      $\overline{H}$ += COM($H_{\text{eff}}, 0, k$)/($k\omega$)
      $\overline{H}$ −= COM($H_{\text{eff}}, -k, k$)/($2k\omega$)
    end if
  end for
  return $\overline{H}$
end function
function AVGHAMTERM(2, $H_{\text{eff}} = \langle k : H_k \rangle$, ω)
  return <See Equation 2.15>
end function
function AVERAGEHAMILTONIAN(order, $H = \langle k : H_k \rangle$, H_{rot}, P, ω)
  $H[0]$ −= H_{rot}
  $H_{\text{eff}} \leftarrow$ EFFECTIVEHAMILTONIAN(H, P)
  $H_{\text{avg}} \leftarrow 0$
  for $k = 0,\ldots,$order **do**
    H_{avg} += AVGHAMTERM($k, H_{\text{eff}}, \omega$)
  end for
  return H_{avg}
end function

nian. If time slicing is manually chosen, decrease the time-step size until the results converge.

2. Simulate the intended application under successive average Hamiltonian approximations until the solution aligns with the exact solution to within a precision suitable for the application.

These steps were performed in Figure 2.2 in the case of evolution under static parameters.

We can also use back-of-the-envelope heuristics as follows. Let's take the example of the first rotating frame, $\omega_{\mu w}\mathbf{S}_{\mathrm{z}}^2$, discussed in Section 2.3. The secular approximation was stated in Equation 2.16. The accuracy of this approximation depends mostly on the size of the first order correction, $\overline{H}_1$. Observing Equation 2.15, we see that terms of $\overline{H}_1$ will have coefficients of order $\frac{cd}{\omega_{\mu w}}$, where c and d are coefficients appearing in $H - H_{\mathrm{rot}}$ whose operators do not commute. The largest coefficient in $H - H_{\mathrm{rot}}$ will often be the field along z, $\omega_e = \gamma_e B_z$, and a term which does not commute with $\mathbf{S}_{\mathrm{z}}$ is $\mathbf{S}_{\mathrm{x}}$. This leads to corrections of order $\Omega\frac{\omega_e}{\omega_{\mu w}}$, where Ω is the Rabi strength, and $\omega_x\frac{\omega_e}{\omega_{\mu w}}$, where $\omega_x = \gamma_e B_x$ is the static field along x. If, say, $\omega_e/\omega_{\mu w} \approx 0.05$ and there is a 1 G stray magnetic field along x, then this latter correction will have strength $\omega_x\frac{\omega_e}{\omega_{\mu w}} \approx 2\pi \cdot 140\,\mathrm{kHz}$, which will be relevant on timescales of $> 1\,\mathrm{us}$ for precision applications. One can use commutators of the corresponding operators to find out whether a given correction will enter as a nutation or a phase shift.

In Appendix A, a reference is provided to the Mathematica package 'nvham' that we have made available online. This package computes average Hamiltonians for the NV system, with a convenient interface for including nearby nuclear spins, and performing coordinate transformations between the PAS and other frames. As a simple example, and for brevity of the output equations, we continue the example of the previous paragraph with no nuclear spins, a static magnetic field along x and z, and a Rabi field along the x axis at frequency $\omega_{\mu w} = \Delta + \omega_e$. Thus in the lab frame the Hamiltonian is given by

$$H = \Delta\mathbf{S}_{\mathrm{z}}^2 + \omega_e\mathbf{S}_{\mathrm{z}} + \omega_x\mathbf{S}_{\mathrm{x}} + 2\Omega\cos(2\pi(\Delta + \omega_e)t)\mathbf{S}_{\mathrm{x}}. \tag{2.28}$$

In Listing 2.1 we analytically compute $\overline{H}_0$ and $\overline{H}_1$ for this example using 'nvham'. The secular output of this code sample is

$$\overline{H}_0 = \begin{pmatrix} 0 & \frac{\Omega}{\sqrt{2}} & 0 \\ \frac{\Omega}{\sqrt{2}} & 0 & \frac{\Omega}{\sqrt{2}} \\ 0 & \frac{\Omega}{\sqrt{2}} & -2\omega_e \end{pmatrix} \tag{2.29}$$

where we see the static Rabi term $\Omega \mathbf{S}_x$ as expected, and where the $m_s = -1$ energy level is off resonance by $2\omega_e$, also as expected. The first order correction is given by

$$\overline{H}_1 = \begin{pmatrix} -\frac{\Omega^2+4\omega_x\Omega-2\omega_x^2}{4(\Delta+\omega_e)} & 0 & -\frac{\Omega^2+4\omega_x\Omega-2\omega_x^2}{4(\Delta+\omega_e)} \\ 0 & \frac{\Omega^2+4\omega_x\Omega-2\omega_x^2}{2(\Delta+\omega_e)} & -\frac{\omega_e(\Omega+2\omega_x)}{\sqrt{2}(\Delta+\omega_e)} \\ -\frac{\Omega^2+4\omega_x\Omega-2\omega_x^2}{4(\Delta+\omega_e)} & -\frac{\omega_e(\Omega+2\omega_x)}{\sqrt{2}(\Delta+\omega_e)} & -\frac{\Omega^2+4\omega_x\Omega-2\omega_x^2}{4(\Delta+\omega_e)} \end{pmatrix} \tag{2.30}$$

We see a leakage nutation error (we have entered a frame targeted at $|0\rangle \leftrightarrow |+1\rangle$ transitions) between $|0\rangle \leftrightarrow |-1\rangle$ at strength $\frac{\omega_e(\Omega+2\omega_x)}{(\Delta+\omega_e)}$. We also have a frequency shift term of strength $-\frac{\Omega^2+4\omega_x\Omega-2\omega_x^2}{4(\Delta+\omega_e)}$ with diagonal operator $\mathrm{diag}(1,2,1)$; the pieces of this with Ω^2 in the numerator are the well-known Bloch-Siegert shift—shifts in resonance frequency induced by the drive field. Finally, there are double quantum nutation errors, with the same strength as the frequency shift error. In Figure 2.2(b) the huge discrepancy between the secular approximation and its first order correction is due to the significant frequency shift induced by the large off-axis static magnetic field—full Rabi flops are not observed in the corrections because the applied microwaves are effectively off-resonance.

Finally, we mention that, for us, the second rotating frame has two purposes. The first, as will be seen in Section 7.2, is to improve the control landscape when pulse finding. The second, as mentioned in the paragraph following Equation 2.24, is that it makes clear which cosine-mixed sinusoids generate which single quantum transitions, which can be useful in simple square or gaussian pulse design. The improvement to simulation efficiency alone is likely not worth the effort of switching into the second rotating frame.

Listing 2.1: Computing the average Hamiltonian terms $\overline{H}_0$ and $\overline{H}_1$ for Equation 2.28.

```
Needs["NVHamiltonian`"];
Bfield = Vector[{Ωx+2Ω Cos[2π(Δ + Ωe)t],0,Ωe}/γe, Cartesian];
H0 = NVAverageHamiltonian[
        0, Δ + Ωe,
        MagneticField → Bfield,
        AngularUnits → True
    ]/(2π) // FullSimplify;
H1 = NVAverageHamiltonian[
        1, Δ + Ωe,
        MagneticField → Bfield,
        AngularUnits → True
    ]/(2π) - H0 // FullSimplify;
```

2.6 Optical Dynamics of the NV Center in Diamond

The previous sections considered dynamics in the optical ground state alone. Such dynamics are usually sufficient to describe experiments because the more complicated optical dynamics are most often only used for initialization and readout, and therefore separated from spin manipulation. Here we briefly summarize the optical dynamics of the NV^-center in diamond, setting up our notation for future sections. However, because of the detail that we will pay to analyzing measurement data from this system, it is necessary for us to summarize these dynamics in detail. Yet more detailed descriptions can be found elsewhere, see, for example, the review article [38] and references therein.

The NV^- center can absorb and emit photons. The energy difference between the optical ground state and the first excited state is 637 nm (red). At room temperature, however, a spontaneously emitted photon is usually accompanied with with the absorption or emission of a phonon of varying energy, so that the emission spectrum of an NV^-center is spread out from about 600nm to 800nm, with a small zero-phonon spike at 637nm [70]. Many NV experiments in literature elect to work at cryogenic temperatures to reduce phonons, thereby enabling spectroscopic distinction between different spin-state decays or applications that require single mode photon output (our experiments, however, are all performed at room temperature). It is experimentally convenient to optically excite the defect from the ground state to the first excited state using light with a wavelength outside of the emission spectrum, for example 532 nm (green). This is possible due to the presence of higher energy states above the first excited state which very quickly decay to the first excited state, while preserving spin populations. Separating red and green allows a confocal microscope to be set up so that optical cycling of a single isolated NV^-center can be studied [70]: incident green light is delivered to a small region inside of a diamond, roughly a cubic micron in volume, and red light is extracted from the same region. Assuming the use of a diamond whose impurities are sparse enough, this region can be chosen to contain a single NV^- center.

The dynamics of the NV^- center are usually described for the spin system alone, with the optical degrees of freedom assumed to be in the ground state. However, since we are interested in the measurement process, we must include both degrees of freedom. We describe the dynamics using a seven level system: three levels for the optical ground state spin system, three levels for an optical excited state spin system, and one level for an optical inter-system-crossing (ISC). It is known that more levels exist [38], but adding them to this discussion will not change our model of the measurement process.

We decompose our Hilbert space as the direct sum

$$\mathcal{H} = \mathcal{H}_{\text{ground}} \oplus \mathcal{H}_{\text{excited}} \oplus \mathcal{H}_{\text{isc}}, \tag{2.31}$$

where $\dim(\mathcal{H}_{\text{ground}}) = \dim(\mathcal{H}_{\text{excited}}) = 3$ and $\dim(\mathcal{H}_{\text{isc}}) = 1$. We define a basis for $\mathcal{H}$ as

$$|\text{g}, +1\rangle = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad |\text{g}, 0\rangle = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad |\text{g}, -1\rangle = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad |\text{e}, +1\rangle = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \tag{2.32}$$

$$|\text{e}, 0\rangle = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad |\text{e}, -1\rangle = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad |\text{s}\rangle = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \tag{2.33}$$

where $\{+1, 0, -1\}$ are spin labels corresponding to the eigenvalues of $\mathbf{S}_{\text{z}} = \text{diag}(1, 0, -1)$, and $(\text{g}, \text{e}, \text{s})$ refer to optical ground, excited and singlet, respectively. In this way we can write, for example, the spin-1 z operator in the optical ground state as $\mathbf{S}_{\text{z}} \oplus 0 \oplus 0 = |\text{g}, +1\rangle\langle\text{g}, +1| - |\text{g}, -1\rangle\langle\text{g}, -1|$.

Note that throughout this thesis when we omit the electronic label from an NV^- ket, as we did in the previous sections of this chapter, our convention is that it refers to the ground state, i.e., $|0\rangle = |\text{g}, 0\rangle$, $|-1\rangle = |\text{g}, -1\rangle$, and $|+1\rangle = |\text{g}, +1\rangle$.

Given an external magnetic field applied along the z-axis with strength ω_e (in angular frequency units), the Hamiltonian of the system is given by

$$H = H_{\text{ground}} \oplus 0 \oplus 0 + 0 \oplus H_{\text{excited}} \oplus 0 \tag{2.34}$$

with

$$\begin{aligned} H_{\text{ground}} &= \Delta_g \mathbf{S}_{\text{z}}^2 + \omega_e \mathbf{S}_{\text{z}} \\ H_{\text{excited}} &= \Delta_e \mathbf{S}_{\text{z}}^2 + \omega_e \mathbf{S}_{\text{z}} \end{aligned} \tag{2.35}$$

The terms $\Delta_g \approx 2\pi \cdot 2.87\,\text{GHz}$ and $\Delta_e \approx 2\pi \cdot 1.4\,\text{GHz}$ denote the zero-field splittings. As the name implies, they are energies that enter into the Hamiltonian without the application of an external field; they result from the couplings between the electrons constituting the NV^- center.

Absorption of a photon takes the system from $\mathcal{H}_{\text{ground}}$ to $\mathcal{H}_{\text{excited}}$, and vice versa for spontaneous photon emission. These processes are known to be spin-conserving. Although coherent optical control is possible, we restrict our attention to the more commonly used dissipative regime. Therefore, we describe excitation and spontaneous emission using spin-conserving Lindblad operators,

$$L_1 = \sqrt{\gamma_{eg}}\left(|\text{g},+1\rangle\langle\text{e},+1| + |\text{g},0\rangle\langle\text{e},0| + |\text{g},-1\rangle\langle\text{e},-1|\right) \tag{2.36}$$

$$L_2 = \sqrt{k\cdot\gamma_{eg}}\left(|\text{e},+1\rangle\langle\text{g},+1| + |\text{e},0\rangle\langle\text{g},0| + |\text{e},-1\rangle\langle\text{g},-1|\right) \tag{2.37}$$

where γ_{eg} is the rate of spontaneous emission, about 77 MHz. Without the additional dynamics described below, this would imply an average excited state lifetime of about $1/\gamma_{ea} = 13\,\text{ns}$. The dimensionless parameter k corresponds to the laser power (as a fraction of the spontaneous emission rate γ_{eg}) and in our experiments is typically on the order of unity when the laser is on. It can be set to 0 in periods where the laser is off.

Spin selective measurement is possible because of an additional decay path that is not spin-conserving and emits photons with a wavelength outside of the 600 nm–800 nm emission spectrum. This route preferentially allows the excited $|\text{e},-1\rangle$ and $|\text{e},+1\rangle$ states to decay to the ground states through the ISC to the state $|\text{s}\rangle$. It can be modeled using the Lindblad dissipaters

$$L_3 = \sqrt{\gamma_{es}/2}\,|\text{s}\rangle\langle\text{e},+1| \tag{2.38}$$

$$L_4 = \sqrt{\gamma_{es}/2}\,|\text{s}\rangle\langle\text{e},-1| \tag{2.39}$$

$$L_5 = \sqrt{\gamma_{sg}/3}\,|\text{g},+1\rangle\langle\text{s}| \tag{2.40}$$

$$L_6 = \sqrt{\gamma_{sg}/3}\,|\text{g},0\rangle\langle\text{s}| \tag{2.41}$$

$$L_7 = \sqrt{\gamma_{sg}/3}\,|\text{g},-1\rangle\langle\text{s}|\,. \tag{2.42}$$

The first two move support on the excited $|\text{e},-1\rangle$ and $|\text{e},+1\rangle$ states to the ISC state. The last three spread ISC population approximately evenly (and incoherently) over the ground space. This may seem counterintuitive given that optical excitation of sufficient duration results in high spin state polarization; the resolution is that the selective decay from $|\text{e},+1\rangle, |\text{e},-1\rangle$ to $|\text{s}\rangle$ is what actually drives polarization, as measured by [146]. The

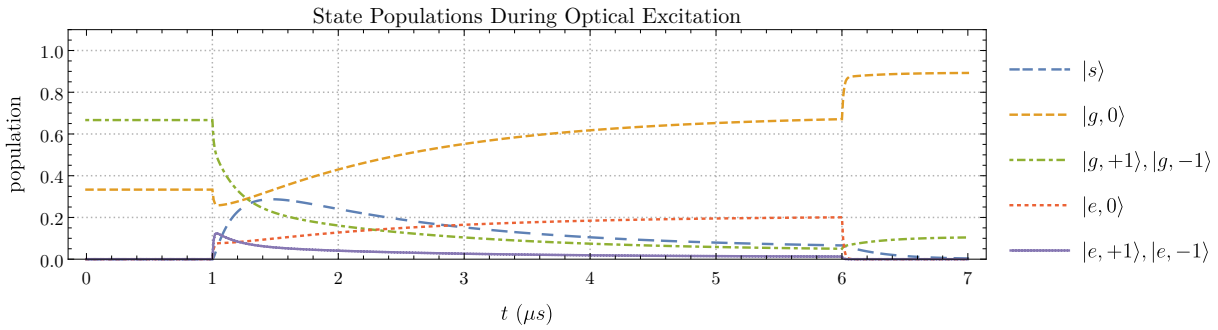

Figure 2.4: The quantum state is initialized to fully mixed state in the optical ground spin-1 manifold, which is approximately the Boltzmann distribution at room temperature and low magnetic field. Therefore at $t = 0$, the combined $|\mathrm{g}, +1\rangle$, $|\mathrm{g}, -1\rangle$ manifold has twice as much population as $|\mathrm{g}, 0\rangle$. A laser pulse of duration 5μs is applied starting at 1μs. The populations of the subspaces spanned by the pure states in the legend are tracked through time. The values used are $k = 0.3$, $\gamma_{eg} = 77\,\mathrm{MHz}$, $\gamma_{es} = 30\,\mathrm{MHz}$, $\gamma_{sg} = 3\,\mathrm{MHz}$, and $\gamma_{01} = 0.5\,\mathrm{MHz}$. At 6μs when the laser pulse is turned off, the excited states quickly decay to the ground states, and the singlet state slowly leaks back to the ground state; normally we wait around 5 or 10 times the characteristic decay time before resuming activity. Note that full polarization will never be reached due to the nonzero mixing rate γ_{01}.

rate of the spin-selective decay is roughly $\gamma_{es} = 30\,\mathrm{MHz}$, and comparing this to γ_{ea}, shows that the excited ± 1 states take the ISC path roughly $1/3^{\mathrm{rd}}$ of the time. The lifetime of the singlet $|\mathrm{s}\rangle$ is quite long, with a decay rate of roughly $\gamma_{sg} = 3\,\mathrm{MHz}$; this is the time scale that will end up dominating the optimal measurement time.

Small non-spin conserving transitions are estimated to have a rate of about $\gamma_{01} = 1\,\mathrm{MHz}$ [38]. They can be modelled as the Lindblad operators

$$\begin{aligned} L_8 &= \sqrt{\gamma_{01}/4}\,|\mathrm{g}, +1\rangle\langle \mathrm{e}, 0| \\ L_9 &= \sqrt{\gamma_{01}/4}\,|\mathrm{g}, -1\rangle\langle \mathrm{e}, 0| \\ L_{10} &= \sqrt{\gamma_{01}/4}\,|\mathrm{g}, 0\rangle\langle \mathrm{e}, +1| \\ L_{11} &= \sqrt{\gamma_{01}/4}\,|\mathrm{g}, 0\rangle\langle \mathrm{e}, -1|\,. \end{aligned} \tag{2.43}$$

Solving the Lindblad master equation (introduced in Section 1.2.3),

$$\frac{\mathrm{d}\rho(t)}{\mathrm{d}t} = -\mathrm{i}[H, \rho(t)] + \sum_{i=1}^{11} \left(L_k \rho(t) L_i^\dagger - (\rho(t) L_i^\dagger L_i + L_i^\dagger L_i \rho(t))/2 \right), \tag{2.44}$$

with some initial state $\rho(0)$ allows us to track the populations and coherences of the quantum system through time. A simulation of this master equation is plotted in Figure 2.4. If we start with spin-state coherences, they will quickly die off due to the mismatch in zero-field splittings between the optical ground and excited states. Since we have no coherence generating terms in our internal Hamiltonian, describing the optical dynamics in a fully quantum setting is overkill. If we were to simultaneously turn on a resonant microwave field and the green laser, the simplifications we will make in Section 2.7.2 would not apply.

2.7 Measurement Dynamics

It is possible to gain information about the spin state of the NV system by simultaneously illuminating it and counting the photons it emits in the process. This works because the ISC is spin selective. States initially with support on the subspace $\mathrm{span}(|\mathrm{g}, -1\rangle, |\mathrm{g}, +1\rangle)$, once excited, have a decay path which does not emit a detectable photon. Therefore, states initially with support in this space, on average, appear dimmer. We now formalize this. A great many symbols were introduced in the previous subsection, and even more are introduced in this section—a comprehensive summary is provided in Table 2.4.

2.7.1 Measurement Description

A single measurement of an NV^- center consists of turning the laser on for some amount of time (on the order of 500 ns) and counting the spontaneously emitted photons in this duration. We assume in our model that no information about arrival time or spectral properties is recorded.

The probability of a quantum system spontaneously emitting a detectable photon within a short duration dt at time t is given by the product of the rate of spontaneous emission, the length of the time window, and probability of being in the excited triplet manifold [168]:

$$\Pr\left(\text{photon}; [t, t+dt]\right) = \gamma_{ea} \operatorname{Tr}\left[P_e \cdot \rho(t)\right] dt, \tag{2.45}$$

Symbol	Value	Description
$0, \pm 1$		Spin-1 labels
e, g, s		Excited ($\|\mathrm{e}, +1\rangle, \|\mathrm{e}, 0\rangle, \|\mathrm{e}, -1\rangle$), ground ($\|\mathrm{g}, +1\rangle \equiv \|+1\rangle, \|\mathrm{g}, 0\rangle \equiv \|0\rangle, \|\mathrm{g}, -1\rangle \equiv \|-1\rangle$), and singlet ($\|\mathrm{s}\rangle$) manifolds
γ_{eg}	77 MHz	Spontaneous decay rate from e to g
γ_{es}	30 MHz	Spontaneous decay rate from e to s (only applies to $\|e, \pm 1\rangle$)
γ_{sg}	3 MHz	Spontaneous decay rate from s to g
γ_{01}	0.5 MHz	Non-spin conserving spontaneous decay rate from e to g
k	~ 1	Laser excitation from g to e, as a unitless fraction of γ_{eg}
L_i		Lindblad jump operators describing measurement process
$\rho(t)$		In context, the density operator at time t when measurement begins at $t = 0$
R		Rate matrix of population dynamics (Equation 2.52)
$\vec{p}(t)$		5D vector describing populations in various subspaces; function of $\rho(t)$
$\mu(t)$	≤ 77 MHz	Instantaneous visible photon emission rate given state $\rho(t)$, sometimes denoted $\mu(t, \rho(t))$
Δt	~ 400 ns	Duration of measurement (see also Figure 3.4)
n_e	$\sim 0 - 20$	Number of visible NV photons emitted during single measurement process of pre-measurement state $\rho(0)$
Γ		Average background photon counts in duration Δt
η	$\sim 10^{-4}$	Fraction of n_e detected by photon detector, on average
n_d		Number of visible NV photons detected during single measurement process of pre-measurement state $\rho(0)$
μ_0	$\sim 10^{-3}$	Expected number of detections in single measurement, assuming pre-measurement state $\rho(0) = \|\mathrm{g}, 0\rangle$
μ_1	$\sim 0.7 \cdot 10^{-3}$	Expected number of detections in single measurement, assuming pre-measurement state $\rho(0) = \|\mathrm{g}, -1\rangle$
q	~ 0.9	Pseudo-pure state preparation purity
μ_0', μ_1'		Modifications of μ_0, μ_1 to account for pseudo-pure preparation
α		Bright reference; expected number of detections when $\rho(0) = \|\mathrm{g}, 0\rangle$, including background, finite visibility, and pseudo-pure preparation. May implicitly be summed over some number of measurement repetitions, N, depending on context
β		Dark reference; same as α, but for $\rho(0) = \|\mathrm{g}, -1\rangle$ (or $\|\mathrm{g}, +1\rangle$)
γ		Signal; same as α, but for $\rho(0)$ of interest
p		$p = \mathrm{Tr}[\|\mathrm{g}, 0\rangle\langle\mathrm{g}, 0\| \rho(0)]$ (distinct from $\vec{p}(t)$)

Table 2.4: Symbols used to describe measurement dynamics. Some values, such as background counts and visibility, may change significantly depending on the experimental setup—in such cases we quote values from our own setup.

where $P_e = |\mathrm{e},+1\rangle\langle\mathrm{e},+1| + |\mathrm{e},0\rangle\langle\mathrm{e},0| + |\mathrm{e},-1\rangle\langle\mathrm{e},-1|$ is the projector onto the excited subspace $\mathcal{H}_{\text{excited}}$ and $\rho(t) \in \mathrm{D}(\mathcal{H})$ is the state of the system at time t. This defines an inhomogeneous Poisson process, where the rate of events is time-dependent, given by

$$\mu(t) = \gamma_{ea} \operatorname{Tr}\left[P_e \cdot \rho(t)\right]. \tag{2.46}$$

This is a generalized version of the more common homogeneous Poisson process, where the event rate λ is constant, and the probability of k events in the time duration t is given by $\Pr(k) = \mathrm{e}^{-\lambda t}(\lambda t)^k/k!$. An inhomogeneous Poisson process has a similar formula given by

$$\Pr\left(n_e|\rho(0)\right) = \frac{\left(\int_0^{\Delta t}\mu(t)dt\right)^{n_e}}{n_e!}\mathrm{e}^{-\int_0^{\Delta t}\mu(t)dt} \tag{2.47}$$

where n_e is the number of emitted photons during the interval $[0, \Delta t]$. The expected number of emitted photons is then

$$\mathbb{E}[n_e|\rho(0)] = \int_0^{\Delta t}\mu(t)dt. \tag{2.48}$$

With our typical parameters, the expected number of emitted photons is on the order of a dozen. Optimal measurement times Δt will be discussed in Section 3.3.6.

Notice that we are conditioning on what we call *the pre-measurement state*, $\rho(0)$. If the parameters in the dynamics are known and fixed, then so too is $\rho(t)$ for $0 \le t \le \Delta t$ given the pre-measurement state. It follows that the inhomogeneous rate function $\mu(t)$ is conditioned upon the pre-measurement state, $\mu(t) = \mu(t|\rho(0))$, but we often omit this for notational simplicity, writing simply $\mu(t)$.

Given the parameters of the Hamiltonian and Lindblad dissipaters, the pre-measurement state, and an integration time, we now have a concrete method of calculating the expected number of photons emitted.

2.7.2 The Rate Equation Simplification

Due to the absence of coherence, a full open quantum system simulation of $\rho(t)$ is unnecessary to compute $\mu(t)$. Instead, we can reduce our dynamics to a rate equation picture without further approximation.

To this end, we define the probabilities

$$p_{g0}(t) = \mathrm{Tr}\left[|\mathrm{g},0\rangle\langle\mathrm{g},0|\,\rho(t)\right] \tag{2.49a}$$
$$p_{g1}(t) = \mathrm{Tr}\left[(|\mathrm{g},-1\rangle\langle\mathrm{g},-1| + |\mathrm{g},+1\rangle\langle\mathrm{g},+1|)\rho(t)\right] \tag{2.49b}$$
$$p_{e0}(t) = \mathrm{Tr}\left[|\mathrm{e},0\rangle\langle\mathrm{e},0|\,\rho(t)\right] \tag{2.49c}$$
$$p_{e1}(t) = \mathrm{Tr}\left[(|\mathrm{e},-1\rangle\langle\mathrm{e},-1| + |\mathrm{e},+1\rangle\langle\mathrm{e},+1|)\rho(t)\right] \tag{2.49d}$$
$$p_{s}(t) = \mathrm{Tr}\left[|\mathrm{s}\rangle\langle\mathrm{s}|\,\rho(t)\right], \tag{2.49e}$$

as well as the vector $\vec{p}(t) = (p_{g0}(t), p_{g1}(t), p_{e0}(t), p_{e1}(t), p_s(t))^{\mathrm{T}} \in [0,1]^5$. Here, $\rho(t)$ is the solution to the Lindblad master equation Equation 2.44. Notice that the components of this vector sum to unity, $\sum_{i=1}^{5} p_i(t) = 1$ because the projectors used in the definitions of its components sum to the identity.

Combining the master equation with the definitions from Equation Equation 2.49, we can compute the time evolution of each component of $\vec{p}(t)$. For example, we have

$$\begin{aligned}
\dot{p}_{g0}(t) &= \mathrm{Tr}[|\mathrm{g},0\rangle\langle\mathrm{g},0|\,\dot{\rho}(t)] \\
&= \mathrm{Tr}\left[|\mathrm{g},0\rangle\langle\mathrm{g},0|\left(-i[H,\rho(t)] + \sum_k \left(L_k\rho(t)L_k^\dagger - (\rho(t)L_k^\dagger L_k + L_k^\dagger L_k\rho(t))/2\right)\right)\right] \\
&= -k\gamma_{eg}p_{g0}(t) + \gamma_{eg}p_{e0}(t) + \frac{\gamma_{sg}}{3}p_s(t)
\end{aligned} \tag{2.50}$$

where we have skipped a few lines of algebra. In this way we end up with a set of coupled linear differential equations involving rates from the Lindblad operators which can be described by the matrix DE

$$\begin{pmatrix} \dot{p}_{g0}(t) \\ \dot{p}_{g1}(t) \\ \dot{p}_{e0}(t) \\ \dot{p}_{e1}(t) \\ \dot{p}_{s}(t) \end{pmatrix} = \begin{pmatrix} -k\gamma_{eg} & 0 & \gamma_{eg} & \gamma_{01}/2 & \gamma_{sg}/3 \\ 0 & -k\gamma_{eg} & \gamma_{01}/2 & \gamma_{eg} & 2\gamma_{sg}/3 \\ k\gamma_{eg} & 0 & -\gamma_{eg}-\gamma_{01}/2 & 0 & 0 \\ 0 & k\gamma_{eg} & 0 & -\gamma_{eg}-\gamma_{es}-\gamma_{01}/2 & 0 \\ 0 & 0 & 0 & \gamma_{es} & -\gamma_{sg} \end{pmatrix} \cdot \begin{pmatrix} p_{g0}(t) \\ p_{g1}(t) \\ p_{e0}(t) \\ p_{e1}(t) \\ p_{s}(t) \end{pmatrix}, \tag{2.51}$$

which we write in condensed notation as

$$\dot{\vec{p}}(t) = R \cdot \vec{p}(t). \tag{2.52}$$

Notice that the columns of R sum to 0 which ensures that $\vec{p}(t)$ remains a probability vector as it evolves. This condensation of the Lindblad master equation into a rate equation of probabilities is possible because, in our special case, the Hamiltonian commutes with

the projectors and the Lindblad dissipaters have no dynamics within these subspaces. This assumption would break if the NV were placed in a magnetic field with off-axis field components comparable to the zero-field splittings, or if near-resonance microwave fields were turned on during the laser pulse.

The solution of Equation Equation 2.52 is $\vec{p}(t) = \mathrm{e}^{tR}\vec{p}(0)$. Thus the inhomogeneous Poisson rate from Equation Equation 2.46, $\mu(t) = \gamma_{ea}\,\mathrm{Tr}\left[P_e \cdot \rho(t)\right]$, can be simplified to $\mu(t) = \gamma_{ea}(p_{e0}(t) + p_{e1}(t))$, or in terms of the initial state,

$$\begin{aligned} &\mu(t) = \gamma_{ea}\vec{m} \cdot \mathrm{e}^{tR}\vec{p}(0) \\ \text{with} \quad &\vec{p}(0) = \left(\mathrm{Tr}[|\mathrm{g},0\rangle\langle\mathrm{g},0|\,\rho(0)], \mathrm{Tr}\left[(|\mathrm{g},-1\rangle\langle\mathrm{g},-1| + |\mathrm{g},+1\rangle\langle\mathrm{g},+1|)\rho(0)\right], 0, 0, 0\right)^{\mathrm{T}} \end{aligned} \tag{2.53}$$

where $\vec{m} = (0,0,1,1,0)^{\mathrm{T}}$ is the projector onto the excited space, and where we have assumed that the pre-measurement state $\rho(0)$ has support only on $\mathcal{H}_{\text{ground}}$. This new expression is much more tractable as it involves just a 5×5 matrix exponential, instead of a 49×49 matrix exponential in superoperator space. It also makes it simpler to derive the following relationship:

$$\begin{aligned} \mu(t|\rho(0)) &= \gamma_{ea}\vec{m} \cdot \mathrm{e}^{tR}\vec{p}(0) \\ &= \gamma_{ea}\vec{m} \cdot \mathrm{e}^{tR}\left(\mathrm{Tr}[|\mathrm{g},0\rangle\langle\mathrm{g},0|\,\rho(0)], \mathrm{Tr}\left[(|\mathrm{g},-1\rangle\langle\mathrm{g},-1| + |\mathrm{g},+1\rangle\langle\mathrm{g},+1|)\rho(0)\right], 0, 0, 0\right)^{\mathrm{T}} \\ &= \mathrm{Tr}[|\mathrm{g},0\rangle\langle\mathrm{g},0|\,\rho(0)]\gamma_{ea}\vec{m} \cdot \mathrm{e}^{tR}(1,0,0,0,0)^{\mathrm{T}} \\ &\quad + \mathrm{Tr}\left[(|\mathrm{g},-1\rangle\langle\mathrm{g},-1| + |\mathrm{g},+1\rangle\langle\mathrm{g},+1|)\rho(0)\right]\gamma_{ea}\vec{m} \cdot \mathrm{e}^{tR}(0,1,0,0,0)^{\mathrm{T}} \\ &= \mathrm{Tr}[|\mathrm{g},0\rangle\langle\mathrm{g},0|\,\rho(0)]\mu(t, |\mathrm{g},0\rangle\langle\mathrm{g},0|) \\ &\quad + \mathrm{Tr}\left[(|\mathrm{g},-1\rangle\langle\mathrm{g},-1| + |\mathrm{g},+1\rangle\langle\mathrm{g},+1|)\rho(0)\right]\mu(t, |\mathrm{g},+1\rangle\langle\mathrm{g},+1|) \\ &= p \cdot \mu(t, |\mathrm{g},0\rangle\langle\mathrm{g},0|) + (1-p) \cdot \mu(t, |\mathrm{g},+1\rangle\langle\mathrm{g},+1|) \end{aligned} \tag{2.54}$$

where $p = \mathrm{Tr}[|\mathrm{g},0\rangle\langle\mathrm{g},0|\,\rho(0)]$, again assuming $\rho(0)$ has support only on $\mathcal{H}_{\text{ground}}$. Note that the choice of using $|\mathrm{g},+1\rangle$ rather than $|\mathrm{g},-1\rangle$ (or any superposition of both) was arbitrary. Therefore the rate of photon emission during measurement given the pre-measurement state $\rho(0)$ is the convex combination of the the photon emission rates of the states $|\mathrm{g},0\rangle$ and $|\mathrm{g},+1\rangle$, where the convex combination parameter is the overlap of $\rho(0)$ with $|\mathrm{g},0\rangle$. It is this relationship that ultimately justifies the notion of taking reference measurements to calibrate the signal measurement.

An immediate corollary of this is that Equation Equation 2.48 can be simplified to

$$\mathbb{E}[n_e|\rho(0)] = p\mathbb{E}[n_e|\,|\mathrm{g},0\rangle\langle\mathrm{g},0|] + (1-p)\mathbb{E}[n_e|\,|\mathrm{g},+1\rangle\langle\mathrm{g},+1|] \tag{2.55}$$

which says that the expected number of emitted photons during measurement for the pre-measurement state $\rho(0)$ is the convex combination of the expected number of photons for the pre-measurement states $|\mathrm{g}, 0\rangle$ and $|\mathrm{g}, +1\rangle$. We already know that n_e is always a Poisson distribution for any pre-measurement state (Section 2.7.1), and that a Poisson distribution is characterized completely in terms of its expected value, therefore once $\mu_0 := \mathbb{E}[n_e| \, |\mathrm{g}, 0\rangle\langle \mathrm{g}, 0|]$, $\mu_1 := \mathbb{E}[n_e| \, |\mathrm{g}, +1\rangle\langle \mathrm{g}, +1|]$, and $p = \mathrm{Tr}[|\mathrm{g}, 0\rangle\langle \mathrm{g}, 0| \, \rho(0)]$ are known, everything about $\Pr(n_e|\rho(0))$ is also known:

$$\Pr(n_e|\rho(0)) = \frac{(p\mu_0 + (1-p)\mu_1)^{n_e}}{n_e!} \mathrm{e}^{-(p\mu_0+(1-p)\mu_1)} \tag{2.56}$$

which is a more tractable version of Equation Equation 2.47. Note that μ_0 and μ_1 will be redefined in the next section to account for visibility and background counts.

2.7.3 Measurement Visibility and Noise

There are two mechanisms that will affect our photon counting: photons can get lost along the way to the detector, or photons can be detected that did not originate from the NV. In Section 2.7.1 we derived the probability of the NV emitting some number of photons during a measurement, $\Pr(n_e)$, and in Section 2.7.2 we provided a simpler formula for the same quantity.

In this section we introduce two new variables, Γ and η. Let Γ be the rate of dark counts per unit time, due both to noise in the photon counter itself and to stray photons. Similarly, let η be the probability that a photon emitted by the NV center will be collected by the detector. This parameter is largely determined by the quality of the confocal microscope and the solid angle of emitted photons in view.

It is useful to define n_{dc} as the random variable representing those detected photons which were dark counts, and n_{td} as the random variable representing those detected photons which originated from the NV, dubbed 'true detections'. We have the relationship

$$n_d = n_{dc} + n_{td} \tag{2.57}$$

where n_d is the random variable representing all detected photons during a single measurement window. Note that n_{dc} and n_{td} are independent random variables, and also, that they are not in practice distinguishable.

Assuming the rate of dark counts is constantly equal to Γ over the a measurement duration Δt we simply have

$$n_{dc} \sim \mathrm{Poisson}(\Gamma \Delta t) \, . \tag{2.58}$$

The true detections are slightly more complicated. Each photon emitted by the NV has an (independent) probability η of arriving at the detector and being detected. We therefore have the conditional distribution

$$n_{td}|n_e \sim \text{Binom}\,(n_e, \eta) \tag{2.59}$$

where we are conditioning on a particular number of photons being emitted by the NV, n_e. Recall that n_e is Poisson distributed with a mean which we label $\mu = \mathbb{E}n_e := \mathbb{E}[n_e|\rho(0)]$ for now. We can therefore use the law of total probability to compute

$$\begin{aligned}
\Pr\left(n_{td} = n\right) &= \sum_{m=0}^{\infty} \Pr\left(n_{td} = n | n_e = m\right) \Pr\left(n_e = m\right) \\
&= \sum_{m=0}^{n-1} 0 \cdot \Pr\left(n_e = m\right) + \sum_{m=n}^{\infty} \binom{m}{n} \eta^n (1-\eta)^{m-n} \frac{\mu^m e^{-\mu}}{m!} \\
&= \frac{\eta^n (1-\eta)^{-n} e^{-\mu}}{n!} \sum_{m=n}^{\infty} \frac{(1-\eta)^m \mu^m}{(m-n)!} \\
&= \frac{(\eta\mu)^n e^{-\eta\mu}}{n!},
\end{aligned} \tag{2.60}$$

which shows that n_{td} is also Poisson distributed, with a mean given by $\eta\mu$. Since the sum of two independent Poisson variables is also Poisson, we conclude that the random varibale of interest, $n_d = n_{td} + n_{td}$, is Poisson distributed,

$$n_d \sim \text{Poisson}\left(\Gamma\Delta t + \eta\mathbb{E}[n_e|\rho(0)]\right). \tag{2.61}$$

This, in combination with Equation 2.55, gives

$$\begin{aligned}
n_d &\sim \text{Poisson}\left(\Gamma\Delta t + \eta(p\mu_0 + (1-p)\mu_1)\right) \\
&= \text{Poisson}\left(p(\Gamma\Delta t + \eta\mu_0) + (1-p)(\Gamma\Delta t + \eta\mu_1)\right)
\end{aligned} \tag{2.62}$$

where $\mu_0 := \mathbb{E}[n_e| \left|\mathrm{g}, 0\right\rangle\!\left\langle \mathrm{g}, 0\right|]$, $\mu_1 := \mathbb{E}[n_e| \left|\mathrm{g}, +1\right\rangle\!\left\langle \mathrm{g}, +1\right|]$, and $p = \text{Tr}[\left|\mathrm{g}, 0\right\rangle\!\left\langle \mathrm{g}, 0\right| \rho(0)]$.

Therefore, just as μ_0 and μ_1 served as the maximum and minimum reference counts for the expected emitted number of photons, $\mathbb{E}[n_e]$, we see that in the case of measurement visibility and noise, the numbers $(\Gamma\Delta t + \eta\mu_0)$ and $(\Gamma\Delta t + \eta\mu_1)$ serve as the new references for $\mathbb{E}[n_d]$.

2.8 State Initialization

Simply illuminating an NV center for a period of time with a laser, and then waiting for the system to settle, results in a highly polarized spin state. This can be shown by analysing the steady state solution to the rate equation. Importantly, the equilibrium state is unique, so that no matter what the initial conditions are to the initialization procedure, the final state is the same.

2.8.1 The Steady State Solution to the Rate Equation

The rate equation derived in Section 2.7.2, while much simpler than the Lindblad master equation it was derived from, is still not analytically solvable unless certain terms like γ_{01} are assumed to be zero. In order to estimate, for example, the time required to initialize the NV center, it will be helpful to consider the steady state solutions to the rate equation.

The rate matrix R from Equation Equation 2.52 has only non-positive eigenvalues, and must have a non-trivial null space. To see this, examine the solution to the rate equation, $\vec{p}(t) = e^{tR}\vec{p}(0)$. If R had positive or complex eigenvalues, the probability vector $\vec{p}(t)$ would either blow up or gain complex entries, neither of which are allowed under a Lindblad master equation. Moreover, if all of the eigenvalues were strictly negative, $\vec{p}(t)$ would asymptote to 0, and would therefore violate conservation of probability, hence at least one of the eigenvalues must be 0.

The null space completely specifies the steady state solution. Indeed, suppose that we have an initial set of populations $\vec{p}(0)$ and decompose this into an eigenbasis $\vec{v}_1, ..., \vec{v}_5$ for R, where it holds that $\vec{v}_1, .., \vec{v}_{i_0}$ all have eigenvalue zero, and $\vec{v}_{i_0+1}, .., \vec{v}_5$ have strictly negative eigenvalues $\lambda_{i_0+1}, ..., \lambda_5$. This gives $\vec{p}(0) = \sum_{i=1}^{5} a_i \vec{v}_i$ for some real coefficients $a_1, ..., a_5$. The evolution is now described by

$$\vec{p}(t) = \sum_{i=1}^{5} a_i e^{Rt}\vec{v}_i = \sum_{i=1}^{5} a_i e^{\lambda_i t}\vec{v}_i = \sum_{i=1}^{i_0} a_i \vec{v}_i + \sum_{i=i_0+1}^{5} a_i e^{\lambda_i t}\vec{v}_i \xrightarrow{t\to\infty} \sum_{i=1}^{i_0} a_i \vec{v}_i, \tag{2.63}$$

which is to say that only the population that was originally within the null subspace can remain in the steady state. This analysis also shows that the non-zero eigenvalue with the smallest absolute value largely determines the rate of decay — populations in this subspace take the longest to die.

As a technical aside, note that the rate matrix R is not normal, so that its eigenspaces are not orthogonal. This means that the population of the null space cannot be naively computed by projecting onto that subspace; a full linear inversion must be performed.

When the laser is off, we have $k = 0$ and the null space is two dimensional, spanned by $(1, 0, 0, 0, 0)^{\mathrm{T}}$ and $(0, 1, 0, 0, 0)^{\mathrm{T}}$, which correspond to the populations of the optical ground state – this is not surprising. The three remaining eigenvalues are $-(\gamma_{eg} + \gamma_{01}/2)$, $-(\gamma_{eg}+\gamma_{es}+\gamma_{01}/2)$, and $-\gamma_{sg}$. The first two die off quickly – these correspond to population escaping from the optically excited states. The last rate, γ_{sg}, is the most important, having an eigenvector $(1/3, 2/3, 0, 0, 1)^{\mathrm{T}}$, corresponds to population slowly seeping out of the singlet state and entering the three optical ground state levels. State initialization is complete once the singlet state is sufficiently depleted. The process is exponential, so one need only wait some small multiple of $1/\gamma_{sg}$ before the population of $|\mathrm{s}\rangle$ is vanishingly small.

When the laser is on, so that $k > 0$, the eigenstructure of the rate matrix is not as tractable, but we can still make progress. First, it can be shown that the null space is now only one dimensional. We denote the probability vector spanning the null subspace as $\vec{p}_{ss} = (p^{ss}_{g0}, p^{ss}_{g1}, p^{ss}_{e0}, p^{ss}_{e1}, p^{ss}_{s})^{\mathrm{T}}$. This is extremely important because it validates the NV$^-$initialization procedure: no matter what the quantum state was before the laser is turned on, if you let it equilibrize under optical illumination, it will reach a unique steady-state.

Under the approximation that $\gamma_{01} = 0$, we have $\vec{p}_{ss} = (1/(1 + k), 0, k/(1 + k), 0, 0)^{\mathrm{T}}$ which has support only on the $m_S = 0$ spin state. We see that if $k = 1$ so that the rate of spontaneous emission is equal to the rate of optical excitation, the ground and excited $m_S = 0$ populations equilibrize to the same value of 1/2. The null eigenspace is much more complicated when $\gamma_{01} > 0$. Though an analytic expression can be derived, it is a large unhelpful mess of divisions and multiplications of the various rates which we have been unable to simplify. To gain some insight, in Figure 2.5 we plot the components of $\vec{p}_{ss}$ as a function of γ_{01}. As γ_{01} increases, two properties are apparent: the steady-state population of the $m_S = \pm 1$ spin states increases, as expected, but also, the steady-state singlet state population increases dramatically. The latter effect occurs because the singlet state's relatively long lifetime makes it a good storage location, and now due to γ_{01}, there is always some population to feed it. Indeed, expanding $\vec{p}_{ss}$ in a γ_{01} power series about 0, the steady state population of the singlet is $p^{ss}_{s} = \frac{3k\gamma_{01}}{2(k+1)\gamma_{sg}} + O(\gamma^2_{01})$.

2.8.2 Imperfect Preparation: The Pseudo-pure State

Due to small but important spin-non-conserving terms in the master equation during continuous optical excitation Equation 2.43, the steady state density matrix must have non-zero support on the states $|\mathrm{g}, -1\rangle$, $|\mathrm{g}, +1\rangle$, $|\mathrm{e}, -1\rangle$, and $|\mathrm{e}, +1\rangle$. This can be seen in

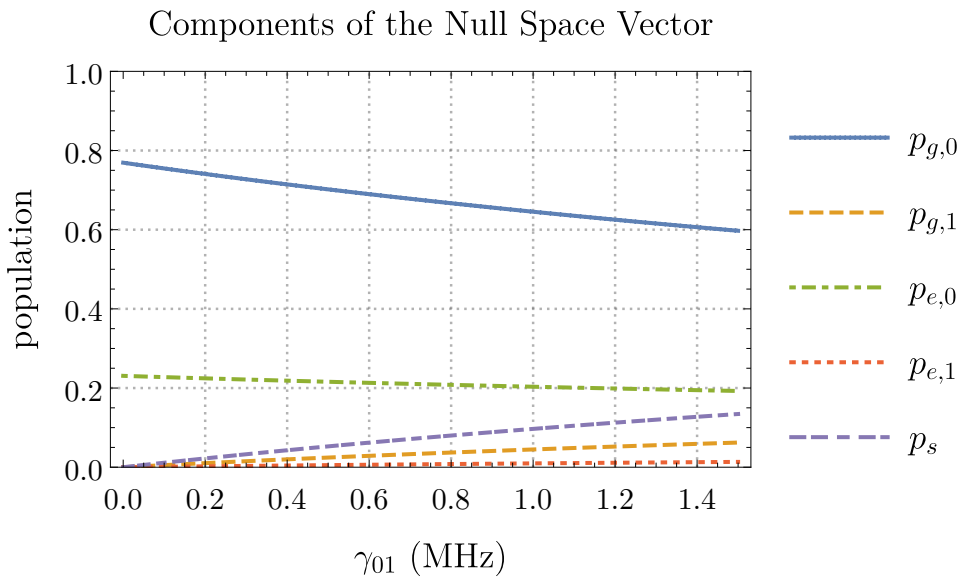

Figure 2.5: The components of $\vec{p}_{ss}$, the vector spanning the one dimensional null space of the rate matrix R, as a function of the spin-flip rate γ_{01}. Population of this vector outside of the $|\mathrm{e}, 0\rangle$, $|\mathrm{g}, 0\rangle$ subspace results in imperfect polarization. The other rate values used are $k = 0.3$, $\gamma_{eg} = 77\,\mathrm{MHz}$, $\gamma_{es} = 30\,\mathrm{MHz}$, and $\gamma_{sg} = 3\,\mathrm{MHz}$.

Figure 2.5. This means that the initialization process does not asymptote to the pure state $|\mathrm{g}, 0\rangle$, but rather to a mixed state ρ_0 which is mostly pure. Since our optical model does not distinguish between $m_s = +1$ and $m_s = -1$, they must have equal population in the pumping steady-state, and therefore ρ_0 can be written as a pseudo-pure state [29],

$$\rho_0 = q\,|\mathrm{g}, 0\rangle\langle \mathrm{g}, 0| + (1-q)\frac{\mathbb{I}}{3}, \tag{2.64}$$

where the purity parameter q depends in a non-trivial way on all of the parameters of the rate equation, but generally decreases as the rates of spin-non-conserving processes increase. It is pseudo-pure in the sense that any unitary (and in general, unital) process only acts on the first term,

$$U\rho_0 U^\dagger = qU\,|\mathrm{g}, 0\rangle\langle \mathrm{g}, 0|\,U^\dagger + (1-q)\frac{\mathbb{I}}{3} \tag{2.65}$$

so that its lack of complete purity, in practice, serves only to limit the contrast of measurement. This is discussed further in Section 7.5, where we consider pseudo-purification of the adjacent nitrogen spin.

To see this, we begin by computing the expected number photons emitted during the measurement of our preparation procedure ρ_0 using Equation Equation 2.55,

$$\mu_0' := \mathbb{E}[n_e|\rho_0] = \frac{1+2q}{3}\mu_0 + \frac{2-2q}{3}\mu_1. \tag{2.66}$$

Next, if prior to measuring ρ_0 we perform an operation $|\mathrm{g}, 0\rangle \mapsto |\mathrm{g}, +1\rangle$, which may be implemented as an adiabatic inversion with a microwave pulse, we have a pre-measurement state

$$\rho_1 = q\,|\mathrm{g}, +1\rangle\langle \mathrm{g}, +1| + (1-q)\frac{\mathbb{I}}{3} \tag{2.67}$$

which when measured emits an expected number of photons

$$\mu_1' := \mathbb{E}[n_e|\rho_1] = \frac{1-q}{3}\mu_0 + \frac{2+q}{3}\mu_1. \tag{2.68}$$

Previously we had considered μ_0 and μ_1 as the quantities that define the reference measurements in the absence of noise. However, given that ρ_0 is the best achievable initial state using standard techniques, it is in practice more convenient to use μ_0' and μ_1'. Indeed, if we prepare the state ρ_0 and perform any unital operation resulting in a pre-measurement state

$$\rho = q\rho_\psi + (1-q)\frac{\mathbb{I}}{3}, \tag{2.69}$$

then some simple algebra shows that

$$\mathbb{E}[n_e|\rho] = p\mu_0' + (1-p)\mu_1' \tag{2.70}$$

where $p = \mathrm{Tr}[\rho_\psi \left|\mathrm{g}, 0\right\rangle\!\left\langle \mathrm{g}, 0\right|]$.

Finally, if we take into consideration finite visibility η and dark count rate Γ as discussed in Section 2.7.3, we may define

$$\begin{aligned} \alpha &:= \mathbb{E}[n_d|\rho_0] = \Gamma\Delta t + \eta\mu_0' \\ \beta &:= \mathbb{E}[n_d|\rho_1] = \Gamma\Delta t + \eta\mu_1' \\ \gamma &:= \mathbb{E}[n_d|\rho]. \end{aligned} \tag{2.71}$$

to arrive at

$$\begin{aligned} \mathbb{E}[n_d|\rho] = \gamma &= p\alpha + (1-p)\beta \\ n_d|\rho &\sim \text{Poisson}\left(p\alpha + (1-p)\beta\right) \end{aligned} \tag{2.72}$$

which is analagous to Equation Equation 2.62 but for our pseudo-pure state preparation. The quantity γ, which we will call the *signal*, is by definition conditioned on the pre-measurement state ρ, and will henceforward generically refer to the expected number of detected photons given the pre-measurement state of interest or, equivalently, the experiment of interest which when performed on the preparation state ρ_0 yields ρ. We will call the quantities α and β our *references* because they bound the expected values of detected photons for arbitrary states of the form in Equation Equation 2.69. More specifically, α is the *bright reference* and β is the *dark reference* since $\alpha > \beta$.

2.9 Experimental Setup

Here we give a brief overview of the experimental setup used for the experiments in this thesis. Giving a complete overview is beyond the scope of this thesis, and we refer the reader to Reference [138], where a full account of our particular setup is outlined. The setup was constructed by Osama Moussa and Om Patange, and maintained primarily by myself while the experiments presented in this thesis were being performed.

The setup consists of three large composite components:

- a custom confocal microscope with micrometer resolution, using 532 nm light for excitation, a single photon counting module for detection, and including a switching component for the laser with a ~ 50 ns risetime,

- a microwave circuit to deliver arbitrarily shaped quadrature pulses to the diamond that can be synced to the laser switch and photon counting module,
- a control computer with accompanying software suite to conveniently acquire images, run microwave experiments, including those specified as a remote command from another computer.

The confocal microscope has four distinct arms. The output of a 100 mW green 532 nm CW laser enters the first arm whose purpose is to switch laser on and off at the end point quickly provided an external digital signal. This is achieved with a double-pass through an acousto-optic modulator (AOM), which, when driven, splits incoming light into multiple spatial modes, one of which is picked off by an iris. The result is input into the second arm which shapes the light back into a Gaussian mode by passing through a single-mode optical fiber. The third arm consists of a pair of electronically controlled galvonometer mirrors which, aided by a pair of lenses, determine the angle of incidence into the main objective lens, thereby allowing optical scanning of the sample. Lower-energy light ($\sim 600-900$nm) that fluoresces from the current region of focus of the objective travels backward along the same path as the incident green light, except that past the galvonemeter mirrors, it passes through a dichroic mirror which reflects green light. This transmitted light is focused through a 5 um pinhole confocal to the to the objective lens' focus, and is subsequently detected by an avalanche photo detector (APD). In addition to being able to steer the focus using the galvo mirrors, the entire diamond can be translated under the objective using a 3-axis stage. A detailed diagram is found in Reference [138].

The microwave circuit, pictured as a part of Figure 2.6, consists primarily of an arbitrary waveform generator (AWG) that shapes two analog profiles with 2 ns resolution, which are then mixed with a microwave tone using an in-phase and quadrature (IQ) modulator. These analog profiles, up to circuit distortions, become the coefficients of $\mathbf{S}_x$ and $\mathbf{S}'_y$ in the NV Hamiltonian. This signal is transmitted to the NV center by placing a 25 um wire nearby carrying the signal, acting as an antenna. The AWG also serves to synchronize the microwave signal with the counter and laser, which are gated by digital outputs.

The experiment is run by a lab computer. It constructs images inside the diamond by rastering the galvo mirrors and recording the number of photons collected at each position of a 2D grid, which become the pixels of a gray-scale image—NV centers are visible as bright spots in a dark background. The lab computer also compiles pulse sequences, specified using a custom text-based syntax, into sequences of voltage levels and digital states at each time-step, which is then sent to the AWG. A photon count is recorded for every instance that this pulse program opens the counting window. This software, implemented in MATLAB, has a rich history—it was originally written by Jonathan Hodges, and subsequently

modified (at times heavily) by at least the following authors: Colm Ryan, Osama Moussa, Chris Granade, Om Patange, Romain Ruhlmann, Madelaine Liddy, Michal Kononenko, and myself. The lab computer is also able to enter a mode where it listens for experiments sent by other computers over the network. Timestamps and photon counts are returned when the experiment is complete. This mode was developed specifically for the adaptive experiments discussed in Chapter 4 with help from Michal Konenenko.

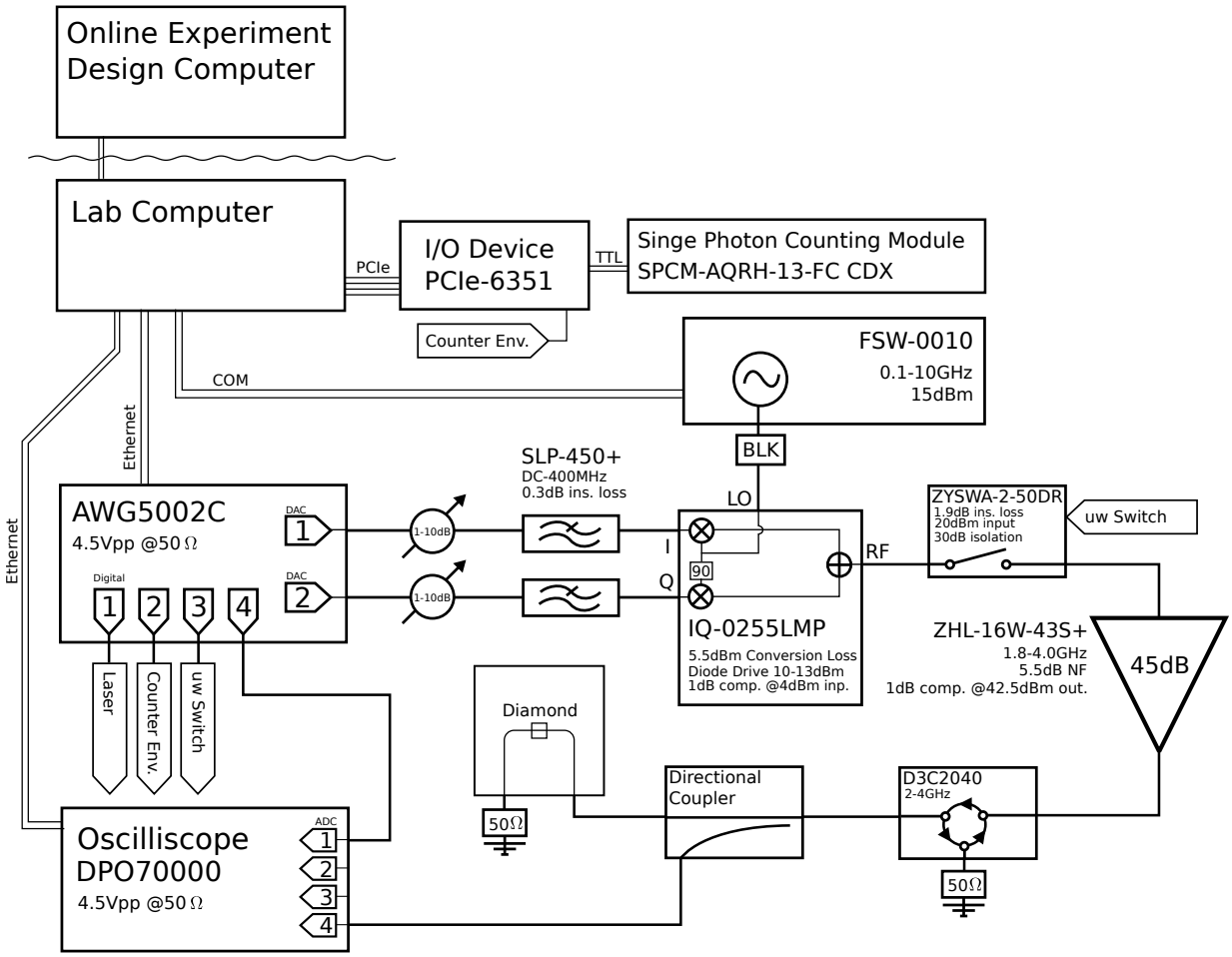

Figure 2.6: The electronic connection diagram of the NV experimental setup, including the microwave circuit and control computers.

Chapter 3

Learning from Quantum Measurements: Hamiltonian Estimates from NV Photon Counts[1]

3.1 Introduction

Having discussed properties and dynamics of the NV center in the previous section, we now begin to consider using this system in practice. Our first step, and the goal of this chapter, is to phrase the measurement process of this system as a statistical inference problem concerning the parameters of interest. We discuss and quantitatively compare two types of estimators, and finish by applying them to experimental data to perform a task known as quantum Hamiltonian Learning (QHL).

The overarching motivation for such a rigorous approach, as opposed to naively applying some canned curve fitting method, is that developing tools for the characterization of quantum systems is an increasingly important problem. As more performance is demanded out of quantum devices, more knowledge about these quantum devices is also required. This applies as much to large-scale multi-qubit quantum information processors as it does to small single-qubit quantum sensors. Importantly, this knowledge must include not

[1]This section is based on Reference [84], work done in collaboration with Chris Granade, who provided statistical and numerical advice. This work represents the first rigorous treatment of the statistics of NV measurement, introducing the Poisson obstructed coin model, which, to our knowledge, has not been considered in estimator theory before. It also represents the first time that fully Bayesian methods have been applied to the quantum Hamiltonian learning problem using experimental data.

only estimates of relevant quantities, but also careful estimates of the uncertainty of these quantities. Indeed, if the application is, for example, metrology, then the very nature of the problem demands that one should be as confident in one's ability to produce meaningful error bars as one's ability to produce the estimate itself. Or, if one is using knowledge about a quantum system to design control sequences (such as unitary gates), then it is important to know how much system parameters are expected to vary through space and time. If the estimate of this parameter distribution is too tight, the control sequence will not meet specifications, and if it is too broad, the control sequence will not have optimal efficiency [144, 140, 87, 78].

Therefore, data from quantum experiments should be analyzed on firm statistical footing. This requires a detailed model that computes the likelihoods of experimental outcomes given a specified set of system parameters or hyperparameters. This is not to say that we need perfect statistical models, but rather, that models and methods should be well enough defined so that rigorous questions can be asked and answered unambiguously. Most of the widespread characterization protocols in quantum information have been described by statistical models. Quantum mechanics is ultimately a statistical theory and so this is usually a natural thing to do. State and process tomography have been studied extensively as matrix-valued inference [92, 14, 61], randomized benchmarking (RB) and derivative protocols are inherently statistical [116, 46, 62, 181], and Hamiltonian parameter learning is often considered from a machine learning perspective [185, 66, 105, 182]. These statistical models of characterization, however, usually stop short of platform dependent considerations; for example, the RB protocol does not tell you how to interpret the noisy voltage measurement of a superconducting qubit. Such divisions are drawn to achieve cross-quantum-platform generality. Specific platforms, therefore, need to consider how to append their own messy details onto these well-established protocols, as we will do in the coming sections.

3.2 Drifting References

In Section 2.8.2, in particular in Equation Equation 2.71, we defined the quantities α, β, and γ, which will figure prominently in this chapter. Respectively, these parameters are the bright reference, the dark reference and the signal. They correspond to the expected number of photons collected (including visibility loss and background counts) in a single repetition of an experiment given that the pre-measurement state was (imperfectly) prepared as $|0\rangle$, $|\pm 1\rangle$, or an arbitrary state ρ, respectively. Due to convenient additive properties of the Poisson distribution, we will later justify the overloading of these symbols

to mean the expected number of collected photons in N repetitions of the same experiment, where the difference between $N = 1$ and $N > 1$ will be obvious from context.

One of the main complications of experimental NV measurement is drifting of the references α and β in time. Though there are many mechanisms which can cause this, for us, the most prevalent is due to relative movement between the NV center and the focal spot of the confocal microscope. This region of focus has a 3D Gaussian profile, and as temperatures or other properties of the lab change in time, movement of the center of this region off of the point-sized NV center causes both a drop-off of delivered laser power, k, and collection efficiency, η. Other mechanisms may include fluctuation of the laser power, and quality of the confocal microscope's alignment, which will affect η, k, and the background count rate Γ. If left unchecked, the drift would eventually cause the NV under study to no longer be in the focal region at all. To avoid this, a *tracking* procedure is periodically run, whose purpose is to recenter the NV with the focal region by taking a series spatial of images and using feedback to realign.

3.2.1 Experiment Ordering

Experiments typically have a set of parameters that are varied. For the sake of concreteness, we choose the specific experiment and repetition ordering described in Figure 3.1; variations of this ordering may require a modified (though similar) analysis to that which follows. We denote the list of experiment parameter configurations as $\vec{a}_1, ..., \vec{a}_S$. For example, in the case of a Ramsey experiment for magnetometry [170], the distance between two $\frac{\pi}{2}$ pulses is varied, so that $\vec{a}_s = (t_s)$ where t_s is the pulse spacing. If additionally the phase of the second Ramsey pulse is varied in proportion to the spacing, we have instead $\vec{a}_s = (t_s, \omega t_s)$ for some angular frequency ω. A parameter configuration $\vec{a} \in \{\vec{a}_1, ..., \vec{a}_S\}$ is fixed and a large number, N, of back-to-back *repetitions* with this parameter configuration is performed before moving on to the next. The choice of N is motivated, for instance, by the time required for experimental control hardware to switch between choices of configuration $\vec{a}$. Each time all parameter configurations have been dealt with, the tracking procedure is run, and then the entire procedure is repeated R times, which we call *averages*.

A signal measurement γ depends only on the current value of α and β, the current parameter configuration $\vec{a}$, and any noise operations acting on $\vec{a}$. We denote the true value of α, β, and γ at the n^{th} repetition of the s^{th} experiment $\vec{a}_s$ in the r^{th} average as $\alpha_{n,s,r}$, $\beta_{n,s,r}$, and $\gamma_{n,s,r}$, for $1 \leq n \leq N$, $1 \leq s \leq S$, and $1 \leq r \leq R$. It holds that $\gamma_{n,s,r} = p_s \alpha_{n,s,r} + (1 - p_s)\beta_{n,s,r}$ where we have used our assumption of the independence of

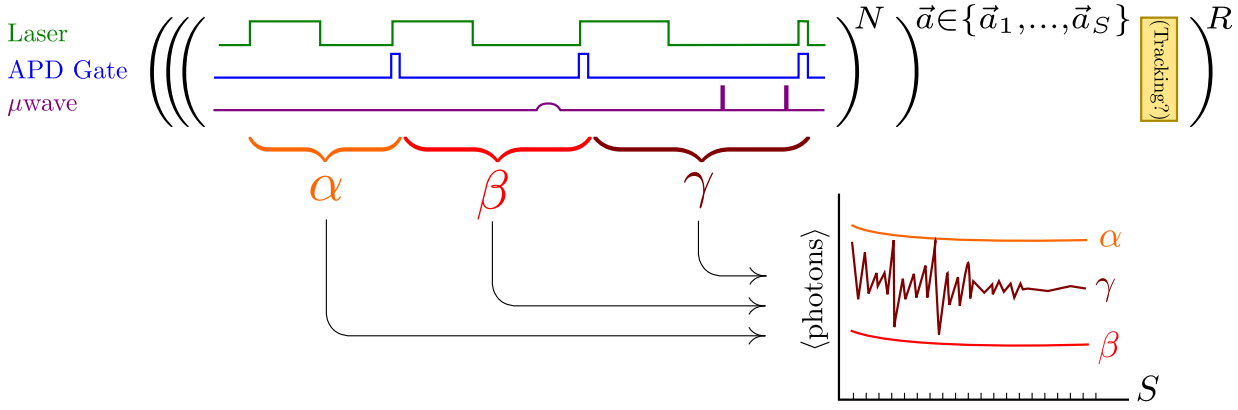

Figure 3.1: The considered experimental ordering. Another popular ordering transposes the inner two levels. Given a particular parameter configuration of the experiment, $\vec{a} \in \{\vec{a}_1, ..., \vec{a}_S\}$ (in the above example, the parameter is the distance between the two last microwave pulses), N repetitions are performed of both the experiment, γ, and the references, α and β. The bright reference α is measured by initializing with a laser pulse, waiting for metastable optical states to decay, and taking a measurement by opening the APD counting gate while the laser is on. The dark reference β is similar, except an inversion pulse is applied prior to measurement. The pulse sequence prior to the reference measurement γ depends on the current parameter $\vec{a}$. Each time N repetitions have been made of all S parameter configurations, the system decides whether to track or not, and this is all repeated R times. A sketch of the resulting data is shown, averaged over both N and R.

p from the drift processes to label p by only s. This yields the random variables

$$\begin{aligned} X_{n,s,r} &\sim \text{Poisson}\,(\alpha_{n,s,r}) \\ Y_{n,s,r} &\sim \text{Poisson}\,(\beta_{n,s,r}) \\ Z_{n,s,r} &\sim \text{Poisson}\,(\gamma_{n,s,r}) \end{aligned} \tag{3.1}$$

with the corresponding variates $(x_{n,s,r}, y_{n,s,r}, z_{n,s,r})_{n,s,r=1}^{N,S,R}$.

3.2.2 Combining Experiments

Due to low visibility, the repetition number N in Figure 3.1 is often quite high. It is usually cumbersome to store the results of the experiment for each individual measurement. We therefore assume that the data is summed over the N repetitions. Because of the additive property of the Poisson distribution, if we define $X_{s,r} := \sum_{n=1}^{N} X_{n,s,r}$, $Y_{s,r} := \sum_{n=1}^{N} Y_{n,s,r}$, and $Z_{s,r} := \sum_{n=1}^{N} Z_{n,s,r}$ we get

$$\begin{aligned} X_{s,r} &\sim \text{Poisson}\,(\alpha_{s,r}) \\ Y_{s,r} &\sim \text{Poisson}\,(\beta_{s,r}) \\ Z_{s,r} &\sim \text{Poisson}\,(\gamma_{s,r}) \end{aligned} \tag{3.2}$$

where $\alpha_{s,r} := \sum_{n=1}^{N} \gamma_{n,s,r}$, $\beta_{s,r} := \sum_{n=1}^{N} \gamma_{n,s,r}$, and $\gamma_{s,r} := \sum_{n=1}^{N} \gamma_{n,s,r}$.

As mentioned earlier, we will often abuse notation so that α, β, and γ, refer to $\alpha_{s,r}$, $\beta_{s,r}$, and $\gamma_{s,r}$ for a particular index (s,r) where some $N \geq 1$ is implicitly understood. Furthermore, in this context δ will refer to $\sum_{n=1}^{N} \Gamma_{n,s,r}\Delta t$, that is, the total contribution to the dark counts. Similarly, ν will refer to $\sum_{n=1}^{N} \eta_{n,s,r}(\mu_0')_{n,s,r}$ and κ to $\frac{1}{\nu}\sum_{n=1}^{N} \eta_{n,s,r}(\mu_1')_{n,s,r}$. (Recall that the quantities μ_0' and μ_1' were defined in Equation 2.66 and Equation 2.66 respectively, reffering to the expected number of photon counts arising from the bright and dark references, accounting for pseudo-pure preparation.) The fraction $0 < \kappa < 1$ represents how much dimmer the reference state ρ_1 is as compared to ρ_0. This results in the relationship

$$\begin{aligned} \alpha &= \delta + \nu \\ \beta &= \delta + \kappa\nu. \end{aligned} \tag{3.3}$$

3.2.3 Correlations of Model Parameters

The three primitive quantities μ_0', μ_1', and η (see Section 2.8.2 for definitions) all depend on the quality of the coupling between the confocal microscope and the NV defect; as the

relative displacement between the defect and the center of focus drifts, η decreases. This could be due to, for example, temperature changes in the lab which expand or contract the components on the optical table. However, μ_0' and μ_1' also change with decreased optical coupling because the preparation state ρ_0 depends on the laser power which is coupling dependent. Therefore we expect correlations between μ_0', μ_1', and η.

Nominally $\Gamma\Delta t$ should be independent of each of the quantities η, μ_0', μ_1', and p. However, this may break if the power of the laser varies in time and a significant portion of the dark counts are caused by unwanted reflections of, or excitations due to the laser; we may end up with correlations between $\Gamma\Delta t$ and each of η, μ_0', and μ_1'.

Finally, and perhaps most importantly, the quantity p will normally be independent of the quantities η, μ_0', μ_1', and $\Gamma\Delta t$. However, it is also possible for this to fail. For example, if the process that takes the preparation state ρ_0 to the pre-measurement state ρ_0 is non-unital, we will not end up with exactly the form $q\rho_\phi + (1-q)\mathbb{I}/3$ (as seen in Section 2.8.2) leading to errors when inferring p from $p\alpha + (1-p)\beta$. Non-unitality could occur due to T_1 relaxation, or leakage of laser light when it is supposed to be off.

3.2.4 A Stochastic Model of Drift

The references α and β at $N = 1$ are best viewed as stochastic processes with autocorrelations in time. As discussed above, they will also be correlated with each other. They will undergo a discontinuous jump every time a tracking operation is performed.

As derived in Section 2.8.2, conditioned on set of parameters, including k, Γ, Δt, η, γ_{es}, γ_{sg}, γ_{eg}, and γ_{01} the references are given by $\alpha = \Gamma\Delta t + \eta\mu_0'$ and $\beta = \Gamma\Delta t + \eta\mu_1$ at a particular instance in time. Given the number variables and unknowns that likely go into the drift process itself (which will in turn affect k, Γ, and η) on top of the already complex conditional model stated above, writing down an analytic model for the stochastic processes α and β would be difficult, if not impossible.

We therefore restrict our attention to simpler effective models which still well describe expected and observed behaviour. We assume that the stochastic process (α, β) is a Gaussian process. This is a weak assumption especially given that in later sections we will only make use of the first two moments.

We may relate this continuous stochastic process to the discrete variables defined in Section 3.2.1 with a standard discretization as follows. Consider the r^{th} average of the experiment and a particular realization of the stochastic process $(\alpha(t), \beta(t))$ during this average. Then we have that $\alpha_{n,s,r} = \alpha(t_{n,s})$ and $\beta_{n,s,r} = \beta(t_{n,s})$ where $t_{n,s}$ is the time of

the n^{th} repetition of the s^{th} experiment relative to the start of the r^{th} average. This gives, for example, the discrete Cox process $X_{n,s,r} \sim \text{Poisson}\left(\alpha(t_{n,s})\right)$.

We take the time value $t = 0$ to mean the time directly after performing a tracking operation. The tracking operation has the effect of drawing the initial values $\alpha(0)$ and $\beta(0)$ from a fixed normal distribution

$$(\alpha(0), \beta(0)) \sim \text{Normal}\left((\alpha_0, \beta_0), \Sigma_0\right) \tag{3.4}$$

where the variances in Σ_0 are determined by the noise and error in the tracking procedure, and the mean is a property of the confocal microscope's quality and the NVs optical properties. Assuming a Gaussian stochastic process leads to the distribution

$$(\alpha(t), \beta(t)) \sim \text{Normal}\left((\alpha_t, \beta_t), \Sigma_t\right) \tag{3.5}$$

at time $t \geq 0$.

We now consider an explicit example of a stochastic model for concreteness. We make use of the Ornstein–Uhlenbeck process, which is one of the simplest continuous-time stochastic processes, consisting of a random walk on a real line (Wiener process) that has been modified with a term that causes reversion back to a fixed location. Such a process can be used, for example, to model the position of a mass attached to a spring undergoing random Gaussian external forces as a function of time. We consider the NV drift model given as

$$\begin{aligned} \alpha_1 &\sim \text{Normal}\left(\alpha_0, \sigma_\alpha\right) \\ \nu &\sim \text{OrnsteinUhlenbeck}\left(0, \sigma_\nu, \theta_\nu, \alpha_1 - \Gamma\Delta t\right) \\ \kappa &\sim \text{OrnsteinUhlenbeck}\left(\kappa_0, \sigma_\kappa, \theta_\kappa, 0\right) \\ (\alpha, \beta) &= (\Gamma\Delta t + \nu, \Gamma\Delta t + \kappa \cdot \nu). \end{aligned} \tag{3.6}$$

Here, $\Gamma\Delta t$ are the expected dark counts in a single measurement, which we have assumed to be deterministic and constant for simplicity. However, the expected number of photons due to the bright state ρ_0, denoted an ν, is an Ornstein–Uhlenbeck process with long-time mean 0, volatility σ_ν, mean reversion speed θ_ν, and initial value $\alpha_1 - \Gamma\Delta t$, where the initial value is marginalized over $\text{Normal}\left(\alpha_0, \sigma_\alpha\right)$ representing imperfections in the tracking process. This implies that, on average, $\alpha(0) = \alpha_0$ and $\alpha(\infty) = \Gamma\Delta t$ with average decay rate μ_ν, where 'shakiness' in getting there is determined by σ_ν. In order to correlate α with β, and also to help enforce the physical constraint $\beta \leq \alpha$, we relate the two components with another Ornstein–Uhlenbeck process κ.

Note that this model does not guarantee, for example, that $\alpha(t) > 0$ or that $0 \leq \kappa(t) \leq 1$. We can only make these scenarios improbable by choosing low volatilities. This is the compromise of having such a simple model in terms of the well known Ornstein-Uhlenbeck Gaussian process.

Solving for the moments of this stochastic process we arrive at the expressions

$$\begin{aligned}
\mathbb{E}[\alpha(t)] &= \Gamma\Delta t + (\alpha_0 - \Gamma\Delta t)\mathrm{e}^{-t\theta_\nu} \\
\mathbb{E}[\beta(t)] &= \Gamma\Delta t + \kappa_0(\alpha_0 - \Gamma\Delta t)\mathrm{e}^{-t\theta_\nu} \\
\mathrm{Var}[\alpha(t)] &= \sigma_\alpha^2\mathrm{e}^{-2t\theta_\nu} + \sigma_\nu^2\frac{(1-\mathrm{e}^{-2t\theta_\nu})}{2\theta_\nu} \\
\mathrm{Var}[\beta(t)] &= \kappa_0^2\sigma_\alpha^2\mathrm{e}^{-2t\theta_\nu} + \sigma_\kappa^2\frac{((\alpha_0 - \Gamma\Delta t)^2 + \sigma_\alpha^2)\mathrm{e}^{-2t\theta_\nu}}{2\theta_\kappa} + \sigma_\nu^2\frac{(2\theta_\kappa\kappa_0^2+1)(1-\mathrm{e}^{-2t\theta_\nu})}{4\theta_\kappa\theta_\nu} \\
\mathrm{Cov}[\alpha(t),\beta(t)] &= \kappa_0\sigma_\alpha^2\mathrm{e}^{-2t\theta_\nu} + \frac{\kappa_0\sigma_\nu^2(1-\mathrm{e}^{-2t\theta_\nu})}{2\theta_\nu}.
\end{aligned} \tag{3.7}$$

These calculations can be found in Section B.1. In Figure 3.2, these moments are plotted along with a single random trajectory of the stochastic process defined above.

The purpose of this section has been to demonstrate that it is possible to meaningfully model the drift process as a stochastic process, which leads to reference count variances and covariances at each time step. These variances will correspond to the moments of the hyperparameter distribution for the references described in Section 3.3.1.

3.3 Statistical Models of Measurement

In this section we state the measurement of an NV center as a statistical inference problem.

3.3.1 The Statistical Model of Measurements

Consider the inner-product space $\mathcal{H} = \mathbb{C}^3$ with the canonical basis $|1\rangle = (1,0,0)^{\mathrm{T}}$, $|0\rangle = (0,1,0)^{\mathrm{T}}$, and $|-1\rangle = (0,0,1)^{\mathrm{T}}$. We call a state of the system prior to the measurement procedure the pre-measurement state. Suppose that the pre-measurement state of interest is given by the density matrix $\rho \in \mathrm{D}(\mathbb{C}^3)$. Define $p = \mathrm{Tr}[\rho P_0]$ where P_0 is the projector onto $|0\rangle$. In the case of a strong quantum measurement we would have access to random

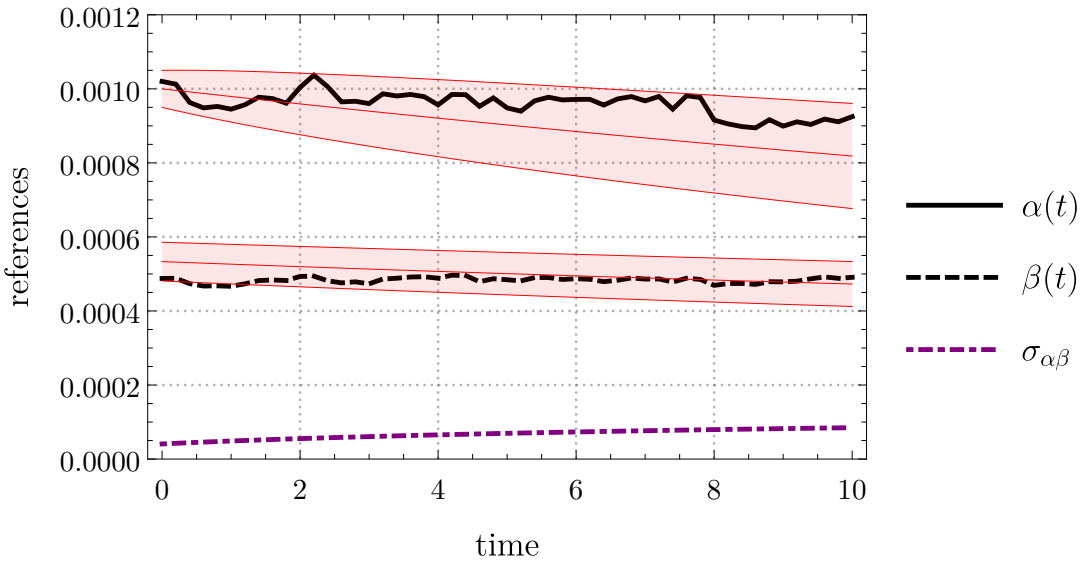

Figure 3.2: Simulation of a severe case of drift. A random instance of the process $(\alpha(t), \beta(t))$ defined in Equation Equation 3.6 is shown on top of their first moments with a shaded single standard deviation. The dot-dashed purple line shows the square root of covariance. The model parameters used are $\alpha_0 = 10^{-3}$, $\sigma_\alpha = 5 \times 10^{-5}$, $\sigma_\nu = 5 \times 10^{-5}$, $\theta_\nu = 0.03$, $\Gamma\Delta t = 3 \times 10^{-4}$, $\kappa_0 = 1/3$, $\sigma_\kappa = 0.01$, and $\theta_\kappa = 0.01$. The time units are arbitrary; scaling the x-axis is equivalent to scaling θ_ν and θ_κ.

variables drawn from the distribution Bernoulli (p). Instead, however, we have access to the random triplet $(X, Y, Z)|\alpha, \beta$ defined by

$$\begin{aligned} X &\sim \text{Poisson}\,(\alpha) \\ Y &\sim \text{Poisson}\,(\beta) \\ Z &\sim \text{Poisson}\,(p\alpha + (1-p)\beta) \equiv \text{Poisson}\,(\gamma) \end{aligned} \tag{3.8}$$

where α and β are the expected number of photons collected in N independent repetitions of a fixed experiment. It is always true that $0 \leq p \leq 1$ and $0 \leq \beta \leq \gamma \leq \alpha$.

The references α and β are in turn random variables drawn from the distribution

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} \sim \text{Normal}\left(\begin{pmatrix} \bar{\alpha} \\ \bar{\beta} \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \sigma_\beta^2 \end{pmatrix} \right) \tag{3.9}$$

with $\sigma_{\alpha\beta} > 0$. The normality of this distribution is typically irrelevant, and is stated as such just to be concrete. We are usually only interested in its first two moments. The 'true' distribution is almost certainly quite complicated, and arises from the stochastic processes described in Section 3.2. In later sections this multinormal distribution will be replaced with a product gamma distribution, or a mixture of product gamma distributions, as discussed in Section 3.5. To be clear, note that when a variate (x, y, z) is sampled from this distribution, all three variates are conditional on the same values of α and β.

This creates a hierarchical model with nuisance hyperparameters $\bar{\alpha}$, $\bar{\beta}$, σ_α, σ_β, and $\sigma_{\alpha,\beta}$. We will see that this second layer is rarely useful, so that the conditional model $(X, Y, Z)|\alpha, \beta$ should usually be used in practice. However, it is important to remember that the second layer exists, because it makes it clear that multiple identical samples cannot be taken from the conditional model. Indeed, the values of α and β will change each time the set of N measurements are made.

3.3.2 Moment Calculations

We first work out the first two moments of the random variables X, Y, and Z, both conditional and unconditional on the hyperparameters α and β.

The conditional moments are trivially given by

$$\begin{aligned} \mathbb{E}[X|\alpha] &= \text{Var}[X|\alpha] = \alpha \\ \mathbb{E}[Y|\beta] &= \text{Var}[Y|\beta] = \beta \\ \mathbb{E}[Z|p,\alpha,\beta] &= \text{Var}[Z|p,\alpha,\beta] = p\alpha + (1-p)\beta \\ \text{and } \text{Cov}[X, Y|\alpha, \beta] &= 0 \end{aligned} \tag{3.10}$$

using basic properties of the Poisson distribution. The law of total expectation can be used to compute

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}_{\alpha,\beta}[\mathbb{E}[X|\alpha]] = \bar{\alpha} \\
\mathbb{E}[Y] &= \mathbb{E}_{\alpha,\beta}[\mathbb{E}[Y|\beta]] = \bar{\beta} \\
\text{and } \mathbb{E}[Z] &= \mathbb{E}_{\alpha,\beta}[\mathbb{E}[Y|p,\alpha,\beta]] = p\bar{\alpha} + (1-p)\bar{\beta},
\end{aligned} \tag{3.11}$$

showing that the variance of α and β do not affect the mean. Similarly, the law of total variance gives

$$\begin{aligned}
\mathrm{Var}[X] &= \mathbb{E}_{\alpha,\beta}[\mathrm{Var}[X|\alpha]] + \mathrm{Var}_{\alpha,\beta}[\mathbb{E}[X|\alpha]] = \bar{\alpha} + \sigma_\alpha^2 \\
\mathrm{Var}[Y] &= \mathbb{E}_{\alpha,\beta}[\mathrm{Var}[Y|\beta]] + \mathrm{Var}_{\alpha,\beta}[\mathbb{E}[Y|\beta]] = \bar{\beta} + \sigma_\beta^2 \\
\mathrm{Var}[Z] &= \mathbb{E}_{\alpha,\beta}[\mathrm{Var}[Z|p,\alpha,\beta]] + \mathrm{Var}_{\alpha,\beta}[\mathbb{E}[Z|p,\alpha,\beta]] \\
&= p\bar{\alpha} + (1-p)\bar{\beta} + p^2\sigma_\alpha^2 + (1-p)^2\sigma_\beta^2 + 2p(1-p)\sigma_{\alpha,\beta}
\end{aligned} \tag{3.12}$$

which shows that the variances have two parts, one due to the usual finite sampling error of a Poisson variable, and one due to the underlying fluctuation of the Poisson parameters.

3.3.3 Three Advantages of the Conditional Model

If the measurement model is not stated as concretely as we have done, then there can be a slight subtlety in the interpretation of random variates. Misunderstanding this point could result in reporting incorrect error bars or confidence intervals/credible regions. In an attempt to be as clear as possible, we illustrate with an example.

Suppose Yves and Zoey together collect R variates of the random variable (X, Y, Z), sampled identically and independently, giving the results $(x_1, y_1, z_1), (x_2, y_2, z_2), ..., (x_R, y_R, z_R)$. They go back to their separate offices and try to analyse the data. Yves calculates the sample mean and variance of $x_1, x_2, ..., x_R$ as $x_{\text{samp}} = \frac{1}{R}\sum_{r=1}^{R} x_r$ and $\sigma^2_{x,\text{samp}} = \frac{1}{R-1}\sum_{r=1}^{R}(x_r - x_{\text{samp}})^2$, respectively. He gets $x_{\text{samp}} = 200$ and $\sigma_{x,\text{samp}} = 20$. Looking at Equation Equation 3.11, and using the standard error of the mean, this informs his rough belief that $\bar{\alpha} \approx x_{\text{samp}} \pm \sigma_{x,\text{samp}}/\sqrt{R} = 200 \pm 20/\sqrt{R}$.

Zoey recalls that each variate x_r was obtained by summing N independent measurements. She wonders why they bothered batching the results into R sets, and why they didn't just take $N \times R$ measurements to begin with. She therefore does a sum to get the new quantity $x_{\text{sum}} = \sum_{r=1}^{R} x_r$. She knows that for each $1 \leq r \leq R$, x_r was drawn from the distribution Poisson (α_r) for some specific but unknown value of α_r. Therefore,

knowing the summative property of Poisson distributions, she correctly deduces that x_{sum} was sampled from Poisson $\left(\sum_{r=1}^{R} \alpha_r\right)$. The Poisson distribution's standard deviation is the square-root of its mean, she therefore adopts the rough belief that $\sum_{r=1}^{R} \alpha_r \approx x_{\text{sum}} \pm \sqrt{x_{\text{sum}}}$, and therefore that $\frac{1}{R}\sum_{r=1}^{R} \alpha_r \approx x_{\text{sum}}/R \pm \sqrt{x_{\text{sum}}/R}) = 200 \pm 14/\sqrt{R}$.

Zoey and Yves get back together and compare their results, and wonder why Zoey is more confident than Yves about the quantity she has estimated, when, naively, it seems that they have estimated the same thing with the same data.

The discrepancy comes down to the fact that they are estimating parameters using different models. Yves is estimating the hyperparameter $\bar{\alpha}$ from the hierarchical model in Section 3.3.1, justified by the moment calculations in Section 3.3.2. Zoey is foregoing the hyperparameter layer of the model and making a direct inference about the sum of the *particular* references $\alpha_1, \alpha_1, ..., \alpha_R$ they happened to draw in their measurements. Neither is wrong, they are simply estimating different but related quantities.

We saw in Section 3.3.2 that $\text{Var}[X|\alpha] = \bar{\alpha}$ and $\text{Var}[X] = \bar{\alpha} + \sigma_\alpha^2$. Yves was assuming he made R independent measurements of X and Zoey was assuming she made a single measurement of $X|\alpha$ with a combined $\alpha = \sum_{r=1}^{R} \alpha_r$. Therefore the difference between their error bars is due to σ_α.

The end goal, of course, is not to estimate the reference α or its mean, but to estimate quantities related to the quantum state like $p = \text{Tr}[\rho P_0]$. However, the better one's accuracy in estimating the references, the better one's accuracy in p. Therefore, given the above discussion, it is apparent that there is no advantage to drawing multiple samples of (X, Y, Z) when it is possible to increase the number of measurements N instead, often by adding samples together as Zoey did. There are, however, special circumstances where drawing multiple samples is desirable. This occurs in cases where p is not fixed shot to shot, but is instead drawn from a distribution on each shot.

There is a second and less obvious advantage to Zoey's method. Yves' model assumes that the multinormal distribution on (α, β) is a good approximation to the moments of the stochastic process governing α and β discussed in Section 3.2.4. This will usually be good enough. But if, for example, there are daily temperature patterns in the lab which significantly affect optical alignment, this could cause the normal approximation to be inaccurate. Zoey's model, however, makes no assumptions about the nature of the drift.

Finally, the third advantage of the conditional model is that it is simpler and therefore more tractable. Just solving for the MLE of the hierarchical model analytically would be difficult or impossible; it is painful enough for the conditional model.

3.3.4 Basic Inference Problem

We state the basic NV measurement inference problem for both the conditional and hierarchical models. For the conditional model, recall that we are, although without much loss of generality, limited to drawing a single sample (x, y, z) from $(X, Y, Z)|\alpha, \beta$. The inference problem is, given the likelihood function

$$\begin{aligned}\mathcal{L}(p, \alpha, \beta|x, y, z) &= \Pr\left(X = x, Y = y, Z = z|p, \alpha, \beta\right) \\ &= \mathrm{pdf}_{\mathrm{Pois}}\left(x; \alpha\right) \cdot \mathrm{pdf}_{\mathrm{Pois}}\left(y; \beta\right) \cdot \mathrm{pdf}_{\mathrm{Pois}}\left(z; p\alpha + (1-p)\beta\right) \\ &= \frac{\alpha^x \mathrm{e}^{-\alpha}}{x!} \cdot \frac{\beta^y \mathrm{e}^{-\beta}}{y!} \cdot \frac{(p\alpha + (1-p)\beta)^z \mathrm{e}^{-(p\alpha+(1-p)\beta)}}{z!},\end{aligned} \tag{3.13}$$

to infer the value of p. Note that α and β are nuisance parameters.

In the case of the hierarchical model, if we take R iid samples (x_r, y_r, z_r) from (X, Y, Z) we have the likelihood function

$$\begin{aligned}&\mathcal{L}(p, \bar{\alpha}, \bar{\beta}, \sigma_\alpha, \sigma_\beta, \sigma_{\alpha,\beta}|(x_1, y_1, z_1), (x_2, y_2, z_2), ..., (x_R, y_R, z_R)) = \\ &\quad \prod_{r=1}^{R} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathrm{pdf}_{\mathrm{Pois}}\left(x_r; \alpha\right) \cdot \mathrm{pdf}_{\mathrm{Pois}}\left(y_r; \beta\right) \cdot \mathrm{pdf}_{\mathrm{Pois}}\left(z_r; \gamma\right) \\ &\qquad \cdot \mathrm{pdf}_{\mathrm{Norm}}\left(\alpha, \beta; \left(\begin{smallmatrix}\bar{\alpha} \\ \bar{\beta}\end{smallmatrix}\right), \left(\begin{smallmatrix}\sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \sigma_\beta^2\end{smallmatrix}\right)\right) \mathrm{d}\alpha\, \mathrm{d}\beta\end{aligned} \tag{3.14}$$

where $\bar{\alpha}$, $\bar{\beta}$, σ_α, σ_β, and $\sigma_{\alpha,\beta}$ are nuisance parameters, and we are still trying to infer the value of p. This integral is generally intractable.

3.3.5 Generalized Inference Problems

In the previous subsection, the inference problem was stated such that the survival probability $p = \mathrm{Tr}[\rho P_0]$ was the quantity of interest. We may of course modify this if some other quantity is preferred.

For example, suppose we are interested in state tomography. We define the ideal unitary operators $\{U_1, ..., U_9\} \in \mathcal{U}(\mathcal{H})$ in such a way that $\{U_n |0\rangle\langle 0| U_n^\dagger\}_{n=1}^9$ is a basis for $\mathrm{L}(\mathcal{H})$. Our scheme is to prepare ρ, implement the gate U_n for some n, and measure the resulting state, so that $p_n = \mathrm{Tr}[U_n \rho U_n^\dagger P_0]$ with corresponding random variables

$$Z_n \sim \mathrm{Poisson}\left(p_n\alpha + (1-p_n)\beta\right) \equiv \mathrm{Poisson}\left(\gamma\right). \tag{3.15}$$

Then our likelihood function becomes

$$\mathcal{L}(\rho,\alpha,\beta|x,y,z_1,z_2,...,z_9)=\mathrm{pdf}_{\mathrm{Pois}}\left(x;\alpha\right)\cdot\mathrm{pdf}_{\mathrm{Pois}}\left(y;\beta\right)\cdot\prod_{n=1}^{9}\mathrm{pdf}_{\mathrm{Pois}}\left(z_n;p_n\alpha+(1-p_n)\beta\right) \tag{3.16}$$

and we are interested in inferring $\rho\in\mathrm{D}(\mathcal{H})$. This will be similar for the hierarchical model. We have taken one set of reference measurements for all $Z_1,...,Z_9$. We could also choose to take one for each, resulting in the data $(x_1,y_1,z_1),...,(x_9,y_9,z_9)$ with a similar likelihood function of the form $\mathcal{L}(\rho,\alpha_1,\beta_1,...,\alpha_9,\beta_9|x,y,z_1,z_2,...,z_9)$. These sorts of details come down to the particulars of the experimental implementation.

It is clear that any measurement inference problem can be stated in a similar way, such as process tomography, and Hamiltonian parameter inference, as will be seen in Section 3.6.

3.3.6 Fisher Information and the Cramér–Rao Bound

The average curvature of the likelihood function provides a measure of how informative data are. Generally, a highly curved (unimodal) likelihood function implies a tight region of support, so that data will tell you a lot about the parameters of interest. This is formalized by Fisher information and the Cramér–Rao bound [30]. Given a likelihood function $\mathcal{L}(\vec{\theta}|\vec{d})$ of parameters $\vec{\theta}$ given data $\vec{d}$, the Fisher information is the average curvature of $\mathcal{L}$, more specifically, it is the negative expected Hessian matrix of the log-likelihood, $I(\vec{\theta})_{i,j}=-\mathbb{E}_{\vec{d}}[\frac{\partial^2\log L(\vec{\theta}|\vec{d})}{\partial\theta_i\partial\theta_j}|\vec{\theta}]$. The Cramér–Rao bound asserts that no unbiased estimator of $\vec{\theta}$ can outperform this intrinsic curvature. Namely, if $\hat{\theta}$ is any unbiased estimator, which takes data $\vec{d}$ and outputs estimates of the true value of $\vec{\theta}$, then its covariance is lower-bounded by the inverse Fisher information matrix,

$$\mathrm{Cov}[\hat{\theta}]\geq I(p,\alpha,\beta)^{-1}. \tag{3.17}$$

As good estimators will often make this inequality nearly tight, the inverse Fisher information sets a benchmark for estimators to aim at. There is a generalized inequality for biased estimators that will be used in Section 3.4.1.

The Fisher information matrix of the conditional model $(X,Y,Z)|\alpha,\beta$ can be computed exactly as

$$I(p,\alpha,\beta)=\begin{pmatrix}\frac{(\alpha-\beta)^2}{p(\alpha-\beta)+\beta} & \frac{p(\alpha-\beta)}{p(\alpha-\beta)+\beta} & \frac{\alpha}{\beta+\alpha p-\beta p}-1\\ \frac{p(\alpha-\beta)}{p(\alpha-\beta)+\beta} & \frac{p^2}{p\alpha-p\beta+\beta}+\frac{1}{\alpha} & -\frac{(p-1)p}{p(\alpha-\beta)+\beta}\\ \frac{\alpha}{p\alpha-p\beta+\beta}-1 & -\frac{(p-1)p}{p(\alpha-\beta)+\beta} & \frac{p\alpha+(p-2)(p-1)\beta}{\beta(p(\alpha-\beta)+\beta)}\end{pmatrix} \tag{3.18}$$

with an inverse matrix given by

$$I(p,\alpha,\beta)^{-1} = \begin{pmatrix} \frac{p(p+1)\alpha+(p-2)(p-1)\beta}{(\alpha-\beta)^2} & \frac{p\alpha}{\beta-\alpha} & \frac{(p-1)\beta}{\alpha-\beta} \\ \frac{p\alpha}{\beta-\alpha} & \alpha & 0 \\ \frac{(p-1)\beta}{\alpha-\beta} & 0 & \beta \end{pmatrix}. \tag{3.19}$$

See Section B.2 for the calculation. The Cramér–Rao bound for the top left entry states that

$$\mathrm{Var}[\hat{p}(x,y,z)] \geq \frac{p(p+1)\alpha + (p-2)(p-1)\beta}{(\alpha-\beta)^2}, \tag{3.20}$$

for any unbiased estimator $\hat{p}$ given the data triple of photon counts (x,y,z). This bound is plotted in Figure 3.3 for the slice $\beta = \alpha/2$. In Section 3.4.5 we will see that this bound is very close to the average error incurred by a few important estimators, including the widely used maximum likelihood estimator, and in many disparate but relevant parameter regimes. This means that this inequality is perhaps better thought of as an approximation, except in exceptionally low contrast regimes.

We can use this result to derive a formula that tells us roughly how much data we will need to collect in order to lower the error bars on p to a specified level. We approximate that $\hat{p} \sim \mathrm{Normal}\left(p, (I(p,\alpha,\beta)^{-1})_{1,1}^{-1/2}\right)$ with α and β known (whereas we will generally only have estimates of them). This gives the $100(1-\zeta)\%$ confidence interval

$$p \pm \Delta p = p \pm c_{\zeta/2}\frac{\sqrt{p(p+1)\alpha + (p-2)(p-1)\beta}}{\alpha-\beta} \tag{3.21}$$

for $\hat{p}$ where $c_{\zeta/2} = \sqrt{2}\,\mathrm{erf}^{-1}(1-\zeta)$ with $\mathrm{erf}(x) = \pi^{-1/2}\int_{-x}^{x} e^{-t^2}\mathrm{d}t$ the error function. Supposing a reference contrast of $C = \frac{\alpha-\beta}{\alpha+\beta}$, we need

$$\alpha \approx \frac{c_{\zeta/2}^2}{2\Delta p^2}(1+1/C)^2 \tag{3.22}$$

to make $\hat{p} \approx p \pm \Delta p$ a $100(1-\zeta)\%$ confidence interval [2]. To derive this formula we have assumed the worst case, $p = 1$. These calculations show that, for example, if we desire a 95% confidence interval of ± 0.01 for p, then we need to do at least enough experiments N so that α is on the order of $\sim 170,000$.

[2] We use this definition for contrast simply because it shows up naturally in Equation 3.28; $\frac{\alpha-\beta}{\alpha}$ is arguably a better choice.

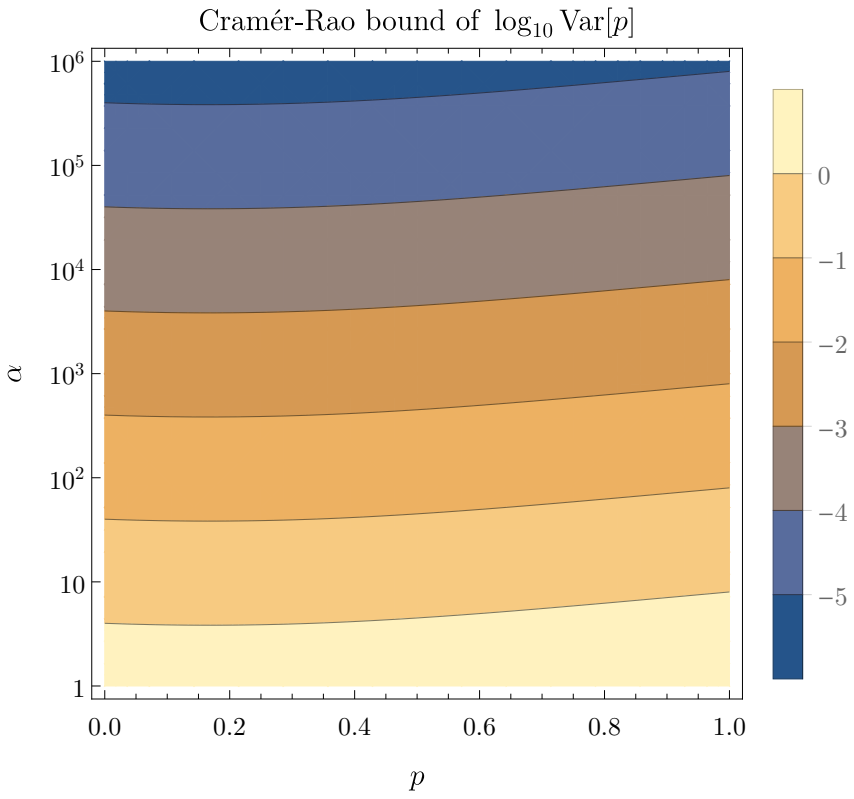

Figure 3.3: With β fixed as $\alpha/2$, the Cramér-Rao bound of $\hat{p}(x, y, z)$ is plotted as a function of α and p. We see that values of p closer to 1 are slightly more difficult to estimate.

This heuristic holds for any optical transition rates and choice of measurement time Δt in the measurement protocol detailed in Section 2.7. However, certain choices of measurement time are better than others. We can use the Cramér–Rao bound to estimate the optimal such time, namely, we wish to choose the measurement time Δt that maximizes the temporal information density of p. Begin by supposing that the total runtime of a fixed experiment, including taking bright and dark reference counts, is $T = N(T_e + 3\Delta t)$, where N is the number of repetitions, and T_e is the amount of time per repetition not spent counting photons (initialization, wait periods, pulse sequences, etc.). We must multiply Δt by 3 to account for all three of the signal, bright reference, and dark reference counting windows. Writing $\alpha = N\overline{\alpha}$ and $\beta = N\overline{\beta}$, with $\overline{\alpha}$ and $\overline{\beta}$ the average per-shot reference values, again at the worst case $p = 1$, gives $\Delta p^2 = \frac{2\alpha}{(\alpha-\beta)^2}$, or rearranging, gives

$$\Delta p\sqrt{T} = \frac{\sqrt{2(T_e + 3\Delta t)\overline{\alpha}(\Delta t)}}{\overline{\alpha}(\Delta t) - \overline{\beta}(\Delta t)}. \tag{3.23}$$

We have written $\overline{\alpha} = \overline{\alpha}(\Delta t)$ and $\overline{\beta} = \overline{\beta}(\Delta t)$ to emphasize their implicit dependence on Δt. These two functions are easily estimated experimentally by sweeping the length of the measurement window. Then for any given T_e, the the quantity $\Delta p\sqrt{T}$ can be be minimized, visualized in Figure 3.4. As the experiment time T_e grows, it becomes increasingly worthwhile to lengthen the duty cycle of measurement. This formula and its units are analogous to widely used magnetometry sensitivity formulas; see [170] or [86] for two examples out of many.

Of note is the steep increase of $\Delta p\sqrt{T}$ as $\Delta t \to 0$. While the slope is relatively gentle as Δt gets larger, causing little harm even if Δt is twice as big as the optimal value for a given T_e, there is a large penalty for choosing a measurement of Δt which is too short. Long refocusing sequences like CPMG are especially at risk of falling into this trap.

3.3.7 Other Measurement Protocols

Our analysis has considered the conventional protocol for NV spin-state measurement. Other measurement protocols exist, such as one that uses multiple laser frequencies to exploit spin-to-charge conversion [158], or another that incoherently stores pre-measurement spin populations on an ancilla nuclear qubit that is subsequently read out several times successively [96], both potentially achieve greater read-out efficiency. The methods we describe here can be applied to these other measurement schemes and will have similar forms—their detailed analyses we leave for future work.

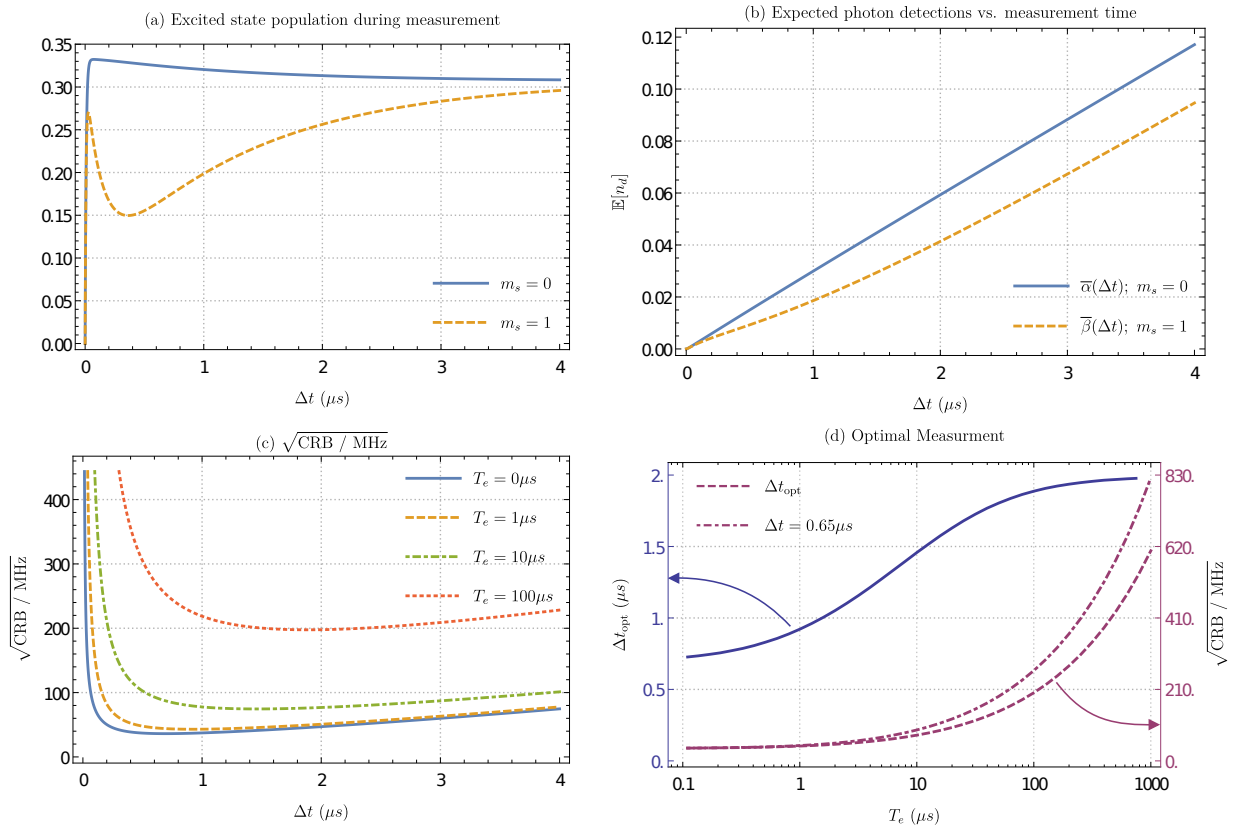

Figure 3.4: Simulated example of optimizing measurement time for a low visibility experiment. The same analysis holds for high visibility experiments. (a) The population of the optical excited state is plotted for two initial states, (b) which results in distinguishable numbers of detected photons given that we average enough repetitions. They are labeled $\overline{\alpha}(\Delta t)$ and $\overline{\beta}(\Delta t)$ in the main body and asymptote to the same slope since they both end up in the same steady state of the master equation. (c) These curves can be used to estimate the standard deviation of p normalized to square-root runtime for various experiment lengths. For example, given $\sqrt{\text{CRB/MHz}} = 400/\sqrt{\text{MHz}}$, a total run time of $100\,\text{s} = 10^8$ us will approximately reduce the uncertainty of Δp to 0.04. (d) As a function of T_e, optimal measurement window length Δt_{opt} is shown (left axis) along with the corresponding $\sqrt{\text{CRB/MHz}}$ values for both the optimal measurement time, and a fixed measurement time of 0.65 us (right axis). It is seen that in this regime the payoff of using the optimal measurement time is rather slim.

3.4 Estimators for the Referenced Coin Model

The previous section discussed inference models of NV measurement in some detail. In this section, we introduce two estimators for the basic inference problem of the conditional model defined in Section 3.3.4. We also compare their relative strengths and weaknesses.

3.4.1 Maximum Likelihood Estimator

The most obvious estimator turns out to be quite a good one, and the one that has been used almost universally in practice. Suppose (x, y, z) is a variate of $(X, Y, Z)|\alpha, \beta$. Equation Equation 3.10 shows that x, y, and z are unbiased estimates of α, β, and $p\alpha + (1-p)\beta$, respectively. We invert the equation $p\alpha + (1-p)\beta$ for p substituting in our estimates above to get the estimator

$$\hat{p}_{\text{MLE}} = \frac{z-y}{x-y} \tag{3.24}$$

for p. Although appearing quite simple, it is difficult to work with this estimator analytically; it is the ratio of two correlated Skellam distributions which does not have many nice properties. However, it can be shown with Lagrange multipliers that this is the maximum likelihood estimator (MLE) of the model, that is, $\left(\frac{z-y}{x-y}, x, y\right)$ is the (unique) maximum of the function $\mathcal{L}(p, \alpha, \beta|x, y, z)$ on the domain $0 \leq p \leq 1$, $0 < \beta \leq \alpha$, for any $x, y, z \geq 0$. See Section B.3.1 for details.

There is always a finite probability that $x = y$, in which case this estimator will divide by zero. With sufficient magnitude of and contrast between α and β this is highly unlikely. It still poses a problem if we wish to prove anything about it, for example, if we wish to find its expectation value. To avoid this situation we define the slightly modified estimator

$$\hat{p}_{\text{MLE},\epsilon} = \frac{z-y}{x-y+\epsilon} \tag{3.25}$$

for some non-integer value of ϵ. One might consider using this estimator instead of the MLE if $x = y$ has a significant probability. In the next section, Section 3.4.2, we show that the bias of $\hat{p}_{\text{MLE},\epsilon}$ is non-zero, a linear function of p, and exactly given by the integral

$$\text{Bias}[\hat{p}_{\text{MLE},\epsilon}] = \text{re}\int_0^\pi \text{i}(-1)^{-\epsilon} e^{-\alpha(1+e^{\text{i}\phi})-\beta(1+e^{-\text{i}\phi})}((\beta+(\alpha-\beta)p)e^{\text{i}\phi\epsilon} + \beta e^{\text{i}\phi(\epsilon-1)})\text{d}\phi - p. \tag{3.26}$$

Taking the limit as ϵ approaches zero gives an expression for the bias of the original estimator,

$$\begin{aligned}\mathrm{Bias}[\hat{p}_{\mathrm{MLE}}] &= \lim_{\epsilon \to 0} \mathrm{Bias}[\hat{p}_{\mathrm{MLE},\epsilon}] \\ &= \int_0^\pi e^{-(\alpha+\beta)(1+\cos\phi)} \left[(\gamma + \beta\cos\phi)\sin((\alpha-\beta)\sin\phi) + (\beta\sin\phi)\cos((\alpha-\beta)\sin\phi)\right] \mathrm{d}\phi - p.\end{aligned} \tag{3.27}$$

This integral can be solved numerically to find the exact bias. If $\alpha + \beta \gg 1$, we can derive an asymptotic approximation to this integral,

$$\mathrm{Bias}[\hat{p}_{\mathrm{MLE}}] \approx \left(p - \frac{\beta}{\alpha+\beta}\right) \frac{\alpha+\beta}{(\alpha-\beta)^2} + \mathcal{O}\left((\alpha+\beta)^{-2}\right), \tag{3.28}$$

which through numerics can be shown to be valid for $\alpha + \beta \gtrsim 300$. Note that the bias vanishes at $p = \frac{\beta}{\alpha+\beta}$, and that if the contrast $C = \frac{\alpha-\beta}{\alpha+\beta}$ is fixed, then the worst case bias scales as $\frac{\alpha+\beta}{(\alpha-\beta)^2} = \mathcal{O}\left((\alpha+\beta)^{-1}\right)$.

Although estimators with no bias are generally preferred, we see that this estimator has the more important property of being consistent, meaning that $\mathrm{Bias}[\hat{p}_{\mathrm{MLE}}] \to 0$ as $\alpha \to \infty$ with fixed contrast. We can make an even stronger statement by using the Cramér–Rao bound for biased estimators, which is a generalization of Equation 3.17 known as the van Trees inequality or the Bayesian Cramér–Rao bound [175], stating that

$$\mathrm{Cov}[\hat{\theta}] \geq J_{\hat{\theta}}(p,\alpha,\beta) I(p,\alpha,\beta)^{-1} J_{\hat{\theta}}^{\mathrm{T}}(p,\alpha,\beta) \tag{3.29}$$

where $J_{\hat{\theta}}(p,\alpha,\beta)$ is the expectation value of the Jacobian matrix of the possibly biased estimator $\hat{\theta}$. Using the approximation from Equation 3.28, this gives us the inequality

$$\mathrm{Var}[\hat{p}_{\mathrm{MLE}}] \geq \frac{p(p+1)\alpha + (p-2)(p-1)\beta}{(\alpha-\beta)^2} + \mathcal{O}\left((\alpha+\beta)^{-2}\right) \tag{3.30}$$

for the maximum likelihood estimator, again assuming that the contrast is fixed as α and β increase. This is approximately the same bound as the unbiased Cramér–Rao bound discussed earlier. Indeed, this is generic, as the van Trees inequality approaches the Cramér–Rao bound for large data sets, such that incorporating prior information can be thought of as an important correction for finite data sets [136].

3.4.2 Bias of the MLE

In this section we derive the bias of the MLE used in the previous section by studying the modified estimator $\hat{p}_{\mathrm{MLE},\epsilon}$ from Equation 3.25. In several places Mathematica is used to evaluate limits and series. We are ultimately interested in the limiting case as $\epsilon \to 0$ since this gives the maximum likelihood estimator. We wish to compute the bias of $\hat{p}_{\mathrm{MLE},\epsilon}$, which is given by

$$\begin{aligned}\mathrm{Bias}[\hat{p}_{\mathrm{MLE},\epsilon}] &= \mathbb{E}_{x,y,z}[\hat{p}_{\mathrm{MLE},\epsilon} - p] \\ &= \sum_{x=0}^{\infty}\sum_{y=0}^{\infty}\sum_{z=0}^{\infty} \frac{z-y}{x-y+\epsilon}\frac{\alpha^x \mathrm{e}^{-\alpha}}{x!}\cdot\frac{\beta^x \mathrm{e}^{-\beta}}{y!}\cdot\frac{\gamma^x \mathrm{e}^{-\gamma}}{z!} - p. \end{aligned} \tag{3.31}$$

This triple sum is not straight forward to compute. Our strategy is to first sum over z and x resulting in

$$\begin{aligned}\mathbb{E}_{x,y,z}[\hat{p}_{\mathrm{MLE},\epsilon}] = \sum_{y=0}^{\infty}\Bigg[&\frac{\gamma e^{-\alpha-\beta}\beta^y(-\alpha)^{y-\epsilon}\Gamma(\epsilon-y)}{\Gamma(y+1)} - \frac{e^{-\alpha-\beta}\beta^y(-\alpha)^{y-\epsilon}\Gamma(\epsilon-y)}{\Gamma(y)} \\ &+\frac{e^{-\alpha-\beta}\beta^y(-\alpha)^{y-\epsilon}\Gamma(\epsilon-y,-\alpha)}{\Gamma(y)} - \frac{\gamma e^{-\alpha-\beta}\beta^y(-\alpha)^{y-\epsilon}\Gamma(\epsilon-y,-\alpha)}{\Gamma(y+1)}\Bigg]\end{aligned} \tag{3.32}$$

where $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\mathrm{d}t$ is the gamma function, and $\Gamma(s,x) = \int_x^\infty t^{s-1}e^{-t}\mathrm{d}t$ is the upper incomplete gamma function. Note that the non-integer value of ϵ allows us to avoid poles of the gamma function. The first two terms of the sum are seen to be purely imaginary. Since the expectation value is known to be real, we may ignore them.

To proceed, we make use of the complex integral form of the incomplete gamma function,

$$\Gamma(\epsilon-y,-\alpha) = \int_{-\alpha}^{\infty} t^{\epsilon-y-1}e^{-t}\mathrm{d}t = \lim_{R\to\infty}\int_{-1}^{R}(\alpha t)^{\epsilon-y-1}e^{-\alpha t}\alpha\mathrm{d}t, \tag{3.33}$$

which holds for any integration path in $\mathbb{C}$ from -1 to R which does not cross the negative real axis. For the third term we get

$$\begin{aligned}\sum_{y=0}^{\infty}\frac{e^{-\alpha-\beta}\beta^y(-\alpha)^{y-\epsilon}\Gamma(\epsilon-y,-\alpha)}{\Gamma(y)} &= \int_{-1}^{\infty}\sum_{y=0}^{\infty}\frac{e^{-\alpha-\beta}\beta^y(-1)^{y-\epsilon}t^{\epsilon-y-1}e^{-\alpha t}}{\Gamma(y)}\mathrm{d}t \\ &= \int_{-1}^{\infty}\beta(-1)^{1-\epsilon}t^{\epsilon-2}e^{-\alpha(1+t)-\beta(1+1/t)}\mathrm{d}t\end{aligned} \tag{3.34}$$

and similarly

$$\sum_{y=0}^{\infty} -\frac{\gamma e^{-\alpha-\beta}\beta^y(-\alpha)^{y-\epsilon}\Gamma(\epsilon-y,-\alpha)}{\Gamma(y+1)} = \int_{-1}^{\infty}\sum_{y=0}^{\infty} -\frac{\gamma e^{-\alpha-\beta}\beta^y(-1)^{y-\epsilon}t^{\epsilon-y-1}e^{-\alpha t}}{\Gamma(y+1)}\mathrm{d}t$$
$$= \int_{-1}^{\infty}\gamma(-1)^{1-\epsilon}t^{\epsilon-1}e^{-\alpha(1+t)-\beta(1+1/t)}\mathrm{d}t \tag{3.35}$$

for the last term. Therefore

$$\mathbb{E}_{x,y,z}[\hat{p}_{\mathrm{MLE},\epsilon}] = \int_{-1}^{\infty}(-1)^{1-\epsilon}e^{-\alpha(1+t)-\beta(1+1/t)}(\gamma t^{\epsilon-1}+\beta t^{\epsilon-2})\mathrm{d}t \tag{3.36}$$

Consider the integration path which is a semi-circle from -1 to 1, and then a straight line to R along the positive real axis, avoiding the singularity at $t = 0$. Note that along the positive real axis, the integrand is purely imaginary. We are therefore only interested in the real part of the integral along the unit semi-circle. Making the change of variables $t = e^{i\phi}$ we are left with

$$\mathbb{E}_{x,y,z}[\hat{p}_{\mathrm{MLE},\epsilon}] = \mathrm{re}\int_0^{\pi}\mathrm{i}(-1)^{-\epsilon}e^{-\alpha(1+e^{\mathrm{i}\phi})-\beta(1+e^{-\mathrm{i}\phi})}(\gamma e^{\mathrm{i}\phi\epsilon}+\beta e^{\mathrm{i}\phi(\epsilon-1)})\mathrm{d}\phi \tag{3.37}$$

from which we conclude that

$$\begin{aligned}\mathbb{E}_{x,y,z}[\hat{p}_{\mathrm{MLE}}] &= \lim_{\epsilon\to 0}\mathbb{E}_{x,y,z}[\hat{p}_{\mathrm{MLE},\epsilon}]\\ &= \mathrm{re}\int_0^{\pi}\mathrm{i}e^{-\alpha(1+e^{\mathrm{i}\phi})-\beta(1+e^{-\mathrm{i}\phi})}(\gamma+\beta e^{-\mathrm{i}\phi})\mathrm{d}\phi\\ &= \int_0^{\pi}e^{-(\alpha+\beta)(1+\cos\phi)}\\ &\qquad\cdot\left[(\gamma+\beta\cos\phi)\sin((\alpha-\beta)\sin\phi)+(\beta\sin\phi)\cos((\alpha-\beta)\sin\phi)\right]\mathrm{d}\phi.\end{aligned} \tag{3.38}$$

From this expression we can see that $\mathbb{E}_{x,y,z}[\hat{p}_{\mathrm{MLE}}]$ is exactly linear with respect to γ, and therefore the bias will be exactly linear with respect to p. This integral is best done numerically. What follows is a closed form approximation which will usually be more practical.

We see that the integrand falls off exponentially with rate $\alpha+\beta$ as ϕ decreases from π. Since it will usually hold that $\alpha+\beta \gg 2$, the integrand will have most of its support in

a region close to π. This justifies a low order series expansion of trigonometric functions about $\phi = \pi$, giving

$$\begin{aligned}
\mathbb{E}_{x,y,z}[\hat{p}_{\text{MLE}}] &\approx \int_0^{\pi} e^{-(\alpha+\beta)(\phi-\pi)^2/2} \\
&\qquad [-(\gamma-\beta)\sin((\alpha-\beta)(\phi-\pi)) - (\phi-\pi)\beta\cos((\alpha-\beta)(\phi-\pi))]\,\mathrm{d}\phi \\
&= \int_{-\pi}^{0} e^{-(\alpha+\beta)\phi^2/2}\,[-(\gamma-\beta)\sin((\alpha-\beta)\phi) - \phi\beta\cos((\alpha-\beta)\phi)]\,\mathrm{d}\phi \\
&\approx \int_{-\infty}^{0} e^{-(\alpha+\beta)\phi^2/2}\,[-(\gamma-\beta)\sin((\alpha-\beta)\phi) - \phi\beta\cos((\alpha-\beta)\phi)]\,\mathrm{d}\phi \\
&= \frac{\gamma-\beta}{\alpha-\beta} + \left(\frac{\gamma-\beta}{\alpha-\beta} - \frac{\beta}{\alpha+\beta}\right)\left(\sqrt{2}\frac{\alpha-\beta}{\sqrt{\alpha+\beta}}F\left(\frac{\alpha-\beta}{\sqrt{2}\sqrt{\alpha+\beta}}\right) - 1\right) \\
&= p + \left(p - \frac{\beta}{\alpha+\beta}\right) f\left(\frac{\alpha-\beta}{\sqrt{\alpha+\beta}}\right)
\end{aligned} \tag{3.39}$$

where $F(x) = e^{-x^2}\int_0^x e^{t^2}\mathrm{d}t$ is the Dawson function and $f(x) = \sqrt{2}xF(x/\sqrt{2}) - 1$. Numerics show that these approximations are accurate to $\mathcal{O}\left((\alpha+\beta)^{-2}\right)$ for $\alpha + \beta \gtrsim 700$. For large x we have the series $f(x) = \frac{1}{x^2} + \frac{3}{x^4} + \mathcal{O}\left(x^{-6}\right)$. It follows that

$$\text{Bias}[\hat{p}_{\text{MLE}}] \approx \left(p - \frac{\beta}{\alpha+\beta}\right)\frac{\alpha+\beta}{(\alpha-\beta)^2} + \mathcal{O}\left((\alpha+\beta)^{-2}\right) \tag{3.40}$$

We see that the estimator is unbiased at the single point $p = \frac{\beta}{\alpha+\beta}$ and that worst bias happens at $p = 0$ or $p = 1$ and scales as $\mathcal{O}\left(\frac{1}{\alpha+\beta}\right)$. Note that we have assumed that the contrast $\frac{\alpha-\beta}{\alpha+\beta}$ stays fixed to make the above asymptotic arguments.

3.4.3 Bias Corrected Estimator

We can use Equation Equation 3.28 to attempt to derive an estimator which is less biased than the MLE by subtracting off an estimate of the bias. This gives

$$\begin{aligned}
\hat{p}_{\text{BCE}} &= \hat{p}_{\text{MLE}} + \left(\hat{p}_{\text{MLE}} - \frac{\hat{\beta}}{\hat{\alpha}+\hat{\beta}}\right)\frac{\hat{\alpha}+\hat{\beta}}{(\hat{\alpha}-\hat{\beta})^2} \\
&= \frac{z((x-y)^2 - x - y) - y((x-y)^2 - 2x - 2y)}{(x-y)^3}
\end{aligned} \tag{3.41}$$

However, recall from Section 3.3.6 that the Cramér–Rao bound gives us the lower bound

$$\mathrm{Var}[\hat{p}_{\mathrm{MLE}}] \geq \frac{p(p+1)\alpha + (p-2)(p-1)\beta}{(\alpha-\beta)^2} \tag{3.42}$$

on the variance of our estimator. If we take the average of this bound at $p = 0$ and $p = 1$ we get $\frac{\alpha+\beta}{(\alpha-\beta)^2}$ which happens to be equal to the worst-case bias derived above. We conclude that when $\alpha + \beta \gtrsim 300$, the bias of the MLE is negligible since it will be of $\mathcal{O}\left((\alpha+\beta)^{-1}\right)$, well contained within a single standard deviation of $\hat{p}_{\mathrm{MLE}}$, $\mathcal{O}\left((\alpha+\beta)^{-\frac{1}{2}}\right)$.

Moreover, the estimate of the bias $\left(\hat{p}_{\mathrm{MLE}} - \frac{\hat{\beta}}{\hat{\alpha}+\hat{\beta}}\right)\frac{\hat{\alpha}+\hat{\beta}}{(\hat{\alpha}-\hat{\beta})^2}$ used above is itself likely to have a variance and bias which outweigh that which it is trying to correct. Therefore, this estimator is not useful in practice.

3.4.4 Bayes Estimator

If we assume our prior knowledge of the parameters (p, α, β) is encoded in the probability distribution $\pi(p, \alpha, \beta)$, then assuming our model is correct, Bayes' theorem will tell us how to best update our beliefs about the parameters after we have measured the variate (x, y, z):

$$\begin{aligned}\pi^*(p,\alpha,\beta) \equiv \Pr(p,\alpha,\beta|x,y,z) &= \frac{\Pr(x,y,z|p,\alpha,\beta)\,\pi(p,\alpha,\beta)}{\int \Pr(x,y,z|p,\alpha,\beta)\,\pi(p,\alpha,\beta)\mathrm{d}p\,\mathrm{d}\alpha\,\mathrm{d}\beta} \\ &= \frac{L(p,\alpha,\beta|x,y,z)\pi(p,\alpha,\beta)}{\mathcal{N}}.\end{aligned} \tag{3.43}$$

Here, L is the likelihood function from Equation Equation 3.13 and $\mathcal{N}$ is a normalization constant.

If we assume a separable prior $\pi(p, \alpha, \beta) = \pi(p)\pi(\alpha, \beta)$, and it would be strange not to, then Bayes' theorem can be applied sequentially. This relies on the conditional independence of X, Y, and Z. We can first update our prior distribution using the datum (x, y) to get

$$\begin{aligned}\Pr(p,\alpha,\beta|x,y) &= \frac{\Pr(x,y|p,\alpha,\beta)\,\pi(p,\alpha,\beta)}{\int \Pr(x,y|p,\alpha,\beta)\,\pi(p,\alpha,\beta)\mathrm{d}p\,\mathrm{d}\alpha\,\mathrm{d}\beta} \\ &= \frac{\Pr(x|\alpha)\Pr(y|\beta)\,\pi(\alpha,\beta)}{\int \Pr(x|\alpha)\Pr(y|\beta)\,\pi(\alpha,\beta)\mathrm{d}\alpha\,\mathrm{d}\beta}\pi(p) \equiv \pi^*(\alpha,\beta)\pi(p)\end{aligned} \tag{3.44}$$

and subsequently

$$\Pr(p,\alpha,\beta|z) = \frac{\Pr(z|p,\alpha,\beta)\,\pi^*(\alpha,\beta)\pi(p)}{\int \Pr(z|p,\alpha,\beta)\,\pi^*(\alpha,\beta)\pi(p)\mathrm{d}p\,\mathrm{d}\alpha\,\mathrm{d}\beta} = \pi^*(p,\alpha,\beta). \tag{3.45}$$

This sequential break down is useful because a conjugate prior can be found for the likelihood $\Pr(x,y|p,\alpha,\beta)$ in a couple of useful cases, meaning the posterior $\pi^*(\alpha,\beta)$ can be computed exactly. Formulas for two different conjugate priors for the references are given in Section 3.5. The full likelihood L, however, almost certainly does not have a conjugate prior.

Given the posterior distribution π^*, there are many choices for the estimator. The most common is the mean square error (MSE) Bayes estimator which simply takes the expectation value of π^*. For the particular parameter p, we have

$$\hat{p}_{\mathrm{Bayes}} = \mathbb{E}[p|x,y,z] = \int_0^1 \int_0^\infty \int_0^\infty p\pi^*(p,\alpha,\beta)\mathrm{d}p\,\mathrm{d}\alpha\,\mathrm{d}\beta. \tag{3.46}$$

This is known to be the estimator which minimizes the expected MSE of the estimate over all possible estimators,

$$(\hat{p}_{\mathrm{Bayes}},\hat{\alpha}_{\mathrm{Bayes}},\hat{\beta}_{\mathrm{Bayes}}) = \mathrm{argmin}_{\hat{\theta}}\,\mathbb{E}[(\hat{\theta}(x,y,z)-(p,\alpha,\beta))^2]. \tag{3.47}$$

3.4.5 Comparing Estimators: Risk

We wish to compare the quality of the above estimators. In general, when one wants to compare several estimators, or even assess the quality of a single estimator, the first step is to decide which properties make an estimator 'good' in the context of the problem at hand. For example, if there are serious consequences to being wrong, then it might be better to err on the side of a cautious estimator, rather than one with lousy worst case performance but better performance on average. In our case, the quantity p will rarely, if ever, be the primary quantity of interest. However, it will always be a gateway, implicitly or explicitly, to the real quantities of interest. If we assume, for the sake of generality, that the size of the average mean-squared-error (MSE) on these parameters is a sufficient figure of merit, then by the standard rules of error propagation, out of a collection of estimators for p, we should prefer the one with the best MSE averaged over random realizations. It is possible that there are situations in which this general principle does not hold, in which case the full and specific problem should be considered.

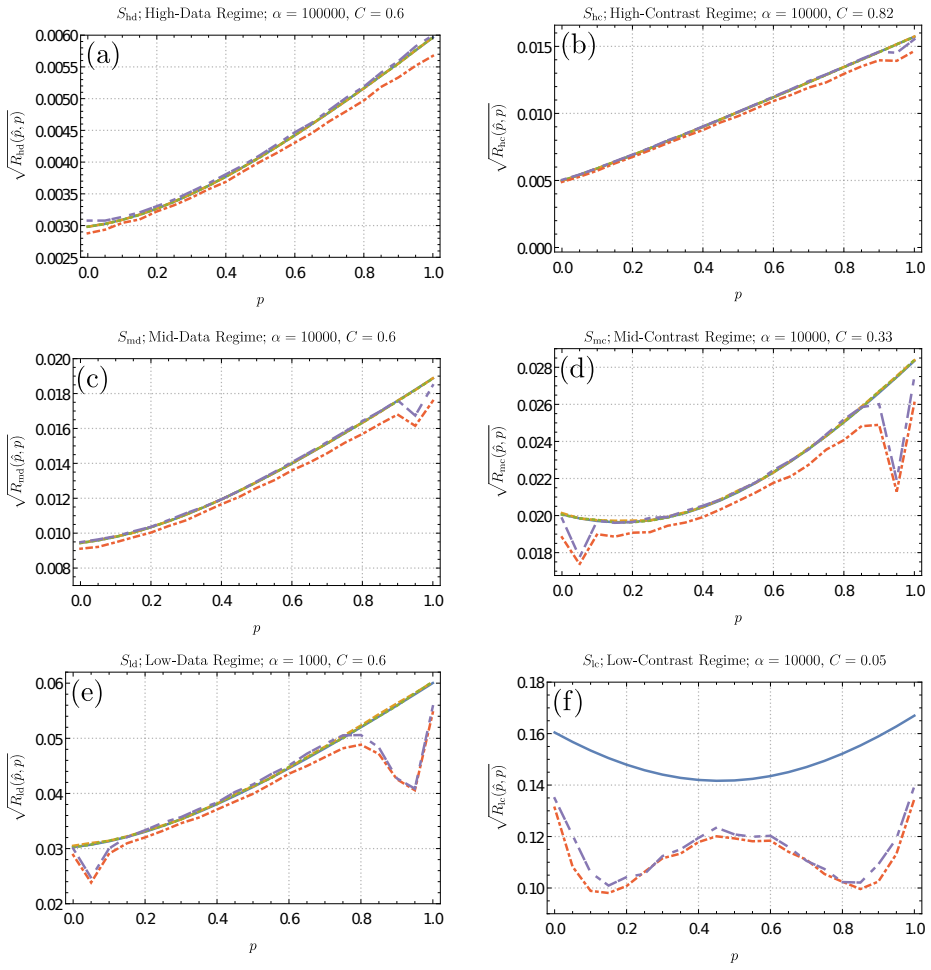

Figure 3.5: The MSE risk for several estimators of p, labeled in the legend, is plotted for six different regimes of experimental setup, (a)-(f). The square root has been taken so that the units of the y-axes have the same units as p. The estimators under study are the maximum likelihood estimator, $\hat{p}_{\mathrm{MLE}}$, the bias corrected estimator (see Section 3.4.3), $\hat{p}_{\mathrm{BCE}}$, and the Bayes estimator, $\hat{p}_{\mathrm{Bayes}}$, with two different priors. These priors are denoted by "Bayes" and "Bayes-10", with the latter being a more conservative prior corresponding to a ten-fold increase in the assumed covariance, as explained in the main body. Sharp peaks for the Bayes estimators are artefacts of the coarse sampling along the x-axis; risk was evaluated at p ranging from 0 to 1 in steps of 0.05. The risks of $\hat{p}_{\mathrm{MLE}}$ and $\hat{p}_{\mathrm{BCE}}$ are much bigger than 1 for the low-contrast regime due to the common occurence of $y > x$, and are therefore not plotted.

To proceed, we accept the idea of a 'true value' of p, and define the MSE loss between the true value and the estimate $\hat{p}(x, y, z)$ as

$$\mathrm{L}_{\mathrm{MSE}}(\hat{p}(x, y, z), p) = (\hat{p}(x, y, z) - p)^2, \tag{3.48}$$

where (x, y, z) is some realization of data drawn from $(X, Y, Z)|p, \alpha, \beta$. Next, it is reasonable to assume that the operators of a given experimental setup will have a rough idea of what to expect as their reference counts. Or, given the results of a set of experiments, all of the reference counts can be pooled to empirically construct a distribution of reference counts. As such, we assume the existence of a probability distribution $P_{\mathrm{S}}(\alpha, \beta) = \Pr(\alpha, \beta|\text{experimental setup S})$ which characterizes a particular setup called S (assuming a fixed number of shots, N, per experiment).

Given an estimator $\hat{p}$, its *risk* with respect to the true values (p, α, β) is defined in the standard way as the average value of the loss,

$$\mathrm{R}(\hat{p}, p, \alpha, \beta) = \mathbb{E}_{x,y,z|p,\alpha,\beta}[\mathrm{L}_{\mathrm{MSE}}(\hat{p}, p)] = \sum_{x,y,z=0}^{\infty} \mathrm{L}_{\mathrm{MSE}}(\hat{p}(x, y, z), p) \Pr(x, y, z|\alpha, \beta, p). \tag{3.49}$$

With a particular setup S we can then quantify its overall risk,

$$\mathrm{R}_{\mathrm{S}}(\hat{p}, p) = \mathbb{E}_{\alpha,\beta}[\mathrm{R}(\hat{p}, p, \alpha, \beta)] = \int \mathrm{R}(\hat{p}, p, \alpha, \beta) P_{\mathrm{S}}(\alpha, \beta) \mathrm{d}\alpha \, \mathrm{d}\beta, \tag{3.50}$$

by marginalizing over our knowledge about it. Note that the units of $\sqrt{\mathrm{R}_{\mathrm{S}}(\hat{p}, p)}$ are the same as those of p, and that, for example, a value of 0.01 has the straight forward interpretation as being the amount that $\hat{p}$ differs from the true value p, averaging over all possible data coming from the given setup, weighted by the likelihood of that data. If we were to additionally marginalize over a distribution of p, this would be the Bayes risk, and an estimator minimizing this quantity would be a Bayes estimator. Therefore, the expression above can be seen as a hybrid between Bayes risk and frequentist risk, where we marginalize over α and β, but not p.

We will treat the prior of a Bayesian estimator as an implicit property of the estimator. We can then compare, for example, $\hat{p}_{\mathrm{Bayes}}$ with itself under different priors. The prior on α and β does not need to have any relationship with $P_{\mathrm{S}}(\alpha, \beta)$. A large part of choosing a prior has to do with assessing one's level of paranoia, and one may be more paranoid about generating a fair risk comparison than about giving the estimator an over-informed prior, or vice versa. However, setting them equal to each other will often be the most sensible thing to do.

We study a few regimes of experimental setups. We consider a high-data regime, S_{hd}, where $\overline{\alpha} = 100,000$, a mid-data regime, S_{md}, where $\overline{\alpha} = 10,000$, and a low-data regime, S_{ld} where $\overline{\alpha} = 1,000$. In these three cases the contrast is the same, $C = \frac{\overline{\alpha}-\overline{\beta}}{\overline{\alpha}+\overline{\beta}} = 0.6$. We additionally consider mid-data regimes of varying contrast, where $\overline{\alpha} = 10,000$ for $C = 0.05, 0.33, 0.82$. This defines the respective low, medium, and high contrast setups S_{lc}, S_{mc}, and S_{hc}. The above setup descriptions only supply the mean values of their respective distributions P_{S}, that is, $\mathbb{E}_{P_{\mathrm{S}}}[(\alpha, \beta)] = (\overline{\alpha}, \overline{\beta})$. To keep things simple, we take these distributions to all be binormal with super-Poisson standard deviations $\sigma_\alpha = 2\sqrt{\overline{\alpha}}$ and $\sigma_\beta = 2\sqrt{\overline{\beta}}$, and covariances defined by $\sigma_{\alpha,\beta} = 1.5\beta$.

In Figure 3.5, $\sqrt{\mathrm{R}_{\mathrm{S}}(\hat{p}, p)}$ is plotted for each of the setups described above, and for each of the estimators $\hat{p}_{\mathrm{MLE}}$, $\hat{p}_{\mathrm{BCE}}$, and $\hat{p}_{\mathrm{Bayes}}$. Two different priors are used for the Bayes estimator on each setup, both are product gamma distributions, discussed further in Section 3.5.1. The first, 'Bayes', uses the same mean value and diagonal covariance elements as P_{S}. The second, 'Bayes-10', uses the same mean value as P_{S}, but standard deviations which are ten times larger than P_{S}, corresponding to a rather uninformative prior. The more sophisticated prior discussed in Section 3.5.2, which allows for correlations between α and β, should in theory be strictly better than the ones used in these calculations, but were found to be too computationally expensive for naive implementations of risk computation. For all setups and estimators, risk is computed by Monte Carlo sampling; for each value of p, many pairs (α, β) are sampled from $P_{\mathrm{S}}(\alpha, \beta)$, for each pair many variates (x,y,z) are drawn from the likelihood distribution, and the loss L_{MSE} is computed for each. The average of these loss values for this value of p forms an estimate of $\mathrm{R}_{\mathrm{S}}(\hat{p}, p)$.

These plots show that the Cramér–Rao bound Equation 3.20 is an excellent estimate of the risk of the MLE in most regimes. Further, under our loss function, the Bayes estimator never has more risk than the MLE, and has superior performance especially near the boundaries of $[0, 1]$, even for the rather uninformative Bayes-10 prior.

3.5 Conjugate Priors for Reference Parameters

Given a likelihood function, its *conjugate prior* is a special family of distributions such that the posterior is also in the same family of distributions. Therefore conjugate priors, when they exist, are very useful at reducing the complexity of applying Bayesian inference.

3.5.1 Uncorrelated Conjugate Prior

Since the likelihood function $L(\alpha, \beta|x, y)$ for the drift parameters is separable into two Poisson distributions,

$$L(\alpha, \beta|x, y) = \frac{\alpha^x \mathrm{e}^{-\alpha}}{x!} \cdot \frac{\beta^y \mathrm{e}^{-\beta}}{y!}, \tag{3.51}$$

the product of Gamma distributions will be a conjugate prior. Indeed, with the prior

$$\pi(\alpha, \beta) = \pi(\alpha)\pi(\beta) = \mathrm{pdf}_{Gamma}(\alpha; a_\alpha, b_\alpha) \cdot \mathrm{pdf}_{Gamma}(\beta; a_\beta, b_\beta) \tag{3.52}$$

where $\mathrm{pdf}_{Gamma}(\xi; a, b) = \frac{b^a \xi^{a-1} \mathrm{e}^{-\xi b}}{\Gamma(a)}$ and Γ is the gamma function, and given the variate (x, y) of $(X, Y)|\alpha, \beta$, the posterior distribution takes the analytic form

$$\begin{aligned} \pi^*(\alpha, \beta) = \Pr[\alpha, \beta|x, y] &= \frac{\Pr[x, y|\alpha\beta]\pi(\alpha, \beta)}{\int \Pr[x, y|\alpha\beta]\pi(\alpha, \beta)\mathrm{d}\alpha\mathrm{d}\beta} \\ &= \mathrm{pdf}_{Gamma}(\alpha; a_\alpha + x, b_\alpha + 1) \cdot \mathrm{pdf}_{Gamma}(\beta; a_\beta + y, b_\beta + 1). \end{aligned} \tag{3.53}$$

This convenient fact means that if we describe our knowledge of the references α and β by the hyperparameters $(a_\alpha, b_\alpha, a_\beta, b_\beta)$, then the hyperparameters describing the posterior are $(a_\alpha + x, b_\alpha + 1, a_\beta + y, b_\beta + 1)$. Note that the mean and variance of the gamma distribution $\mathrm{Gamma}(a, b)$ are given by $\mu = \frac{a}{b}$ and $\sigma^2 = \frac{a}{b^2}$, respectively. These equations can be uniquely inverted as $a = \frac{\mu^2}{\sigma^2}$ and $b = \frac{\mu}{\sigma^2}$. Therefore we can equivalently, but more intuitively, describe our prior with the hyperparameters $(\mu_\alpha, \sigma_\alpha^2, \mu_\beta, \sigma_\beta^2)$ which give posterior hyperparameters

$$\left(a_\alpha^* = x + \frac{\mu_\alpha^2}{\sigma_\alpha^2}, b_\alpha^* = 1 + \frac{\mu_\alpha}{\sigma_\alpha^2}, a_\beta^* = y + \frac{\mu_\beta^2}{\sigma_\beta^2}, b_\beta^* = 1 + \frac{\mu_\beta}{\sigma_\beta^2}, \right) \tag{3.54}$$

or, in terms of mean and variance,

$$\left(\mu_\alpha^* = \frac{\mu_\alpha^2 + \sigma_\alpha^2 x}{\mu_\alpha + \sigma_\alpha^2}, (\sigma_\alpha^*)^2 = \frac{\sigma_\alpha^2 (\mu_\alpha^2 + \sigma_\alpha^2 x)}{(\mu_\alpha + \sigma_\alpha^2)^2} \mu_\beta^* = \frac{\mu\beta^2 + \sigma\beta^2 y}{\mu\beta + \sigma\beta^2}, (\sigma_\beta^*)^2 = \frac{\sigma\beta^2 (\mu\beta^2 + \sigma\beta^2 y)}{(\mu\beta + \sigma\beta^2)^2} \right). \tag{3.55}$$

3.5.2 Correlated Conjugate Prior

The prior introduced in the previous subsection assumes that the parameters α and β are uncorrelated. This will rarely if ever be true in practice; a large positive correlation is expected. Therefore, in addition to the hyperparameters $(\mu_\alpha, \sigma_\alpha^2, \mu_\beta, \sigma_\beta^2)$ we would like to add a fifth hyperparameter, $\sigma_{\alpha,\beta}$, which describes the covariance of α and β.

To this end we consider a bivariate Poisson model inspired by Equation Equation 3.3. In that equation it is clear that variations in δ will cause correlations between α and β. A bivariate Poisson random variable is defined as $(A, B) \sim \mathrm{BP}(\theta_0, \theta_1, \theta_2)$ where $A = C_0 + C_1$ and $B = C_0 + C_2$ with $C_i \sim \mathrm{Poisson}(\theta_0)$, $i = 0, 1, 2$. This produces the probability density

$$\mathrm{pdf}_{BP}(x, y; \theta_0, \theta_1, \theta_2) = \frac{\mathrm{e}^{-(\theta_0+\theta_1+\theta_2)}\theta_1^x\theta_2^y}{x!y!} \sum_{i=0}^{\min(x,y)} \binom{x}{i}\binom{y}{i} i! \left(\frac{\theta_0}{\theta_1\theta_2}\right)^i. \tag{3.56}$$

This distribution has marginal distributions

$$A \sim \mathrm{Poisson}(\theta_0 + \theta_1) \quad \text{and} \quad \sim \mathrm{Poisson}(\theta_0 + \theta_1) \tag{3.57}$$

and covariance $\mathrm{Cov}(A, B) = \theta_0$. As discovered by Karlis and Tsiamyrtzis [98], it has an exact family of conjugate priors given by the mixture distributions

$$\sum_{j=0}^{r} w_j G(\theta_0; a_0 + j, b_0) \cdot G(\theta_1; a_1 - j, b_1) \cdot G(\theta_2; a_2 - j, b_2) \tag{3.58}$$

where G is the probability density of the gamma distribution, $a_i, b_i > 0$, $r \in \{0, 1, 2, ...\}$, $0 \le w_j \le 1$, and $\sum_{j=0}^{r} w_j = 1$.

For the prior that interests us, fixing $r = 0$ and $b_0 = 1$ will suffice; this leaves us with five hyperparameters $(a_0, a_1, a_2, b_0, b_1)$ which we will bijectively map onto the more intuitive hyperparameters $(\mu_\alpha, \sigma_\alpha^2, \mu_\beta, \sigma_\beta^2, \sigma_{\alpha,\beta})$.

Indeed, make the change of variables $\alpha = \theta_0 + \theta_1$ and $\beta = \theta_0 + \theta_2$ and consider the prior

$$\pi(\alpha, \beta, \theta_0) = G(\alpha - \theta_0; a_1, b_1) \cdot G(\beta - \theta_0; a_2, b_2) \cdot G(\theta_0; a_0, 1) \tag{3.59}$$

where

$$\begin{aligned} a_1 &= \frac{(\mu_\alpha - \sigma_{\alpha,\beta})^2}{\sigma_\alpha^2 - \sigma_{\alpha,\beta}} & b_1 &= \frac{\mu_\alpha - \sigma_{\alpha,\beta}}{\sigma_\alpha^2 - \sigma_{\alpha,\beta}} \\ a_2 &= \frac{(\mu_\beta - \sigma_{\alpha,\beta})^2}{\sigma_\beta^2 - \sigma_{\alpha,\beta}} & b_2 &= \frac{\mu_\beta - \sigma_{\alpha,\beta}}{\sigma_\beta^2 - \sigma_{\alpha,\beta}} \\ a_0 &= \sigma_{\alpha,\beta}. \end{aligned} \tag{3.60}$$

It then holds that, for example,

$$\begin{aligned}\mathbb{E}[\alpha^2] &= \int_0^\infty \mathrm{d}\theta_0 \int_{\theta_0}^\infty \mathrm{d}\alpha \int_{\theta_0}^\infty \mathrm{d}\beta \alpha^2 \pi(\alpha,\beta,\theta_0) \\ &= \int_0^\infty \mathrm{d}\theta_0 \int_0^\infty \mathrm{d}\theta_1 (\theta_0^2+\theta_1^2+2\theta_1\theta_3) G(\theta_1;a_1,b_1)\cdot G(\theta_0;a_0,1) \\ &= \left(\frac{a_0}{1^2}+\frac{a_0^2}{1^2}\right)+\left(\frac{a_1}{b_1^2}+\frac{a_1^2}{b_1^2}\right)+2\frac{a_1}{b_1}a_0 \\ &= \sigma_\alpha^2+\mu_\alpha^2 \end{aligned} \tag{3.61}$$

so that $\mathrm{Var}[\alpha]=\sigma_\alpha^2$. Similarly we get $\mathbb{E}[\alpha]=\mu_\alpha$, $\mathbb{E}[\beta]=\mu_\beta$, $\mathrm{Var}[\alpha]=\sigma_\alpha^2$, $\mathrm{Var}[\beta]=\sigma_\beta^2$, and $\mathrm{Cov}[\alpha,\beta]=\sigma_{\alpha,\beta}$.

Since this prior is conjugate to the likelihood function of

$$(X,Y)\sim \mathrm{BP}\left(\theta_0,\alpha-\theta_0,\beta-\theta_1\right), \tag{3.62}$$

if we receive the iid data $(x_1,y_1),...,(x_n,y_n)$ sampled from this distribution, the posterior will have the form of Equation Equation 3.58 with updated parameters. Let $x=\sum_{k=1}^n x_k$ and $y=\sum_{k=1}^n y_k$. First we have the simple updated posterior parameters $a_0^*=a_0$, $a_1^*=a_1+x$, $a_2^*=a_2+y$, $b_i^*=b_i+n$ for $i=0,1,2$, which is very similar to the uncorrelated case of the previous section. The new weights w_j^* of the posterior (recall we had a single weight $w_0=1$ in the prior) also has a closed form, but it is cumbersome to write down. Defining $s_i=\min(x_i,y_i)$ and $S_n=\sum_{i=1}^n s_i$, then for each $0\le k\le S_n$ we have $w_k=\bar{p}_k/\sum_{m=0}^{S_n}\bar{p}_k$ where

$$\bar{p}_k = c_k^{(n)}\frac{b_1^{a_1}b_2^{a_2}}{\Gamma(a_1)\Gamma(a_2)\Gamma(a_0)}\Gamma(a_1-k+x)\Gamma(a_2-k+y)\Gamma(a_0+k)\left(\frac{(n+b_1)(n+b_2)}{n+b_0}\right)^k. \tag{3.63}$$

and the quantities $c_k^{(n)}$ are defined recursively as

$$c_k^{(n)} = \sum_{r=\max(0,k-s_n^*)}^{\min(k,s_n^*)} v_r^{(n)} c_{k-r}^{(n-1)} \tag{3.64}$$

where $v_r^{(m)}=((x_m-r)!(y_m-r)!r!)^{-1}$, $c_k^{(1)}=v_k^{(1)}$, and $s_m^*=\min(s_m,S_{m-1})$. For generality, we have provided these formulas given n random samples, when in practice, because of the

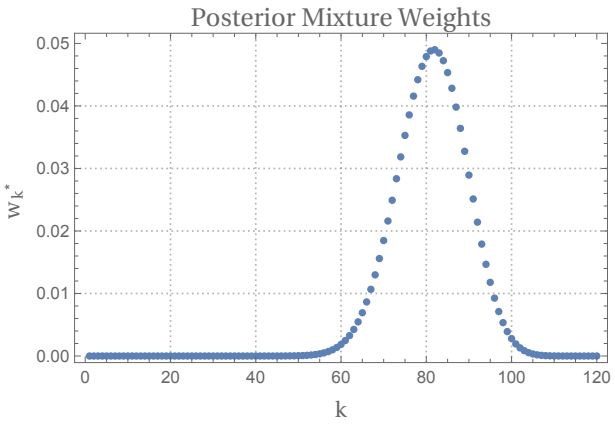

Figure 3.6: An example of posterior mixture weights w_k^* for $k = 0, ..., S_1$. Parameters used were $\mu_\alpha = 200$, $\mu_\beta = 140$, $\sigma_\alpha = 40$, $\sigma_\beta = 15$, $\sigma_{\alpha,\beta} = 90$. The prior was updated with a single sample $x = 220$ and $y = 120$.

drift discussed in the main body of this article, only one random sample will usually be taken. With $n = 1$, the recursive definition is unnecessary and the weights are given by

$$w_k^* = \frac{x!y!\sin(\pi(\alpha_1 + x))\sin(\pi(\alpha_2 + y))}{{}_3\tilde{F}_2\left(\{-x, -y, \alpha_0\}, \{1 - x - \alpha_1, 1 - y - \alpha_2\}, \frac{1}{2}(\beta_1 + 1)(\beta_2 + 1)\right)} \cdot \frac{\Gamma(k + \alpha_0)\Gamma(x - k + \alpha_1)\Gamma(y - k + \alpha_2)}{\pi^2 k!(x - k)!(y - k)!\,\Gamma(\alpha_0)} \left(\frac{(1 + b_1)(1 + b_2)}{2}\right)^k \tag{3.65}$$

where ${}_p\tilde{F}_q(\{c_1, ..., c_p\}, \{d_1, ..., d_q\}, z) = \frac{{}_pF_q(\{c_1,...,c_p\},\{d_1,...,d_q\},z)}{\Gamma(d_1)\cdots\Gamma(d_q)}$ is the regularized generalized hypergeometric function and ${}_pF_q(\{c_1, ..., c_p\}, \{d_1, ..., d_q\}, z)$ is the generalized hypergeometric function. However, the only factors in Equation 3.65 which are are relevant to numerical implementations are those which involve k. Everything else can be implicitly calculated by demanding the normalization $\sum_{k=0}^{S_n} w_k^* = 1$. An example of posterior mixture weights is shown in Figure 3.6.

With these definitions, the posterior is given by the exact distribution

$$\pi^*(\alpha, \beta, \theta_0) = \sum_{j=0}^{S_n} w_j^* \big[G(\theta_0; a_0 + j, b_0 + n) \cdot G(\alpha - \theta_0; a_1 + x - j, b_1 + n) \cdot G(\beta - \theta_0; a_2 + y - j, b_2 + n)\big]. \tag{3.66}$$

Depending on the size of S_n, the posterior may be expensive to compute the value of at a given coordinate. However, once the posterior weights ω_k^* have been computed, drawing a sample is essentially the same cost as drawing a sample from three Gamma distributions.

3.6 Example: Quantum Hamiltonian Learning

One of the primary advantages of using a Bayesian approach to NV measurement is that it can be used as an overlay model on other estimation problems. This results in seamless propagation of error bars to the final quantities of interest. In frequentist settings, it is usually a pain to justifiably propagate error bars near the boundaries of an interval like $[0, 1]$ because the usual normality approximations are dubious. To illustrate this Bayesian approach, in this section we provide a thorough example of Quantum Hamiltonian Learning (QHL) [185] using experimental data. QHL has been shown to be a powerful method of characterizing quantum systems [182].

3.6.1 Hamiltonian Model

We assume a secular form of the spin-1 Hamiltonian (in the optical ground state, see Section 2.3) of the form

$$H_{m_{\mathbf{I}}} = \delta\Delta\mathbf{S}_z^2 + (\omega_e + m_{\mathbf{I}}A_N)\mathbf{S}_z + \Omega(t)\mathbf{S}_x \tag{3.67}$$

where we are in a frame rotating near the ground state ZFS, 2.87 GHz. Here, $\delta\Delta$ is the mismatch between the applied microwave frequency and the ZFS, ω_e is the projection of the static magnetic field onto the z-axis, A_N is the hyperfine splitting due to 14-Nitrogen, $m_{\mathbf{I}}$ is the spin number of the nitrogen atom, and Ω is the microwave nutation strength. All of the parameter units are angular frequency, $2\pi\cdot$MHz, except $m_{\mathbf{I}}$ which is unitless. Our goal is to learn these parameters, as well as the T_2 decay time. In particular, we would like good error bars on ω_e since this is the quantity of interest in magnetometry, considering other parameters as nuisances.

At room temperature, the nitrogen atom is equally likely to be in each of its three energy states $|m_{\mathbf{I}} \in \{-1, 0, +1\}\rangle$. Since the nitrogen T_1 is much longer than a single experiment, we assume that for each experiment, the state of the nitrogen is fixed. Hence in the Hamiltonian above, we simply treat the axial hyperfine coupling between the $NV^-$and the nitrogen as a small $m_{\mathbf{I}}$ dependent shift in the static magnetic field.

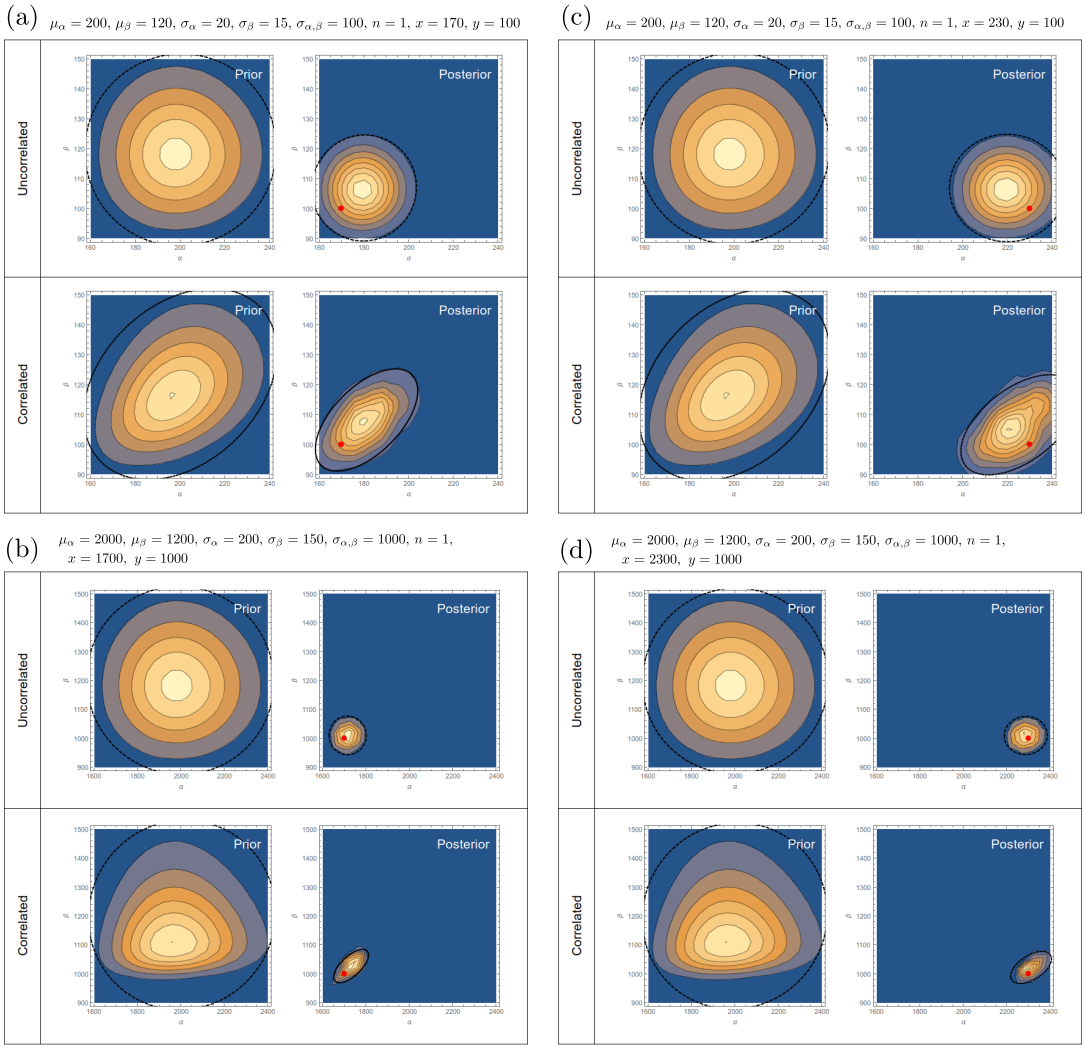

Figure 3.7: Examples of Bayesian updates showing contours of the prior and posterior probability distributions, where the update data is depicted by red dots. In each of the four cases, the same data is given to both the correlated and uncorrelated priors described in Section 3.5.1 and Section 3.5.2, respectively. Black dashed ellipses represent 90% confidence regions; their centers are at the mean of the distribution, and their eccentricity matrix is equal to 4.6 times the covariance matrix of the distribution. Cases (a) and (b) represent a low data scenario, whereas cases (c) and (d) represent a high data scenario. Cases (a) and (c) represent correlated measurement data, whereas cases (b) and (d) represent anti-correlated measurement data.

With an initial state $\rho_i = |\mathrm{g},0\rangle\langle\mathrm{g},0|$, at time t the density matrix is described by $\rho(t) = \frac{1}{3}\sum_{m_\mathbf{I}} \mathcal{S}(t)_{m_\mathbf{I}}[\rho_i]$ where $S_{m_\mathbf{I}}(t)$ is the evolution superoperator under the Hamiltonian $H_\mathbf{I}$, along with a single dephasing Lindblad term $L = \sqrt{T_2}\mathbf{S}_\mathrm{z}$. The superoperators $\mathcal{S}_{m_\mathbf{I}}$ can be computed by exponentiating the supergenerator derived from the Lindblad master equation. We will only consider constant or piecewise constant values of $\Omega(t)$, which simplifies simulation (finite rise-times are ignored). This convex combination approach is valid because we will be summing over many trials of the same experiment, and hence will see an equal mixture of all three nitrogen states on average.

3.6.2 Experiment Choices and Data

The time dependence of the nutation envelope Ω is controlled by the experimentalist. To learn the parameters of this system, we choose to do two types of experiments. The first is the Rabi experiment, where Ω is finite and constant for a period t_r and the state is subsequently measured. Rabi experiments are primarily sensitive to nutation frequency. The second is the Ramsey experiment, where there is a wait period with $\Omega = 0$ of length t_w between two identical unitary gates which are created by turning Ω on at full power for a duration t_p. Ramsey experiments are sensitive to fields along the z-axis. These two experiments are sensitive to roughly orthogonal regions of parameter space.

These considerations are only heuristics, and due to the continuum of possible experiment types, it is intractable to discover which ones are globally optimal for learning some set of parameters. Some improvement in learning efficiency could be almost certainly be achieved by considering additional pulse sequences which might further decorrelate and/or isolate parameters of interest. However, such considerations are beyond the scope of the present example where our goal is a proof of principle of fully Bayesian methods.

Experiments were performed on a microscope with relatively poor optical characteristics; for a single repetition ($N = 1$) we had an average number of detected bright reference photons $\alpha \approx 0.006$, and an average number of detected dark reference photos $\beta \approx 0.004$, giving a contrast value of 0.2. The static magnetic field acting on the NV^-center was just the ambient stray field in the laboratory; some combination of Earth's field, building characteristics, and nearby electronics. We took $R = 400$ averages of $N = 30000$ repetitions for both the Rabi and Ramsey experiments. Rabi flops were sampled at 100 linearly spaced points between $t_r = 8$ns and $t_r = 800$ns. The Ramsey wait times were sampled at 200 linear spaced points between $t_w = 0.01\mu$s and $t_w = 2\mu$s, with a pulse time $t_p = 44$ns. Raw (summed) data is plotted in Figure 3.10.

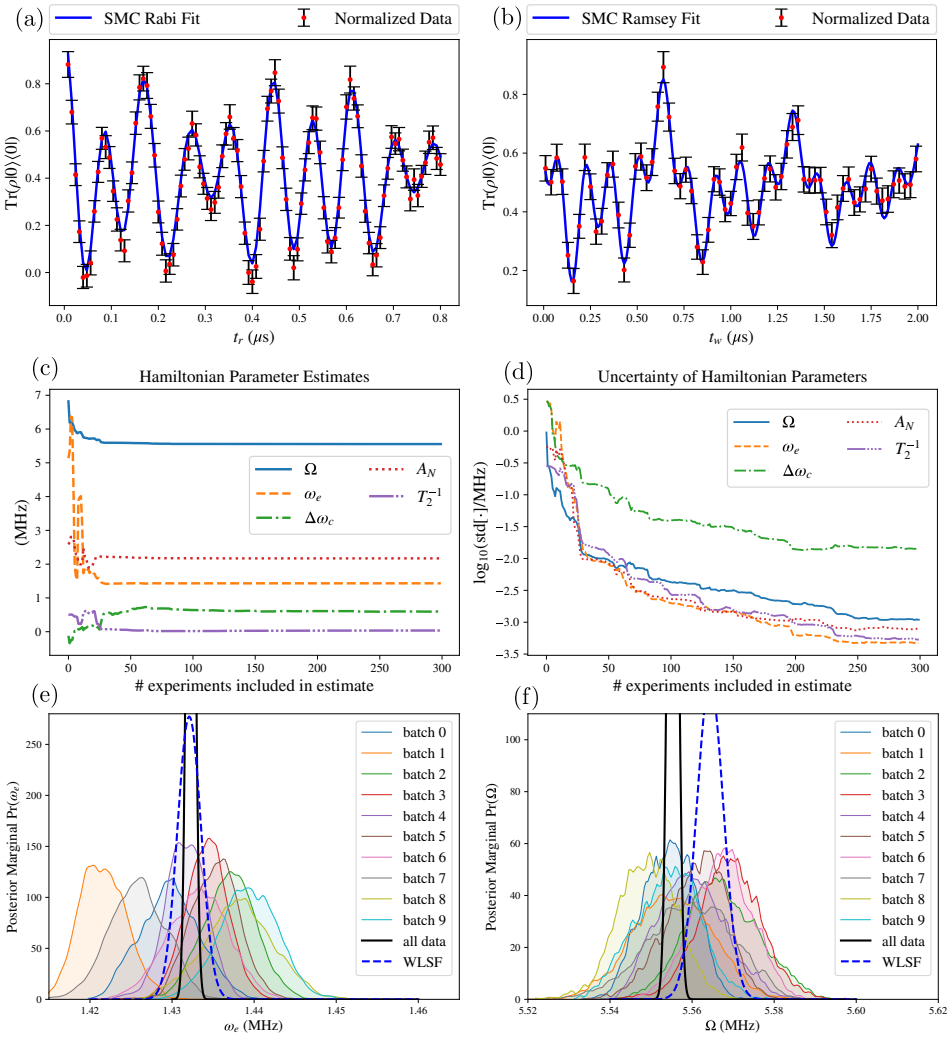

Figure 3.8: In (a-b), the expectation value of the SMC posterior, $\mathbb{E}_{\pi^*}[\vec{x}] = (\hat{\omega}_e, \delta\hat{\Delta}, \hat{\Omega}, \hat{A}_N, \hat{T}_2^{-1})$, is is used in a simulation of the Hamiltonian model (Section 3.6.1), and shown on top of the normalized raw data. The raw data was normalized using the MLE in Equation Equation 3.24, and the 95% error bars are computed with Equation Equation 3.21 for comparison. In (c-d), the expectation and variance of the SMC posterior is shown as a function of the number of Bayes' update steps in SMC. In (e) and (f) posterior marginal distributions are shown for the parameters ω_e and Ω, respectively. The broad shaded curves come from the same data-processing algorithm run on disjoint subsets of experimental data. The full amalgamated dataset results are shown for both SMC (black, solid) and weighted least-squares (blue, dashed).

3.6.3 QHL Likelihood Function

As we did for the tomography model sketched in Section 3.3.5, we now write down a model for our QHL problem. We label a given experimental configuration as $\vec{c} = (t_r, t_w, t_p, k)$, where the three timing parameters were defined in Section 3.6.1 and $k \in \{\text{RABI}, \text{RAMSEY}\}$ selects the experiment types explained in Section 3.6.2. Similarly, we denote a hypothetical parameter set as a vector $\vec{x} = (\omega_e, \delta\Delta, \Omega, A_N, T_2^{-1})$. A specific pair, $(\vec{x}, \vec{c})$, provides enough information to do a full quantum simulation of the spin-1 manifold, resulting in the probability of a projective $|0\rangle$ measurement given by

$$p_{\vec{x},\vec{c}} = \text{Tr}\left(|0\rangle\langle 0|\, \mathcal{S}_{\vec{x},\vec{c}}(|0\rangle\langle 0|)\right). \tag{3.68}$$

Here, $\mathcal{S}_{\vec{x},\vec{c}}$ is the solution to the Lindblad master equation under the Hamiltonian model described in Section 3.6.1. This yields the conditional model $X, Y, Z|\alpha, \beta, \vec{x}; \vec{c}$ with $X \sim \text{Poisson}(\alpha)$, $Y \sim \text{Poisson}(\beta)$, and $Z \sim \text{Poisson}(\beta + p_{\vec{x},\vec{c}}(\alpha - \beta))$. The goal of the inference problem is to deduce the true values of $\vec{x}$, and in particular, ω_e, given a dataset of photon counts.

3.6.4 Bayesian Inference with Sequential Monte Carlo

We begin with a prior distribution $\pi(\vec{x})$ describing our knowledge of the system before any measurements, given by the following product distribution:

$$\omega_e \sim \text{Unif}(0\,\text{MHz}, 10\,\text{MHz}) \tag{3.69}$$
$$\delta\Delta \sim \text{Unif}(-5\,\text{MHz}, 5\,\text{MHz}) \tag{3.70}$$
$$\Omega \sim \text{Unif}(0\,\text{MHz}, 10\,\text{MHz}) \tag{3.71}$$
$$A_N \sim \text{Unif}(1.5\,\text{MHz}, 3.5\,\text{MHz}) \tag{3.72}$$
$$T_2^{-1} \sim \text{Unif}\left((100\,\mu\text{s})^{-1}, (1\,\mu\text{s})^{-1}\right). \tag{3.73}$$

Additionally, we empirically choose a prior for the references α and β by computing the sample moments of the experimental reference count data, multiplying the standard deviations by 4 to be conservative, and choosing a product gamma distribution with these moments. This distribution is discussed in Section 3.5.1. We label this distribution as $\pi_k(\alpha, \beta)$ where $k \in \{\text{RABI}, \text{RAMSEY}\}$.

The prior distribution is now sequentially updated through Bayes' law one triple $(x_{\vec{c}}, y_{\vec{c}}, z_{\vec{c}})$ at a time. Here, $x_{\vec{c}} = \sum_{n=1,r=1}^{N,R} x_{n,\vec{c},r}$, $y_{\vec{c}} = \sum_{n=1,r=1}^{N,R} y_{n,\vec{c},r}$, and $z_{\vec{c}} = \sum_{n=1,r=1}^{N,R} z_{n,\vec{c},r}$ are the

photon counts for a particular experiment $\vec{c} = (t_r, t_w, t_p, k)$ summed over all repetitions and averages (see Section 3.2.2). The distribution $\Pr(\alpha, \beta, \vec{x})$ is stored as a so-called *particle approximation* consisting of a finite list of hypothetical values, called particles, labeled as $\{(\alpha_i, \beta_i, \vec{x}_i)\}$ with corresponding weights $\{w_i\}$, $\sum_i w_i = 1$. Typically on the order of 10000 particles are used for numerical stability.

The distribution for the references α and β is reset to $\pi_k(\alpha, \beta)$ before each triple of data is used. As discussed in Section 3.4.4, we may process the reference pair $(x_{\vec{c}}, y_{\vec{c}})$ first and subsequently process the signal count $z_{\vec{c}}$. The reference pair is processed by replacing the reference prior with the analytically derived posterior, discussed in Section 3.5.1. This is done by replacing the (α, β) coordinates of each particle $(\alpha_i, \beta_i, \vec{x}_i)$ with a random variate drawn from the posterior $\pi^*(\alpha, \beta | x_{\vec{c}}, y_{\vec{c}})$. The signal is processed using the Sequential Monte Carlo (SMC) algorithm, as implemented by the software package QInfer [63].

While Bayes' update rule is agnostic to the order in which we enter data since each data triple is statistically independent, the order is relevant to the numerical implementation. An intuitive explanation is as follows. Suppose one is interested in determining the frequency of a cosine wave using amplitude data sampled at various time points, and that we assign a flat prior distribution over a wide range of frequencies. If we first update our prior with the data from a late time point, the posterior will have many peaks because every divisor of the measurement time will correspond to a period consistent with the observed data. Subsequent updates will eventually inform us about which peak contains the true frequency. However, if we first update our prior with data from a time point early enough that we know (according to our prior) that less than a full period has had time to take place, then the posterior will be a very broad but unimodal. Subsequent chronological Bayes updates will tend to shift and narrow this peak. Since the SMC algorithm — specifically, the Liu–West resampler [113] — implicitly assumes unimodality, the first approach will usually fail and the second approach will usually succeed, assuming the ansatz that there is only one true value in parameter space. Given this data processing constraint, we fed the data to the Bayes updater in strictly increasing times t_w and t_r, shuffling the Rabi and Ramsey data together randomly. Alternatively, an algorithm without a unimodality constraint could be considered [64].

One nice feature of the SMC algorithm is that it typically heralds its own failure through the effective sample size criterion [11]. Such failures can result from multi-modalities, as discussed above. Another common failure path is through overly-informative data, where a single Bayes update causes only a handful of particles to remain relevant. We mitigate against this partly by using a conjugate prior for the reference indices, as discussed in Section 3.4.4, and also by a technique called *bridging the transition*, discussed in the next subsection.

3.6.5 Bridged Updater for the Referenced Poisson Model in SMC

For us, there are two main mechanisms that can cause the SMC algorithm to become unreliable. Both have to do with the finite particle approximation and its reliance on having enough effective particles near the true model parameter values, known as importance sampling. The first mechanism is that periodicities in the model have the effect of creating temporary multimodalities in the posterior as we analyze the data. If certain early data happen to cause disproportionate support on an incorrect mode, we lose particles where we need them, and this might lead to a runaway effect where all particle weights become zero. We mitigate against this by processing the data in ascending order, as discussed the previous subsection.

The second cause of instability is that some data points are overly informative. From a learning perspective, informative data are great. However, from SMC's perspective, very informative data have the tendency to drastically reduce the weight of most particles, causing the effective particle count,

$$n_{\text{eff}} = 1/\sum_{n=1}^{N} w_n^2, \tag{3.74}$$

to become a tiny fraction of the actual particle count, N. For example, it is not unusual (nor is it common) for data, with our QHL model, to reduce the effective particle count to below 100 while there are 16000 actual particles. Storing a multi-dimensional distribution (seven dimensions in our current QHL example) on 100 particles is a bad idea, and causes the remainder of the inference to become suspect. A costly solution to this problem is to, for example, increase to 160000 particles, so that the effective particle count does not dip much below 1000. The purpose of this section is to discuss a less costly solution.

The idea is as follows. It is easy to detect when an update *would* cause the effective particle count to drop below some threshold, say 1000; one simply need compute Equation 3.74 on the posterior weights before overwriting the current weights. If this flag is raised, instead of performing the update (i.e. overwriting the current weights), we instead perform a sequence of less informative updates that do not correspond to actual observations, but that do result in the same posterior. This technique is called *bridging the transition* [2], and it gives a chance for the updater to resample the particles "mid-update" so that particles can be relocated to where they are actually needed. This technique is only amenable to certain likelihood functions, and thankfully, owing to some nice properties of the Poisson distribution, ours is one of them.

To see this, consider the generic NV^- model at some particular step of SMC:

$$s \sim \pi(s) \tag{3.75a}$$
$$\alpha, \beta \sim \pi(\alpha, \beta) \tag{3.75b}$$
$$X|\alpha \sim \text{Poisson}\,(\alpha) \tag{3.75c}$$
$$Y|\beta \sim \text{Poisson}\,(\beta) \tag{3.75d}$$
$$Z|s, \alpha, \beta \sim \text{Poisson}\,(p(s)\alpha + (1 - p(s))\beta)\,. \tag{3.75e}$$

Here, s is a set of parameters we wish to learn, $\pi(s)$ is our current particle distribution describing these parameters, $\pi(\alpha, \beta)$ is our prior on the next data point's references, $p(s)$ is the function that takes the parameters of interest and returns $p = \operatorname{Tr} \rho P_0$, and (X, Y, Z) is the next data point triplet. On obtaining the variate (x, y, z), the next SMC update step should result in the posterior distribution $\Pr(s, \alpha, \beta|x, y, z)$, which, using Bayes' rule, is proportional to

$$\Pr(s, \alpha, \beta|x, y, z) \propto \Pr(x, y, z|s, \alpha, \beta)\pi(\alpha, \beta)\pi(s) \tag{3.76a}$$
$$= \Pr(z|s, \alpha, \beta) \Pr(x|\alpha) \Pr(y|\beta)\pi(\alpha, \beta)\pi(s) \tag{3.76b}$$
$$\propto \Pr(z|s, \alpha, \beta)\pi^*_{x,y}(\alpha, \beta)\pi(s). \tag{3.76c}$$

As noted in Section 3.4.4, this formula shows that our update consists of first updating α and β analytically using a conjugate prior, and subsequently updating s, α, β using $\pi^*_{x,y}(\alpha, \beta)\pi(s)$ as a prior. The proportionalities allow us to neglect those factors which depend only on x, y, and/or z but not any of s, α, or β; all particles see the same values of x, y, and z, so any such factors will be canceled out when enforcing the normalization condition of the particle weights.

To bridge this transition, we start with the particles in the state $\pi^*_{x,y}(\alpha, \beta)\pi(s)$ and notice that

$$\Pr(z|s, \alpha, \beta) \propto (p(s)\alpha + (1 - p(s))\beta)^z e^{-p(s)\alpha - (1-p(s))\beta} \tag{3.77a}$$
$$= \left(p(s)\frac{\alpha}{m} + (1 - p(s))\frac{\beta}{m}\right)^{m\frac{z}{m}} m^z e^{-mp(s)\frac{\alpha}{m} - m(1-p(s))\frac{\beta}{m}} \tag{3.77b}$$
$$\propto \text{pdf}_{\text{Pois}}\left(\frac{z}{m}; p(s)\frac{\alpha}{m} + (1 - p(s))\frac{\beta}{m}\right)^m \tag{3.77c}$$

so that the update can be achieved instead with m Poisson updates with reduced count data z/m and reference particle locations for α and β also reduced by m. This is illustrated in Figure 3.9. Note that our fictional data can have fractional photon counts; this is

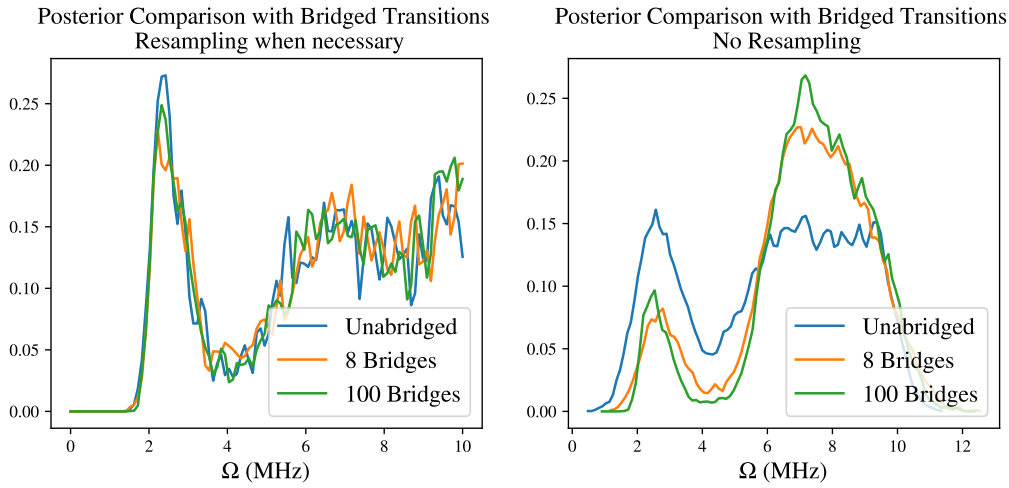

Figure 3.9: A particle distribution was initialized to the prior of Equation 3.73 with 16000 particles, and separately updated with the data from a single point of a Rabi experiment in six different ways. We show a slice through the posterior for each case. On the left are bridged and un-bridged updates with no resamples allowed, The final effective particle count was about 1800 for all three of these updates. This demonstrates the bridging technique works in practice. On the right are bridged and un-bridged updates with resamples taken whenever the distribution was detected to have fewer than 8000 effective particles. These two bridge cases maintained at least 8000 effective particles at all times. Since the posterior is far from normal, we can expect the resampler to introduce distortions.

not a problem, the Poisson mass function is well defined for non-integer data, assuming the factorial is implemented using the gamma function, which it will be in any modern programming language.

3.6.6 Results and Validation

Bayesian inference with the SMC algorithm was run on the entire dataset to obtain a posterior distribution (two-parameter marginals are plotted in Figure 3.11). Recall that our entire dataset consists of 400 averages of 30000 repetitions for each of the 300 different experimental configurations, corresponding to roughly 24 hours of experiment time given our particular optical efficiency. This number of averages was chosen to be large to allow for more convincing validation of our techniques. To this end, the 400 averages were divided into ten disjoint and chronological batches of 40 averages each, and the SMC algorithm

was run independently on each batch. Since each batch has strictly less data than the entire dataset (effectively lower values of α and β), wider posteriors are expected for these than for the entire data set.

The main results are shown in Figure 3.8 and Table 3.1. The top two plots, Figure 3.8(a-b), show that the SMC posterior corresponds to a sensible traditional fit of the data; the posterior is used to obtain a point estimate of each of the parameters, and these parameters are then used in a simulation spanning the experimental configurations. Since the posterior distribution is tight enough that simulations from randomly sampled values are visually indistinguishable, these fits can be interpreted as a visual posterior predictive check, where data simulated according to the posterior is compared with actual data [58]. The middle two figures show convergence properties of the SMC algorithm. Finally, the bottom two figures show that the disjoint data sets result in posteriors that are consistent with each other, and consistent with the posterior of the amalgamated data set. Keep in mind that the parameters could be fluctuating slightly over long time scales.

For comparison, we also analyzed the data using a weighted least-squares fit (WLSF) of of the parameters $\vec{x} = (\omega_e, \delta\Delta, \Omega, A_N, T_2^{-1})$. This was done by using the SciPy [97] function optimization. curve_fit to minimize the quantity

$$\Phi(\vec{x}) = \sum_{\vec{c}} \left(\frac{\hat{p}_{\vec{c}} - p_{\vec{x},\vec{c}}}{\sigma_{\vec{c}}} \right)^2$$
$$\hat{p}_{\vec{c}} = \frac{z_{\vec{c}} - y_{\vec{c}}}{x_{\vec{c}} - y_{\vec{c}}}, \quad \sigma_{\vec{c}}^2 = \frac{\hat{p}_{\vec{c}}(\hat{p}_{\vec{c}} + 1)x_{\vec{c}} + (\hat{p}_{\vec{c}} - 2)(\hat{p}_{\vec{c}} - 1)y_{\vec{c}}}{(x_{\vec{c}} - y_{\vec{c}})^2} \tag{3.78}$$

where the sum is taken over all experiment configurations $\vec{c} = (t_r, t_w, t_p, k)$ that were performed, $p_{\vec{x},\vec{c}}$ is the simulation of hypothesis $\vec{x}$ under conditions $\vec{c}$ defined in Equation 3.68, $\hat{p}_{\vec{c}}$ is the MLE of p given the data $(x_{\vec{c}}, y_{\vec{c}}, z_{\vec{c}})$, and the formula for the estimated variance $\sigma_{\vec{c}}^2$ is derived from the Cramér–Rao bound Equation 3.20. For simplicity, the initial guess of the WLSF function was taken to be the SMC point estimate. The WLSQ fit is shown in Figure 3.8(e-f) on top of the SMC marginal posteriors. Table 3.1 provides a more comprehensive comparison, where WLSQ fits are also performed on each of the ten batches. Visual checks for each batch are plotted in Figure 3.12. For the smaller batch sizes, the WLSQ confidence intervals are more comparable in size to the SMC credible regions. We suspect this is because SMC did not have enough data to significantly reduce posterior correlations between parameters, especially between $\delta\Delta$ and ω_e.

		ω_e		$\delta\Delta$		Ω		A_N		T_2^{-1}	
		$\mathbb{E}$	σ	$\mathbb{E}$	σ	$\mathbb{E}$	σ	$\mathbb{E}$	σ	$\mathbb{E}$	σ
all data	SMC	1432	0.5	597	13.9	5555	1.1	2171	0.8	35	0.5
	WLSF	1432	1.4	158	54.7	5564	3.3	2169	1.8	43	1.7
batch 0	SMC	1430	3.6	639	115.0	5554	6.2	2181	4.7	44	2.8
	WLSF	1424	4.0	773	127.8	5551	9.5	2176	5.2	49	4.8
batch 1	SMC	1422	2.8	-463	211.0	5555	9.6	2172	3.1	40	3.6
	WLSF	1425	3.5	-222	172.7	5562	8.4	2171	4.6	41	4.0
batch 2	SMC	1437	3.2	-323	279.6	5567	8.9	2163	4.9	48	4.2
	WLSF	1433	4.0	-188	153.3	5566	8.6	2164	4.9	49	4.5
batch 3	SMC	1434	2.6	30	247.2	5569	7.6	2166	3.3	31	2.0
	WLSF	1433	3.3	153	140.4	5569	8.3	2164	4.3	36	3.9
batch 4	SMC	1432	2.5	-120	233.3	5560	8.0	2173	4.2	40	2.3
	WLSF	1431	3.6	-132	133.3	5563	7.8	2175	4.8	46	4.3
batch 5	SMC	1435	2.9	125	147.8	5561	7.3	2177	4.2	40	2.7
	WLSF	1435	3.7	163	140.1	5566	8.4	2178	4.8	45	4.2
batch 6	SMC	1433	3.7	-41	248.9	5567	7.1	2169	4.5	44	3.7
	WLSF	1433	3.5	0	6962.6	5571	8.1	2166	4.5	44	4.0
batch 7	SMC	1426	3.4	390	499.5	5559	10.4	2168	4.3	40	3.3
	WLSF	1429	3.3	410	127.2	5562	7.8	2165	4.1	40	3.8
batch 8	SMC	1438	4.3	203	348.1	5549	7.2	2168	3.8	41	2.9
	WLSF	1440	3.7	511	144.9	5550	8.9	2168	4.6	43	4.3
batch 9	SMC	1439	3.5	226	650.5	5554	7.8	2167	3.6	45	2.9
	WLSF	1439	3.9	538	140.1	5554	8.8	2169	4.8	46	4.3

Table 3.1: Results of the SMC and WLSF fits are shown for for the entire data sat, as well as for each batch. All units are kHz. Both the point estimate ($\mathbb{E}$) and the marginal standard deviation (σ) are displayed for each of the five fit parameters, $(\Omega, \omega_e, \delta\Delta, A_N, T_2^{-1})$.

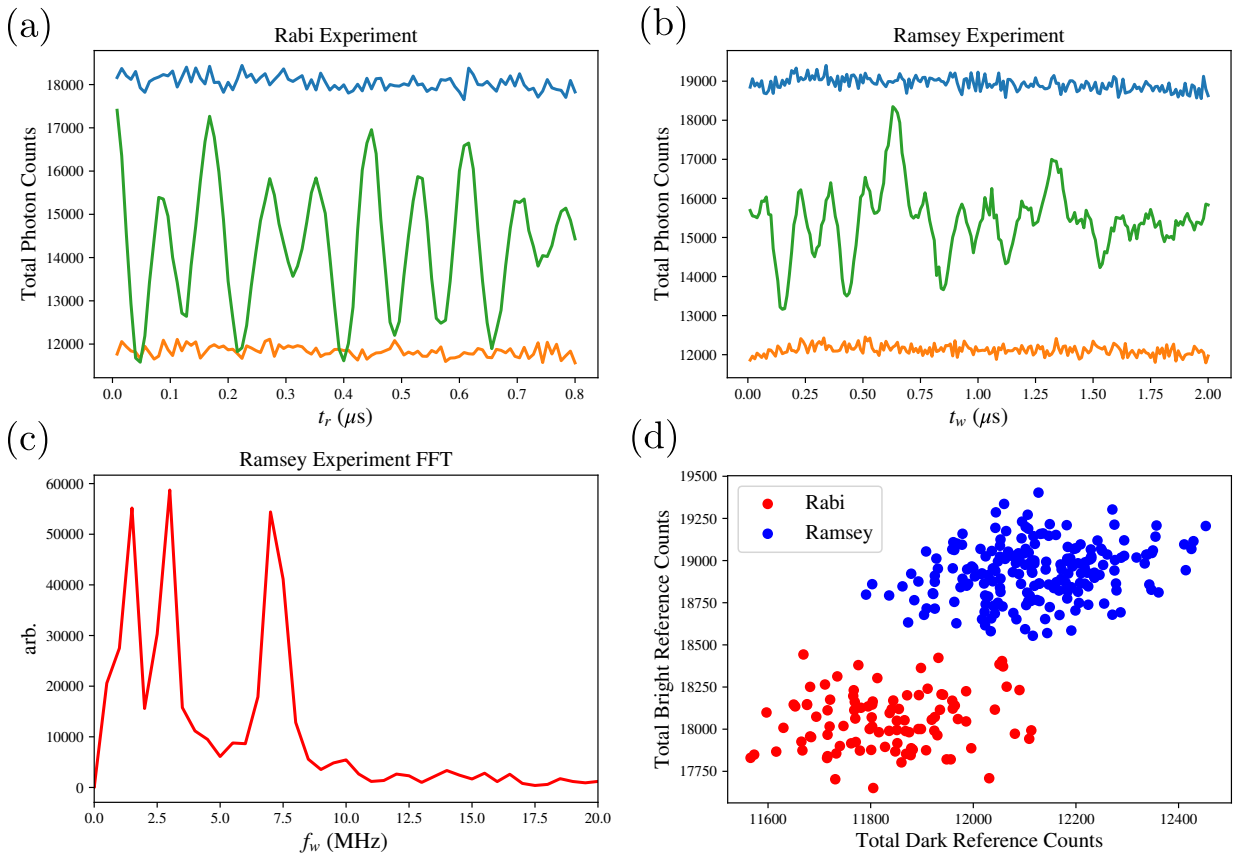

Figure 3.10: (a),(b) Time domain data from Rabi and Ramsey experiments. Photon counts are summed over all 400×30000 repetitions at each experiment parameter on the x-axis. Bright and dark references are shown in addition to the signal of interest. (c) The discrete Fourier transform of the Ramsey experiment. (d) Scatter plot of the summed reference counts for both experiments. Each point represents a different experiment configuration, the discrepancy between distributions is due to performing the experiments on different days of the week.

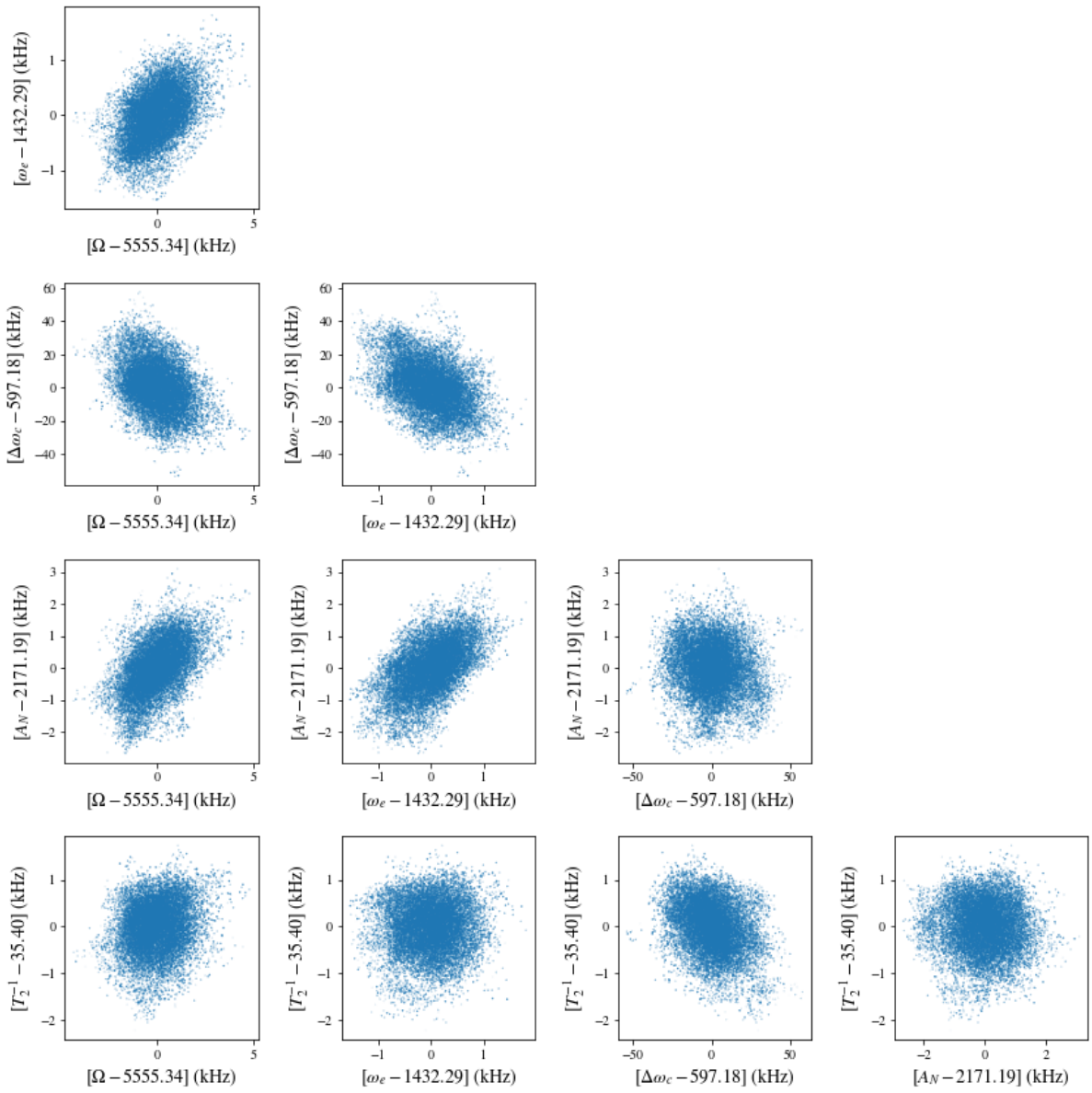

Figure 3.11: Two-parameter marginals of the QHL posterior distribution, where each dot is a member of the particle approximation projected onto the corresponding axes. The plots have been centered around the mean value of the distribution, the components of which are specified in the axis labels.

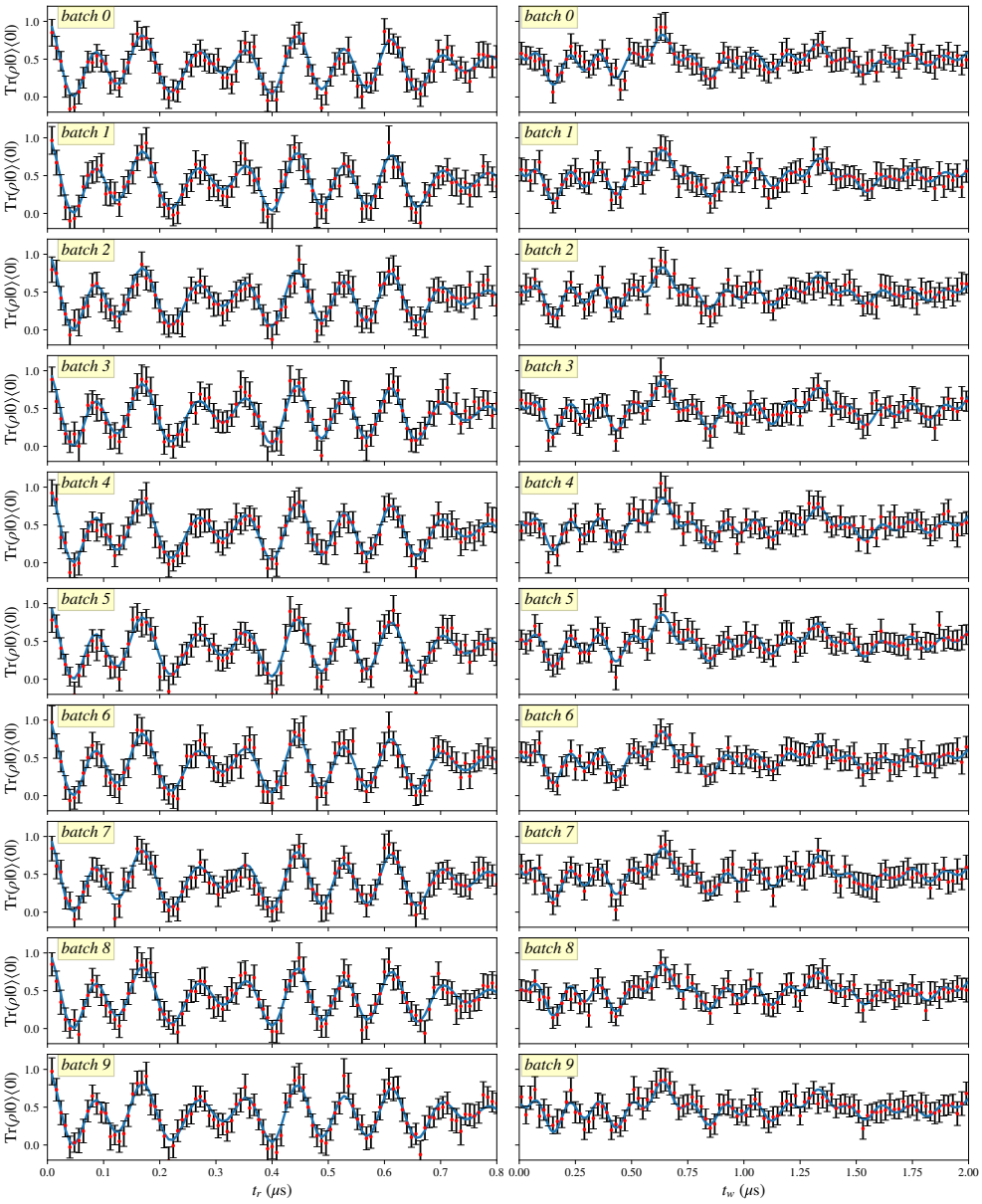

Figure 3.12: Fits to the data from each of the 10 batches of 40 averages. The left column contains the Rabi experiments, and the right column contains the Ramsey experiments. The points are the normalized data used in the corresponding SMC algorithm, with error bars calculated using Equation 3.20.

Chapter 4

Improving Hamiltonian Estimates with Adaptive Experiment Design[1]

4.1 Introduction

Characterizing quantum devices efficiently is an increasingly important problem. In the case of quantum processors, knowing system properties and error processes is helpful for designing robust high-fidelity control. If system parameters drift in time, they will need to be periodically recharacterized, which reduces uptime. Or, in the case of metrology, certain properties of the quantum system are themselves the quantities of interest, and so more efficient characterization leads to higher sensitivities.

Quantum system characterization is typically done by performing a set of predetermined experiments and subsequently processing statistics of the resulting data. This was the approach taken in the Hamiltonian learning example of the previous chapter. While there is nothing wrong with this—and indeed, in some cases, this strategy can even be tuned to have near optimal performance—it has long been known that *online* (also called *adaptive*

[1]This chapter is based on Reference [83], done in collaboration with Thomas Alexander, Michal Kononenko, and Benjamin Soloway. Experiments and analysis were performed by myself, with some consultation from Thomas Alexander. Thomas Alexander was responsible for help with the Bayes risk numerical implementation, see Appendix A 'nv-adaptive'. Michal Konenenko was responible for setting up a server to communicate experiments between the experiment-choosing computer with the experiment-performing computer. Benjamin Soloway helped with the theory of effective strong measurements. The third paragraph of this section puts this work in the context of adaptive quantum literature, both theoretical and experimental—in particular, this reference contains the first adaptive quantum Hamiltonian learning experiments that have been performed.

by some authors) experiment design is generally capable of outperforming predetermined experiment sweeps [23, 79]. As its name implies, online experiment design allows the next experiment choice to depend somehow on what has already been learned. The reason for the advantage is obvious—online experiments can potentially avoid executing experiments that are expected to be uninformative by using information that was initially unavailable.

Online experiment design has a long history in quantum systems. Almost five decades ago, it was used to reduce the time required to determine relaxation rates in NMR spin systems [56], and later to speed up inversion recovery T_1 measurements [169]. In recent decades, it has been studied extensively, both in theory and experiment, in the context of quantum phase estimation [187, 10, 80, 9, 79, 191, 195, 26] and quantum state tomography [92, 106, 50, 165, 167, 61, 143]. Online experiment design has been suggested for sequence length choices in randomized benchmarking experiments[62], and adaptive protocols to generate control pulses for quantum systems have been proposed [44, 52, 147]. Here we build on online experiment design applied to quantum Hamiltonian estimation [155, 66, 51, 184, 166, 164], where a Hamiltonian form (or set of forms) is specified, and unknown coefficients of Hamiltonian terms are sought.

The purpose of this chapter is to study online Bayesian experiment design, with Hamiltonian estimation as the inference problem of choice, using experimental data and noise on a system with slightly non-trivial dynamics. By non-trivial we mean that there are more than one or two relevant inference parameters (we ultimately use 10, including nuisance parameters describing optical drift), that quantum state evolution does not admit a nice closed form solution, and that we allow the ability to turn on and off the control field within an experiment. In doing so we hope to pave the way for similar experiments in yet more complex systems. To this end we interface a sequential Bayesian inference engine with an experimental setup that controls the qutrit manifold of a single Nitrogen Vacancy (NV) defect in diamond.

4.2 Inference of Quantum Devices

We begin by defining some notation while reviewing parameter estimation as applied to quantum devices.

Information about a quantum device can be encoded into a list of real values, which we call *model parameters*, labeled x. For example, in the case of Hamiltonian learning, these values parameterize the Hamiltonian operator of the quantum system, or in the case of state tomography, the entries of a density operator. This set of parameters includes both

parameters of interest, which one is interested in learning, and nuisance parameters, which are not of principle interest, but are still necessary to sufficiently describe the system.

Quantum devices are controlled by some collection of classical knobs that adjust various settings such as power, timings, carrier frequencies, and so on. We refer to a specific assignment of all of these settings as an *experiment configuration*, sometimes called the control variables, which we label e. Then an *experiment* consists of a quantum measurement (or set of quantum measurements) made using this fixed experiment configuration. For example, in this nomenclature, a standard Rabi curve would be constructed by making a set of experiments, each one defining—among other fixed parameters—a pulsing time in its experimental configuration, $e = (\ldots, t_{\text{pulse}}, \ldots)$.

An experiment returns a datum d. This might be a photon count over a known time interval, a time series of voltages, or a number of 'down' strong measurement results out of N repetitions, and so on.

Generally, the goal of statistical inference is to learn the parameters x given a data set $d_1, \ldots, d_n$ with respective configurations $e_1, \ldots, e_n$. This requires us to additionally specify a model for the system—something which connects the model parameters to the experiment configurations and data. This is done through a likelihood function,

$$\mathcal{L}(x; d_{1:n}, e_{1:n}) = \Pr(d_{1:n}|x, e_{1:n}), \tag{4.1}$$

which returns the probability of receiving a given dataset conditioned on a hypothetical configuration x. Here, and throughout this chapter, we use subscripted index-range notation, where, for example, $d_{1:n} = \{d_1, ..., d_n\}$. Note that multiple models can be considered and compared—known as model selection—if the true model is not known. For quantum systems, these likelihood models come naturally through quantum system evolution formulas in conjunction with Born's rule.

One popular inference choice is to maximize the likelihood function with respect to x, producing the maximum likelihood estimate (MLE) $\hat{x}_{\text{MLE}} := \text{argmax}_x \mathcal{L}$. Confidence regions of this estimate can be constructed with statistical derivations, or more generally, through techniques like bootstrapping. Least-squared curve fitting is often used as a proxy for the MLE (with confidence intervals arriving from assuming a linearized model) since it is exactly equal to the MLE for linear models and normal likelihood functions.

The MLE is one example of an estimator in a vast literature on estimator theory. In the present work, we limit ourselves to the use of Bayesian inference because of its natural integration with online experiments, discussed below. In short, in the paradigm of (sequential) Bayesian inference, one maintains the most current state of knowledge about

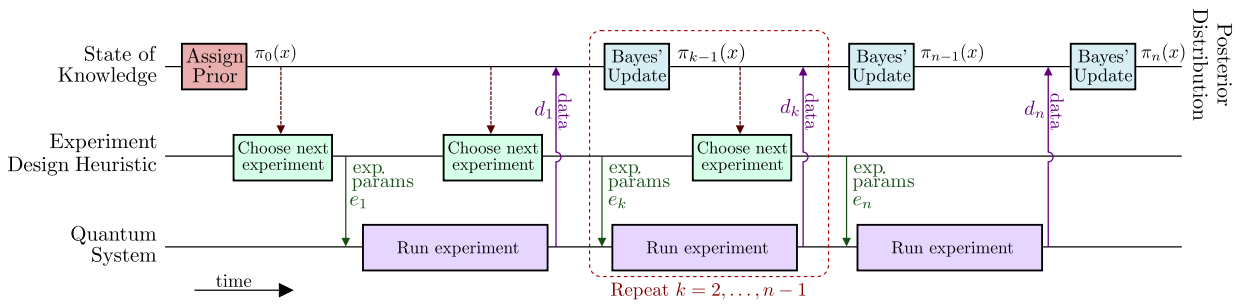

Figure 4.1: Timing diagram of online Bayesian learning. The role of the experiment design heuristic is to pick the next experiment configuration e_{n+1}, possibly based on the current state of knowledge, $\pi_n(x)$, resulting in the new data point d_{n+1}. This choice of experiment be computationally expensive, and is therefore run concurrently with quantum experiments.

the model parameters x, encoded as a probability distribution $\pi_n(x) = \Pr(x|d_{1:n}, e_{1:n})$, where $n = 1, 2, 3, ...$ indexes the state of knowledge when the first n data points $d_{1:n}$ have been collected and processed from the first n experiments $e_{1:n}$. We write $\pi_0(x)$ to denote the distribution prior to all measurements. The update from π_{n-1} to π_n is done through Bayes' law,

$$\pi_n(e) = \frac{\Pr(d_n|x, e_n)\pi_{n-1}(x)}{\Pr(d_n|e_n)}, \tag{4.2}$$

so that our knowledge is improved sequentially as each datum arrives. Note that the chain rule of conditional probabilities can be used to expand this equation into $\pi_n(e) = \Pr(d_{1:n}|x, e_{1:n})\pi_0(x)/\Pr(d_{1:n}|e_{1:n})$.

4.3 Bayesian Experimental Design

An *experiment design heuristic* is simply a function that determines the next experiment configuration to use. We say such a heuristic is *online* if it explicitly uses the results of preceding experiments, and we call it *offline* otherwise. An experiment design timing diagram is shown in Figure 4.1. Conventionally, as an example, Rabi curves are generated with offline heuristics, where the next experiment is chosen by increasing the pulse time by a fixed duration in each experiment. The number of experiments and pulse time increments are usually chosen through Nyquist considerations based on prior implicit beliefs about the frequencies and relaxation times of the system.

We restrict our online design heuristics to Bayesian designs, summarized in the following framework. Let $U_n(x, d, e)$ be the utility of collecting the datum d under configuration e given the hypothetical model parameters x and the current state of knowledge $\pi_n(x)$, where a large value is good. Using the Bayesian maxim of marginalizing over unknown quantities, the average utility of observing d at step $n+1$ under the possible experiment configuration e is

$$U_n(d, e) = \int \tilde{\pi}_{n,d,e}(x) U_n(x, d, e) \mathrm{d}x. \tag{4.3}$$

where $\tilde{\pi}_{n,d,e}(x) \propto \mathcal{L}(x; d, e)\pi_n(x)$ is the hypothetical posterior at step $n+1$ assuming d will be observed. Since we do not know *a priori* which d will occur, the average utility of the possible configuration e as a whole is

$$U_n(e) = \int \Pr(d|e) U_n(d, e) \mathrm{d}d. \tag{4.4}$$

where $\Pr(d|e) = \int \Pr(d|x, e)\pi_n(x)\mathrm{d}x$ is the predictive distribution[2]. Based on this quantity we can choose the next experiment to be the one that maximizes the utility,

$$e_{n+1} = \operatorname{argmax}_e U_n(e), \tag{4.5}$$

with the maximum taken over some space of possible experiments. If computed numerically, we might only hope to find local maxima.

One can consider different choices of utility function U. When the application is inference of a non-linear system, such as ours, it is common to choose a utility based on mean-squared error [23]. In particular, we choose $U_n = -r_{n,Q}$ where

$$r_{n,Q}(x, d, e) = \operatorname{Tr}\left[Q(x - \hat{x}_{n,d,e})^{\mathrm{T}}(x - \hat{x}_{n,d,e})\right] \tag{4.6}$$

where Q is a positive semi-definite weighting matrix. Here, $\hat{x}_{n,d,e} = \int x\tilde{\pi}_{n,d,e}(x)\mathrm{d}x$ is the Bayes estimator of x. In this case, $r_{n,Q}(e)$ has the simple interpretation of being the expected posterior covariance matrix weighted against Q,

$$r_{n,Q}(e) = \operatorname{Tr}\left[Q\mathbb{E}_d[\mathrm{Cov}_{\tilde{\pi}}[x|d, e]]\right], \tag{4.7}$$

a quantity known as the Q-weighted mean-squared-error *Bayes risk* (some numerical implementation details are outlined in Section B.5).

[2]Note that $\tilde{\pi}_{n,d,e}(x)\Pr(d|e) = \Pr(x, d|x_{1:n}, d_{1:n}, e)$, and therefore $U_n(e)$ is the joint average over x and d of $U_n(x, d, e)$ given the current state knowledge—this is a description some may prefer to the two-step description involving the intermediate quantity $U_n(d, e)$ provided in the main-body.

4.4 Nitrogen Vacancy System Model

The quantum system used in our experiment is a nitrogen vacancy (NV) center, introduced in Chapter 2. Our goal for this section is to explicitly define model parameters, experiment configurations, and a likelihood function for this system. Once this is achieved, we will be able to employ sequential Bayesian inference and online experiment design.

Recall that in the rotating frame $\omega_{\mu w}\mathbf{S}_{\mathrm{z}}^2$, with the rotating wave and secular approximations, the Hamiltonian of the optical ground state is given by

$$H/2\pi = (D - \omega_{\mu w})\mathbf{S}_{\mathrm{z}}^2 + (\omega_e + A\mathbf{I}_{\mathrm{z}})\mathbf{S}_{\mathrm{z}} + \Omega_1(t)\mathbf{S}_{\mathrm{x}} \tag{4.8}$$

where $(\mathbf{S}_{\mathrm{x}}, \mathbf{S}_{\mathrm{y}}, \mathbf{S}_{\mathrm{z}})$ are the spin-1 operators, $\omega_{\mu w}$ is the applied microwave frequency, $\Omega_1(t)$ is the microwave drive strength, A is the hyperfine splitting due to the adjacent nitrogen-14 atom, and $\mathbf{I}_{\mathrm{z}}$ is the nitrogen spin-1 operator along z. Along with the T_2^* decoherence time that introduces the Lindblad operator $L = \sqrt{1/T_2^*}\mathbf{S}_{\mathrm{z}}$, these parameters are sufficient to simulate the experiments that we perform. Therefore, the model parameters of our spin system (a few more nuisance parameters will be added later) are given by

$$x = (\Omega, \omega_e, \delta D, A, (T_2^*)^{-1}) \tag{4.9}$$

where $\delta D = D - 2.87\,\mathrm{GHz}$. Here, Ω is the maximum possible value that $\Omega_1(t)$ can take, so that we can write $\Omega_1(t) = a(t)\Omega$ using the unitless pulse-profile function $a(t) : [0, t_{\mathrm{e}}] \to [-1, 1]$ of duration t_{e}.

A general experiment configuration is then specified by

$$e = (a(t), \omega_{\mu w}, N) \tag{4.10}$$

where $a(t)$ pulse profile, $\omega_{\mu w}$ is the applied microwave frequency, and N is the number of repetitions of this experiment[3]. In this paper, we restrict our attention to two special cases of this general form, depicted in Figure 4.2, given by

1. Rabi experiments, $e_{\mathrm{Rabi}} = (t_{\mathrm{p}}, \omega_{\mu w}, N)$, $a(t) = 1$ for all $0 \le t \le t_{\mathrm{e}} = t_{\mathrm{p}}$; and
2. Ramsey experiments, $e_{\mathrm{Ramsey}} = (t_{\mathrm{p}}, t_{\mathrm{w}}, \omega_{\mu w}, N)$,

$$a(t) = \begin{cases} 0 & t_{\mathrm{p}} < t < t_{\mathrm{p}} + t_{\mathrm{w}} \\ 1 & \text{else} \end{cases}$$

for all $0 \le t \le t_{\mathrm{e}} = 2t_{\mathrm{p}} + t_{\mathrm{w}}$.

[3]The experiment configuration must also specify values for each of the timings labeled in Figure 4.2, but as they are calibrated independently from the experiment of interest, we omit them here for simplicity.

Given a hypothetical set of model parameters x and an experiment configuration e, the superoperator (in column-stacking convention) is given by the solution to the Lindlad master equation,

$$S(x,e) = \mathcal{T}\mathrm{e}^{\int_0^{t_e}(C[H(t)]+D[L])\mathrm{d}t}, \quad \text{where} \tag{4.11a}$$

$$C[H(t)] = -i(\mathbb{I}\otimes H(t) - \overline{H(t)}\otimes\mathbb{I}) \quad \text{and} \tag{4.11b}$$

$$D[L] = \overline{L}\otimes L - (\mathbb{I}\otimes L^\dagger L + \overline{L^\dagger L}\otimes\mathbb{I})/2, \tag{4.11c}$$

and where $\mathcal{T}$ is Dyson's time ordering operator. This results in the measurement probability

$$p(x,e) = \langle\langle P_0|\, S(x,e)\,|\rho_0\rangle\rangle\,, \tag{4.12}$$

where our initial state is $\rho_0 = |0\rangle\langle 0|\otimes\mathbb{I}/3$ and the measurement projector is $P_0 = 3\rho_0$.

Further recall that the standard measurement protocol of the NV system at room temperature does not have direct access to strong measurements, see Chapter 3. Instead, the probability $p(x,e)$ is obstructed by three Poisson rates, so that data is in the form of a triple $d = (X,Y,Z)$ where

$$X|\alpha \sim \text{Poisson}\,(N\alpha) \tag{4.13a}$$

$$Y|\beta \sim \text{Poisson}\,(N\beta) \tag{4.13b}$$

$$Z|x,e,\alpha,\beta \sim \text{Poisson}\,(N(\beta + p(x,e)(\alpha-\beta))) \tag{4.13c}$$

with α and β, the number of expected photons for the bright and dark references in a single shot with a given measurement duration t_m, satisfying $0<\beta<\alpha$. The values α and β are nuisance parameters which we must append to our model parameters, giving

$$x = (\Omega, \omega_e, \delta D, A, (T_2^*)^{-1}, \alpha, \beta). \tag{4.14}$$

The likelihood function (see Equation 4.1) for a single experiment is then given by

$$\begin{aligned}\mathcal{L}(x;d,e) = {} & f(X, N\alpha)\cdot f(Y, N\beta) \\ & \times f(Z, N(\beta + p(x,e)(\alpha-\beta))\end{aligned} \tag{4.15}$$

where f is the probability mass function of the Poisson distribution, $f(Q,\lambda) = \mathrm{e}^{-\lambda}\lambda^Q/Q!$. Some example risk plots (Equation 4.7) of this model are shown in Figure 4.3.

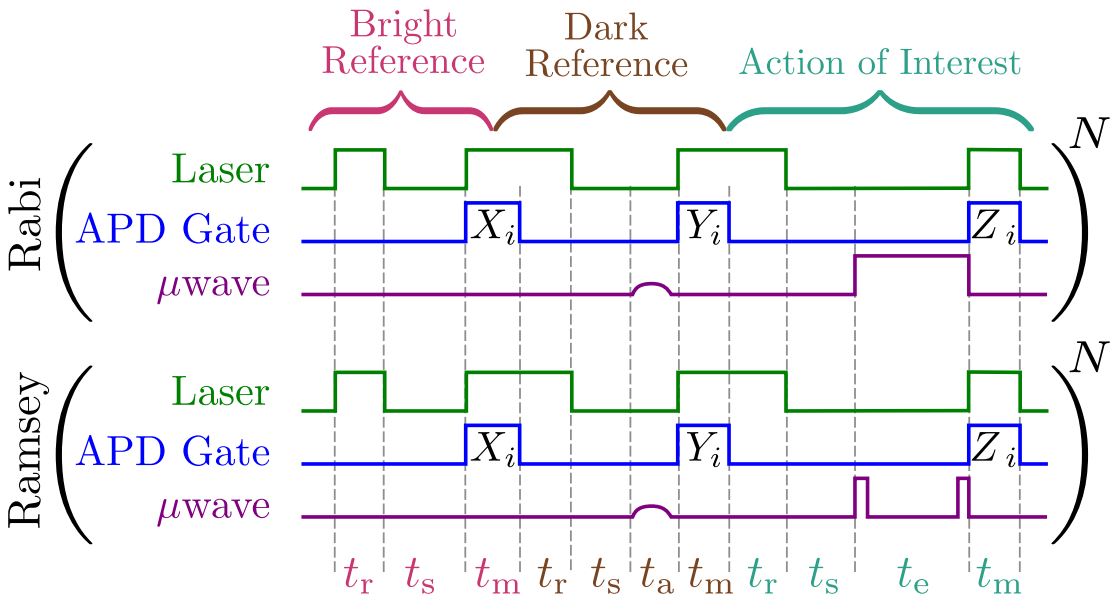

Figure 4.2: Pulse timing diagrams for Rabi (top) and Ramsey (bottom) experiments. An experiment has three control lines: whether the laser is on or off, whether the APD is counting photons or not, and the microwave amplitude profile. The pulse sequence is repeated N times, collecting photon counts (X_i, Y_i, Z_i) for $i = 1, ..., N$ for the bright reference, dark reference, and experiment, respectively, and finally summing them each over i to produce the data point $d = (X, Y, Z)$. Initial states are prepared by lasing for time t_r and letting the system settle for time t_s. Measurements consist of detecting photons for durations of length t_m while lasing. The dark reference includes an adiabatic pulse of length t_a which causes the state transfer $|0\rangle \rightarrow |+1\rangle$. The action of interest implements the microwave envelope $\Omega_1(t)$ of duration t_e. Relative timing is not to scale in this diagram.

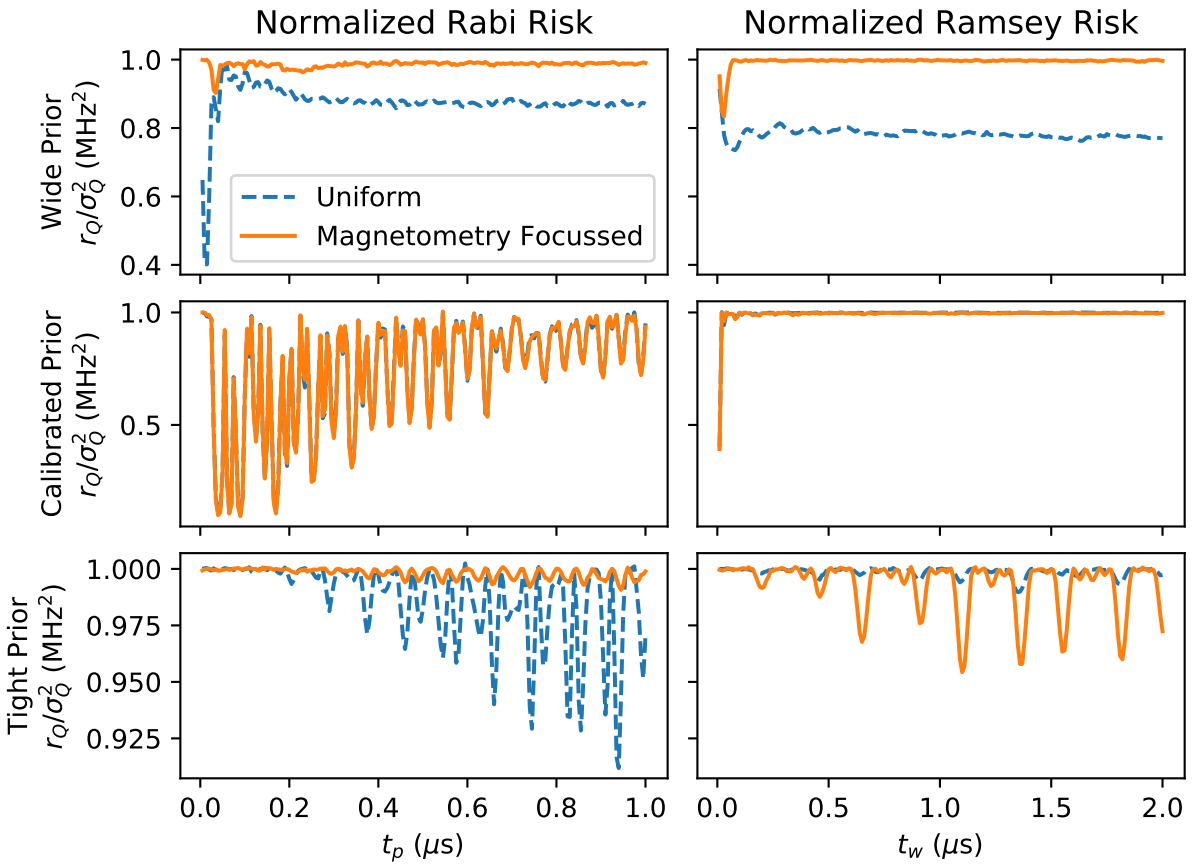

Figure 4.3: Calculation of risk for three different prior distributions (rows) and for both Rabi and Ramsey type experiments (columns). The dashed blue lines use a uniform weight matrix $Q = \text{diag}(1, 1, 1, 1, 1)$, and the solid orange lines use a weight matrix focused only on ω_e, $Q = \text{diag}(0, 1, 0, 0, 0)$. Values have been normalized against $\sigma_Q^2 = \text{Tr}(Q\,\text{Cov}_\pi[x])$ where $\text{Cov}_\pi[x]$ is the covariance matrix of a prior distribution π, so that, for example, a value of $r_Q(e)/\sigma_Q^2 = 0.95$ for a given experiment e implies a 5% expected improvement in weighted covariance. The wide prior (top row) is defined in Equation 4.20, the calibrated prior (middle row) is defined in Equation 4.21, and the tight prior (bottom row) is the same as the calibrated prior, but without widening the ω_e parameter. Note that the Rabi and Ramsey experiments share a y-axis on each row. We see that, among these examples, the only beneficial setting to perform a Ramsey experiment is with the tight prior when ω_e is the parameter of interest.

4.5 Computation and Hardware

For all experiment design heuristics, offline and online, we use the sequential Monte Carlo (SMC) [40] method to numerically compute sequential posteriors using the Python library QInfer[63]. In this algorithm, the state of knowledge about the model parameters, $\pi_n(x)$, is approximated as a finite list of weighted hypothetical values (which are called *particles*),

$$\pi_n(x) = \sum_{i=1}^{K} w_{n,i}\delta(x - x_{n,i}), \tag{4.16}$$

where $w_{n,i} \geq 0$ with $\sum_{i=1}^{K} w_{n,i} = 1$, and where $\delta(\cdot)$ is the delta mass distribution centered at 0. The particle-approximated prior, $\pi_0(x)$, is generated by sampling K initial particles $x_{0,i}$ from the prior distribution and setting uniform weights $w_{0,i} = 1/K$. Given the new datum d_{n+1} under experiment configuration e_{n+1}, Bayes update can be implemented with the simple multiplication

$$w_{n+1,i} \propto w_{n,i} \cdot \mathcal{L}(x_{n,i}; d_{n+1}, e_{n+1}) \tag{4.17}$$

which requires K simulations of the quantum system to compute the likelihoods (Equation 4.15), and where the constant of proportionality is chosen so that $\sum_{i=1}^{K} w_{n+1,i} = 1$. We use the scheme of Liu and West [113] to resample particle locations, triggered by a threshold in the effective particle count, $n_{\text{eff}} := 1/\sum_{i=1} w_{n,i}^2$ [63]. We also use the bridged-updating trick discussed in Reference [84].

We note that the expensive stage of this algorithm is embarrassingly parallel—simulations under the various model parameters $x_{n,i}$ can be performed independently. All of our processing was run on a desktop computer with simulations parallelized over the 12 cores on a pair of Intel Xeon X5675 CPUs. In this configuration, our updates took on the order of 2 seconds with $K = 30000$ particles. In principle, simulations could instead be run on quantum simulators, as was recently demonstrated [182].

For online heuristics, the Bayes risk (Equation 4.7) is calculated by noting that the particle approximation turns all integrals, which includes expectations and covariances, into finite sums—see Section B.5 for details. Some risk calculations for the NV model are plotted in Figure 4.3. As seen in the timing diagram in Figure 4.1, these calculations (along with the Bayes updates) are performed concurrently with experiments so that they do not add to experiment cost[4]. This causes the side-effect where the next experiment is selected

[4]Of course, this is only possible so long as the experiment repetition count is large enough compared to the parallelized simulation cost. In our setup, at our count rates, we landed naturally in this regime with CPU computation a single desktop computer.

using information that is one cycle out-of-date; however, in our simulations at our data collection rates, we found that this did not have a noticeable effect on learning rates. A new experiment configuration e_{n+1} having been decided, by whatever heuristic, the processing computer sends e_{n+1} to the computer which controls experiments. The experiment is run, and the datum $d_{n+1} = (X, Y, Z)$ is returned to the processing computer. This process is iterated until some stopping criterion is met—for example, in our experiments, we chose to stop after 200 experiments had been performed.

In our setup, the processing computer and the experiment computer communicate over ethernet with TCP. A diagram of electrical connections is found in Figure 2.6. We use a custom built confocal microscope to isolate an individual NV center in bulk diamond. All of our experiments were performed on the same NV center. We use a caching strategy, where the experiment computer uses a hash table to check if the desired experiment already exists in the AWGs memory, avoiding data transfer costs when possible.

4.6 Effective strong measurements and drift tracking

The amount of information provided by a measurement of Z (see Equation 4.13) depends on the values of α and β. Their magnitudes, relative contrast, and uncertainty all contribute to this information content. We quantify this idea by introducing what we call the number of *effective strong measurements* (ESM), defined as the number of two-outcome strong measurements one would (hypothetically) have to do to gain the equivalent amount of information about $p(x, e)$, averaged uniformly over $p \in [0, 1]$. This works out to

$$\text{ESM} = \frac{(\hat{\alpha} - \hat{\beta})^2}{3(\hat{\alpha} + \hat{\beta}) + 2\left(\sigma_\alpha^2 + \sigma_\beta^2\right)}. \tag{4.18}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are our current estimates of α and β, and σ_α and σ_β are standard deviation uncertainties in these estimates. See Section B.4 for details. We choose the number of repetitions in the next experiment, N, such that the expected value of ESM is constant—see Figure 4.4(b-c). This is especially important for the purpose of this chapter, which is to compare experiment design heuristics. In this way, certain heuristics are not artificially improved because of favorable lab conditions on a certain day of the week.

The true specific values of the references α and β depend not only on the optical dynamics of the quantum system itself, but also on the quality of the microscope's alignment. As the temperature of the lab changes, for instance, one can expect the values of α and β to drift as the location of the NV center moves with respect to the focal spot of the

microscope. To account for this, a tracking operation is performed periodically, where the focus of microscope is repositioned based on a new set of images taken with the microscope.

A model that assumes these reference values are constant in time can lead to inaccurate results, or even failure. To account for this drift, we append a Gaussian random walk model for the parameters α and β to the static model defined in Section 4.4. Specifically, we assume that immediately prior to a particle update (Equation 4.17) the reference indices of the each model parameter particle undergo a resampling step defined as

$$\begin{pmatrix} \alpha_{n,i} \\ \beta_{n_i} \end{pmatrix} \sim \text{Normal}\left(\begin{pmatrix} \alpha_{n,i} \\ \beta_{n_i} \end{pmatrix}, \Delta t \begin{pmatrix} \sigma_\alpha^2 & \sigma_{\alpha,\beta} \\ \sigma_{\alpha,\beta} & \sigma_\beta^2 \end{pmatrix} \right), \tag{4.19}$$

where Δt is the amount of time elapsed since the last update. The hyper-parameters σ_α, σ_β, and $\sigma_{\alpha,\beta}$ are treated as unknown; they are appended to the model parameters, and co-learned along with the parameters defined in Equation 4.14. We use a wide inverse Wishart distribution as the prior with a degrees-of-freedom parameter $\nu = 30$ and a scale matrix Ψ such that the mean value of the prior corresponds to $\sigma_\alpha = \sigma_\beta = \sigma_{\alpha,\beta}/0.7 = 0.036\,/\text{hour}$. We use an empirical prior on α and β, where before the actual experiments take place, a reference-only experiment is performed with $N = 300000$ repetitions, and the prior is set as $\alpha \sim \Gamma\mu = X/N\sigma = 3\sqrt{X}/N$ and $\beta \sim \Gamma\mu = Y/N\sigma = 3\sqrt{Y}/N$. When a tracking operation is performed, the distribution of α and β is resampled from the prior $\pi_0(x)$, with all other parameters of the model held fixed. We chose to perform tracking operation at the start of each trial, and each time our estimate of α dipped below our prior estimate of α minus five times the standard deviation of our prior for α.

4.7 Heuristics Used and Data Collection

There are many choices to be made, even for this small system. For example, we have already limited ourselves to Rabi and Ramsey experiments. Put differently, and given that our free evolution commutes with both our initial state and measurement, we have limited ourselves to bang-bang control with a maximum of two pulses. This is to ease simulations (bang-bang), and to reduce the search space for online heuristics (two pulses or fewer). We simplify the situation further by choosing to work in the low field regime, say $\lesssim 3\,\text{G}$. This saves us from having to adaptively modify the synthesizer frequency $\omega_{\mu w}$; we keep a fixed value of $\omega_{\mu w} = 2.87\,\text{GHz}$ for all experiments. It also prevents us from having to make decisions about the relative phase between the two Ramsey pulses, to which we are almost entirely insensitive at low field and with linearly polarized microwaves. These particular choices are by no means necessary, but serve as a starting place to explore the landscape.

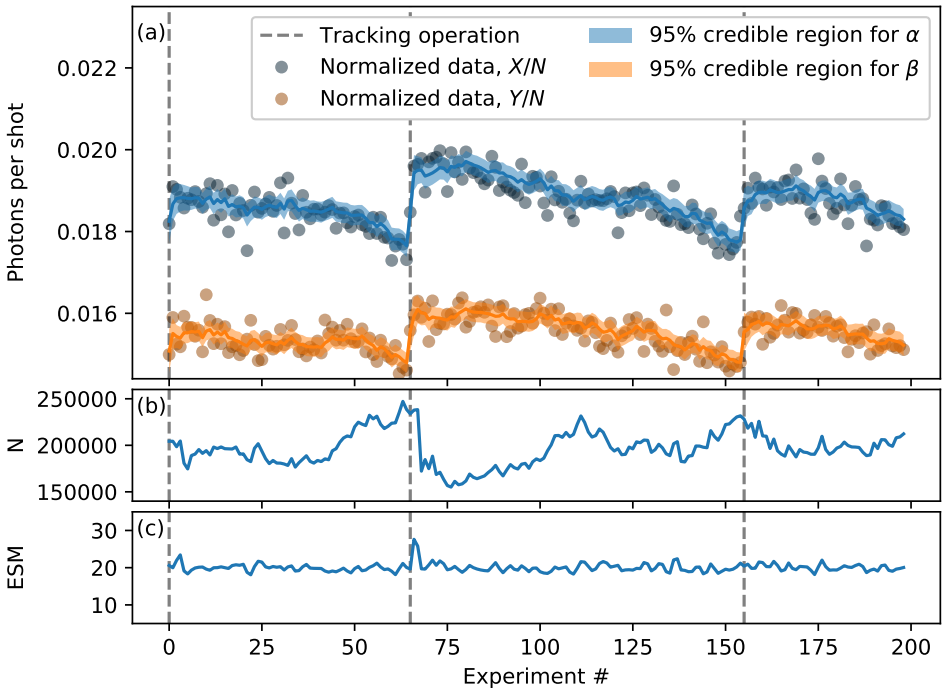

Figure 4.4: An NV drift tracking example, where tracking operations take place at the vertical dashed lines. (a) Sub-poissonian 95% credible regions are shown on top of data normalized by the experiment repetition count, N. (b) The repetition count was chosen online to maintain a constant ESM value of 20, which is plotted in (c). Several hundred trials were searched through to find this extreme but illustrative example—references are typically quite flat.

Heuristic	**Definition**
Alternating Linear	Offline; Sequential alternation between elements of the experiment sets $E_{\text{Rabi}}(500\,\text{ns}, 100)$ and $E_{\text{Ramsey}}(\hat{t}_{\text{p,best}}, 2\,\text{us}, 100)$
Ramsey Sweeps	Offline; Two back-to-back sweeps through the experiment set $E_{\text{Ramsey}}(\hat{t}_{\text{p,best}}, 2\,\text{us}, 100)$
Uniformly Weighted Risk	Online; $e_{n+1} = \underset{e\in E}{\text{argmax}}\,(r_{n,Q}(e))$ where $Q = \text{diag}(1,1,1,1,1)$ and $E = E_{\text{Rabi}}(500\,\text{ns}, 100) \cup E_{\text{Ramsey}}(\hat{t}_{\text{p,best}}, 2\,\text{us}, 100)$
Magnetometry Weighted Risk	Online; $e_{n+1} = \underset{e\in E}{\text{argmax}}\,(r_{n,Q}(e))$ where $Q = \text{diag}(0,1,0,0,0)$ and $E = E_{\text{Rabi}}(500\,\text{ns}, 100) \cup E_{\text{Ramsey}}(\hat{t}_{\text{p,best}}, 2\,\text{us}, 100)$

Table 4.1: Summary of heuristics used to choose experiments. The best Ramsey tip time is defined by $\hat{t}_{\text{p,best}} = 1/(4\hat{\Omega})$ (rounded to the nearest 2 ns), where $\hat{\Omega}$ is the current Bayes estimate of the microwave drive amplitude. $E_{\text{Rabi}}(t_{\max}, m)$ denotes a set of Rabi experiments with pulse times $t_{\text{p}} = t_{\max}/m, 2t_{\max}/m, \ldots, t_{\max}$, and $E_{\text{Ramsey}}(t_{\text{p}}, t_{\max}, m)$ denotes a set of Ramsey experiments with wait times $t_{\text{w}} = t_{\max}/m, 2t_{\max}/m, \ldots, t_{\max}$ and pulse times t_{p}. The components of weight matrices Q correspond to the Hamiltonian parameters $(\Omega, \omega_e, \delta D, A, (T_2^*)^{-1})$, with zeros for reference parameters.

From the perspective of metrology, these choices amount to studying the efficiency of DC magnetometry at low field with the NV system using the double quantum manifold.

In our first comparison between experiment design heuristics, we use a wide prior on the Hamiltonian parameters given by

$$\Omega/\text{MHz} \sim \text{Unif}\left([0, 20]\right), \tag{4.20a}$$
$$\omega_e/\text{MHz} \sim \text{Unif}\left([0, 10]\right), \tag{4.20b}$$
$$\delta D/\text{MHz} \sim \text{Unif}\left([-5, 5]\right), \tag{4.20c}$$
$$A/\text{MHz} \sim \text{Unif}\left([1.5, 3.5]\right), \tag{4.20d}$$
$$T_2^*/\mu\text{s} \sim \text{Unif}\left([1, 20]\right). \tag{4.20e}$$

along with the reference priors discussed in Section 4.6. We implement the offline heuristic *Alternating Linear* and the online heuristics *Uniformly Weighted Risk* and *Magnetometry*

Weighted Risk defined in Table 4.1. The offline heuristic is motivated by standard DC magnetometry, where, intuitively, Rabi experiments are used to determine the pulse length that causes $|0\rangle \mapsto \frac{|+1\rangle+|-1\rangle}{\sqrt{2}}$, and Ramsey experiments subsequently exploit this superposition state to measure the relative phase accumulation between $|+1\rangle$ and $|-1\rangle$, which is proportional to $t_{\mathrm{w}}\omega_e$. Note that, unconventionally, this heuristic alternates between Rabi and Ramsey experiments, as was done in [84]—this improves numerical stability of the SMC sampler; as different experiments are statistically independent, alternation does not affect the overall information content. The two online experiments differ only in the weighting matrix Q that is used—the first weights all quantum system parameters equally, and the second projects risk onto only one parameter, ω_e.

Results of this first comparison are shown in Figure 4.5(a-c). Here, it is seen that both online heuristics outperform the offline heuristic, with a final gap of a bit more than two orders of magnitude in the median (over trials) posterior variance of ω_e after 4000 ESM. In the histograms we see that the magnetometry focused online heuristic uses almost all Ramsey experiments, and the uniformly weighted online heuristics uses almost all Rabi experiments, which agrees with the risk profiles plotted in Figure 4.3. We see also that the offline heuristic has a much larger spread in posterior variances across trials (area of shaded regions), where some trials perform almost as well as the online heuristics, but many perform significantly worse. In this sense, in addition to tighter posteriors on average, these online heuristics have the extra advantage of being more reliable. Our guess is that offline heuristics require luckily informative data at certain key experiments to perform well, whereas online experiments can simply repeat these key experiments. Finally, note that the magnetometry focused online heuristic slightly outperforms the evenly weighted online heuristic—this is unsurprising as we happen to be plotting the variance of the magnetometry parameter, ω_e.

In the context of magnetometry, it is unrealistic to assume such a wide prior as given in Equation 4.20. More likely, one has already calibrated the quantum device and wants to learn only the value of ω_e. For example, one might be constructing a magnetic image [118, 68, 149], and each pixel of the image requires a new field measurement. Therefore, for our second comparison, we place a prior that is tight in all Hamiltonian parameters except ω_e, given as

$$\omega_e/\mathrm{MHz} \sim \mathrm{Unif}\left([0, 10]\right), \tag{4.21a}$$

$$(\Omega, \delta D, A, (T_2^*)^{-1})/\mathrm{MHz} \sim \mathrm{Normal}\left(\mu_{\mathrm{cal}}, \Sigma_{\mathrm{cal}}\right), \tag{4.21b}$$

where μ_{cal} and Σ_{cal} are taken

from the posterior obtained by processing two trials of the *Alternating Linear* heuristic,

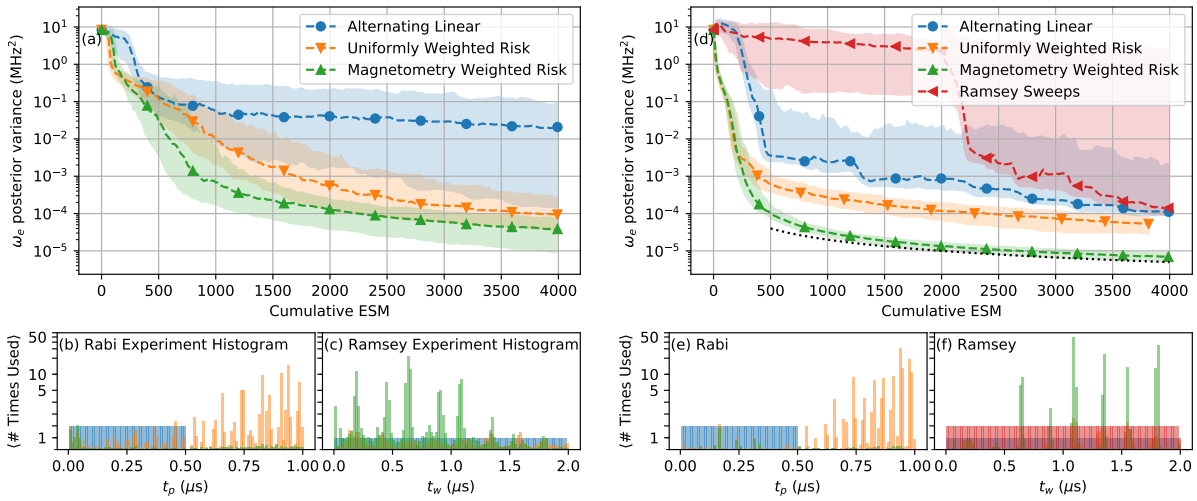

Figure 4.5: Comparison of experiment design heuristics (see Table 4.1) where each heuristic was run with 100 independent trials using 200 experiments per trial. The left figures (a-c) use the wide prior of Equation 4.20, and the right figures (d-f) use the calibrated prior of Equation 4.21. (a,d) For the parameter ω_e, the median posterior variance over 100 trials is plotted (dashed lines), and regions between the 10% and 90% percentiles are shaded. The x-axes display ESM (effective strong measurements), where roughly 20 effective bits of data are collected per experiment, see Section 4.6. The black dotted line scales as ESM^{-1}. In (b-c,e-f), histograms of which experiments each heuristic uses are shown, normalized to represent the average number of times used per trial. Note that the y-axis between histograms is shared, that the scaling switches from linear to logarithmic at $y = 5$, and that all four subfigures contain 100 histogram bins.

for a total of 400 experiments, and roughly 8000 ESM. Explictly, these are given by

$$\mu_{\text{cal}} = \begin{pmatrix} 11.55 \\ -0.86 \\ 2.18 \\ 0.35 \end{pmatrix} \text{MHz} \tag{4.22a}$$

$$\Sigma_{\text{cal}} = \begin{pmatrix} 2.56 \times 10^{-5} & 1.02 \times 10^{-3} & 7.67 \times 10^{-7} & 3.80 \times 10^{-5} \\ 1.02 \times 10^{-3} & 1.06 \times 10^{-1} & 1.97 \times 10^{-4} & 2.50 \times 10^{-3} \\ 7.67 \times 10^{-7} & 1.97 \times 10^{-4} & 7.51 \times 10^{-5} & -1.02 \times 10^{-4} \\ 3.80 \times 10^{-5} & 2.50 \times 10^{-3} & -1.02 \times 10^{-4} & 1.01 \times 10^{-3} \end{pmatrix} \text{MHz}^2 \tag{4.22b}$$

for the ordered parameters $(\Omega, \delta D, A, (T_2^*)^{-1})$, where ω_e and nuisance parameters have been marginalized over.

In our study of this second prior, in addition to the three heuristics used above, we consider another heuristic called *Ramsey Sweeps* that uses only Ramsey experiments, since they are the *de facto* method for measuring static Hamiltonian terms along z. Results for this second prior are shown in Figure 4.5(d-f). There are a few interesting features. The first is that it is clearly visible where the *Ramsey Sweeps* heuristic finishes one sweep and starts the next, at 2000 ESM. The second is that all three of the heuristics that were also used for the wide prior (Equation 4.20) have significantly less spread under the calibrated prior. The third is that the magnetometry weighted online heuristic has a much clearer advantage over the uniformly weighted online heuristic than in the case of the wide prior comparison. Finally, notice in the histograms, that the uniformly weighted online heuristic again chooses Rabi experiments almost exclusively.

In the summarizing figure Figure 4.5, only the learning rates of ω_e are reported. In Figure 4.6 and Figure 4.7, all learning rates are shown. Posteriors are shown in Figure 4.8 and Figure 4.9, where the first trial from each heursitic is used as a representative.

Our online learning rates appear to be at the standard quantum limit (SQL) once transient behavior has settled down; the dotted line in Figure 4.5(d) guides the eye with a curve $\propto \text{ESM}^{-1}$. The transient behavior prior to the SQL regime looks qualitatively exponential as a function of ESM. This does not violate the Heisenberg limit ($\sigma^2 \propto \text{ESM}^{-2}$) because experiment times, t_e, are able to exponentially increase, too [155]. Exponential-into-SQL scaling is consistent with previous Hamiltonian estimation research, where the coherence time of the system controls the transition location—ideally we would perform Ramsey experiments with arbitrarily long wait times, but finite T_2^* makes such experiments uninformative [51].

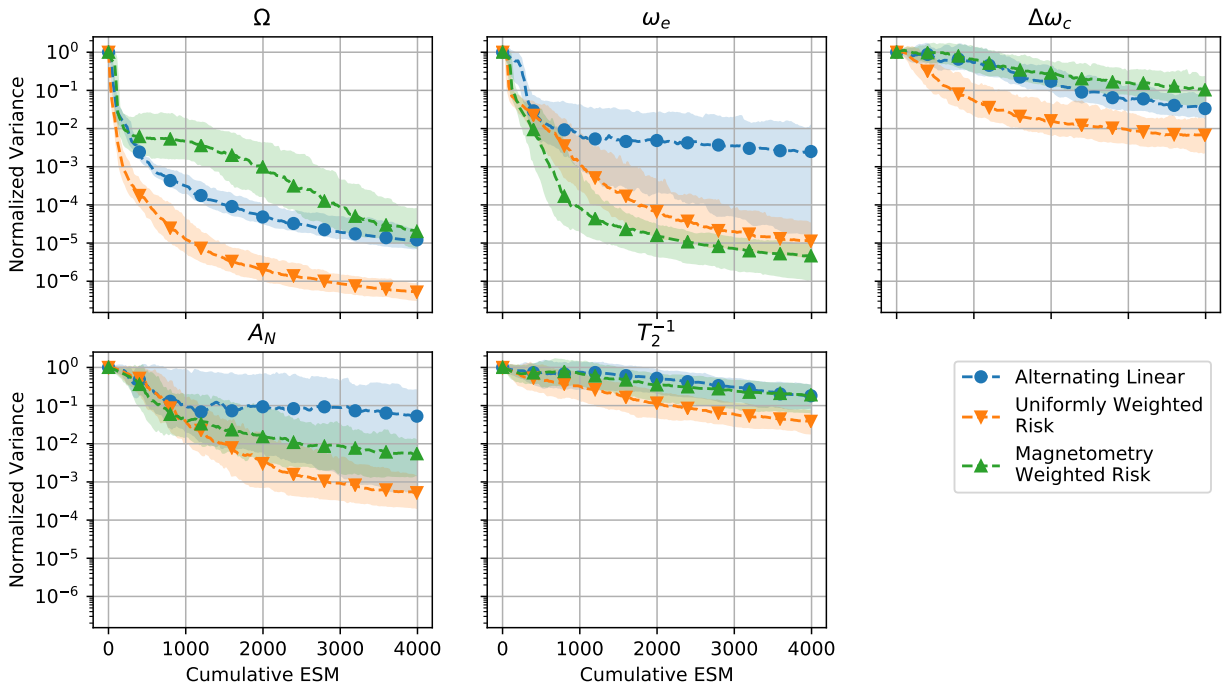

Figure 4.6: An extension of Figure 4.5(a-c) that shows learning rates of all parameters relevant to the quantum dynamics of the system.

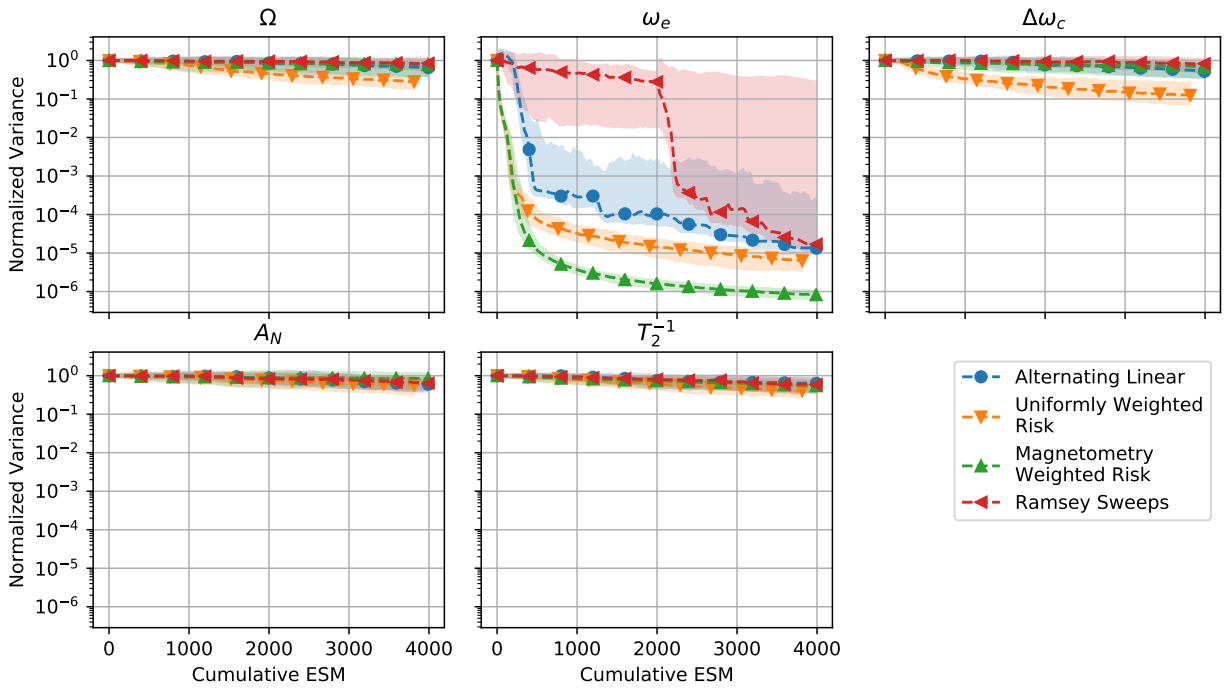

Figure 4.7: An extension of Figure 4.5(d-f) that shows learning rates of all parameters relevant to the quantum dynamics of the system.

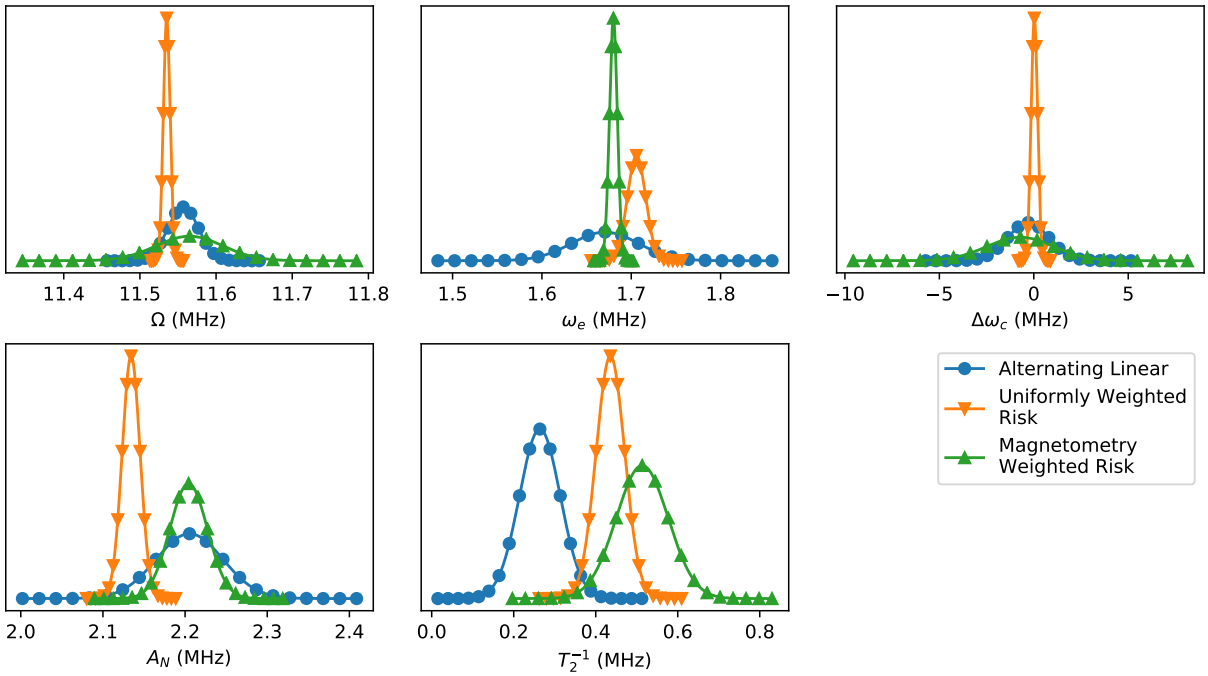

Figure 4.8: For each heuristic in Figure 4.5(a-c), posterior marginal distributions are plotted for the first (of 100) trials on each parameter relevant to the quantum dynamics of the system.

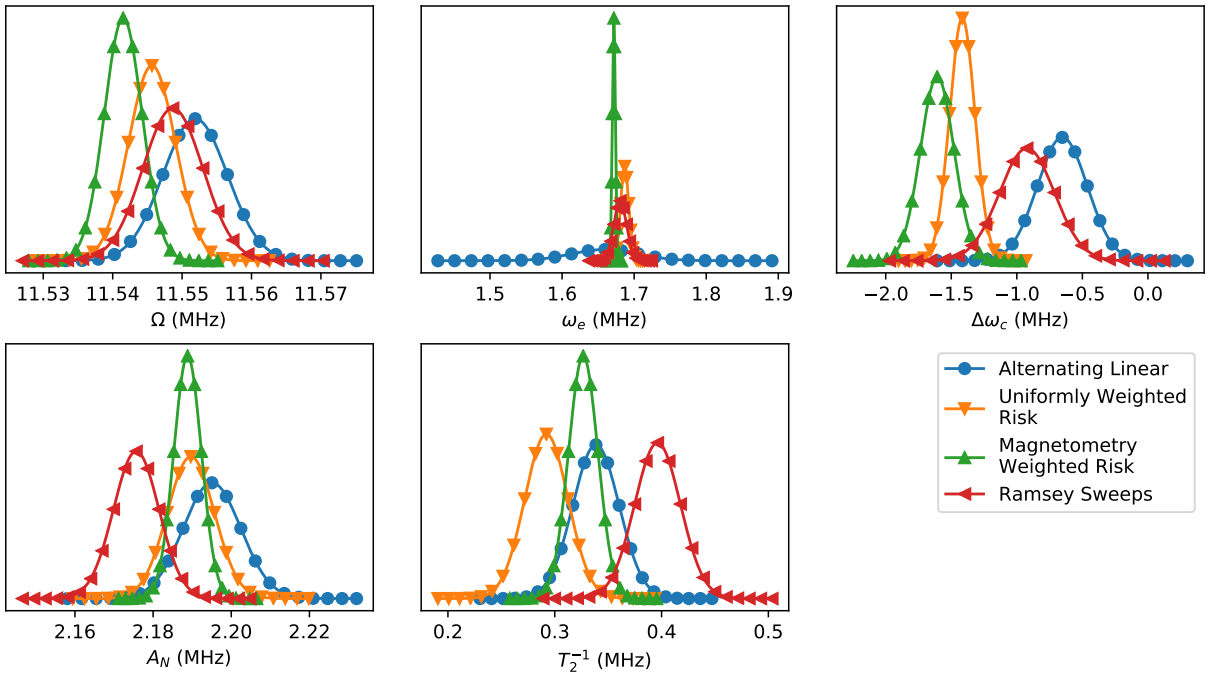

Figure 4.9: For each heuristic in Figure 4.5(d-f), posterior marginal distributions are plotted for the first (of 100) trials on each parameter relevant to the quantum dynamics of the system.

4.8 Summary of Results

We compared the ability of several experiment design heuristics to experimentally learn the electronic ground state Hamiltonian of an NV defect in diamond. Some of our heuristics were offline—using experiment sweeps that were predetermined, and some of heuristics were online—using knowledge gained from previous experiments to choose the next experiment adaptively. The heuristics we used are summarized in Table 4.1. All data analysis was done with sequential Bayesian inference, and all online heuristics were based on minimizing the weighted Bayes risk over a collection of possible experiments. Heuristics were compared by running 100 independent trials of each, and comparing the reduction in posterior variance of certain parameters as a function of the number of experiments performed.

We found that our online heuristics outperformed our offline heuristics; results are summarized in Figure 4.3. In particular, in the case of a very wide prior on all parameters (Figure 4.5(a-c)), we found that the median posterior variance of the parameter ω_e—which is proportional to the external magnetic field's projection onto the z-axis—is over two orders of magnitude smaller after 200 experiments (comprising 200 effective strong measurements per experiment) for the online heuristic called *Magnetometry Weighted Risk* than it is for the offline heuristic called *Alternating Linear*. Next, in the case of a prior that is tight on all paramaters except ω_e (Figure 4.5(d-f)), we found about an order of magnitude of improvement between the best online heuristic and the best offline heuristic. The use case of this prior is when one wants to use a calibrated NV device to measure many magnetic fields.

Consistent with intuition, we found that when online experiments are weighted to improve ω_e alone, they tend to choose Ramsey experiments almost exclusively, rather than Rabi experiments, see Figure 4.5(b-c,e-f).

In addition to faster decrease in variance, we also found that variance decreases more predictibly for online heuristics than it does for predetermined heuristics. This is seen in the tighter 80% percentile regions of Figure 4.5(a,d) for online experiments. For example, the difference in the final posterior variance of the parameter ω_e varies by as much as four orders of magnitude between independent trials for the *Ramsey Sweeps* heuristic, whereas it always varies by less than one order of magnitude for all online heuristics.

Studies of the sort presented here necessarily suffer from having to make choices—in the end we had to choose a small number of heuristics to compare, which types of experiments heuristics should be allowed to perform, what the hyper-parameters of each heuristic should be, what the initial prior over parameters should be, and so on. Though these choices are ultimately arbitrary, we attempted to make them reasonable, with the

end goal of comparing a fully brute-force Bayesian scheme against what have historically been the *de facto* methods of characterization. While we would not be surprised to find a less computationally expensive experiment design heuristic for this particular problem with similar performance (for example, see the heuristic policies in [164]), the advantage of a full-risk based approach is that it doesn't require an expert to design a heuristic for every particular combination of system and protocol. Indeed, minimizing Bayes risk, if computationally feasible either with classical or quantum resources, is a sensible approach for practically any characterization protocol, from tomography to randomized benchmarking.

Chapter 5

High Fidelity Control of Real-World Quantum Devices[1]

5.1 Introduction

The ability to coherently control the dynamics of quantum systems with high fidelity is a critical component of the development of modern quantum devices, including quantum computers [108], actuators [88, 15], and sensors [19, 120, 170] that push beyond the capabilities of classical computation and metrology. In recent years, quantum computation has presented a compelling application for quantum control, as high-fidelity control is essential to implement quantum information processors that may achieve fault-tolerance [60, 55, 72]. As quantum devices continue to grow in size and complexity, the requirements of classical control hardware also increase. This will more frequently produce situations with a significant trade-off between hardware response simplicity and overall hardware capability.

The performance of numerically optimized quantum gates in laboratory applications strongly depends on the response of the classical electronics used to apply the control

[1] This chapter is based on Reference [82], done in collaboration with Chris Granade, and Troy Borneman. Troy Borneman was responsible for providing the non-linear circuit model, putting its importance in context of existing literature, and helping to interpret its results. Chris Granade was responsible for co-developing the theory of distortion operators and making significant contributions to the code (see Appendix A, 'QuantumUtils'). The novel contributions of this work are to introduce a comprehensive framework for integrating distortions into numerical pulse-finding, with many special cases provided in detail, and also to numerically demonstrate the feasibility of controlling a quantum system using a nonlinear resonator.

sequence. In this article, we provide a formalism for including arbitrary classical hardware models into pulse finding algorithms such that the produced control sequences are tailored to work robustly for the intended hardware controllers. A novelty of our general framework is the ability to natively incorporate nonlinear and non-invertible hardware behaviour. Importantly, it also naturally allows for robustness against uncertainties and errors in parameters describing the hardware, in contrast to previous methods which dealt only with Hamiltonian parameters such overall power and offset frequency [173, 186, 151].

It has been demonstrated how a model of linear distortions of the control sequence, such as those arising from finite bandwidth of the classical control hardware, may be integrated into optimal control theory (OCT) algorithms [16, 162, 93, 128]. We generalize and extend these methods to admit hardware models which are non-invertible or non-linear, allowing the experimenter to maximize control efficiency and measurement sensitivity by driving hardware performance to its limits without sacrificing the ability to perform robust, high-fidelity quantum control.

Our framework includes a complete integration of the system-apparatus dynamics and models hardware components explicitly, such that their effect on a quantum system can be computed and compensated for using numerical OCT [139] algorithms to optimize control sequences. Control sequences designed using OCT algorithms, such as the GRadient Ascent Pulse Engineering (GRAPE) [99] algorithm, can be made robust to a wide variety of field inhomogenities, pulse errors and noise processes [102, 104, 17]. These methods are also easily extended [122, 121, 154, 59] to other applications and may be integrated into other protocols [44].

We begin developing our method generally, without making assumptions about the device of interest, so that our results may be broadly applicable to a wide range of quantum devices. We briefly discuss how our theory is easily applied to any linear distortion, and then in more detail, demonstrate with numerics how nonlinearities in control hardware may be included. As an example, we derive high-fidelity control pulses for strongly-driven superconducting resonators exhibiting non-linear kinetic inductance [114, 126, 31], that are robust to uncertainty in the amount of non-linearity present. While our methods apply generally to a wide range of quantum control modalities, superconducting resonators serve well as an illustrative test-bed of our method's utlity, having found significant recent application in pulsed electron spin resonance (ESR) [8, 119, 159] and circuit QED to increase induction measurement sensitivity and provide an interface for microwave photon quantum memories[107, 192].

5.2 Review of GRAPE

As mentioned previously, we are concentrating on the GRAPE algorithm for concreteness. This algorithm was proposed in Reference [99]. In this section we briefly review its standard treatment to make way for our generalization in the next section.

Suppose that we are given a system Hamiltonian

$$H(t) = H_0 + \sum_{l=1}^{L} q_l(t) H_l \tag{5.1}$$

acting on a Hilbert space of dimension d, where H_0 is the fixed internal Hamiltonian, $\{H_l\}_{l=1}^{L}$ are the fixed control Hamiltonians, and $\{q_l : \mathbb{R} \to \mathbb{R}\}_{l=1}^{L}$ are amplitude functions caused by an experimenter. The question we are interested in is as follows: how do we choose the envelopes $\beta(t) = \{q_l(t)\}_{l=1}^{L} \in \mathbf{L}_1\left(\mathbb{R}, \mathbb{R}^K\right)$ such that at time T we effect the desired total unitary $U_{\text{target}} \in \mathcal{U}(d)$? That is, if we compute the effect that a pulse $\beta \in \mathbf{L}_1\left(\mathbb{R}, \mathbb{R}^K\right)$ has on the quantum system by solving Shrödinger's equation,

$$U(\beta) = \mathcal{T} \exp(-i \int_0^T H(t) \mathrm{d}t), \tag{5.2}$$

then we desire that $U(\beta) = U_{\text{target}}$, up to a global phase. Here we are using the $\mathbf{L}_1$ function space to broadly emphasize that we are restricted to a finite amount of power.

A well-known result in quantum control theory tells us that such an envelope β exist for any target unitary U_{target} whenever the span of the nested commutations of the anti-hermitian operators $\{\mathrm{i}H_l\}_{l=0}^{L}$ is at least $d^2 - 1$ [32].

Theorem 5. *(Operator Controllability) Let* $\{H_0, \ldots, H_L\}$ *be as in Equation 5.1, let* $\mathrm{Lie}(S)$ *denote the Lie algebra generated by the matrices* $S \subset \mathrm{L}(d)$*, and let* $U_{target} \in \mathcal{U}(d)$ *be any* $d \times d$ *unitary matrix. If* $\dim \mathrm{Lie}(\{\mathrm{i}H_0, \ldots, \mathrm{i}H_L\}) \geq d^2 - 1$ *then there exists some* $T > 0$*, and some* $\beta \in \mathbf{L}_1\left(\mathbb{R}, \mathbb{R}^K\right)$ *such that* $U(\beta) = \mathrm{e}^{\mathrm{i}\phi} U_{target}$ *for some* $\phi \in \mathbb{R}$*.*

In this theorem, $\mathrm{Lie}(A_0, A_1, \ldots, A_L)$ denotes the smallest matrix Lie-algebra (vector space of matrices closed under multiplication and commutation) containing $A_0, A_0, \ldots, A_L$. Its proof is not constructive in a practical sense, and therefore our stated problem is more-or-less a constructive demand for this result. However, as it turns out, there is no known closed-form solution, and so we must fall back on either numerics or case-by-case specialized tricks. We now consider the numerical algorithm GRAPE, one of many tools used for this problem. Begin by discretizing candidate control functions β with the following mid-point sampler.

Definition 5.1. Given a *slicing interval* δt and a step count $M \in \mathbb{N}$, the *discretization operator* is defined as

$$\begin{aligned} f_1 : \mathbf{L}_1\left(\mathbb{R}, \mathbb{R}^L\right) &\to \mathbb{R}^M \otimes \mathbb{R}^L \\ \beta &\mapsto \vec{q} = (\beta(\delta t \cdot (1/2)), \ldots, \beta(\delta t \cdot (M - 1/2))). \end{aligned} \tag{5.3}$$

Here we are interpreting members of $\mathbb{R}^M \otimes \mathbb{R}^L$ as a vector of vectors, written $\vec{q} = (q_1, \ldots, q_M)$, where each $q_m \in \mathbb{R}^L$ is a vector of control amplitudes, and $1 \leq m \leq M$ indexes time.

Then, in lieu of Equation 5.2, for any $\vec{q} \in \mathbb{R}^M \otimes \mathbb{R}^L$ we can use a time-sliced version, which is more ammenable to numerics,

$$U(\vec{q}) = \prod_{m=M}^{1} U_m(\vec{q}), \tag{5.4}$$

where the product is ordered such that U_1 is on the far right, and where

$$U_m(\vec{q}) = \exp\left(-\mathrm{i}\delta t \left[H_0 + \sum_{l=1}^{L} q_{m,l} H_l\right]\right). \tag{5.5}$$

Formalizing our language a bit for clarity in future sections, we have the following definition.

Definition 5.2. A *continuous control pulse* is a function $\beta \in \mathbf{L}_1\left(\mathbb{R}, \mathbb{R}^L\right)$ for some $L \in \mathbb{N}$, *the number of controls*. A *discretized control pulse* (or just *pulse*) is a vector $\vec{q} \in \mathbb{R}^M \otimes \mathbb{R}^L$ for some $M \in \mathbb{N}$, the *number of control steps*. A discretized control pulse can be generated from a continuous control pulse by choosing a slicing interval δt, a control step number $M \in \mathbb{N}$ and applying the discretization operator, $\vec{q} = f_1(\beta)$. The *action due to pulse* $\vec{q}$ is given by $U(\vec{q})$, defined in Equation 5.4.

Keeping in mind our end goal $U(\vec{q}) = U_{\text{target}}$, we define a function that describes how close we are to achieving it. Note that instead of using a some norm distance between the target and the action of a pulse, a trace overlap is used. The primary reason is to allow for for a global phase difference; up to a sign, additive constant, and a global phase, the trace overlap is equal to the square Frobenious distance for unitaries,

$$\begin{aligned} \|U - V\|_F^2 &= \operatorname{Tr}\left((U - V)(U - V)^\dagger\right) \\ &= \|U\|_F^2 + \|V\|_F^2 - \operatorname{Tr} UV^\dagger - \operatorname{Tr} VU^\dagger \\ &= 2d^2 - 2\operatorname{re}\operatorname{Tr}(UV^\dagger). \end{aligned} \tag{5.6}$$

An added benefit is that overlap is easy to compute—distinct from many measures of matrix distance, we don't require singular values. We end up with the following utility function, which is labeled 'bare' to distinguish it from a generalized utility function defined in the next section.

Definition 5.3. Suppose that $\vec{q} \in \mathbb{R}^M \otimes \mathbb{R}^L$ is a discretized control pulse and U_{target} is a $d \times d$ unitary matrix. Then the *bare utility function* $\Phi : \mathbb{R}^M \otimes \mathbb{R}^L \to [0, 1]$ is defined by

$$\Phi[\vec{q}] = \left| \text{Tr} \left(U_{\text{target}}^\dagger U(\vec{q}) \right) \right|^2 / d^2, \tag{5.7}$$

where $U(\cdot)$ is the unitary effected by $\vec{q}$, defined in Equation 5.4.

Remark 5.1. It holds that $\Phi[\vec{q}] = 1$ iff $U_{\text{target}} = \mathrm{e}^{\mathrm{i}\phi} U(\vec{q})$ for some global phase $\phi \in \mathbb{R}$. Therefore, we are generally interested in maximizing Φ.

Penalties can be added to this utility function in order to demand that the solution admit certain properties. For instance, penalty functions have been used to ensure robustness to control noise and limited pulse fluence [87, 160, 104, 103] or to ensure that undesired subspaces are avoided [137, 132].

The GRAPE algorithm is a gradient ascent of this (or similar) utility function, where the $M \cdot L$ values of $\vec{q}$ are the search space. Naively, it might seem that we would need to evaluate Φ on the order of M times to compute the gradient vector. However, the main insight of the algorithm—and indeed the reason it merits its own acronym, not having been realized before Reference [99]—is that the cost of computing the gradient of the utility function is roughly the same cost as computing the utility function itself.

Indeed, the cost of computing $\Phi(\vec{q})$ is dominated by the cost of evaluating the M matrix exponentials $U_1, \ldots, U_M$, even in the case of very small values of d. Once $U_1, \ldots, U_M$ have been computed, however, we can use them both to compute $\Phi(\vec{q})$, and also to compute the quantities $P_m := \left(\prod_{i=m+1}^{M} U_i^\dagger \right) U_{\text{target}}$ and $X_m := \prod_{i=m}^{1} U_i$ that can then be used to compute all of the components of the gradient $\nabla_{\vec{q}} \Phi$ as

$$\frac{\partial \Phi}{\partial q_{m,l}} = -\,\text{re}\left[2\mathrm{i}\delta t \,\text{Tr}\left(P_m^\dagger H_l X_m \right) \text{Tr}\left(X_m^\dagger P_m \right) \right]. \tag{5.8}$$

A standard (conjugate) gradient ascent method can be used in conjunction with Equation 5.7 and Equation 5.8. The initial value of $\vec{q}$ is typically chosen to approximate a random smooth function. For example, random values for every 10^{th} timestep of $\vec{q}$ are chosen from a uniform distribution, and intermediate values are chosen by sampling along a spline.

5.3 Theory of Distortions

Recall our Hamiltonian system

$$H(t) = H_0 + \sum_{l=1}^{L} q_l(t) H_l \tag{5.9}$$

from the previous section. The functions $\{q_l(t)\}_{l=1}^{L}$ seen by the quantum system Hamiltonian represent a distorted version of what was actually input to the classical hardware. Effects such as circuit transfer functions, mixer imbalance, noise, amplifier non-linearity, and cross talk, will all contribute to this distortion acting on the input pulse. Such distortions are encompassed in the following definition.

Definition 5.4. A *continuous distortion operator* is a function

$$f : \mathbf{L}_1\left(\mathbb{R}, \mathbb{R}^K\right) \to \mathbf{L}_1\left(\mathbb{R}, \mathbb{R}^L\right)$$
$$\alpha(t) \mapsto \beta(t)$$

where, by convention, we write component functions with the symbols $\alpha(t) = \{p_1(t), ..., p_K(t)\}$ and $f[\alpha(t)] = \beta(t) = \{q_1(t), ..., q_L(t)\}$. Here, $K, L \in \mathbb{N}$ are respectively known as the number of *input and output control fields*, and α, β are respectively known as the *input and output continuous control pulses*.

Note that it will typically hold that $K = L$, though it is not hard to find examples where $K \neq L$ is useful to have. In the case of on-resonant quadrature control of a qubit, $K = L = 2$. Since we are ultimately interested in doing numerics, we will prefer to work with a discretized version of this operator.

Definition 5.5. A *discretized distortion operator* (or sometimes just *distortion operator*) is a triplet $(g, dt, \delta t)$ where g is a function

$$g : \mathbb{R}^N \otimes \mathbb{R}^K \to \mathbb{R}^M \otimes \mathbb{R}^L$$
$$\vec{q} \mapsto \vec{p},$$

and where $dt > 0$ is the duration of each of the $N \in \mathbb{N}$ input time steps, and $\delta t > 0$ is the duration of each of the $M \in \mathbb{N}$ output time steps. Here, $K, L \in \mathbb{N}$ are respectively known as the number of *input and output control fields*, and $\vec{q}$, $\vec{p}$ are respectively known as the *input and output discretized control pulses*, or just the *input and output pulses*.

Remark 5.2. As before, we write $\vec{p} = (\vec{p}_1, ..., \vec{p}_N)$, and similar for $\vec{q}$; a vector of vectors, where $\vec{p}_k \in \mathbb{R}^K$ corresponds to the amplitudes of the controls at the k^{th} timestep. Further, we will usually abuse notation and refer to g as the distortion operator.

The fact that input and output timesteps have uniform duration is merely for notational convenience—generalizing to $dt_1, \ldots, dt_N$ and $\delta t_1, \ldots, \delta t_M$ does not intruduce any significant complications. The condition $M \cdot \delta t = N \cdot dt$ need not hold, for example, $M \cdot \delta t > N \cdot dt$ will be useful when the distortion has a finite ringdown time. It is not strictly necessary for a discretized distortion operator to be derived from a continuous distortion operator, but in our examples, we will find that it is usually convenient to do so. This is achieved using the discretization operator from Definition 5.1 along with the following operator that has the opposite effect.

Definition 5.6. Given a *time-step duration* dt, the *dediscretization operator* is given by

$$\begin{aligned} f_2 : \mathbb{R}^N \otimes \mathbb{R}^L &\to \mathbf{L}_1\left(\mathbb{R}, \mathbb{R}^L\right) \\ (\vec{p}_1, ..., \vec{p}_N) &\mapsto \sum_{n=1}^{N} \vec{p}_n \cdot \text{Top}(t - dt \cdot (n - 1/2)) \end{aligned} \tag{5.10}$$

where Top is the L-dimensional top hat function,

$$T(t) = \begin{cases} (1, 1, ..., 1) & 0 \leq t < dt \\ (0, 0, ..., 0) & \text{else.} \end{cases} \tag{5.11}$$

Definition 5.7. Given a continuous distortion operator f, the *corresponding discretized distortion operator* is given by the composition $g = f_1 \circ f \circ f_2$.

The action of a distortion operator is depicted in Figure 5.1, and in the following example we give a concrete, useful transformation from a continuous distortion operator to a discretized distortion operator.

Example 5.1. Consider the continuous distortion operator given by the convolution with an $L \times K$ kernel $\phi(t)$,

$$\beta(t) = f(\alpha)(t) = (\phi \star \alpha)(t) = \int_{-\infty}^{\infty} \phi(t - \tau) \cdot \alpha(t) d\tau. \tag{5.12}$$

The convolution kernel ϕ models any distortion that can be described by a linear differential equation, such as a simple exponential rise time, control line crosstalk, or the transfer

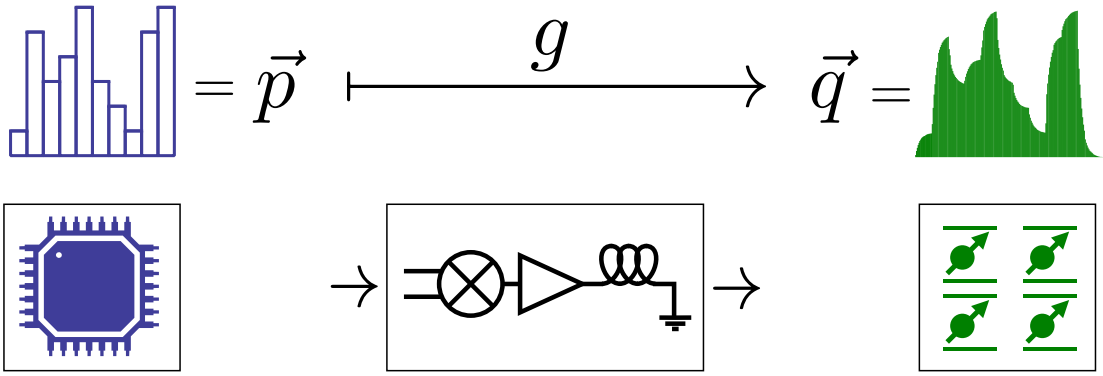

Figure 5.1: A cartoon depicting the action of the distortion operator g on the input pulse $\vec{p}$.

function of the control hardware [162, 145, 16, 7]. Splitting the integral into N parts and applying f_1 and f_2 gives the output values

$$q_{m,l} = \sum_{n=1,k=1}^{N,K} \left(\int_{(n-1)dt}^{ndt} \phi_{l,k}((m-1/2)\delta t - \tau) \mathrm{d}\tau \right) p_{n,k} \tag{5.13}$$

of the corresponding discretized distortion operator. Though straight-forward, explicit steps to derive this are provided in Section 5.7.2. This can be written more compactly as

$$\vec{q} = g(\vec{p}) = \tilde{\phi} \cdot \vec{p}, \tag{5.14}$$

where we are contracting over the n and k indices, and the components of the tensor $\tilde{\phi}$ are given by the integrals

$$[\tilde{\phi}]_{m,l,n,k} = \int_{(n-1)dt}^{ndt} \phi_{l,k}((m-1/2)\delta t - \tau) \mathrm{d}\tau. \tag{5.15}$$

Being a linear operator, the Jacobian matrix is simply given by $J_{\vec{p}}(g) = \tilde{\phi}$ which is independent of the pulse $\vec{p}$. ⊣

We can now incorporate the distortion operator g into standard methods from optimal control theory. In particular, for the case of GRAPE defined in the previous section, we have the following modified utility function.

Definition 5.8. Suppose that $\vec{p} \in \mathbb{R}^N \otimes \mathbb{R}^K$ is a discretized input pulse, U_{target} is a $d \times d$ unitary matrix, and $g : \mathbb{R}^N \otimes \mathbb{R}^K \to \mathbb{R}^M \otimes \mathbb{R}^L$ is a discretized distortion operator. Then the *utility function relative to distortion* g, denoted $\Phi_g : \mathbb{R}^N \otimes \mathbb{R}^K \to [0, 1]$, is defined by

$$\Phi_g[\vec{p}] = \Phi \circ g(\vec{p}). \tag{5.16}$$

where Φ is the bare distortion operator, see Definition 5.3

This gives us a measure of the quality of an input gate $\vec{p}$ including whatever distortions are added by our classical circuit. We ascend this utility function to the nearest local maxima starting with a random initial guess and choosing an uphill direction based on the gradient of Φ_g with respect to the components of $\vec{p}$. In the examples that follow in later sections, we use a standard adaptive step-size conjugate-gradient routine implemented in the QuantumUtils for Mathematica library [188]. In practice, as with standard GRAPE, a surprising fraction of local maxima are globally optimal when using experimentally relevant Hamiltonians and parameters.

Using the multivariable chain rule, we compute the gradient of Φ_g to be

$$\nabla_{\vec{p}}(\Phi_g) = \nabla_{g(\vec{p})}(\Phi) \cdot J_{\vec{p}}(g) \tag{5.17}$$

$$[J_{\vec{p}}(g)]_{m,l,n,k} = \frac{\partial g_{m,l}}{\partial p_{n,k}}, \tag{5.18}$$

thus separating the utility function derivatives into the derivatives of the distortion operator, and the bare utility function. Here, the dot represents a contraction over the indices m and l, and $J_{\vec{p}}(g)$ is the Jacobian of g at $\vec{p}$.

The remaining challenge is to compute the Jacobian $J_{\vec{p}}(g)$. This task depends entirely on the specific nature of the discrete distortion operator g. For instance, if g is linear, then computing the Jacobian tensor is in principle trivial, and is independent of the input pulse $\vec{p}$—we saw this in Example 5.1. In the examples that follow, the components of this Jacobian tensor will be worked out in detail.

Since the cost of evaluating g will typically not grow more than polynomially with the number of qubits, the computational cost of the optimization effectively remains unchanged from standard GRAPE, as it is still dominated by the cost of computing M matrix exponentials. We include a pseudo-code implementation of our modification to GRAPE in Algorithm 4.

Although we have singled out the GRAPE algorithm as our routine to optimize the utility function, this choice is based largely on the favourable convergence properties of

the algorithm [127], and does not prevent the use of a different routine. In particular, GRAPE is a greedy algorithm which attempts to find an optimum closest to the initial value by choosing a direction related to the steepest uphill slope. Global optimizers such as Nelder-Mead, genetic algorithms, or hybrid gradient algorithms [134, 44, 65, 127] could be used without modification by substituting the usual utility function, Φ, with the distortion modified utility function, Φ_g. Such methods are useful in cases where the control landscape is known to be saturated with suboptimal maxima. Gradient free methods may be advantageous in cases where it is difficult or overly expensive to compute the Jacobian tensor of Φ_g.

5.4 Including Robustness Distributions

In practice, it is unreasonable to assume that the Hamiltonians $H_0, \ldots, H_1$ in Equation 5.1 are perfectly known, or static across the temporal and/or spatial quantum ensemble. For example, we may

- know a certain Hamiltonian parameter only up to some error, i.e., $\omega \in [a, b]$ with 99% confidence;
- be studying many nearly-identical systems in parallel with Hamiltonian parameters that vary spatially, a common situation for magnetic resonance systems;
- or, have system parameters that drift slowly in time throughout the day as a result of lab conditions.

These quasi-static-like examples are the focus of this section—more sophisticated methods are needed for fluctuations that occur within a quantum system during a single experiment, such as those that require refocusing.

The standard approach to this problem is to parameterize all uncertain quantities in a vector $\vec{a}$ with an associated probability distribution $\Pr(\vec{a})$. We write

$$H_0[\vec{a}], \ldots, H_L[\vec{a}] \tag{5.19}$$

to denote the internal and control Hamiltonians conditional on a particular set of parameters $\vec{a} \in \mathbb{R}^A$. As a result, the action of a discretized pulse $\vec{q}$ may also be dependent on $\vec{a}$:

$$U(\vec{q}; \vec{a}) = \prod_{m=M}^{1} U_m(\vec{q}; \vec{a}) \quad \text{where} \quad U_m(\vec{q}; \vec{a}) = \exp\left(-\mathrm{i}\delta t \left[H_0[\vec{a}] + \sum_{l=1}^{L} q_{m,l} H_l[\vec{a}]\right]\right). \tag{5.20}$$

Example 5.2. The canonical example is a qubit with an off-resonance and control power distribution. In this case we have $A = 2$, $\vec{a} = (\delta\omega, \kappa)$, where $\delta\omega$ is the difference between the frequency of the control circuit and the frequency of the qubit, and $\kappa > 0$ is the multiplicitive error of the nutation frequency. This results in the Hamiltonians

$$\begin{aligned} H_0[\vec{a}] &= \delta\omega \mathbf{S}_\mathrm{z} \\ H_1[\vec{a}] &= (1+\kappa)\mathbf{S}_\mathrm{x} \\ H_2[\vec{a}] &= (1+\kappa)\mathbf{S}_\mathrm{y}. \end{aligned}$$

⊣

Similarly, we should allow distortions to depend on $\vec{a}$, denoted

$$\vec{q} = g[\vec{a}](\vec{p}) \tag{5.21}$$

conditional on the parameter vector $\vec{a}$.

Definition 5.9. Suppose that $\vec{p} \in \mathbb{R}^N \otimes \mathbb{R}^K$ is a discretized input pulse, U_target is a $d \times d$ unitary matrix. Suppose also that $\Pr(\vec{a})$ is a probability distribution on $\mathbb{R}^A$ and that for any $\vec{a} \in \mathbb{R}^A$, $g[\vec{a}] : \mathbb{R}^N \otimes \mathbb{R}^K \to \mathbb{R}^M \otimes \mathbb{R}^L$ is a discretized distortion operator. Then the *bare utility function conditional on* $\vec{a}$ is denoted $\Phi_{\vec{a}} : \mathbb{R}^M \otimes \mathbb{R}^L \to [0,1]$ and defined by

$$\Phi_{\vec{a}}(\vec{q}) = \left| \mathrm{Tr}\left(U_\text{target}^\dagger U(\vec{q};\vec{a}) \right) \right|^2 / d^2, \tag{5.22}$$

where $U(\cdot)$ is the unitary effected by $\vec{q}$ conditional on $\vec{a}$, defined in Equation 5.20. Next, the *utility function relative to distortion* g *conditional on* $\vec{a}$ is denoted $\Phi_{g,\vec{a}} : \mathbb{R}^N \otimes \mathbb{R}^K \to [0,1]$, is defined by

$$\Phi_{g,\vec{a}}(\vec{p}) = \Phi_{\vec{a}} \circ g[\vec{a}](\vec{p}). \tag{5.23}$$

This finally yields the *expected utility function relative to distortion* g, defined as

$$\begin{aligned} \Phi_g :& \mathbb{R}^N \otimes \mathbb{R}^K \to [0,1] \\ \Phi_g(\vec{p}) &= \int \Pr(\vec{a}) \Phi_{g,\vec{a}}(\vec{p}) \mathrm{d}\vec{a}. \end{aligned} \tag{5.24}$$

Remark 5.3. It is common to use a particle approximation for the distribution $\Pr(\vec{a})$, writing

$$\Pr(\vec{a}) = \sum_{s=1}^{S} p_s \delta(\vec{a} - \vec{a}_s) \tag{5.25}$$

where $p_s > 0$, $\sum_s p_s = 1$, in which case we have the more friendly expected utility function expression

$$\Phi_g(\vec{p}) = \sum_{s=1}^{S} p_s \Phi_{g,\vec{a}_s}(\vec{p}). \tag{5.26}$$

Having slightly modified the utility function to include this distribution over parameters $\vec{a}$, pulse optimization proceeds as normal.

Example 5.3. Suppose that the convolution matrix $\phi(t)$ from Example 5.1 is parameterized by a value $\vec{a}$, which we write as $\phi[\vec{a}](t)$. If the distribution $\Pr(\vec{a})$ is independent between the part of $\vec{a}$ that affects $\phi[\vec{a}]$ and the part that affect the Hamiltonians, then the following shortcut can be taken, which results in a new effective convolution. First observe that the continuous distortion operator acts as

$$\beta(t) = f(\alpha)(t) = \int \Pr(\vec{a}) \, (\phi[\vec{a}] \star \alpha)(t) \mathrm{d}\vec{a} \tag{5.27}$$

where $\Pr(\vec{a})$ is the probability of realization $\vec{a}$, then linearity helps us out and we simply have

$$\tilde{\phi} = \int \Pr(\vec{a}) \, \tilde{\phi}[\vec{a}] \mathrm{d}\vec{a}. \tag{5.28}$$

If, on the other hand, there is correlation between $\phi[\vec{a}]$ and the Hamiltonians $H_0[\vec{a}], ..., H_L[\vec{a}]$, then the integration should be done at the level of the utility function, as in Definition 5.9. ⊣

5.5 A Generalized Utility Function

The utility function of Definition 5.3 assumes that the goal of the optimization is to find a pulse which generates some unitary U_{target} on the full Hilbert space. In this section we introduce a simple, but very useful, generalization of this utility function that can assign different unitary targets to orthogonal subspaces, allowing arbitrary relative phases between them. We will see that pure state-to-state transfers, mixed state-to-state transfers, and global unitary targets, are all captured by this framework.

5.5.1 Motivating Example

Suppose $\mathcal{H} = \mathbb{C}^3 \otimes \mathbb{C}^2$ and that we wish to flip two levels of the first spin (the qutrit) depending on the state of the second spin (the qubit). Therefore, the following would be sufficient:

$$U^1_{\text{target}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \otimes \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = \left(\begin{array}{cc|cc|cc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right).$$

However, suppose that throughout our experiment we never plan on having population within the subspace $|-1\rangle \otimes \mathbb{C}^2$, except perhaps when a control gate is taking place. Then anything of the form

$$U^2_{\text{target}} = \left(\begin{array}{cc|cc|cc} \multicolumn{2}{c|}{\multirow{2}{*}{U}} & 0 & 0 & 0 & 0 \\ & & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & e^{i\phi} & 0 \\ 0 & 0 & 0 & e^{i\phi} & 0 & 0 \\ \hline 0 & 0 & e^{i\phi} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & e^{i\phi} \end{array}\right)$$

would suffice, where U is any 2×2 unitary, and ϕ is any real number which is effectively a global phase. This example will be useful in Chapter 7 when designing gates for a qubit manifold of the NV center.

5.5.2 The Utility Function

In a Hilbert space $\mathcal{H}$ of dimension d, suppose we have a collection of groups of vectors, $A_i = \{|\psi_{i,1}\rangle, ..., |\psi_{i,n_i}\rangle\} \subseteq \mathcal{H}$ for $1 \leq i \leq s$, which we want mapped to a second collection of groups of vectors, $B_i = \{|\phi_{i,1}\rangle, ..., |\phi_{i,n_i}\rangle\} \subseteq \mathcal{H}$ for $1 \leq i \leq s$. Suppose moreover that $\cup_{i=1}^s A_i$ is an orthonormal set, and that $\cup_{i=1}^s B_i$ is an orthonormal set. Viewed differently, when each A_i and B_i are (in a slight abuse of notation) treated as $d \times n_i$ matrices with columns $\{|\psi_{i,1}\rangle, ..., |\psi_{i,n_i}\rangle\}$ and $\{|\phi_{i,1}\rangle, ..., |\phi_{i,n_i}\rangle\}$ respectively, then $X_1, \ldots, X_s$ is a set of pairwise orthogonal isometries, and likewise so is $Y_1, \ldots, Y_s$.

We desire a $d \times d$ unitary matrix such that

$$U \left|\psi_{i,j}\right\rangle = \mathrm{e}^{\mathrm{i}\alpha_i} \left|\phi_{i,j}\right\rangle \quad \forall \quad 1 \leq j \leq n_i, 1 \leq i \leq s \tag{5.29}$$

for some $\alpha_i \in \mathbb{R}$, $1 \leq i \leq s$. This is written more compactly as

$$UA_i = e^{i\alpha_i} B_i \quad \forall \quad 1 \leq i \leq s. \tag{5.30}$$

Claim 1. *A U satisfies Equation 5.29 if and only if* $\sum_{i=1}^{s} \frac{\left|\mathrm{Tr}\, U^\dagger B_i A_i^\dagger\right|^2}{s n_i^2} = 1$.

Proof. Given a unitary $d \times d$ unitary U, we have

$$\begin{aligned}
&\exists \vec{\alpha} \in \mathbb{R}^s \text{ s.t. } UA_i = e^{i\alpha_i} B_i \quad \forall \quad 1 \leq i \leq s \\
\iff &\exists \vec{\alpha} \in \mathbb{R}^s \text{ s.t. } \left\| UA_i - e^{i\alpha_i} B_i \right\|^2 = 0 \quad \forall \quad 1 \leq i \leq s \\
\iff &\exists \vec{\alpha} \in \mathbb{R}^s \text{ s.t. } \mathrm{Tr}(A_i^\dagger A_i) + \mathrm{Tr}(B_i^\dagger B_i) - 2\,\mathrm{re}(e^{i\alpha_i} \mathrm{Tr}(UA_i)^\dagger B_i) = 0 \quad \forall \quad 1 \leq i \leq s \\
\iff &\exists \vec{\alpha} \in \mathbb{R}^s \text{ s.t. } n_i + n_i - 2\,\mathrm{re}(e^{i\alpha_i} \mathrm{Tr}(UA_i)^\dagger B_i) = 0 \quad \forall \quad 1 \leq i \leq s \\
\iff &\exists \vec{\alpha} \in \mathbb{R}^s \text{ s.t. } \mathrm{re}(e^{i\alpha_i} \mathrm{Tr}\, U^\dagger B_i A_i^\dagger) = n_i \quad \forall \quad 1 \leq i \leq s \\
\iff &\left|\mathrm{Tr}\, U^\dagger B_i A_i^\dagger\right|^2 = n_i^2 \quad \forall \quad 1 \leq i \leq s \\
\iff &\sum_{i=1}^{s} \frac{\left|\mathrm{Tr}\, U^\dagger B_i A_i^\dagger\right|^2}{s n_i^2} = 1
\end{aligned}$$

□

This claim justifies the (bare) utility function

$$\Phi(\vec{q}) = \sum_{i=1}^{s} \frac{\left|\mathrm{Tr}\, U(\vec{q})^\dagger B_i A_i^\dagger\right|^2}{s n_i^2}. \tag{5.31}$$

Partial derivatives straight-forwardly derived from Equation 5.8, given by

$$\frac{\partial \Phi}{\partial q_{m,l}} = -2\mathrm{i}\delta t \sum_{i=1}^{s} \mathrm{re}\left[\mathrm{Tr}\left(P_{m,i}^\dagger H_l X_{m,i}\right) \mathrm{Tr}\left(X_{m,i}^\dagger P_{m,i}\right)\right] \tag{5.32}$$

where $P_{m,i} := \left(\prod_{i=m+1}^{M} U_i^\dagger\right) B_i A_i^\dagger$ and $X_{m,i} := \prod_{i=m}^{1} U_i$, with U_i as in Equation 5.5. Note that in the case where $s = 1$ and $n_1 = 1$, then this is just the overlap between the initial state and the final state. Also note that in the case where $s = 1$, $n_1 = d$, $X = \mathbb{I}_d$ (the computation basis), and $Y = U_{\text{target}}$ (an orthonormal basis for $\mathcal{H}$), then this objective function is the standard GRAPE objective function.

Example 5.4. A pure state-to-state transfer between $|\psi\rangle \in \mathcal{H}$ and $|\phi\rangle \in \mathcal{H}$ is achieved trivially with $s = 1$ and $n_1 = 1$. ⊣

Example 5.5. A full unitary target U_{target} is achieved with $s = 1$, $X_1 = \mathbb{I}$, and $Y_1 = U_{\text{target}}$. ⊣

Example 5.6. A mixed state-to-state transfer between ρ and σ is achieved by first decomposing into the eigenforms $\rho = \sum_{i=1}^{s} p_i |\psi_{i,1}\rangle\langle\psi_{i,1}|$ and $\sigma = \sum_{i=1}^{s} p_i |\phi_{i,1}\rangle\langle\phi_{i,1}|$, so that each $n_i = 1$, $1 \leq i \leq s \leq d$. Note that they must both have the same eigenvalues or else a unitary transfer would not be possible. ⊣

Example 5.7. The example from Section 5.5.1 is achieved by taking $s = 1$, $X_1 = \mathbf{0}_2 \oplus \mathbb{I}_4$, and $Y_1 = \mathbf{0}_2 \oplus \text{CNOT}$. ⊣

5.6 In-depth Nonlinear Circuit Example

As a more involved example than the convolution case, Example 5.1, we consider a quantum system being controlled by a tuned and matched resonator circuit [6] with nonlinear circuit elements (Figure 5.2). We emphasize that while we have picked a relatively simple circuit for this demonstration, it has a general enough form to accurately describe the majority of resonators used in spin resonance experiments, including the non-linear resonator described in Reference [126]. Moreover, arbitrarily complex circuits with additional poles could just as easily be incorporated by finding their circuit equations with a standard application of Kirchhoff's laws, resulting in a higher order equation in place of Equation (5.34).

Nonlinear superconducting resonators are used in a variety of applications, including circuit QED for quantum information processing and quantum memories [153, 142, 67], microwave kinetic inductance detectors for astronomy [34], and pulsed electron spin resonance [3, 8, 119, 159]. Often, however, these devices are operated in their linear regime to avoid complications resulting from nonlinearity. Avoiding nonlinearities requires reducing input power, leading to longer control sequences that reduce the number of quantum operations that can be performed before the system decoheres. Additionally, limiting input power removes the natural robustness of high-power sequences to uncertainties in the environment achieved by strongly modulating the quantum system [54, 140].

If the circuit were linear, the distortion could be modelled as a convolution $\phi \star$ as discussed above. However, with nonlinear circuit elements present we must numerically solve the circuit's differential equation every time we wish to compute the distorted pulse [114].

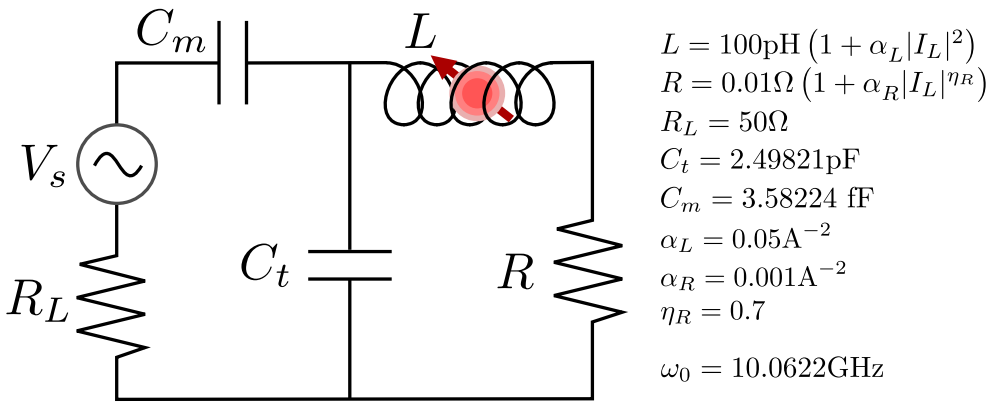

Figure 5.2: A quantum system being controlled by the magnetic field produced by the inductor of a nonlinear resonator circuit. The ideal voltage source $V_s(t)$ is specified by the input undistorted pulse $\vec{p}$, and the resulting current through the inductor, $I_L(t)$, is computed. The inductance and the resistance are both functions of the current passing through them. The form of the nonlinearity is chosen to be consistent with kinetic inductance.

5.6.1 Overview and Results

It is most natural to demonstrate with a qubit system. This is both because it lets us more clearly isolate the change in control landscape induced by the non-linear distortion operator, and because it is known that control landscapes generally scale well with Hilbert space dimension [144]. Our qubit is a near-resonance spin system whose Hamiltonian in the rotating frame, after invoking the rotating wave approximation, as in Example 5.2, is given by

$$H = \frac{\delta\omega}{2}\sigma_z + (1+\kappa)\left(\frac{\omega_x(t)}{2}\sigma_x + \frac{\omega_y(t)}{2}\sigma_y\right) \tag{5.33}$$

where $\delta\omega$ and κ represent off-resonance and control power errors, respectively.

The time evolution of the circuit shown in Figure 5.2 is governed by the third order differential equation

$$\frac{d}{dt}\begin{bmatrix} I_L \\ V_{C_m} \\ V_{C_t} \end{bmatrix} = \begin{bmatrix} -\frac{R}{L} & 0 & \frac{1}{L} \\ 0 & \frac{-1}{R_L C_m} & \frac{1}{R_L C_m} \\ \frac{-1}{C_t} & \frac{-1}{R_L C_t} & \frac{1}{R_L C_t} \end{bmatrix}\begin{bmatrix} I_L \\ V_{C_m} \\ V_{C_t} \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{V_s(t)}{R_L C_m} \\ \frac{V_s(t)}{R_L C_t} \end{bmatrix} \tag{5.34}$$

where the nonlinearities arise when the inductance, L, and resistance, R, are functions of the current passing through them [114, 172]. In the case of kinetic inductance, these nonlinearities take on the form

$$\begin{aligned} L &= L(I_L) = L_0(1 + \alpha_L |I_L|^2) \\ R &= R(I_R) = R_0(1 + \alpha_R |I_R|^\eta) \end{aligned} \tag{5.35}$$

where α_L, α_R and η are constants [31, 126]. Kinetic inductance leads to a reduction in the circuit resonance frequency, coupling, and quality factor with increasing power, as shown in Figure 5.3(a-b).

Since our Hamiltonian in Equation 5.33 is written in a frame rotating at the circuit resonance frequency in the linear-regime, it is convenient to write our differential equation in this frame, too. To this end, with the differential Equation 5.34 shorthanded as $\dot{\vec{y}}(t) = B(\vec{y}(t))\vec{y}(t) + V_s(t)\vec{b}$, we introduce the complex change of variables $\vec{x}(t) = e^{-i\omega_0 t}\vec{y}(t)$. In this new frame, since $B(\vec{y}(t)) = B(\vec{x}(t))$, our dynamics become

$$\begin{aligned} \dot{\vec{x}}(t) &= \left(B(\vec{x}(t)) - i\omega_0 \mathbb{I}\right)\vec{x}(t) + \tilde{V}_s(t)\vec{b} \\ &\equiv A(\vec{x}(t))\vec{x}(t) + \tilde{V}_s(t)\vec{b} \end{aligned} \tag{5.36}$$

where we have invoked the rotating wave approximation, and $\tilde{V}_s(t) = e^{-i\omega_0 t}V_s(t)$ is the rotating version of $V_s(t)$. Details of this frame change are described in Section 5.6.2. Now the real and imaginary parts of the complex current in the rotating frame, $\tilde{I}_L(t) = e^{-i\omega_0 t}I_L(t)$, are proportional via a geometric factor to the control amplitudes appearing in the Hamiltonian,

$$\omega_x(t) \propto \mathrm{re}[\tilde{I}_L(t)] \quad \text{and} \quad \omega_y(t) \propto \mathrm{im}[\tilde{I}_L(t)]. \tag{5.37}$$

To compute the distortion $\vec{q} = g(\vec{p})$ caused by the resonator, we set the circuit's input voltage $\tilde{V}_s(t)$ to be the piecewise constant function with amplitudes coming from $\vec{p}$. To improve stiffness conditions, a small finite risetime may be added to the forcing term $\tilde{V}_s(t)$, which is equivalent to adding a low-pass filter to the ideal voltage source in the circuit. We can now solve the Equation 5.36 for $\tilde{I}_L(t)$ using the NDSolve function in Mathematica 10, interpolate the results, and resample at a rate δt to determine the distorted pulse $\vec{q}$.

Since our distortion is nonlinear, the Jacobian of g will not be constant with respect to the input pulse $\vec{p}$. However, we may compromise the accuracy of the Jacobian in

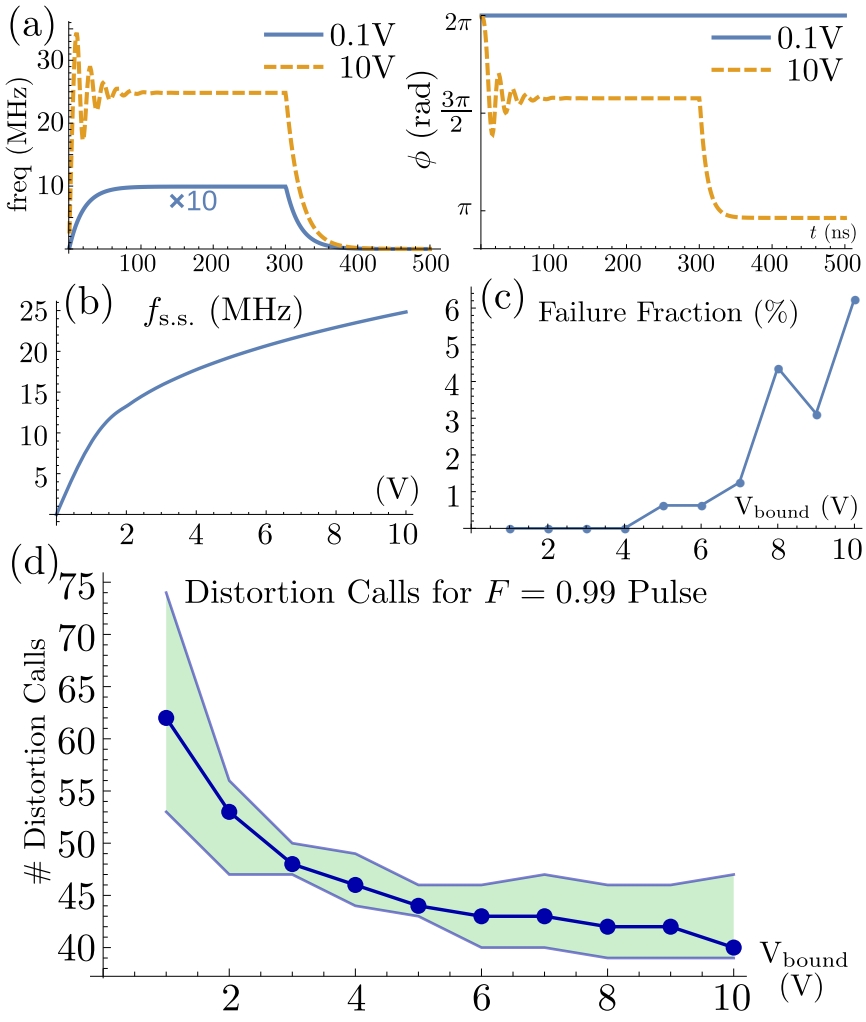

Figure 5.3: (a) Response from the same resonator to a top-hat input pulse of length 300ns with an amplitude in both a linear (0.1V) and nonlinear (10V) regime. The amplitude of the 0.1V pulse is multiplied by 10 to make it visible. (b) The steady state driving frequency as seen by the spins as a function of the voltage input to the resonator. (c) Out of 160 pulses searched for at each of 10 voltage bounds, V_{bound}, with corresponding total pulse length $T_{\text{pulse}} = \frac{0.25}{f_{\text{s.s.}}}$, the fraction that failed to reach $F = 0.99$ before the step size was effectively zero, and (d) the median number of calls made to the distortion function g along with the 16% and 84% quantiles during the gradient ascent for those pulses which did reach $F = 0.99$.

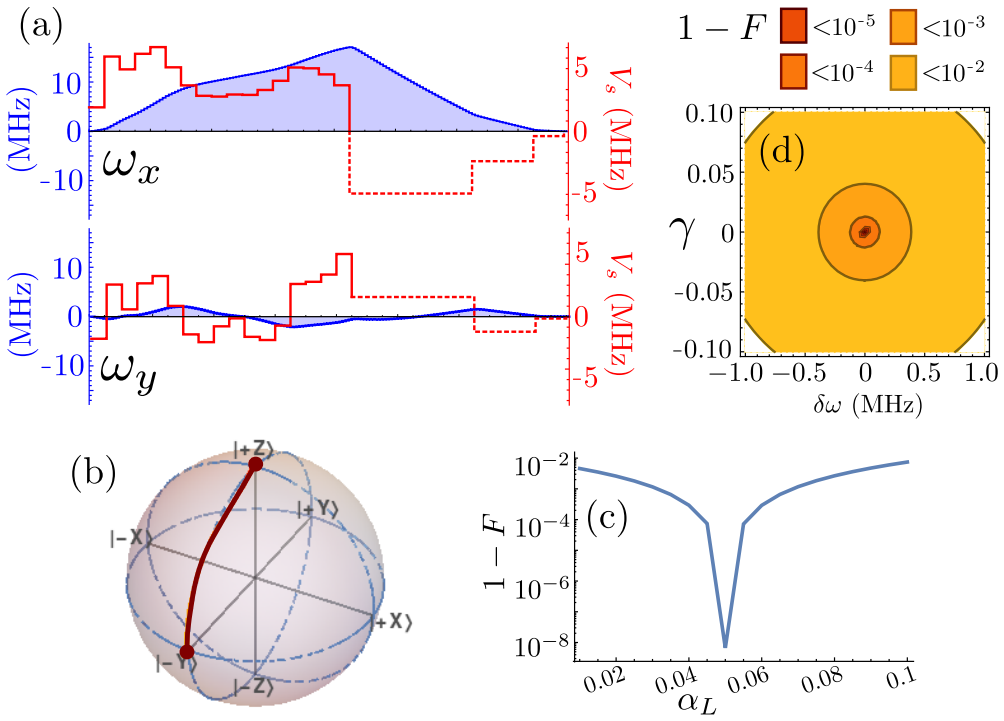

Figure 5.4: (color online) (a) Example of a $\pi/2)_x$ pulse generated for the matched nonlinear resonator circuit. The driving term ($\vec{p}$) is shown in red, while the distorted pulse ($\vec{q}$) is shown in blue. The dashed segments are the ringdown compensation steps. (a) The trajectory of the state $|0\rangle$ under this pulse is shown on the Bloch sphere, and (c-d) the average fidelity is plotted for different values of α_L, $\delta\omega$, and γ.

favour of taking a larger number of ascent steps that are still generally uphill by using the approximation

$$\left.\frac{\partial g_{m,l}}{\partial p_{n,k}}\right|_{\vec{p}} \approx \left[g(\epsilon\vec{e}_{n,k})/\epsilon\right]_{m,l}. \tag{5.38}$$

These quantities may be precomputed prior to gradient ascent and therefore only add a constant to the computation time. Exact partial derivatives may be computed for a cost that scales as $K \cdot N$ and whose implementation can be parallelized, as derived in Section 5.6.3.

In Figure 5.4, we show an example of a GRAPE-optimized pulse for $U = \frac{\pi}{2})_x$, with the circuit of Figure 5.2 used as a distortion operator. There are 16 times steps of length 0.5ns shown as a solid red step function. The pulse has been made to be robust to static uncertainty in the Hamiltonian parameters $\delta\omega$ and γ and the nonlinearity parameter α_L. Since the circuit has a high quality factor, it would take many times the length of the pulse for the ringdown tail to decay to zero. We therefore utilize an active ringdown suppression scheme with three compensation steps of lengths 4ns, 2ns, and 1ns. This is a generalization of ringdown suppression in linear circuits [16, 125, 91] with details presented in Section 5.6.4. Further, a pseudo-code implementation including ringdown suppression is listed in Algorithm 4.

Having demonstrated the ability to find a robust gate in the presence of our nonlinear distortion operator, we now study the effect it has on the control landscape. It would perhaps be anticipated that, in presence of a non-trivial distortion operator, finding optimal solutions would become more expensive, measured in the number of steps taken by the optimizer. Therefore, a trade-off between computational cost and gate time length could reasonably be expected. We perform a numerical study to examine this relationship.

We bound the allowed input power to the resonator used by the GRAPE algorithm by 10 different voltages, 1 V to 10 V, where 1 V is on the edge of the linear regime, and 10 V is highly nonlinear. In analogy to the numerical control landscape experiments performed in Reference [127], for each of these bounds, we attempt to compute a fidelity $F = 0.99$ $\frac{\pi}{2})_x$ pulse 160 times, with a different random initial guess each time. The total length of the pulse is set to $T_{\text{pulse}} = \frac{0.25}{f_{\text{s.s.}}}$ where $f_{\text{s.s.}}$ is the steady state driving frequency of the resonator at the corresponding voltage bound. The number of time steps is held constant at $N = 16$ for each trial. The gradient approximation from (5.38) is used. On each trial, we count the number of times the distortion function g is called. The results are shown in Figure 5.3 where it seen that the number of calls actually tends to decrease as the allowed nonlinearity is increased, indicating that the control landscape does not become more difficult to navigate.

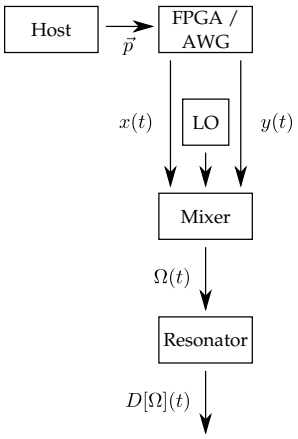

Figure 5.5: Configuration of the microwave mixing components in relation to pulse distortion operators.

5.6.2 The Rotating Frame of a Circuit

A spin in a large static magnetic field $\gamma B_0 = \omega_0$ with a transverse time dependent field $\gamma B_1(t) = 2\Omega(t)$ will evolve under the Hamiltonian

$$H = \frac{\omega_0}{2}\sigma_z + 2\frac{\Omega(t)}{2}\sigma_x. \tag{5.39}$$

Since ω_0 is taken to be the dominant term, in analogy to a wide range of experimental settings, it is helpful to enter the rotating frame generated by $H_{\text{rot}} := \omega_r\sigma_z/2$, where $\omega_r := [\omega_0 + \delta\omega]$. In doing so, we will suppose that

$$\begin{aligned}\Omega(t) &= \omega_1(t)\cos(\omega_r t + \phi(t)) \\ &= \omega_1(t)\cdot\frac{1}{2}\left(\mathrm{e}^{\mathrm{i}t(\omega_r t+\phi(t))} + \mathrm{e}^{-\mathrm{i}t(\omega_r t+\phi(t))}\right),\end{aligned}$$

representing that Ω is produced by mixing a modulating signal with an oscillator at $\omega_0+\delta\omega$ (see Figure 5.5). We will later relate this model to the in-phase and quadrature control fields.

In the frame of H_{rot}, the effective Hamiltonian H_{eff} is given by

$$\begin{aligned}H_{\text{eff}}(t) &= \mathrm{e}^{+\mathrm{i}H_{\text{rot}}t}H(t)\mathrm{e}^{-\mathrm{i}H_{\text{rot}}t} - H_{\text{rot}} \\ &= \Omega(t)\mathrm{e}^{+\mathrm{i}H_{\text{rot}}t}\sigma_x\mathrm{e}^{-\mathrm{i}H_{\text{rot}}t} - \delta\omega\,\sigma_z \\ &= \Omega(t)[\cos(\omega_r t)\sigma_x - \sin(\omega_r t)\sigma_y] - \frac{\delta\omega}{2}\,\sigma_z.\end{aligned}$$

Discarding the terms which oscillate at $2\omega_r$ (that is, the rotating wave approximation), we can rewrite this in terms of ω_1 and ϕ instead,

$$H_{\text{eff}}(t) = \frac{\omega_1(t)}{2}[\cos\phi(t)\sigma_x + \sin\phi(t)\sigma_y] - \frac{\delta\omega}{2}\,\sigma_z.$$

Since Ω is often produced by mixing, as noted above, we can also represent the rotating frame control using two real control fields $\omega_x(t)$ and $\omega_y(t)$,

$$\omega_x(t) = \text{re}[\omega_1(t)\text{e}^{\text{i}\phi(t)}] \qquad \omega_y(t) = \text{im}[\omega_1(t)\text{e}^{\text{i}\phi(t)}]. \tag{5.40}$$

We can now relate to the circuit dynamics. The nutation frequency of the spins is proportional to the magnetic field generated by the inductor via the gyromagnetic ratio, which is in turn proportional to the current passing through the inductor, thus

$$\Omega(t) = \eta I_L(t), \tag{5.41}$$

where the exact value of η depends on the relevant geometry. This along with Equation 5.40 leads us to the relationships

$$\begin{aligned}\omega_x(t) = \text{re}[\omega_1(t)\text{e}^{\text{i}\phi(t)}] = \text{re}[\Omega(t)\text{e}^{-i\omega_r t}] = \eta\,\text{re}[I_L\text{e}^{-i\omega_r t}] = \eta\,\text{re}[\tilde{I}_L]\\ \omega_y(t) = \text{im}[\omega_1(t)\text{e}^{\text{i}\phi(t)}] = \text{im}[\Omega(t)\text{e}^{-i\omega_r t}] = \eta\,\text{im}[I_L\text{e}^{-i\omega_r t}] = \eta\,\text{im}[\tilde{I}_L],\end{aligned} \tag{5.42}$$

where $\tilde{I}_L(t) = \text{e}^{-\text{i}\omega_1 t}I_L(t)$ and we have ignored pieces rotating at $2\omega_r$ in the calculation. We recover the expression in the main body, such that by solving the differential equation $\dot{\vec{x}}(t) = A(\vec{x})\vec{x} + \tilde{V}_s(t)\vec{b}$ for a complex driving function $\tilde{V}_s(t)$, we can find the rotating-frame unitary action $U(t) = \mathcal{T}\exp(-\text{i}\int_0^t H_{\text{eff}}(t))$ of that pulse.

5.6.3 The Distortion Operator of a Nonlinear Resonator Circuit

We define the distortion operator $g : \mathbb{R}^N \otimes \mathbb{R}^2 \to \mathbb{R}^M \otimes \mathbb{R}^2$ corresponding to the non-linear resonator circuit shown in the main body. Note that the input to g will have units of volts, and the output of g will have units of Hz. We use a uniform input discretization time dt and a uniform output discretization time δt. Given an input pulse $\vec{p} \in \mathbb{R}^N \otimes \mathbb{R}^2$, we define the complex vector $\tilde{p} \in \mathbb{C}^N$ by $\tilde{p}_n = p_{n,1} + ip_{n,2}$. Setting $n(t) = \left\lceil \frac{t}{dt} \right\rceil$, we define

$$\alpha(t) = \tilde{p}_{n(t)} \tag{5.43}$$

in the case where we add no rise time to the forcing term, or

$$\alpha(t) = \tilde{p}_{n(t)-1} + (\tilde{p}_{n(t)} - \tilde{p}_{n(t)-1})(1 - e^{-\frac{t-n(t)dt}{\tau_r}}). \tag{5.44}$$

in the case where we include a finite rise time on the forcing with timescale τ_r. Note that we must have $\tau_r \ll dt$ for the function α to be (approximately) continuous. This limitation could easily be overcome with a more sophisticated definition of α, for example, by using a convolution operator. Also note we are using the convention $\tilde{p}_i = 0$ for $i < 1$ or $i > N$.

Now we solve the vector differential equation

$$\dot{x} = A(x)x + \alpha(t)b, \tag{5.45}$$

where

$$x = \begin{bmatrix} \tilde{I}_L \\ \tilde{V}_{C_m} \\ \tilde{V}_{C_t} \end{bmatrix} \qquad A(x) = \begin{bmatrix} -\frac{R}{L} & 0 & \frac{1}{L} \\ 0 & \frac{-1}{R_L C_m} & \frac{1}{R_L C_m} \\ \frac{-1}{C_t} & \frac{-1}{R_L C_t} & \frac{1}{R_L C_t} \end{bmatrix} - i\omega_r \mathbb{I} \qquad b = \begin{bmatrix} 0 \\ \frac{1}{R_L C_m} \\ \frac{1}{R_L C_t} \end{bmatrix} \tag{5.46}$$

with Mathematica 10's function NDSolve. By default, this function dynamically chooses the step size and switches between solvers, of both the implicit and explicit time stepping variety, to ensure that the solution is accurate and stable. An interpolating function for $\tilde{I}_L(t)$ is returned. Recalling Equation 5.42, we sample the real and imaginary parts of $\tilde{I}_L(t)$ at a rate δt to obtain $\vec{q}$:

$$\begin{aligned} q_{m,1} &= \eta \,\mathrm{re}\, \tilde{I}_L(\delta t(m - 1/2)) \\ q_{m,2} &= \eta \,\mathrm{im}\, \tilde{I}_L(\delta t(m - 1/2)) \end{aligned} \tag{5.47}$$

To populate elements of the Jacobian tensor $J_{\vec{p}}(g)$, we are interested in approximating partial derivatives of the form

$$\frac{\partial g_{m,l}}{\partial p_{n,k}} \tag{5.48}$$

where g is the distortion corresponding to the non-linear resonator circuit. The most straight forward way of approximating such would be to use a central difference formula

$$\frac{\partial g_{m,l}}{\partial p_{n,k}} \approx \left[\frac{g(\vec{p} + \epsilon \vec{e}_{n,k}) - g(\vec{p} - \epsilon \vec{e}_{n,k})}{2\epsilon} \right]_{m,l}, \tag{5.49}$$

where $\vec{e}_{n,k}$ is the unit vector in the $\{n, k\}$ direction, and $\epsilon > 0$ is a small number that is greater than the precision of the DE solver. Such an approximation would require $2NK$ calls to the DE solver. It is also numerically unstable as it involves the difference of two numerical DE solutions whose forcing terms are only slightly different; ϵ would have to be very carefully tuned and may have no reliable value at all, especially when searching for high fidelity pulses.

If we consider the approximation $g(\vec{p} \pm \epsilon\vec{e}_{n,k}) \approx g(\vec{p}) \pm g(\epsilon\vec{e}_{n,k})$ the central difference reduces to

$$\frac{\partial g_{m,l}}{\partial p_{n,k}} \approx [g(\epsilon\vec{e}_{n,k})/\epsilon]_{m,l}\,. \tag{5.50}$$

which is the approximation quoted in the main body. Importantly, this approximation does not depend on the current pulse $\vec{p}$ and can therefore be precomputed eliminating the $2NK$ calls to g (i.e. DE solver calls) per ascension step.

An exact method to compute these partial derivatives is derived below, which will take $N * K + 1$ calls to the DE solver to compute the entire Jacobian matrix. Begin with the resonator differential (Equation 5.45)

$$\dot{x} = A(x)x + \alpha(t)b. \tag{5.51}$$

As discussed, we have

$$\begin{aligned}[g(\vec{p})]_{m,1} &= \eta\,\mathrm{re}\,\tilde{I}_L(t_m) \equiv h_1(x(t_m)) \\ [g(\vec{p})]_{m,2} &= \eta\,\mathrm{im}\,\tilde{I}_L(t_m) \equiv h_2(x(t_m))\end{aligned} \tag{5.52}$$

where $t_m = (m - 1/2)\delta t$. Thus it is clear that the difficult part of computing $\frac{\partial g_{m,l}}{\partial p_{n,k}}$ is computing $\frac{\partial \tilde{I}_L}{\partial p_{n,k}}$, or more generally $\frac{\partial x}{\partial p_{n,k}}$.

We derive a set of $K * N = 2N$ secondary partial differential vector equations whose time sampled solutions produce the necessary partial derivatives. To do this we just take the partial derivative $\frac{\partial}{\partial p_{n,k}}$ of Equation 5.45, which gives as the l^{th} component of the $(n,k)^{\text{th}}$ equation

$$\frac{\partial}{\partial p_{n,k}}\frac{\partial x_l}{\partial t} = \frac{\partial A_{l,l'}}{\partial x_{l''}}\frac{\partial x_{l''}}{\partial p_{n,k}}x_{l'} + [A(x)]_{l,l'}\frac{\partial x_{l'}}{\partial p_{n,k}} + T_{n,k}b_l \tag{5.53}$$

where Einstein summation notation is used and (in the case $\tau_r = 0$)

$$T_{n,k}(t) = \begin{cases} 0 & 0 \le t \le dt \\ \vdots & \\ \delta_{1,k} + i\delta_{2,k} & (n-1)dt \le t \le ndt\,. \\ \vdots & \\ 0 \le t \le Ndt & \end{cases} \tag{5.54}$$

Denote

$$y_{n,k}(t) = \frac{\partial x}{\partial p_{n,k}}(t)$$
$$[A'(x)]_{l,l''} = \frac{\partial A_{l,l'}}{\partial x_{l''}} x_{l'} \tag{5.55}$$

and commuting the partial derivatives, the components Equation 5.53 can be rewritten as the non-linear vector PDE

$$\dot{y}_{n,k} = [A'(x) + A(x)]y_{n,k} + T_{n,k}(t)b \tag{5.56}$$

where $x(t)$ is the solution to Equation 5.45. Therefore once $x(t)$ has been computed, we can plug it into each of the DEs for $y_{n,k}$, solve them with the initial condition $y_{n,k}((n-1)dt) = 0$ (by causality $y_{n,k} = 0$ for $t < (n-1)dt$) and we arrive at the exact formula

$$\begin{aligned}\frac{\partial g_{m,l}}{\partial p_{n,k}} &= \left.\frac{\partial h_l(x(t))}{\partial p_{n,k}}\right|_{t=t_m} \\ &= \left.\frac{\partial h_l}{\partial x_{l'}}\frac{\partial x_{l'}}{\partial p_{n,k}}\right|_{t=t_m} \\ &= \frac{\partial h_l}{\partial x_{l'}}[y_{n,k}(t_m)]_{l'}\end{aligned} \tag{5.57}$$

where h_l was defined implicitly in Equation 5.52 and each $\frac{\partial h_l}{\partial x_{l'}}$ is easy to compute.

If we take the Taylor series of $A(x)$ about $x = 0$, we have

$$A(x) = A_0 + A_1(x) + A_2(x) + \dots \tag{5.58}$$

where each A_p is a matrix polynomial in the coordinates of x with all terms having order exactly p. The 0^{th} order approximation of Equation 5.56 gives

$$\dot{y}_{n,k} = A_0 y_{n,k} + T_{n,k}(t)b. \tag{5.59}$$

In this form we see that $y_{n,k}$ is just the same as x where the DE for x, Equation 5.56, has been linearized and the forcing is the top hat $T_{n,k}$: $y_{n,k} = x|_{A=A_0,\alpha=T_{n,k}}$. The linearization condition $A = A_0$ is approximately the same as the guarantee $\|A(x) - A_0\| \ll 1$, which can be met by setting $\alpha = \epsilon T_{n,k}$ with ϵ chosen so that $\left\|A(\frac{\epsilon\|b\|}{\|A_0\|}) - A_0\right\| \ll 1$. Therefore the zeroth order approximation to the Jacobian is

$$\frac{\partial g_{m,l}}{p_{n,k}} \approx \frac{g(\epsilon e_{n,k})}{\epsilon}. \tag{5.60}$$

which is a somewhat more satisfying derivation of Equation 5.50.

5.6.4 Ringdown Compensation

A resonator or cavity with a large quality factor Q will store energy for times that are long compared to the time steps that are used in pulse design. If this effect is not included in optimization by integrating the distortion differential equation for a sufficient period, then the integrated action of the pulse on the quantum system will not be accurate. This can be dealt with by defining the image of the distortion operator to represent a longer time interval than the domain, but this is inconvenient in experimental practice, where we would like to turn off a pulse quickly. Thus, a better alternative is to actively compensate for the ringdown introduced by large Q, and to demand that the distorted pulse goes to zero at a given time step.

For a resonator with only linear elements, this problem has been solved [16] by appealing to the transfer function $h : \mathbb{R}^M \to \mathbb{R}^K$,

$$g[\vec{p}] = f_1[f_2(\vec{p}) \star h] \tag{5.61}$$

where $\star$ is the convolution operator. For the case $M = K = 1$, the transfer function takes on the simple form

$$h(t) = A\mathrm{e}^{-t/\tau_c} \tag{5.62}$$

for some amplitude A and where $\tau_c = Q/\omega_0$ is a time constant. In this case, it is easy to append an additional pulse segment of amplitude

$$p_{K+1} = -A\frac{g[\vec{p}]_m}{e^{\delta t/\tau_c} - 1}, \tag{5.63}$$

where m is a time step index such that $t_m = t_K$.

In the nonlinear case, Q, ω_0 and A are not constant, but depend on $\vec{p}$, and so more attention is required. One solution is to modify the performance continuous to include the demand that the ringdown go to zero by defining

$$\Phi'_g(\vec{p}) := \Phi_g(\vec{p}) - \Omega_g(\vec{p}) = (\Phi - \Omega) \circ g. \tag{5.64}$$

For ringdown compensation,

$$\Omega_{\mathrm{RD}} := \sum_{m=m_0}^{M} |p_m|^2, \tag{5.65}$$

where m_0 is the time step index at which we start demanding that the solution goes to zero. The derivatives of this function are easily found, such that $\vec{\nabla}\Phi'$ is easy to compute given

$\vec{\nabla}\Phi$ and $J(g)$. Since a solution that both has high fidelity with a unitary target and admits ringdown compensation can be hard to find, we use the ringdown-compensation method found in the next section to generate *initial guesses* which result in a small penalty $\Phi'_g(\vec{p})$.

Another solution is to include ringdown suppression in the distortion operator g itself. That is, given an input pulse $\vec{p}$, the forcing term α now includes not only steps taken directly from $\vec{p}$, but also additional steps which are chosen (according to the results from the next section) to eliminate the energy from the cavity in a short period of time. This was the method employed in the example above.

5.6.5 Eliminating energy from a non-linear resonator

Here, we derive a scheme to calculate the values of compensation steps to append to a pulse which act to remove the energy from a resonator on a timescale shorter than the ringdown time.

Write the equation of the circuit as

$$\dot{x} = Ax + \alpha b \tag{5.66}$$

where x is a vector of state variables for the circuit, A is a matrix describing the circuit without forcing, b is the forcing direction of the circuit, and α is a controllable scalar which sets the magnitude of the forcing. We assume that we have already entered the frame rotating at the resonance frequency so that all quantities are complex, where real quantities correspond to in-phase components, and imaginary quantities correspond to quadrature components. Note that for a non-linear circuit, A will depend on the state of the system, that is, $A = A(x)$. Moreover, α can be time dependent, $\alpha = \alpha(t)$.

Our goal is as follows: start with an undistorted pulse $\vec{p}_0$ and append n_{rd} steps of length dt_{rd} to form the undistorted pulse $\vec{p} = [\vec{p}_0, \vec{p}_{\text{rd}}]$ which cause the distorted pulse $g(\vec{p})$ to have near zero amplitude at the end of the last time step. To simplify our task, we make the approximation that A remains constant during each of the compensation steps, taking on a value corresponding to the state x at the end of the previous time step.

The general solution to Equation 5.66 is given by

$$x(t) = e^{tA}x_0 + \int_0^t \alpha(s)e^{(t-s)A}b\ ds. \tag{5.67}$$

Substituting our continuous forcing solution from Equation 5.44 and translating the time coordinate so that $t = 0$ corresponds to the transition from the $(n-1)^{\text{th}}$ to the n^{th} gives

the solution

$$\begin{aligned} x(t) &= e^{tA}x_0 + e^{tA}\left[\int_0^t e^{-sA}\left(\tilde{p}_{n-1} + (\tilde{p}_n - \tilde{p}_{n-1})(1 - e^{-s/\tau_r})\right) ds\right] b \\ &= e^{tA}x_0 + \left[\tilde{p}_n A^{-1}(e^{tA} - \mathbb{I}) - (\tilde{p}_n - \tilde{p}_{n-1})\left(A + \mathbb{I}/\tau_r\right)^{-1}\left(e^{tA} - e^{-t/\tau_r}\mathbb{I}\right)\right] b \end{aligned} \tag{5.68}$$

in the region $t \in [0, dt_{\text{rd}}]$. We wish to drive the state of the system, x, to 0. Therefore, let's try to demand that at time $t = dt_{\text{rd}}$, x becomes some fraction of its value at the end of the $(n-1)^{\text{th}}$ step, so that $x(dt_{\text{rd}}) = rx_0$ for some $r \in [0, 1]$. We refrain from setting $r = 0$ when x is large because if x changes too much in the time span dt_{rd} our approximation of constant A will break down. Since all we can do is change the value of $\tilde{p}_n$, the equality $x(dt_{\text{rd}}) = rx_0$ won't in general be achievable. We therefore instead minimize the quantity

$$\beta(\tilde{p}_n) = \|P(x(dt_{\text{rd}}) - rx_0)\|_2 \tag{5.69}$$

where P is a positive semi-definite matrix which relates the importance of minimizing certain state variables over others. This quantity can be rewritten as

$$\begin{aligned} \beta(\tilde{p}_n) &= \|w - \tilde{p}_n v\|_2 \\ w &= P\left[(e^{tA} - r\mathbb{I})x_0 + \tilde{p}_{n-1}\left(A + \mathbb{I}/\tau_r\right)^{-1}\left(e^{tA} - e^{-t/\tau_r}\mathbb{I}\right) b\right] \\ v &= P\left[\left(A + \mathbb{I}/\tau_r\right)^{-1}\left(e^{tA} - e^{-t/\tau_r}\mathbb{I}\right) - A^{-1}(e^{tA} - \mathbb{I})\right] b \end{aligned} \tag{5.70}$$

In this form it is clear that $\beta(\tilde{p}_n)$ is minimized when $\tilde{p}_n$ is chosen to be the complex projection amplitude of the vector w onto v:

$$\tilde{p_n} = \frac{\langle v, w\rangle}{\langle v, v\rangle}. \tag{5.71}$$

For reference, note that in the limit $\tau_r \to 0$, the vectors v and w simplify to

$$\begin{aligned} w &= P(e^{tA} - r\mathbb{I})x_0 \\ v &= -PA^{-1}(e^{tA} - \mathbb{I})b. \end{aligned} \tag{5.72}$$

5.7 Further Examples of Distortions

In this section we provide a range of useful distortions and some useful formulas along the way.

5.7.1 Composition

First, it is worth noting the (perhaps obvious) fact that composing distortion operators is easily implemented in this framework. Suppose we have characterized the first half of our classical hardware with the discrete distortion operator $g_1 : \mathbb{R}^N \otimes \mathbb{R}^K \to \mathbb{R}^{N'} \otimes \mathbb{R}^{K'}$ and the second half with the operator $g_2 : \mathbb{R}^{N'} \otimes \mathbb{R}^{K'} \to \mathbb{R}^M \otimes \mathbb{R}^L$. We have been careful to make the domain of g_2 compatible with the range of g_1. Then the total distortion operator is given by the composition

$$g = g_2 \circ g_1 : \mathbb{R}^N \otimes \mathbb{R}^K \to \mathbb{R}^M \otimes \mathbb{R}^L \tag{5.73}$$

$$\vec{p} \mapsto g_2(g_1(\vec{p})). \tag{5.74}$$

To find the Jacobian matrix of g at point $\vec{p}$ we just need to use the multivariate chain rule,

$$J_{\vec{p}}(g) = J_{g_1(\vec{p})}(g_2) \cdot J_{\vec{p}}(g_1), \tag{5.75}$$

or in terms of indices,

$$[J_{\vec{p}}(g)]_{m,l,n,k} = \sum_{n'=1}^{N'} \sum_{k'=1}^{K'} [J_{g_1(\vec{p})}(g_2)]_{m,l,n',k'} [J_{\vec{p}}(g_1)]_{n',k',n,k}. \tag{5.76}$$

5.7.2 Transfer Functions and Convolutions

Linear electronic systems can be fully described by a *transfer function* $\Phi(\omega)$. This function gives a simple relationship between an input tone $X(\omega)$ at frequency ω and the resulting output tone $Y(\omega)$, namely $Y(\omega) = \Phi(\omega)X(\omega)$. The magnitude of $H(\omega)$ represents the gain or attenuation, and the argument represents the phase shift. Taking the inverse Fourier transform of this equation yields the convolution, $y(t) = (\phi \star x)(t)$, where we have chosen the non-unitary angular frequency convention for the Fourier transform, that is,

$$\mathcal{F}\left[\phi(t)\right](\nu) = \Phi(\nu) = \int_{-\infty}^{\infty} \phi(t) \mathrm{e}^{-i\nu t} dt. \tag{5.77}$$

The transfer function may be measured experimentally [162, 71, 145], or may be computed if a good model of the system is known. In the main body, the formula for a discrete convolution operator arising from a time domain transfer function ϕ is shown. Here, we derive it in slightly more detail.

To begin, the time domain version of the transfer function $\phi(t)$ results in the distortion operator f defined as

$$\beta(t) = f(\alpha)(t) = (\phi \star \alpha)(t) = \int_{-\infty}^{\infty} \phi(t-\tau) \cdot \alpha(t) d\tau. \tag{5.78}$$

Note that here $\phi(t)$ is a function whose values are $L \times K$ matrices. In the usual context where $K = L$ and the $k'th$ output channel is mostly a distorted version of the $k'th$ input channel, the the diagonals of $\phi(t)$ represent channel-wise distortions, and the off-diagonals represent cross contamination between channels.

Discretizing the input we get

$$f(f_2(\vec{p}))(t) = \sum_{n=1}^{N} \int_{(n-1)dt}^{ndt} \phi(t-\tau) \cdot \vec{p}_n \, \mathrm{d}\tau \tag{5.79}$$

$$\equiv \sum_{n=1}^{N} \phi_n(t) \cdot \vec{p}_n, \tag{5.80}$$

which we then discretize the output of, to get

$$[(f_1 \circ f \circ f_2)(\vec{p})]_{m,l} = \sum_{n=1}^{N} \sum_{k=1}^{K} [\phi_n((m-1/2)\delta t)]_{l,k} p_{n,k} \tag{5.81}$$

$$\equiv \sum_{n=1}^{N} \sum_{k=1}^{K} \phi_{m,l,n,k} p_{n,k}, \tag{5.82}$$

for all $1 \leq m \leq M$ and $1 \leq l \leq L$ where

$$\phi_{m,l,n,k} = [\phi_n((m-1/2)\delta t)]_{l,k} \tag{5.83}$$

$$= \int_{(n-1)dt}^{ndt} [\phi((m-1/2)\delta t - \tau)]_{l,k} \, \mathrm{d}\tau \tag{5.84}$$

Letting $\tilde{\phi} \in \mathbb{R}^M \otimes \mathbb{R}^L \otimes \mathbb{R}^N \otimes \mathbb{R}^K$ be the tensor with entries $\phi_{m,l,n,k}$ gives

$$g(\vec{p}) = (f_1 \circ f \circ f_2)(\vec{p}) = \tilde{\phi} \cdot \vec{p} \tag{5.85}$$

as a compact representation of the discretized distortion operator, where the dot represents a contraction over the indices n and k.

The elements of the Jacobian matrix $J(g)$ are now easily computed as

$$[J(g)]_{m,l,n,k} = \frac{\partial(g(\vec{p}))_{m,l}}{\partial p_{n,k}} = \frac{\partial(\tilde{\phi}\cdot\vec{p})_{m,l}}{\partial p_{n,k}} = \frac{\partial \sum_{n'=1}^{N}\sum_{k'=1}^{K}\phi_{m,l,n',k'}p_{n',k'}}{\partial p_{n,k}} = \phi_{m,l,n,k} \tag{5.86}$$

so that

$$J(f_\phi) = \tilde{\phi}. \tag{5.87}$$

5.7.3 Transfer Functions in the Rotating Frame

In the previous subsection we discussed convolutions of the form

$$\beta(t) = f(\alpha)(t) = (\phi \star \alpha)(t) \tag{5.88}$$

and provided relevant derivations of the discrete distortion operator. Since it is often the case that we consider the quantum dynamics of a quantum device in a rotating frame rather than in the lab frame, it is useful to additionally go through how a transfer function applied in the lab frame translates into a distortion in the rotating frame.

We consider a qubit whose Hamiltonian is given by

$$H = \frac{\omega_0}{2}\sigma_z + 2\frac{\Omega(t)}{2}\sigma_x \tag{5.89}$$

where ω_0 is the dominant energy and $\Omega(t) = (K \star \Lambda)(t)$ with $\Lambda(t) = \omega_1(t)\cos(\omega_{\text{rf}}t + \phi(t))$. $K(t)$ is our time domain transfer function and our control fields have a modulation frequency ω_{rf}, with modulation amplitude and phase given by $\omega_1(t)$ and $\phi(t)$, respectively. Note that since $K(t)$ is purely real, $\tilde{K}(\nu)$ must be conjugate symmetric, $\overline{\tilde{K}(-\nu)} = \tilde{K}(\nu)$.

As usual, we enter the rotating frame given by $H_{\text{rot}} = \frac{\omega_{\text{rf}}}{2}\sigma_z$. This produces an effective Hamiltonian

$$\begin{aligned} H_{\text{eff}}(t) &= \mathrm{e}^{itH_{\text{rot}}}(H - H_{\text{rot}})\mathrm{e}^{-itH_{\text{rot}}} \\ &= \frac{\omega_0 - \omega_{\text{rf}}}{2}\sigma_z + \Omega(t)\begin{pmatrix} 0 & \mathrm{e}^{it\omega_{\text{rf}}} \\ \mathrm{e}^{-it\omega_{\text{rf}}} & 0 \end{pmatrix}. \end{aligned} \tag{5.90}$$

Examining the time dependent control pieces in Fourier domain yields

$$\begin{aligned}
\mathcal{F}\left[\mathrm{e}^{it\omega_{\mathrm{rf}}}\Omega(t)\right](\nu) &= \mathcal{F}\left[\Omega(t)\right](\nu-\omega_{\mathrm{rf}}) \\
&= \tilde{K}(\nu-\omega_{\mathrm{rf}})\cdot\mathcal{F}\left[\frac{\omega_1(t)}{2}\left(\mathrm{e}^{i\phi(t)}\mathrm{e}^{it\omega_{\mathrm{rf}}}+\mathrm{e}^{-i\phi(t)}\mathrm{e}^{-it\omega_{\mathrm{rf}}}\right)\right](\nu-\omega_{\mathrm{rf}}) \\
&= \tilde{K}(\nu-\omega_{\mathrm{rf}})\cdot\mathcal{F}\left[\frac{\omega_1(t)}{2}\mathrm{e}^{i\phi(t)}\right](\nu-2\omega_{\mathrm{rf}})+\tilde{K}(\nu-\omega_{\mathrm{rf}})\cdot\mathcal{F}\left[\frac{\omega_1(t)}{2}\mathrm{e}^{-i\phi(t)}\right](\nu) \\
&= \tilde{K}(\nu-\omega_{\mathrm{rf}})\cdot\mathcal{F}\left[\frac{\omega_x(t)-i\omega_y(t)}{2}\right](\nu) \\
\implies \mathrm{e}^{it\omega_{\mathrm{rf}}}\Omega(t) &= \frac{1}{2}\left[\left(K_-\star\omega_x\right)(t)-i\left(K_-\star\omega_y\right)(t)\right]. \qquad (5.91)
\end{aligned}$$

In this derivation, we dropped components of $\omega_1\mathrm{e}^{i\phi}$ at the frequency $\nu-2\omega_{\mathrm{rf}}$ since they are far off resonance; this is the usual rotating wave approximation. We further defined the shifted transfer function $\tilde{K}_-(\nu)=\tilde{K}(\nu-\omega_{\mathrm{rf}})$ and the quadrature control components $\omega_x(t)=\omega_1(t)\cos(\phi(t))$ and $\omega_y(t)=\omega_1(t)\sin(\phi(t))$. A similar calculation results in

$$\mathrm{e}^{-it\omega_{\mathrm{rf}}}\Omega(t)=\frac{1}{2}\left[\left(K_+\star\omega_x\right)(t)+i\left(K_+\star\omega_y\right)(t)\right] \qquad (5.92)$$

where $\tilde{K}_+(\nu)=\tilde{K}(\nu+\omega_{\mathrm{rf}})$. It is important to realize that while $K(t)$ is a real function, in general, neither $K_+(t)$ nor $K_-(t)$ will be. However, it is straight forward to show that $\overline{K_+(t)}=K_-(t)$ using the conjugate symmetry of $\tilde{K}(\nu)$ noted above. Alternatively, this could be shown by observing that our effective Hamiltonian must remain Hermitian. In any case, our derivation is concluded by rewriting our effective Hamiltonian as

$$\begin{aligned}
H_{\mathrm{eff}}(t) &= \frac{\omega_0-\omega_{\mathrm{rf}}}{2}\sigma_z+\Omega(t)\begin{pmatrix}0 & \mathrm{e}^{it\omega_{\mathrm{rf}}}\\ \mathrm{e}^{-it\omega_{\mathrm{rf}}} & 0\end{pmatrix} \\
&= \frac{\omega_0-\omega_{\mathrm{rf}}}{2}\sigma_z+\frac{1}{2}\begin{pmatrix}0 & \overline{K_+\star(\omega_x+i\omega_y)(t)}\\ K_+\star(\omega_x+i\omega_y)(t) & 0\end{pmatrix} \\
&= \frac{\omega_0-\omega_{\mathrm{rf}}}{2}\sigma_z+\frac{1}{2}\left(\mathrm{re}[K_+\star\omega_x](t)-\mathrm{im}[K_+\star\omega_y](t)\right)\sigma_x+\frac{1}{2}\left(\mathrm{re}[K_+\star\omega_y](t)+\mathrm{im}[K_+\star\omega_x](t)\right)\sigma_y \\
&= \frac{\omega_0-\omega_{\mathrm{rf}}}{2}\sigma_z+\frac{\omega_x'(t)}{2}\sigma_x+\frac{\omega_y'(t)}{2}\sigma_y \qquad (5.93)
\end{aligned}$$

where

$$\omega_x'(t)=\mathrm{re}[K_+\star\omega_x](t)-\mathrm{im}[K_+\star\omega_y](t) \qquad (5.94)$$

$$\omega_y'(t)=\mathrm{re}[K_+\star\omega_y](t)+\mathrm{im}[K_+\star\omega_x](t) \qquad (5.95)$$

.

If we let $\alpha(t) = \begin{pmatrix} \omega_x(t) \\ \omega_y(t) \end{pmatrix}$ and $\beta(t) = \begin{pmatrix} \omega_x'(t) \\ \omega_y'(t) \end{pmatrix}$ then in the notation of Equation 5.78 we have

$$\phi(t) = \begin{pmatrix} \operatorname{re} K_+(t) & -\operatorname{im} K_+(t) \\ \operatorname{im} K_+(t) & \operatorname{re} K_+(t) \end{pmatrix} \tag{5.96}$$

and the results from the same section can be directly applied.

5.7.4 Discrete Distortion Operators for Linear Circuits

Applying the distortion operator due to a nonlinear circuit—see Section 5.6—involved solving the circuit equations given a piecewise constant forcing term in the time domain every time the effect on a new pulse needed to be calculated. For linear circuits, that is, circuits where all circuit elements behave linearly, this should not be required. Linear circuits yield a linear differential equation, which is equivalent to a convolution, and therefore, given a linear circuit we expect to be able to cast it in the form

$$\beta(t) = f(\alpha)(t) = (\phi \star \alpha)(t) = \int_{-\infty}^{\infty} \phi(t-\tau) \cdot \alpha(\tau) d\tau. \tag{5.97}$$

as described in Equation 5.78, where the convolution tensor $\phi(t)$ will be some function of the circuit.

To begin, let's consider a circuit which produces the linear differential equation

$$\frac{dx}{dt} = Ax(t) + f(t)b \tag{5.98}$$

where $x(t) : \mathbb{R} \to \mathbb{C}^c$ is the state of the circuit consisting of a list of voltages, currents, fluxes, etc, A is a $c \times c$ matrix describing the circuit, $b \in \mathbb{C}^c$ is the direction of the forcing term, and $f(t) : \mathbb{R} \to \mathbb{C}$ is the amplitude of the forcing. Note that we are in the rotating frame, so that real and imaginary values are interpreted as in-phase and quadrature components, respectively. In the case of a simple matched resonator we have

$$x = \begin{bmatrix} \tilde{I}_L \\ \tilde{V}_{C_m} \\ \tilde{V}_{C_t} \end{bmatrix} \qquad A(x) = \begin{bmatrix} -\frac{R}{L} & 0 & \frac{1}{L} \\ 0 & \frac{-1}{R_L C_m} & \frac{1}{R_L C_m} \\ \frac{-1}{C_t} & \frac{-1}{R_L C_t} & \frac{1}{R_L C_t} \end{bmatrix} - i\omega_r \mathbb{I} \qquad b = \begin{bmatrix} 0 \\ \frac{1}{R_L C_m} \\ \frac{1}{R_L C_t} \end{bmatrix} \tag{5.99}$$

so that $c = 3$ and $f(t)$ has units of (complex) volts. This will be our motivating example, and though we will try to work generally, we will still have some limitations. Mainly, we

will consider only quadrature control drawn from the real and imaginary parts of a single element of $x(t)$.

Thus, in terms of our continuous input and output $\alpha(t)$ and $\beta(t)$ we have the relationships

$$f(t) = W(\alpha(t)) \quad \text{where} \quad W\left(\begin{bmatrix}\alpha_1\\ \alpha_2\end{bmatrix}\right) = \alpha_1 + i\alpha_2 \tag{5.100}$$

$$\beta(t) = V(x(t)) \quad \text{where} \quad V\left(\begin{bmatrix}x_1\\ x_2\\ \vdots\\ x_c\end{bmatrix}\right) = \begin{bmatrix}\operatorname{re} x_1\\ \operatorname{im} x_1\end{bmatrix} \tag{5.101}$$

where we have presumed that the state variable of interest is x_1 in the definition of V. Note that V and W are linear functions (when the scalar field is the real numbers).

The general solution to Equation 5.98 is given by

$$x(t) = e^{tA}x_0 + \int_0^t e^{(t-s)A}bf(s)\ ds. \tag{5.102}$$

and assuming the initial state is $x_0 = 0$, we have in terms of α and β

$$\beta(t) = V\left[\int_0^t e^{(t-s)A}bW[\alpha(s)]\ ds\right] \tag{5.103}$$

which we can hopefully put into the form of Equation 5.78. To this end we define the real functions $d_{ij}(t)$ implicitly by

$$e^{tA}b = \begin{bmatrix}d_{11}(t)\\ d_{21}(t)\\ \vdots\\ d_{c1}(t)\end{bmatrix} + i\begin{bmatrix}d_{12}(t)\\ d_{22}(t)\\ \vdots\\ d_{c2}(t)\end{bmatrix} \tag{5.104}$$

and therefore

$$\begin{aligned}
\beta(t) &= V\left[\int_0^t e^{(t-s)A} bW[\alpha(s)]\ ds\right] \\
&= V\left[\int_0^t \left(\begin{bmatrix} d_{11}(t-s) \\ d_{21}(t-s) \\ \vdots \\ d_{c1}(t-s) \end{bmatrix} + i \begin{bmatrix} d_{12}(t-s) \\ d_{22}(t-s) \\ \vdots \\ d_{c2}(t-s) \end{bmatrix}\right) (\alpha_1(s) + i\alpha_2(s))\ ds\right] \\
&= V\Bigg[\int_0^t \Bigg(\begin{bmatrix} d_{11}(t-s)\alpha_1(s) - d_{12}(t-s)\alpha_2(s) \\ d_{21}(t-s)\alpha_1(s) - d_{22}(t-s)\alpha_2(s) \\ \vdots \\ d_{c1}(t-s)\alpha_1(s) - d_{c2}(t-s)\alpha_2(s) \end{bmatrix} \\
&\qquad + i \begin{bmatrix} d_{12}(t-s)\alpha_1(s) + d_{11}(t-s)\alpha_2(s) \\ d_{22}(t-s)\alpha_1(s) + d_{21}(t-s)\alpha_2(s) \\ \vdots \\ d_{c2}(t-s)\alpha_1(s) + d_{c1}(t-s)\alpha_2(s) \end{bmatrix} \Bigg)\ ds\Bigg] \\
&= \int_0^t \begin{bmatrix} d_{11}(t-s)\alpha_1(s) - d_{12}(t-s)\alpha_2(s) \\ d_{12}(t-s)\alpha_1(s) + d_{11}(t-s)\alpha_2(s) \end{bmatrix}\ ds \\
&= \int_0^t \begin{pmatrix} d_{11}(t-s) & -d_{12}(t-s) \\ d_{12}(t-s) & d_{11}(t-s) \end{pmatrix} \cdot \begin{bmatrix} \alpha_1(s) \\ \alpha_2(s) \end{bmatrix}\ ds. \qquad (5.105)
\end{aligned}$$

This is very similar to the desired Equation 5.78, except that the integration bounds are different. Since we have $\alpha(t) = 0$ for $t < 0$, we can move the lower bound from 0 to $-\infty$ without issue. The upper integration bound is slightly more subtle. To fix it, we must make the definitions

$$\tilde{d}_{11}(t) = \begin{cases} d_{11}(t) & t \geq 0 \\ 0 & t < 0 \end{cases} \quad \text{and} \quad \tilde{d}_{12}(t) = \begin{cases} d_{12}(t) & t \geq 0 \\ 0 & t < 0 \end{cases} \qquad (5.106)$$

which results in

$$\begin{aligned}
\beta(t) &= \int_{-\infty}^t \begin{pmatrix} d_{11}(t-s) & -d_{12}(t-s) \\ d_{12}(t-s) & d_{11}(t-s) \end{pmatrix} \cdot \begin{bmatrix} \alpha_1(s) \\ \alpha_2(s) \end{bmatrix}\ ds \\
&= \int_{-\infty}^{\infty} \begin{pmatrix} \tilde{d}_{11}(t-s) & -\tilde{d}_{12}(t-s) \\ \tilde{d}_{12}(t-s) & \tilde{d}_{11}(t-s) \end{pmatrix} \cdot \begin{bmatrix} \alpha_1(s) \\ \alpha_2(s) \end{bmatrix}\ ds \\
&= \int_{-\infty}^{\infty} \phi(t-s) \cdot \alpha(s) ds \qquad (5.107)
\end{aligned}$$

where

$$\phi(t) = \begin{cases} \begin{pmatrix} [\mathrm{re}\, e^{tA}b]_1 & -[\mathrm{im}\, e^{tA}b]_1 \\ [\mathrm{im}\, e^{tA}b]_1 & [\mathrm{re}\, e^{tA}b]_1 \end{pmatrix} & t \geq 0 \\ 0 & t < 0 \end{cases}. \tag{5.108}$$

Now according to Equation 5.84, the discretized kernel is given by

$$[\tilde{\phi}]_{m,l,n,k} = \int_{(n-1)dt}^{ndt} [\phi((m-1/2)\delta t - \tau)]_{l,k}\, \mathrm{d}\tau. \tag{5.109}$$

Given a particular value of m and n, there are three cases: $(m-1/2)\delta t < (n-1)dt$, $(m-1/2)\delta t \geq ndt$, or $(n-1)dt \leq (m-1/2)\delta t < ndt$. In the first case, we are guaranteed that the integrand is 0. The last two cases can be combined as

$$\begin{aligned} [\tilde{\phi}]_{m,1,n,1} &= [\tilde{\phi}]_{m,2,n,2} \\ &= \int_{(n-1)dt}^{\min(ndt,(m-1/2)\delta t)} [\mathrm{re}\, e^{((m-1/2)\delta t-\tau)A}b]_1 \mathrm{d}\tau \\ &= \mathrm{re}\left[-A^{-1} e^{((m-1/2)\delta t-\tau)A} \Big|_{(n-1)dt}^{\min(ndt,(m-1/2)\delta t)} b \right]_1 \\ &= -\,\mathrm{re}\left[A^{-1} \left(e^{-[\min(ndt,(m-1/2)\delta t)-ndt]A} - e^{dtA} \right) e^{[(m-1/2)\delta t-ndt]A} b \right]_1 \end{aligned} \tag{5.110}$$

which results in the total formula

$$\begin{aligned} [\tilde{\phi}]_{m,1,n,1} &= [\tilde{\phi}]_{m,2,n,2} \\ &= \begin{cases} -\,\mathrm{re}\left[A^{-1} \left(\mathbb{I} - e^{dtA} \right) e^{[(m-1/2)\delta t-ndt]A} b \right]_1 & (m-1/2)\delta t \geq ndt \\ -\,\mathrm{re}\left[A^{-1} \left(\mathbb{I} - e^{[(m-1/2)\delta t-(n-1)dt]A} \right) b \right]_1 & (n-1)dt \leq (m-1/2)\delta t < ndt \\ 0 & (m-1/2)\delta t < (n-1)dt \end{cases} \end{aligned} \tag{5.111}$$

and similarly

$$\begin{aligned} [\tilde{\phi}]_{m,2,n,1} &= -[\tilde{\phi}]_{m,1,n,2} \\ &= \begin{cases} -\,\mathrm{im}\left[A^{-1} \left(\mathbb{I} - e^{dtA} \right) e^{[(m-1/2)\delta t-ndt]A} b \right]_1 & (m-1/2)\delta t \geq ndt \\ -\,\mathrm{im}\left[A^{-1} \left(\mathbb{I} - e^{[(m-1/2)\delta t-(n-1)dt]A} \right) b \right]_1 & (n-1)dt \leq (m-1/2)\delta t < ndt \\ 0 & (m-1/2)\delta t < (n-1)dt \end{cases} \end{aligned} \tag{5.112}$$

This lends itself to an efficient implementation of computing $\tilde{\phi}$ since A only needs to be diagonalized once.

Note that A will have eigenvalues whose real parts are negative, otherwise conservation of energy would be broken. Therefore, for numerical stability, it is critical that whenever A appears in a matrix exponent, the time appearing before it must be positive.

5.7.5 Finite Rise Times

A simple special case of the general convolution discussed in Section 5.7.2 is a rise time acting independently on each control channel. This will cause the rising edge of a square input pulse to be smoothed over with an exponential of time constant τ, and the trailing edge to decay back to zero with an exponential of the same time constant. Such a transfer function arises, for example, from a simple RL circuit, where the time constant is given by $\tau = L/R$.

Given a rise time τ_c^k acting independently on each of the $1 \leq x \leq K = L$ control channels gives the time-domain transfer function as

$$\phi_{l,k}(t) = \begin{cases} \frac{1}{\tau_c^k} e^{-t/\tau_c^k} & l = k \text{ and } t \geq 0 \\ 0 & \text{else} \end{cases} \tag{5.113}$$

which, when the integral from Equation 5.84 is performed and simplified (by Mathematica in this case), results in the discrete convolution

$$[\tilde{\phi}]_{m,l,n,k} = \begin{cases} \delta_{l,k}(e^{dt/\tau_c^k} - 1)e^{\frac{t_{n-1}-t'_m}{\tau_c^k}} & t_n < t'_m \\ \delta_{l,k}\left(1 - e^{\frac{t_{n-1}-t'_m}{\tau_c^k}}\right) & (t_n = t'_m) \vee ((t_n > t'_m) \wedge (n = 1 \vee t_{n-1} < t'_m)) \\ 0 & \text{else} \end{cases} \tag{5.114}$$

for each $1 \leq k \leq K = L$, where $t_n = ndt$ and $t'_m = (m - 1/2)\delta t$. This is illustrated in Figure 5.6 with $K = L = 1$.

Example 5.8. We consider designing a CNOT gate for two qubits with an internal Hamiltonian

$$H = \frac{\omega_1}{2}\sigma_z^1 + \frac{\omega_2}{2}\sigma_z^2 + \frac{J}{4}(\sigma_x^1\sigma_x^2 + \sigma_y^1\sigma_y^2 + \sigma_z^1\sigma_z^2) \tag{5.115}$$

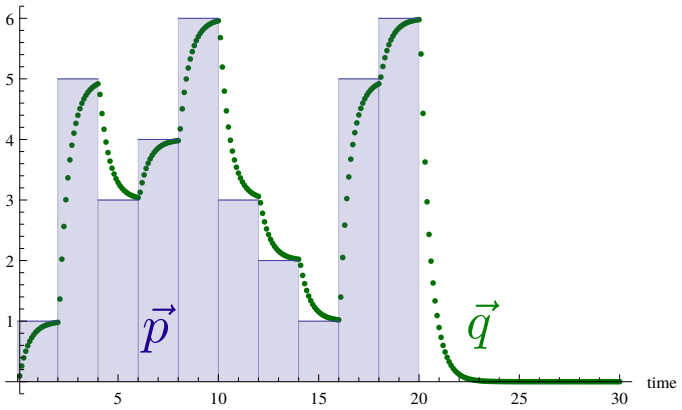

Figure 5.6: An example of the application of a discrete convolution distortion g to an input pulse with $N = 10$ time steps. We have $dt = 2$, and the output space has 20 time steps per input time step, thus $\delta t = 0.1$.

and control Hamiltonians

$$\left\{H_{\mathrm{x}} = \sigma_x^1 + \sigma_x^2, \quad H_{\mathrm{y}} = \sigma_y^1 + \sigma_y^2\right\} \tag{5.116}$$

where $\omega_1 = -2\pi \cdot 15$, $\omega_2 = +2\pi \cdot 15$, $J = 2\pi \cdot 50$, and the amplitudes of the control Hamiltonian are bounded by $2\pi \cdot 50$. This is the style of Hamiltonian found in liquid state NMR homonuclear samples [111]. We use $N = 30$ input time steps of length $dt = 0.005$, and $M = 2N + \lceil 10\tau/dt \rceil$ output time steps of length $dt/2$. Here, τ is the characteristic exponential rise time of both control channels, as defined in Section 5.7.5.

To make the resulting pulse sequence robust against the value τ, we set the utility function of the optimization problem to be a convex combination of utility functions, each with a different value of τ, as explained in Section 5.4. The results are shown in Figure 5.7, where it is seen that $F > 0.99$ is achieved in a region about $\pm 7\%$ around a nominal value of $\tau = 0.005$.

⊣

5.7.6 Crosstalk

Crosstalk is the phenomenon where a signal sent along one control channel is overheard by other control channels. As alluded to earlier, this can be fully accounted for (in the

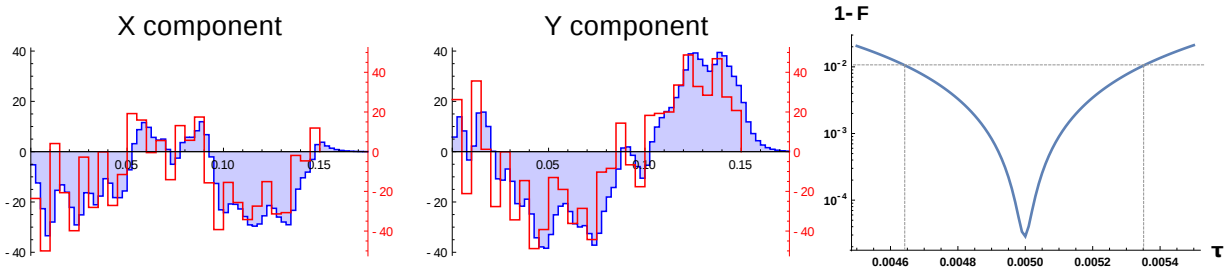

Figure 5.7: The pulse envelope of a CNOT gate at an exponential rise time value of $\tau = 0.005$. The shaded curves (blue) show the distorted output pulse, and the empty curves (red) show the undistorted input pulse. The robustness curve, in terms of one minus average fidelity, is shown to the right as a function of τ.

case of linear controllers) by the off-diagonal elements of the transfer function ϕ. This may be overkill as crosstalk can often be accurately modelled as one control line seeping into each of the other control lines with attenuation factors that are constant in time. See Reference [7] for example, where crosstalk between five coupled superconducting qubits is observed.

To model this situation, we consider that our quantum system has I subsystems (or qubits) each with $L = K$ control channels. The Hamiltonian is given by

$$H = H_{\text{int}} + \sum_{i=1}^{I} \sum_{l=1}^{L} q_{i,l}(t) H_{i,l} \tag{5.117}$$

where H_{int} is the internal Hamiltonian, containing all coupling terms, and $q_{i,l}(t)$ and $H_{i,l}$ are the l^{th} control envelope and Hamiltonian of the i^{th} system. There are now a total of $I \cdot L$ controls, and so instead of indexing the control indeces by single numbers, k and l, we index them by tuples, (i,k) and (i,l). Since this distortion is independent of time, we set $\delta t = dt$ and $M = N$. Ideally, we would have $p_{n,(i,k)} = q_{n,(i,k)}$ representing the fact that the $(i,k)^{\text{th}}$ control signal is sent exactly to the $(i,k)^{\text{th}}$ Hamiltonian at each time step n. With crosstalk included, the $(i,k)^{\text{th}}$ Hamiltonian actually sees a linear combination of each of every control line,

$$q_{n,(i,l)} = \sum_{j=1}^{I} \sum_{k=1}^{K} \chi_{(i,l),(j,k)} p_{n,(j,k)}, \tag{5.118}$$

where $\chi_{(i,l),(j,k)}$ is the fraction of the $(j,k)^{\text{th}}$ control line seen on the $(i,l)^{\text{th}}$ control. More compactly,

$$\vec{q} = g(\vec{p}) = \chi \cdot \vec{p} \tag{5.119}$$

where the dot in this case represents contraction over the indices j and k.

The ideal χ tensor is $\chi_{(i,l),(j,k)} = \delta_{i,j}\delta_{l,k}$. As an example, if there are 4 qubits each with two controls, x and y, then in matrix format the ideal tensor reads

$$\chi_{\text{ideal}} = \begin{array}{cc} & \\ \text{Q1} & \begin{matrix} x \\ y \end{matrix} \\ \text{Q2} & \begin{matrix} x \\ y \end{matrix} \\ \text{Q3} & \begin{matrix} x \\ y \end{matrix} \\ \text{Q4} & \begin{matrix} x \\ y \end{matrix} \end{array} \begin{array}{c} \begin{matrix} Q1 & & Q2 & & Q3 & & Q4 & \end{matrix} \\ \begin{matrix} x & y & x & y & x & y & x & y \end{matrix} \\ \left(\begin{array}{cc|cc|cc|cc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right) \end{array} \tag{5.120}$$

If we add a crosstalk term between adjacent qubits, where an x control only talks to adjacent x controls and similar for y controls, the tensor might look like

$$\chi_{\text{nearest neighbour}} = \begin{array}{cc} & \\ \text{Q1} & \begin{matrix} x \\ y \end{matrix} \\ \text{Q2} & \begin{matrix} x \\ y \end{matrix} \\ \text{Q3} & \begin{matrix} x \\ y \end{matrix} \\ \text{Q4} & \begin{matrix} x \\ y \end{matrix} \end{array} \begin{array}{c} \begin{matrix} Q1 & & Q2 & & Q3 & & Q4 & \end{matrix} \\ \begin{matrix} x & y & x & y & x & y & x & y \end{matrix} \\ \left(\begin{array}{cc|cc|cc|cc} 1 & 0 & 0.2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0.3 & 0 & 0 & 0 & 0 \\ \hline -0.1 & 0 & 1 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0.15 & 0 & 1 & 0 & 0.4 & 0 & 0 \\ \hline 0 & 0 & -0.2 & 0 & 1 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & -0.2 & 0 & 1 & 0 & 0.3 \\ \hline 0 & 0 & 0 & 0 & 0.23 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.7 & 0 & 1 \end{array}\right) \end{array} \tag{5.121}$$

As is clear from Equation 5.119, the Jacobian of this distortion operator is simply given by

$$J_{\vec{p}}(g) = \chi. \tag{5.122}$$

A pulse could be designed to be robust against errors in the crosstalk tensor by including a distribution over crosstalk tensors, or perhaps just a distribution over some of its values, dependently or independently, using the method described in Section 5.4.

Example 5.9. We design a $\pi/2$ gate about x on the third of four qubits arranged in a line, with the other three qubits performing the identity operation. Each qubit has an x and y control, $\{H_{i,x} = \sigma_x^i, H_{i,y} = \sigma_y^i\}$, and the internal Hamiltonian is given by

$$H = \sum_{|i-j|=1} \omega_{ij} \sigma_z^i \sigma_z^j \tag{5.123}$$

where $\omega_{ij} = 2\pi \cdot 20\text{MHz}$ and the input control amplitudes are limited to $2\pi \cdot 40\text{MHz}$. We use a crosstalk tensor

$$\chi = \begin{array}{cc} & \\ \text{Q1} & x \\ & y \\ \text{Q2} & x \\ & y \\ \text{Q3} & x \\ & y \\ \text{Q4} & x \\ & y \end{array} \begin{array}{c} \begin{array}{cccccccc} Q1 & & Q2 & & Q3 & & Q4 & \\ x & y & x & y & x & y & x & y \end{array} \\ \left(\begin{array}{cccccccc} 1 & 0 & 0.3 & 0.001 & 0.05 & 0 & 0.001 & 0 \\ 0 & 1 & 0 & 0.1 & 0 & 0.01 & 0 & 0.001 \\ \hline 0.25 & 0 & 1 & 0 & 0.3 & -0.005 & 0.04 & 0 \\ 0 & 0.2 & 0 & 1 & 0 & 0.4 & 0 & 0 \\ \hline 0 & 0 & 0.2 & 0 & 1 & 0 & -0.2 & 0 \\ 0 & -0.04 & 0 & 0.2 & 0 & 1 & 0 & 0.3 \\ \hline 0.001 & 0 & 0.04 & 0 & 0.3 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0.07 & 0 & -0.3 & 0 & 1 \end{array}\right) \end{array} \tag{5.124}$$

Using this crosstalk distortion tensor, a pulse with average fidelity $F = 0.9999$ was found and is shown in Figure 5.8.

⊣

5.8 Pseudocode for Modified GRAPE

In this Section, we list our modifications to the GRAPE algorithm, including ringdown compensation (though this step may be ignored if irrelevant), in Algorithm 4.

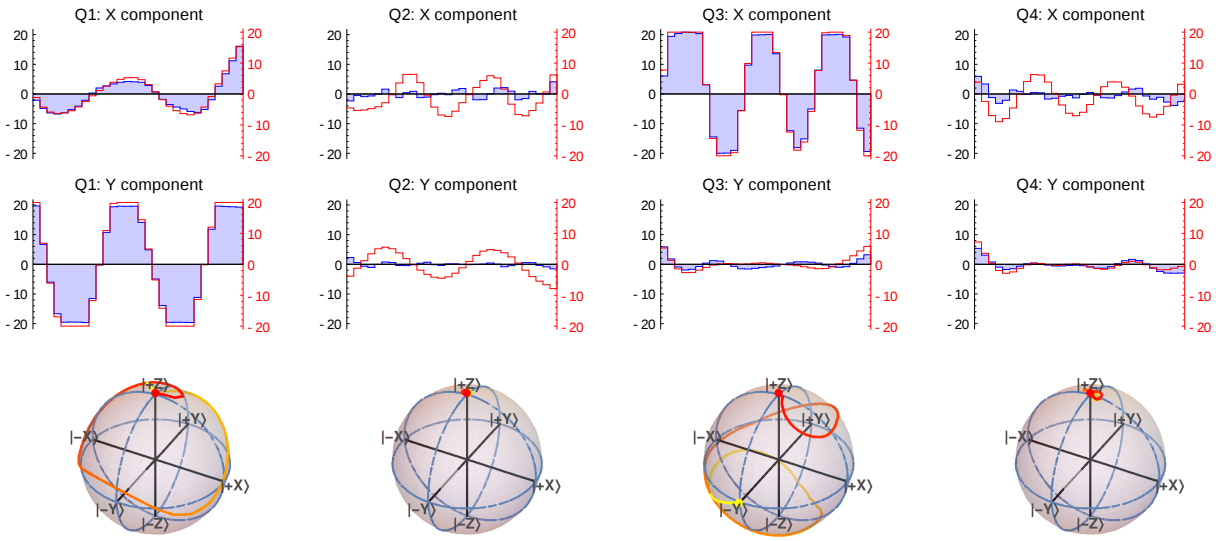

Figure 5.8: (color online) The pulse envelopes and Bloch sphere trajectories of a $\pi/2)_x$ gate on the third qubit. The (unfilled) red curves represent the input pulse, and the (filled) blue curves represent the output pulse seen by the quantum system.

Algorithm 3 Ringdown compensation function used in Algorithm 4.

function RingdownCompensate($\vec{q}$, x_0, n_{steps}, τ_r, r)
  for $i_{\text{step}} \in \{1, \ldots, n_{\text{steps}}\}$ **do**
    $q_0 \leftarrow$ last step in $\vec{q}$
    $w \leftarrow P\left[(e^{tA} - r\mathbb{I})x_0 + q_0\,(A + \mathbb{I}/\tau_r)^{-1}\left(e^{tA} - e^{-t/\tau_r}\mathbb{I}\right)b\right]$
    $v \leftarrow P\left[(A + \mathbb{I}/\tau_r)^{-1}\left(e^{tA} - e^{-t/\tau_r}\mathbb{I}\right) - A^{-1}(e^{tA} - \mathbb{I})\right]b$
    append $\langle v, w\rangle / \langle v, v\rangle$ to $\vec{q}$
  end for
  return $\vec{q}$
end function

Algorithm 4 Modified GRAPE algorithm with ringdown compensation (see Algorithm 3). This pseudocode implements its own conjugate descent for completeness, though in practice, it is likely more sensible to use an gradient ascent method from a scientific library.

Input: Target unitary U_{target}, target fidelity Φ_{target}, distortion operator g, initial pulse $\vec{p}_{\text{init}}$, ringdown compensation steps n_{steps}, ringdown compensation step width τ_r, ringdown compensation ratio $r \in [0, 1]$, [optional] particle approximated distribution $\{(p_s, \vec{a}_s)\}_{s=1}^{S}$.

Output: Pulse $\vec{p}$ such that that $\Phi_g[\vec{p}] \geq \Phi_{\text{target}}$, or $\overline{\Phi}_g[\vec{p}] \geq \Phi_{\text{target}}$ if a list of samples is given.

function BareUtil($\vec{q}$, $\{(p_s, \vec{a}_s)\}_{s=1}^{S}$)
  return $\sum_{s=1}^{S} p_s \Phi_{\vec{a}_s}[\vec{q}]$
end function
function FindPulse(U_{target}, Φ_{target}, g, $\vec{p}$, n_{steps}, τ_r, r [, $\{(p_s, \vec{a}_s)\}_{s=1}^{S}$])
  if distribution samples $\{p_s, \vec{a}_s\}_{s=1}^{S}$ are not given **then**
    $\{(p_s \vec{a}_s)\} \leftarrow \{(1, \vec{0})\}$ ▷ Use a single sample if no samples are given.
  end if
  $\beta \leftarrow 0$
  $\vec{d}' \leftarrow 0$
  $\overline{g} \leftarrow \sum_{s=1}^{S} p_s g[\vec{a}_s]$
  $J_g \leftarrow J(\overline{g})$ ▷ Precalculate the Jacobian of the distortion operator g.
  $u \leftarrow 0$
  while $u \leq \Phi_{\text{target}}$ **do**
    $\vec{q}, x_0 \leftarrow \overline{g}[\vec{p}]$ ▷ Distort the pulse, keeping the final state x_0 of the distortion.
    $\vec{q} \leftarrow$ RingdownCompensate$\left(\vec{q}, x_0, n_{\text{steps}}, \tau_r, r\right)$
    $u \leftarrow$ BareUtil$(\vec{q}, \{(p_s, \vec{a}_s)\}_{s=1}^{S})$
    $\vec{d} \leftarrow \sum_{i=1}^{n} \vec{\nabla}_{\vec{q}} \Phi[\vec{q}|\vec{x}_i] \cdot J_g$ ▷ Use [99] to calculate $\vec{\nabla}_{\vec{q}} \Phi$.
    $\Delta\vec{d} \leftarrow \vec{d} - \vec{d}'$ ▷ Find the conjugate gradient direction.
    $\beta \leftarrow \max\{0, \vec{d} \cdot \Delta\vec{d} / \vec{d}' \cdot \vec{d}'\}$
    $\vec{s} \leftarrow \vec{d} + \beta \vec{d}'$
    $\alpha = \text{argmax}_\alpha$ BareUtil$(\overline{g}[\vec{p} + \alpha\vec{s}], \{(p_s, \vec{a}_s)\}_{s=1}^{S})$ ▷ Perform a line search.
    $\vec{p} \leftarrow \vec{p} + \alpha\vec{s}$ ▷ Update the pulse by the step $\alpha\vec{s}$.
    $\vec{d}' \leftarrow \vec{d}$ ▷ Set the previous gradient to the current and prepare for the next iteration.
  end while
  return $\vec{p}$
end function

Chapter 6

Benchmarking Quantum Devices: Theory and Data Analysis[1]

6.1 Introduction

Accurately characterizing the performance of both large and small quantum devices is vital to ensure that, for example, quantum information processors are reliable and metrology devices are accurate. For critical applications, the reliability of confidence intervals or credible regions for figures of merit is more important than a single-point estimate as there might be practical consequences to over-reporting the performance of a device.

Currently, the only known scalable protocols for characterizing discrete quantum logic gates are randomized benchmarking (RB) [45, 101, 115, 116] and variants thereof, collectively referred to as RB+ (see Table 6.1 for some variants). The standard RB protocol works by applying random sequences of gates that ideally compose to the identity, where the gates form a unitary 2-design [33]. Measuring in the basis of any initial state after applying a random sequence therefore gives an estimate of the survival probability conditioned upon that random sequence. The survival probability averaged over all random sequences of a fixed length decays exponentially with the length, where the decay rate is a

[1] This work is based on Reference [85], done in collaboration with Joel Wallman, Chris Ferrie, and Chris Granade. Joel Wallman was responsible for helping to develop the framework of RB+ protocols, and gave advice on which models to numerically study. Chris Granade and Chris Ferrie independently developed a similar Bayesian model (our efforts were eventually joined), and provided the theory of experiment selection for second moment estimators. This work represents the first complete treatment of the rigorous statistical analysis of RB+ experiment data.

linear function of the average gate fidelity of the overall noise channel. Members of RB+ all have similar structure, modified to suit different goals. RB has been experimentally implemented on a large variety of quantum platforms [18, 129, 176, 7, 131, 190, 157, 49, 77], and is so ubiquitous that its results are often reported with little detail within the context of a larger purpose.

However, these experimental implementations make different ad-hoc statistical assumptions because previous theoretical treatments of RB+ have typically neglected data analysis. The analysis of RB+ experiments is complicated by three factors:

1. every random sequence in a protocol gives rise to a different survival probability, giving rise to a survival distribution for each sequence length;

2. in low- to mid-data regimes, assuming Gaussian errors on either the estimates of the individual survival probabilities or on the mean of the survival distribution through the central limit theorem is dubious; and

3. applying hard physical constraints violates the assumptions of standard statistical fitting routines.

This chapter presents a Bayesian data-processing method that overcomes these difficulties, and that can be applied to all members of RB+. As with any Bayesian approach, the output is a joint posterior distribution over all parameters relevant to the problem. Joint distributions over the parameter(s) of interest can be obtained by marginalizing over nuisance parameters, enabling straight-forward statements like 'under this protocol's model with this prior knowledge, there is a 95% probability that such-and-such parameter is greater than 0.999'. If a point estimate is required for some parameter, the Bayes estimate is just a sum and division away.

6.2 The Framework of RB+

In this section we provide a general framework to rapidly understand and compare the various protocols related to randomized benchmarking. The framework consists of the following six elements, exemplified in Table 6.1:

1. $\mathbb{G}$, Gate Set: the set of R gates $\mathbb{G} = \{G_1, G_2, ..., G_R\}$ used in the protocol, where this set might satisfy specific conditions such as being a group and a unitary 2-design[2];

[2] As a point of practicality, note that gates from $\mathbb{G}$ are often physically implemented by compiling gates

2. $\mathfrak{E}$, Experiment Types: labels for protocols that combine data from multiple sub-protocols, possibly including specification of multiple configurations of preparation and measurement (SPAM), denoted with ρ and E respectively[3];

3. M, the Sequence Length: a positive integer, where $\ell(M, e)$ denotes the exact number of gates from $\mathbb{G}$ needed to construct a sequence at length M under experiment type $e \in \mathfrak{E}$;

4. $J_{M,e}$, Allowable Sequences: a discrete distribution whose sample space is the set of gate-indexing tuples $\{1, ..., R\}^{\ell(M,e)}$, typically uniform on a subset thereof;

5. $\vec{x}_T$, Tying Parameters: the set of parameters that can be learned from the protocol; and,

6. T, Tying Functions: the known dependence of the parameters on the statistics of the measurement data.

For a given sequence of gate indices $\vec{j} = (j_1, ..., j_K)$, define the corresponding ideal gate as

$$\mathcal{G}_{\vec{j}} = \mathcal{G}_{j_K} \cdots \mathcal{G}_{j_2} \mathcal{G}_{j_1}, \tag{6.1}$$

where we use the convention that the scripted version of a letter denoting a unitary operator is the quantum channel which conjugates by that unitary, that is, $\mathcal{G}(\rho) = G\rho G^\dagger$. We write the imperfect implementations of ρ, $\mathcal{G}_{\vec{j}}$, and E as $\tilde{\rho}$, $\tilde{\mathcal{G}}_{\vec{j}}$, and $\tilde{E}$ respectively. The following procedure is then performed experimentally, possibly in a random order to prevent experimental drifts from causing a systematic error:

for each sequence length $M \in \mathbb{M}$ **do**
  for each experiment type $e \in \mathfrak{E}_M$ **do**
    for each sequence $i = 1, 2, 3, ..., I$ **do**
      $\vec{j} \leftarrow \mathrm{RV}\,(J_{M,e})$
      $Q_{M,e,i} \leftarrow \mathrm{RV}\left(\mathrm{Binom}\left(N, \mathrm{Tr}\,\tilde{E}_e \tilde{\mathcal{G}}_{\vec{j}}(\tilde{\rho}_e)\right)\right)$
    end for
  end for
end for

from a smaller generating set of gates that need not share any special properties required by $\mathbb{G}$. Unitary 2-designs are defined in Definition 7.1.

[3]Rather than including SPAM configurations as experiment types, sometimes protocols may instead compile SPAM configurations into the allowable sequences.

where $\mathbb{M} \subset \mathbb{N}$ is some choice of sequence lengths, and (ρ_e, E_E) is the SPAM configuration specified by experiment type e. Here, $\mathrm{RV}(\cdot)$ denotes sampling a random variate from the given distribution, so that $\mathrm{RV}(J_{M,e})$ denotes choosing a random allowable sequence, and $\mathrm{RV}\left(\mathrm{Binom}\left(N, \mathrm{Tr}\,\tilde{E}_e\tilde{\mathcal{G}}_{\vec{j}}(\tilde{\rho}_e)\right)\right)$ corresponds to repeating this experiment N times and summing the resulting 0s and 1s. This binomial model assumes strong measurement with two outcomes. This condition can be loosened, as is done in Section 7.4.2.

In principle the number of random sequences I can depend on M and e, and the number of repetitions N can depend on M, e, and i, and so on, but we avoid this to maintain subscriptural sanity (though our methods will work nonetheless on such ragged structures). For the same reason, we omit any indices which are not relevant to some specific protocol. Generically, this protocol produces the dataset

$$D = (Q_{M,e,i})_{M\in\mathbb{M},e\in\mathfrak{E}_M,1\leq i\leq I}. \tag{6.2}$$

As a concrete example, consider standard RB. Then $\mathbb{G}$ is a unitary 2-design which is also a group. There is only one type of experiment for every sequence length, so $\mathfrak{E} = \{0\}$, with a fixed SPAM configuration $E_0, \rho_0 \approx |0\rangle\langle 0|$. We note that our notation allows, however, for formalizing modifications in which two different final measurements are used to decorrelate preparation and measurement errors [53]. For sequence length M we require $\ell(M, 0) = M + 1$ gates from $\mathbb{G}$, where the extra gate corresponds to the final inversion: the allowable sequences at sequence length M are a uniform distribution of all length $M+1$ gate indices that ideally produce the identity gate, $J_{M,0} = \mathrm{Unif}\left(\{\vec{j} \in \{1, ..., R\}^{M+1} | \mathcal{G}_{\vec{j}} = \mathbb{I}\}\right)$.

Interleaved randomized benchmarking has a similar structure except, for example, that we may have $\mathfrak{E}_M = \{0, 3\}$, where $e = 0$ represents no interleaving, and $e = 3$ represents interleaving the 3^{rd} gate in $\mathbb{G}$. As with standard RB, we have $\ell(M, 0) = M+1$. Interleaved experiments add a fixed gate for every random gate giving us $\ell(M, 3) = 2M + 1$. See Table 6.1 for more examples of RB+ protocols as described by our framework.

6.3 Tying functions

The quantity

$$S_{\vec{j},e} = \mathrm{Tr}\,\tilde{E}_e\tilde{\mathcal{G}}_{\vec{j}}(\tilde{\rho}_e) \in [0, 1] \tag{6.3}$$

is called the *survival probability* of the sequence $\vec{j} \sim J_{M,e}$ at sequence length M for experiment type $e \in \mathfrak{E}$. For a specific noise model and any protocol described by the previous

section, we can consider the discrete *survival distribution* for sequences of length M and experiment type e given by

$$S_{M,e}(q) = \sum_{\vec{j}} \Pr(\vec{j})\delta(q - S_{\vec{j},e}) \tag{6.4}$$

where $\delta(\cdot)$ is the delta mass distribution centered at 0, the sum is over all sequences of the right length, $\vec{j} \in \{1, ..., R\}^{\ell(M,e)}$, and $\Pr(\vec{j})$ is the probability of picking sequence $\vec{j}$ according to the protocol. This distribution has support lying in the unit interval $[0, 1]$.

Such survival distributions depend heavily on the noise model. Complications to the noise model can be introduced successively. See Epstein et al. [46] for a wide set of examples, or Ball et al. [5] for simulations of non-Markovian noise model survival distributions in particular. Letting $\mathcal{E}_{(\cdot)}$ denote a CPTP noise channel and $\vec{j} = (j_1, \ldots, j_K)$ a specific gate sequence, starting with the simplest, the broad categories of noise models are

- Gate-independent noise: For every $\mathcal{G}_r \in \mathbb{G}$ we have $\tilde{\mathcal{G}}_r = \mathcal{E}\mathcal{G}_r$ so that $\tilde{\mathcal{G}}_{\vec{j}} = \mathcal{E}\mathcal{G}_{j_K} \cdots \mathcal{E}\mathcal{G}_{j_1}$.
- Gate-dependent noise: For every $\mathcal{G}_r \in \mathbb{G}$ we have $\tilde{\mathcal{G}}_r = \mathcal{E}_r\mathcal{G}_r$ so that $\tilde{\mathcal{G}}_{\vec{j}} = \mathcal{E}_{j_K}\mathcal{G}_{j_K} \cdots \mathcal{E}_{j_1}\mathcal{G}_{j_1}$.
- Gate- and position- dependent noise: For $\mathcal{G}_r \in \mathbb{G}$ appearing at time k we have $\tilde{\mathcal{G}}_r = \mathcal{E}_{r,k}\mathcal{G}_r$ so that $\tilde{\mathcal{G}}_{\vec{j}} = \mathcal{E}_{K,j_K}\mathcal{G}_{j_K} \cdots \mathcal{E}_{1,j_1}\mathcal{G}_{j_1}$.

We can fine-grain these categories further by specifying the types of channels the errors $\mathcal{E}$ can take, for example, depolarizing, extremal, or unitary rotations. We can also, as a matter of preference, move gate noise to the right side of the ideal operator, or consider both left and right noise. Non-markovian noise models obeying causality are also reasonable to study,

- Non-markovian gate dependent noise: For every $\mathcal{G}_r \in \mathbb{G}$ we have $\tilde{\mathcal{G}}_r = \mathcal{E}\mathcal{G}_r$ where $\mathcal{E}$ depends on both r and the gates preceding $\mathcal{G}_r$, so that $\tilde{\mathcal{G}}_{\vec{j}} = \mathcal{E}_{j_1,\ldots,j_K}\mathcal{G}_{j_K} \cdots \mathcal{E}_{j_2,j_1}\mathcal{G}_{j_2}\mathcal{E}_{j_1}\mathcal{G}_{j_1}$.

The set of allowable sequences $J_{M,e}$ typically grows exponentially with the sequence length M, and numerical evidence suggests that it is reasonable to approximate the survival distribution by a continuous distribution.

RB+ protocols have the shared property of *tying together moments of survival distributions* to extract parameters of interest. For example, the gate-independent noise model

ties the first moments of RB survivals distributions through the relationship[4]

$$\mathbb{E}_{S_M}[q] = (A - B)p^M + B \tag{6.5}$$

where the average gate fidelity of the error map $\mathcal{E}$ is $p + (1-p)/d$, $A = \mathrm{Tr}\,\tilde{E}_0\mathcal{E}(\tilde{\rho}_0)$, and $B = \mathrm{Tr}\,\tilde{E}_0\mathcal{E}(\mathbb{I}/d)$. Note that we have chosen a slightly different parameterization than that of [116], such that the range of valid SPAM parameters is given by $(A, B) \in [0, 1]^2$.

More generally, every protocol will have a function T which ties together the t^{th} moments of the survival distributions through

$$\mathbb{E}_{S_{M,e}}[q^t] = T(t, M, e, \vec{x}_T). \tag{6.6}$$

for some subset of all moments. We call T the *tying function*. Here, $\vec{x}_T$ is a vector of parameters required by the tying function, for instance, $\vec{x}_T = (p, A, B)$ in the case of standard RB. As of this writing, the unitarity protocol is the only protocol which ties together moments past the first [177].

6.4 The Likelihood Function

Let's start with the standard RB protocol in what is known as the 0^{th} order model, as written in Equation 6.5. The parameter of interest is p since it is related to the average gate fidelity of the average error map. Given a dataset D, as defined in Equation 6.2, we are interested in inferring the value of p, with A and B treated as nuisances.

Any inference starts with writing down the likelihood function of the parameter of interest [163], along with nuisance parameters, conditioned on the collected data. The total likelihood will be a product over all sequences lengths and sequence draws. Consider just the factor for the i^{th} draw of length-M, resulting in the binomial outcome $d = Q_{M,e,i} \in \{0, \ldots, N\}$. The likelihood of this outcome, conditional on drawing the particular sequence $\vec{j}$, is given by

$$\mathrm{L}(p, A, B|d, \vec{j}) = \binom{N}{d} q^d (1-q)^{N-d} \tag{6.7}$$

where $q = S_{\vec{j}}$ is the survival probability of sequence $\vec{j}$. The conditional is removed by marginalizing q over the survival distribution,

$$\mathrm{L}(p, A, B|d, M) = \mathbb{E}_{S_M}\left[\binom{N}{d} q^d (1-q)^{N-d}\right]. \tag{6.8}$$

[4]Recall that we omit some indices, here in $S_{M,e}$, for notational convenience. In this case because there is only one experiment type and SPAM setting. Also, the notation $\mathbb{E}_{S_M}[q]$ is the expectation value of the random variable (arbitrarily called q) drawn according to the distribution defined by S_M in Equation 6.4.

Protocol	Parameter	Symbol	Value
RB [116, 28]	Gate Set	$\mathbb{G}$	Group and unitary 2-design, R members
	Experiment Types	$\mathfrak{E}$	$\mathfrak{E}_M = \{0\}$ with SPAM $\rho_0, E_0 \approx \|0\rangle\langle 0\|$
	Allowable Sequences	$J_{M,e}$	Unif $\left(\{\vec{j} \in \{1,...,R\}^{M+1} \| \mathcal{G}_{\vec{j}} = \mathbb{I}\}\right)$
	Tying Parameters	$\vec{x}_T$	(p, A, B)
	Tying Functions	T	$T(1, M, e, \vec{x}_T) = (A - B)p^M + B$
Interleaved RB [117]	Gate Set	$\mathbb{G}$	Group and unitary 2-design, R members
	Experiment Types	$\mathfrak{E}$	$\mathfrak{E}_M = \{0, r\}$ for some $1 \leq r \leq \|\mathbb{G}\|$, with SPAM $\rho_e, E_e \approx \|0\rangle\langle 0\|$
	Allowable Sequences	$J_{M,e}$	$J_{M,0} = \text{Unif}\left(\{\vec{j} \in \{1,...,R\}^{M+1} \| \mathcal{G}_{\vec{j}} = \mathbb{I}\}\right)$
			$J_{M,r} = \text{Unif}\left(\{\vec{j} \in \{1,...,R\}^{2M+1} \| \mathcal{G}_{\vec{j}} = \mathbb{I}, \vec{j}_{\text{even}} = r\}\right)$
	Tying Parameters	$\vec{x}_T$	(p_0, p_r, A, B)
	Tying Functions	T	$T(1, M, e, \vec{x}_T) = (A - B)p_e^M + B$
Unitarity [177]	Gate Set	$\mathbb{G}$	Group and unitary 2-design, R members
	Experiment Types	$\mathfrak{E}$	$\mathfrak{E}_M = \{0\}$ with SPAM $\rho_0, E_0 \approx \|0\rangle\langle 0\|$
	Allowable Sequences	$J_{M,e}$	Unif $\left(\{1,...,R\}^M\right)$
	Tying Parameters	$\vec{x}_T$	(u, A, B)
	Tying Functions	T	$T(2, M, \vec{x}_T) = A + Bu^{M-1}$
Leakage RB [189]	Gate Set	$\mathbb{G}$	Group and unitary 2-design with R members acting on $\mathcal{X}_1$, $\mathcal{X} = \mathcal{X}_1 \oplus \mathcal{X}_2$ with $\dim \mathcal{X} = d_1 + d_2$
	Experiment Types	$\mathfrak{E}$	$\mathfrak{E}_M = \{0, ..., d_1 - 1\}$ with SPAM $E_e \approx \|e\rangle\langle e\|, \rho_e \approx \|0\rangle\langle 0\|$
	Allowable Sequences	$J_{M,e}$	Unif $\left(\{\vec{j} \in \{1,...,R\}^{M+1} \| \mathcal{G}_{\vec{j}} = \mathbb{I}\}\right)$
	Tying Parameters	$\vec{x}_T$	$(L_1, L_2, \mu_1, p_l, \{A_e\}_{e\in\mathfrak{E}}, \{B_e\}_{e\in\mathfrak{E}}, \{C_e\}_{e\in\mathfrak{E}},)$
	Tying Functions	T	$T(1, M, e, \vec{x}_T) = (L_2 A_e + L_1 B_e)/(L_1 + L_2)$ $+\left(\frac{L_1}{L_1+L_2} - p_l\right)(A_e - B_e)(1 - L_1 - L_2)^M$ $+(1 - p_l)(C_e - A_e)(\mu_1(1 - L_1))^M$
Dihedral Benchmarking [21]	Gate Set	$\mathbb{G}$	$\langle Z_j = \mathrm{e}^{\mathrm{i}\pi Z/j}, X\rangle \subseteq \mathcal{SU}(2)$ for some $j \in \mathbb{N}$, R total members
	Experiment Types	$\mathfrak{E}$	$\mathfrak{E}_M = \{X, Z\}$ with SPAM $E_e, \rho_e \approx (I + e)/2$
	Allowable Sequences	$J_{M,e}$	Unif $\left(\left\{\vec{j} \in \{1,...,R\}^{M+1} \| \mathcal{G}_{\vec{j}} \in \{I, e\}\right\}\right)$
	Tying Parameters	$\vec{x}_T$	(p_X, p_Z, A, B_X, B_Z)
	Tying Functions	T	$T(1, M, e, \vec{x}_T) = A + B_e p_e^M$

Table 6.1: Description of some RB+ protocols within our framework.

At this point we have run into a very serious problem. This expression cannot be simplified, even in principle, unless we know more about the survival distribution S_M. There is one exception, however, first explicitly pointed out in an appendix of Granade et al. [62]: if $N = 1$, then the expectation's integrand is linear in q for both values of d and so only the first moment of S_M matters; we get

$$\mathrm{L}(p, A, B|d = 1, M) = (A - B)p^M + B \tag{6.9}$$

for standard RB, or more generally,

$$\mathrm{L}(\vec{x}_T|d = 1, M, e) = T(1, M, e, \vec{x}_T) \tag{6.10}$$

for any protocol whose first moments are tied together. This fact was exploited to great effect by those authors. The same argument shows that the first N moments of $S_{M,e}$ are potentially relevant to the likelihood function for any protocol, and therefore some characterization of them should be appended to the list of nuisance parameters.

Alternatively, one might argue to simply enforce the constraint $N = 1$. This is a reasonable suggestion, and is explored in Section 6.6 where it is shown that $N = 1$ should be considered best-practice for protocols which only tie together their first moments, and whose implementations are quick at switching between random sequences. For some experimental setups, however, switching the sequence every experiment would dominate the duty cycle. The way around this is through fast logic near the quantum system [152], such as was recently demonstrated by [77] in the case of a transmon qubit coupled to an oscillator-encoded logical qubit. Or perhaps, even more seriously, some systems are not capable of strong measurement, and so a binomial model with $N = 1$ is not physically possible. In this case we can still write down a likelihood function, no longer conditionally binomial (as seen, for example, in Section 7.4.2), but one that will involve higher moments by necessity. Finally, in some cases, the second moment is the moment of interest, as in the unitarity protocol, so that $N = 1$ is completely insensitive to the quantity of interest.

In any case, a great deal of RB+ experiments have been performed with $N > 1$ and so it behooves us to devise a statistically rigorous approach for analysing such data.

6.5 Constructing Agnostic Models

In the last section we noted that for a repetition value of N, to fully specify the likelihood function of an RB or related protocol, we require at least N parameters per sequence length and experiment type, in addition to the parameters of the tying function. These

extra parameters correspond to moments of the survival distributions. We will write $\vec{x}_S$ to denote these new parameters, whatever they end up being, distinguishing them from the parameters of the tying function, $\vec{x}_T$. One must tread carefully in any analysis that follows this observation. The goal of this section to develop a framework where we treat these nuisance parameters in a principled yet practical way, while at the same time remaining as agnostic about their structure as possible.

6.5.1 Parameterizations

A Bayesian, by instinct, may be tempted to throw all of the unknown moments of the survival distributions into an inference engine as nuisance hyperparameters. In principle there is nothing wrong with this. However, it would lead to a huge number of parameters for even modest values of N. Care would be required in restricting the domains of these moments, for example, the variance σ^2 of a distribution with support on $[0, 1]$ and expectation value μ must always satisfy $0 \leq \sigma^2 \leq \mu(1-\mu)$.

One might suggest next to truncate the number of moments to be included as hyperparameters down to some tractable, empirically motivated constant. But even in this case, one must specify the higher moments somehow. For example, one might choose to set them all to zero. This would effectively restrict the space of allowed survival distributions to some strange, unmotivated family of distributions. Instead, one might make the moments above the truncation cutoff sure functions of those below in some sensible way.

At this point, we have basically argued for the use of parameterized families of probability distributions; any family of probability distributions, like the Gaussian or gamma families, can be defined as a rule that specifies all moments of a given member in terms of a few parameters. For us, the most natural starting point is the beta distribution family. This family is conjugate to the binomial distribution, and is the canonical family of continuous distributions with support on the unit interval. A member with parameters $\alpha, \beta > 0$ is written $\text{Beta}(\alpha, \beta)$, and has a density function defined by

$$\text{pdf}_{\text{Beta}}(q) = \frac{q^{\alpha-1}(1-q)^{\beta-1}}{\text{B}(\alpha, \beta)} \tag{6.11}$$

where the normalization constant $\text{B}(\alpha, \beta)$ is the beta function. Its first and second central moments are given by $\mu = \frac{\alpha}{\alpha+\beta}$ and $\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, respectively. These equations can be uniquely inverted as

$$\alpha = \mu^2(1-\mu)/\sigma^2 - \mu \tag{6.12a}$$

$$\beta = \mu(1-\mu)^2/\sigma^2 - (1-\mu), \tag{6.12b}$$

which provides an alternate parameterization of the family. In a slight abuse of notation, we write $\text{Beta}(\mu, \sigma)$ for a member written in the new coordinates. Alternate parameterizations and their transforms are provided in Section C.3, and we similarly abuse notation for these other coordinates, writing, for example, $\text{Beta}(\mu, r)$ where $\sigma^2 = r\mu^2(1-\mu)^2$.

This family can produce quite a wide variety of shapes even though it only has two parameters. Setting $\alpha = \beta = 1$ results in the uniform distribution on $[0, 1]$. Fixing any mean $\frac{\alpha}{\alpha+\beta} \in (0, 1)$ while increasing α and β decreases the variance, and the distribution approaches a normal shape. On the other hand, decreasing α and β while the mean is kept fixed increases the variance toward $\mu(1-\mu)$; the probability density at first spreads out over the whole interval $[0, 1]$, and when this is no longer able to keep increasing the variance, the mass begins to build up at the end points, approaching a weighted mixture of two delta functions.

Using this family, for a first order tying function, every sequence length, experiment type, and measurement operator would add one parameter to the likelihood model, so that $\vec{x}_S = \{\sigma_{M,e}\}_{M \in \mathbb{M}, e \in \mathfrak{E}_M}$, or some other parameterization thereof. In the case of any protocol which ties together only first moments, we get the hierarchical model

$$\vec{x}_T \sim \pi(\vec{x}_T) \tag{6.13a}$$

$$\mu_{M,e} | \vec{x}_T = T(1, M, e, \vec{x}_T) \tag{6.13b}$$

$$\sigma_{M,e} \sim \pi(\sigma_{M,e}) \tag{6.13c}$$

$$q_{M,e,i} | \mu_{M,e}, \sigma_{M,e} \overset{\text{iid}}{\sim} \text{Beta}\left(\mu_{M,e}, \sigma_{M,e}\right) \tag{6.13d}$$

$$Q_{M,e,i} | q_{M,e,i} \overset{\text{iid}}{\sim} \text{Binom}\left(N, q_{M,e,i}\right) \tag{6.13e}$$

for the dataset D. The horizontal line is a visual aid to separate the prior from the likelihood distribution, and $\pi(\cdot)$ refers to the prior distribution of the given parameters. The quantities $q_{M,e,i}$ are latent random variables representing survival probabilities — they can be analytically integrated out of the model if desired, resulting in a beta-binomial distribution instead. This set of sampling statements, which are sequentially dependent on previous variables, is an example of a probabilistic program. It is a convenient way of specifying the joint distribution of the prior and the likelihood, which is proportional to the posterior distribution.

Models for higher-order tying functions are just as easy to write down. Note, however, that the beta distribution only has two parameters, so that if both of the first two moments are tied together, there is no more uncertainty in the survival distributions (conditional

on a specific value of $\vec{x}_T$). This can be solved by using a larger family of distributions, or through a nonparametric approach, as discussed in the following section.

6.5.2 Nonparameterizations

The assertion that every survival distribution is approximately beta distributed may sometimes be too strong. In this section we would like to loosen this restriction. One viable path is to use a bigger family, such as the generalized beta family with five parameters [123]. Even more generally, we can resort to Bayesian nonparametrics. This is the approach that we take, and in particular, we use Dirichlet process mixtures, which are distributions of distributions[5].

Let $\mathrm{DP}_K(\alpha, G_0)$ denote a Dirichlet process with a concentration parameter $\alpha > 0$ and a base distribution G_0 that has support on the parameter space Ω, and that is truncated to K modes[6]. We would like to replace the draw of $q_{M,e,i}$ from a beta distribution (see Equation 6.13) to a draw from a random distribution G, such as $G \sim \mathrm{DP}_K(\alpha, G_0)$. The Dirichlet process has two shortcomings that prevent us from directly using it for this purpose. The first is that its variates are not continuous distributions, and the second is that the moments of its draws are random, whereas we would like the ability to (conditionally) fix some of them according to the tying functions.

To overcome these problems we modify the Dirichlet process into a new nonparametric family that we call *constrained Dirichlet process beta mixtures* (CDPBM), denoted $\mathrm{CDPBM}_K(\alpha, G_0, \cdot)$, whose definition is motivated in Section C.2.2. In short, if the desired mean value of our random distributions is $0 < \mu_1 < 1$, then the random distribution $G \sim \mathrm{CDPBM}_K(\alpha, G_0, \mu_1)$ is drawn as follows:

$$\sum_{k=1}^{K} w_k \delta_{(\nu_k^*, r_k)} \sim \mathrm{DP}_K\left(\alpha, G_0\right) \tag{6.14a}$$

$$\nu_k = \frac{1}{1 + \mathrm{e}^{-\nu_k^* - h}} \text{ with } h \text{ such that } \sum_{k=1}^{K} w_k \nu_k = \mu_1 \tag{6.14b}$$

$$G = \sum_{k=1}^{K} w_k \,\mathrm{Beta}(\nu_k, r_k). \tag{6.14c}$$

[5] We provide a brief introduction to Dirichlet processes and Dirichlet process mixtures in Section C.2.1.

[6] As a brief bit of context, recall that G_0 is the mean value of $\mathrm{DP}_K(\alpha, G_0)$, and that α can be interpreted as the number of 'prior observations' from samples of $\mathrm{DP}_K(\alpha, G_0)$; it scales inversely with the variance of $\mathrm{DP}_K(\alpha, G_0)$.

Here, the Dirichlet process sample space is $(\nu_k^*, r_k) \in \mathbb{R} \times (0,1) = \Omega$, upon which the base distribution G_0 is defined, and we are using the (μ, r) parameterization of the beta family (see Section C.3). This procedure ensures that $\mathbb{E}[G] = \mu_1$, and that the support of G lies within $[0,1]$. We typically choose $G_0 = \mathrm{N}(0, 1.9) \times \mathrm{Unif}(0,1)$ as a broad prior, and assign a hyper-prior α $\mathrm{Gam}(1,1)$.

With this defined, our nonparametric model for analyzing RB+ data is a straightforward modification of Equation 6.13, given by

$$\vec{x}_T \sim \pi(\vec{x}_T) \tag{6.15a}$$
$$\mu_{M,e}|\vec{x}_T = T(1, M, e, \vec{x}_T) \tag{6.15b}$$
$$\alpha_{M,e} \overset{\text{iid}}{\sim} \mathrm{Gam}(1,1) \tag{6.15c}$$
$$G_{M,e}|\alpha_{M,e}, \mu_{M,e} \overset{\text{ind}}{\sim} \mathrm{CDPBM}_K\left(\alpha_{M,e}, G_0, \mu_{M,e}\right) \tag{6.15d}$$

$$q_{M,e,i}|G_{M,e} \overset{\text{ind}}{\sim} G_{M,e} \tag{6.15e}$$
$$Q_{M,e,i}|q_{M,e,i} \overset{\text{ind}}{\sim} \mathrm{Binom}(N, q_{m,i}). \tag{6.15f}$$

A slight modification is needed for protocols which tie together higher moments, which we omit for brevity; see Section C.2.2.

6.5.3 Frequentist Approaches

Though we are primarily concerned with a Bayesian approach, we are also interested in comparing to frequentist methods. To date, the *de facto* frequentist inference tool for RB+ data (with exceptions) has been least-squares fitting (LSF) to exponential decay models. Generally, the justification for LSF is that it is equal to the maximum likelihood estimator (MLE) in the case of Gaussian noise on the data.

There are a couple of reasons to be cautious when using estimates and confidence regions based on LSF in the case of RB+. One is that the distribution of the data is not Gaussian, except approximately in the high data regime, and therefore the MLE is not being reported, but some sort of approximation thereof. Another is that weights need to be chosen for weighted LSF (WSLF)—using uniform weights implicitly makes assumptions about the nature of the noise model and should always be avoided.

It is non-trivial to choose appropriate weights for WLSF. One may be tempted to use sample variances as weights, but there is a subtle issue that these variances do not

directly represent the uncertainty of the quantities of interest at a given sequence length and experiment type; they partially contain unnecessary weight due to finite sampling statistics. Even if this is corrected for, one must also make sure that weights are assigned consistently. Additionally, one needs a heuristic for assigning a non-zero weight in the case of no variance in the outcomes at a given sequence length of a protocol.

For these reasons our preferred frequentist method for analyzing data from RB+ models containing many sequence lengths is to look directly at the MLE. This can be done by using a likelihood function that assumes that survival distributions are beta distributed—see the second half of Equation 6.13. The log-likelihood of this model is easily and reliably maximized with gradient-based numerical methods. We avoid having to assign weights at every sequence length since they are now treated as nuisances of the global fit. Confidence intervals for this estimator can be constructed through standard bootstrapping techniques (see for example the survey article of DiCiccio and Efron [35]). We construct bootstrap distributions of the tying parameters by computing the MLE on random data replications drawn from the empirical (non-parametric) distribution of the data, or by sampling the likelihood distribution at the MLE of the data (parametric). Samples are always drawn on a per-sequence-length basis, so that the shape of the bootstrapped data is the same as that of the original data. Confidence intervals are constructed with the simplest bootstrap-t procedure. That is, we look directly at the CDF of these bootstrap distributions.

Occasionally we will also consider the WLSF for the sake of interest. In such cases, we set weights equal to the sample variances of the binomial data normalized by N. We do this because it has been a popular approach historically.

6.6 Sequence Re-Use

Thus far we have only talked about data analysis. In this section we discuss which experiments to perform in the first place. Specifically, we address the question of how many times a fixed random sequence from an RB+ protocol should be reused. In Section 6.4 we hinted at the fact that every random sequence should, ideally, only be used once. Here, we qualify and quantify this idea.

With all of the heavy lifting of getting to the survival distribution out of the way, we can cast the problem of sequence re-use as one of pure statistics. Or, we can think of a concrete and conceptually simple isomorphic problem—we can think of a survival distribution as a bag of coins with different biases. Suppose this bag has a mean bias $\overline{q}$ and a standard deviation of biases σ (or, equivalently characterized by the second moment μ_2). We want

to estimate these unknown quantities from selecting coins from the bag, at random, and flipping them. The isomorphism is that the statistical conclusions of flipping the same coin more than once are the same as repeating a given gate sequence in RB.

So, by considering the trade-off in the number of repetitions of flips using the same coin versus selecting a new coin, we can understand the optimal experimental design policy in RB.

6.6.1 First moment estimators

Naturally, we start with the first moment. With protocols that tie only first moments, we only care, by necessity, about inferring values which depend on $\overline{q}$, but none of the higher moments of the bag. Conditional on picking a coin with bias q, if we perform N Bernoulli trials and add them up, we have the conditional random variable

$$Q|q \sim \text{Binom}(N, p) \tag{6.16}$$

with conditional cumulants $\mathbb{E}[Q|q] = Np$ and $\text{Var}[Q|q] = Nq(1-q)$.

This gives $\text{Var}[Q] = N\overline{q}(1-\overline{q}) + N(N-1)\sigma^2$ through the law of total variance. If we independently and identically repeat the process of drawing a different coin I times and perform N Bernoulli trials on each, we end up with

$$\text{Var}\left[\sum_{i=1}^{I} \frac{Q_i}{N \cdot I}\right] = \frac{1}{I}\left(\frac{\overline{q}(1-\overline{q})}{N} + \frac{N-1}{N}\sigma^2\right) \tag{6.17}$$

as the variance of the scaled quantity $\sum_{i=1}^{I} Q_i/(N \cdot I)$ whose mean value is $\bar{q}$. The take-away from this formula is that the variance approaches 0 as we increase the number of coins (sequences) we use, but asymptotes to the finite value σ^2 if we fix I and increase the binomial parameter N (re-use of the same sequence). If we consider instead the total number of flips of all coins to be fixed, $I \cdot N = \text{const.}$, we can see at once that the variance is minimized when $N = 1$ by completely eliminating the contribution from σ.

We have looked at the variance formula above because it has a simple derivation and gets the point across. However, a better quantity to consider is the Fisher information and the resulting Cramér–Rao bound of $\overline{q}$, because it gives a rigorous bound on how well any (unbiased) estimator of $\overline{q}$ can do. Supposing that we explicitly choose our bag to have a beta distribution with mean value $\overline{q}$ and variance $\sigma^2 = t\overline{q}(1-\overline{q})$ for some $0 < t < 1$, then our likelihood distribution is Beta-Binom(N, μ, t) and the two-by-two Fisher information

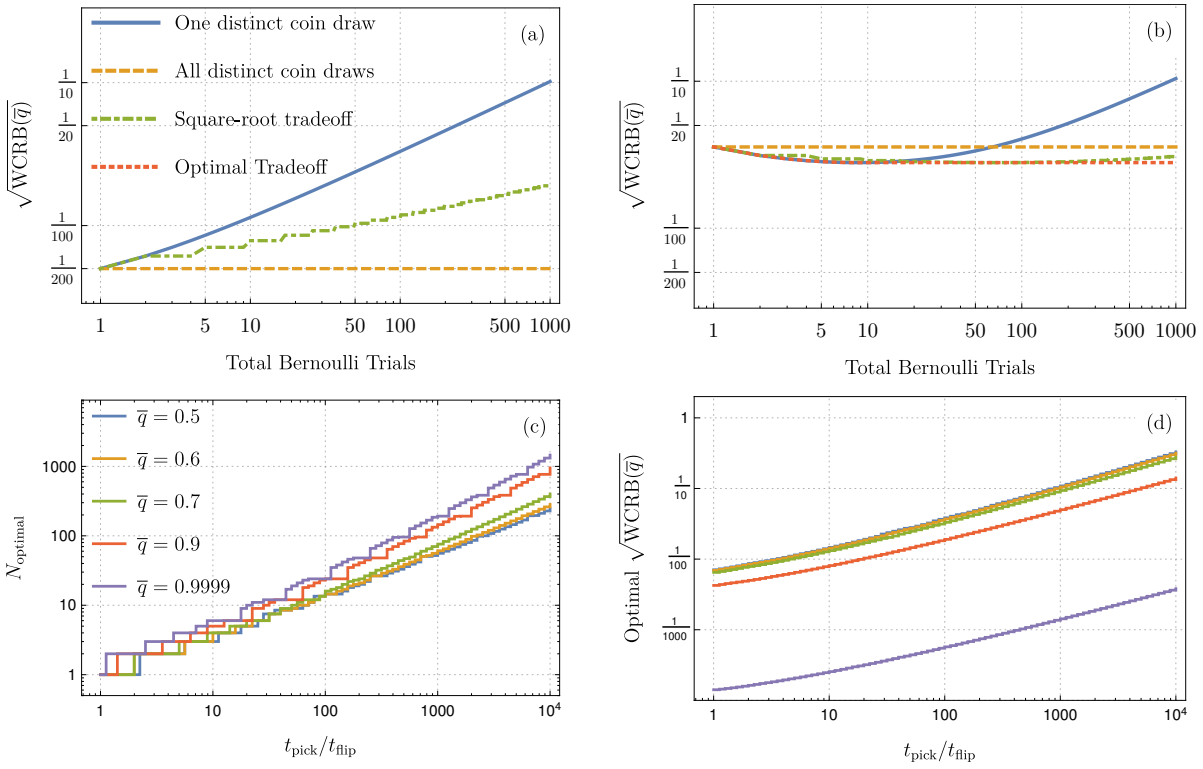

Figure 6.1: (a-b) Supposing a fixed budget of Bernoulli trials for the bag-of-coins experiment, the WCRB (Equation 6.19) of the mean coin bias $\overline{q}$ is shown, normalized to the time it takes to complete the full experiment. The true parameters of the bag are $\overline{q} = t = 0.5$, a single coin flip takes $t_{\text{flip}} = 100\,\text{us}$, and switching coins takes $t_{\text{pick}} = 0$ and $t_{\text{pick}} = 5\,\text{ms} = 50t_{\text{flip}}$ for (a) and (b), respectively. We see choosing all coins to be different is no longer the best strategy when $t_{\text{pick}} > 0$. To explore this, in (c-d), given a ratio $t_{\text{pick}}/t_{\text{flip}}$, we compute the optimal number of coin flips N and the resulting optimal WCRB for $t_{\text{flip}} = 100\,\text{us}$, $t = 0.5$, and various values of $\overline{q}$. (In these final two plots, there is no longer a fixed budget of trials; global minima were found with respect to N.)

matrix, $J(\overline{q}, t)$, is given by the negative expected value of the Hessian of the log-likelihood function. By virtue of our choice of parameterization $(\overline{q}, t)$, the Fisher information matrix happens to be diagonal, and so the the Cramer-Rao bound reads

$$\mathrm{Var}[\hat{\overline{q}}] \geq \frac{1}{I \cdot J(\overline{q})} \tag{6.18}$$

where $J(\overline{q}) = J(\overline{q}, t)_{1,1}$ and $\hat{\overline{q}}(Q_1, ..., Q_I)$ is any unbiased estimator of $\overline{q}$ that depends on I iid samples from the likelihood.

So far we have neglected any cost associated with picking a new coin from our analysis, which is the main reason why experimentalists re-use sequences. We can include this cost by considering the Fisher information per unit time, $J(\overline{q})/T$, where T is the time it takes to collect the data. Suppose that it takes time t_{pick} to pick a new coin and time t_{flip} to flip a coin once. Then we have $T = I(t_{\text{pick}} + N t_{\text{flip}})$, and the CRB weighted by experiment cost is

$$\mathrm{Var}[\hat{\overline{q}}]/\mathrm{Hz} \geq \frac{(t_{\text{pick}} + N t_{\text{flip}})}{J(\overline{q})} \equiv \mathrm{WCRB}(\overline{q}) \tag{6.19}$$

where we have assumed T is in units of seconds. Note that if we take the square root of both sides we get the usual units for sensitivity. This figure of merit is explored in Figure 6.1.

6.6.2 Second moment estimators

As before, we draw a coin I times and perform N Bernoulli trials on each. This time, however, we estimate the *second moment* via summing the squares of the number of successes. This estimator is biased, but not asymptotically so:

$$\mathbb{E}\left[\sum_{i=1}^{I} \frac{Q_i^2}{I \cdot N^2}\right] = \mu_2 + \frac{1}{N}(\overline{q} - \mu_2). \tag{6.20}$$

That is, as the number of repetitions N increases, this estimator becomes less biased.

Due to this bias, the Cramér–Rao cannot tell us much about this estimator. But, we can directly calculate the mean squared error. As before, though, we consider a fixed total number of measurements $T = N \cdot I = \text{const.}$ and calculate $T \cdot$ MSE.

Since the MSE involves the square of the second moment, we need to calculate

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=1}^{I}\frac{Q_i^2}{I\cdot N^2}\right)^2\right] &= \frac{1}{I^2N^4}\sum_{j,k=1}^{I}\mathbb{E}[Q_j^2Q_k^2] \\ &= \frac{1}{I^2N^4}\left(\sum_{k=1}^{I}\mathbb{E}[Q_k^4]+\sum_{j\neq k=1}^{I}\mathbb{E}[Q_j^2]\mathbb{E}[Q_k^2]\right) \\ &= \frac{1}{I^2N^4}\left(I\mathbb{E}[Q^4]+I(I-1)\mathbb{E}[Q^2]^2\right). \end{aligned} \tag{6.21}$$

The fourth moment of the Beta-Binomial Beta-Binom(N,μ,μ_2) is simple yet still too messy to usefully reproduce here.

We calculate the optimal repetition rate by averaging the total cost over a uniform prior on the domain of validity in the parameterization of (μ,μ_2). The final answer for the optimal value of N is

$$N_{\text{opt}} = \left(\frac{16}{40+32\ln(2)-3\ln(3)}\right)^{\frac{1}{3}}T^{\frac{1}{3}}+O\left(\frac{1}{T^{\frac{1}{3}}}\right), \tag{6.22}$$

or roughly $0.65T^{\frac{1}{3}}$. A ball-park amount of data usually taken at each sequence length in randomized benchmarking is about a kilobyte. This corresponds to about $N=13$ repetitions per sequence and $I=615$ difference sequences.

It is also of interest to consider the case when $\mu\in(l,1)$ for some lower bound l. For example, suppose we are fairly confident that our fidelity is above 90%. In this case, we still have

$$N = C(l)T^{\frac{1}{3}}+O\left(\frac{1}{T^{\frac{1}{3}}}\right), \tag{6.23}$$

for some $C(l)<C(0)$. For example, taking $l=0.9$, we have $N=0.39T^{\frac{1}{3}}$.

Finally, we generalize the calculation of Equation 6.22 to include the effects of finite switching costs $\tau := t_{\text{pick}}/t_{\text{flip}}$. In doing so, we proceed numerically, as the series expansion obtained in Equation 6.22 is much less useful for $\tau>0$. We plot the results in Figure 6.2, noting that even for $\tau=30$, the optimal sequence lengths found do not deviate substantially from the case where there is no switching cost. Thus, $N_{\text{opt}}\approx 0.65T^{\frac{1}{3}}$ remains a useful heuristic in this case, even if it is no longer a rigorous approximation.

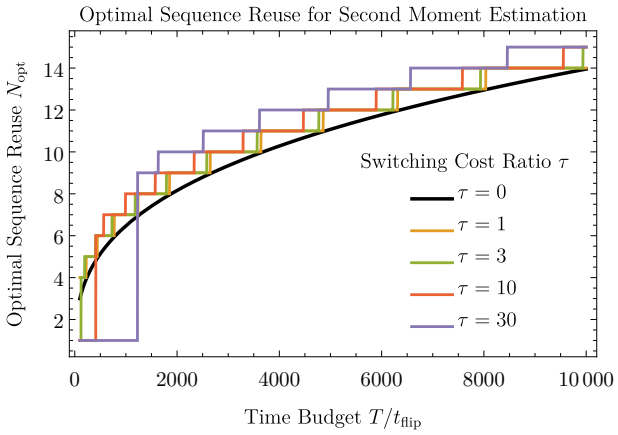

Figure 6.2: The optimal sequence reuse N_{opt} for second moment estimation (as used, for example, in the unitarity protocol), plotted as a function of the total time budget allowed T, for each of several choices of the switching cost ratio $\tau := t_{\text{pick}}/t_{\text{flip}}$.

6.7 Numerical Results

In this section we explore our Bayesian model with a collection of numerical examples, using various protocols and error models.

As with most Bayesian models, analytic formulae for posterior distributions are intractable. Our posterior in the examples throughout this section are therefore computed with numerical techniques. In particular, we use the Hybrid Monte Carlo (HMC) sampler using the No-U-Turns (NUTS) heuristic [42, 89]. This is a type of Markov chain Monte Carlo (MCMC) sampler that has gained widespread use due to its lack of tuning parameters, fast mixing rate, and ability to handle large numbers of parameters. More details about our sampling strategies are outlined in Section C.1.

6.7.1 RB with Various Noise Models

As a first example, we consider the standard RB protocol on a qubit under three noise models. We use an order 12 subgroup of the usual 24 member Clifford group as our gateset. This subgroup is still a 2-design and can be generated as $\mathbb{G} = \langle Z, \sqrt{Z}H \rangle$, where

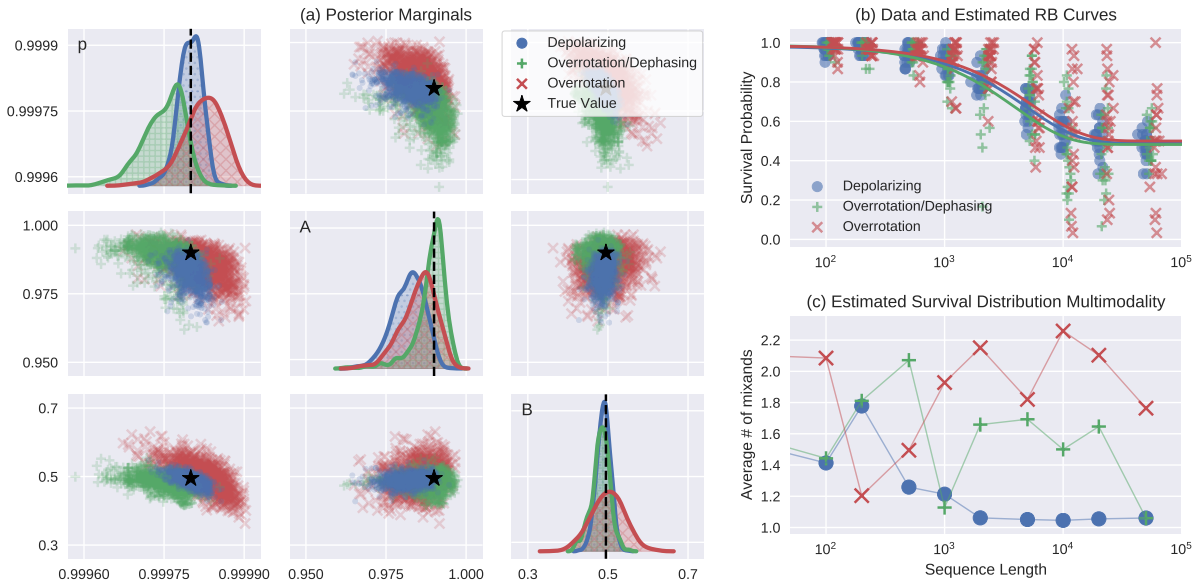

Figure 6.3: (a) Single and joint posterior marginals of the parameters p, A, and B are shown for each of the three noise models defined in Section 6.7.1 of the main text for the standard RB protocol. (b) Using Bayes' estimate for these three parameters, the curve $(A-B)p^M + B$ is plotted for each model, displayed on top of the normalized data used in the inference. The unusual shape is due to the log-linear scale, and jitter in the x-axis on the data points was added for visual appeal — for all three models $I = 20$ random sequences were used with $N = 30$ repetitions each at each of the sequence lengths $\mathbb{M} = \{1, 100, 200, 500, 1000, 2000, 5000, 10000, 20000, 50000\}$, where the maximum sequence length is chosen according to the $M_{\max} = 1/(1-F)$ heuristic [62]. (c) The posterior shown in (a) was calculated using the model in Equation 6.15, which describes each survival distribution as a mixture of beta distributions, and so finally, we plot the posterior mean of $1/\sum_{k=1}^{K} w_k^2$ for each survival distribution, where the weights w_k are defined in Equation 6.14. This quantity ranges between 1 and K and quantifies the estimated number of relevant mixands in each survival distribution. The low values justify our CDPBM truncation at $K = 10$.

$H = \left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right)/\sqrt{2}$ and $Z = \left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$. Our three noise models are defined as

$$\mathcal{E}_r^1 = \Lambda_{s_1} \tag{6.24a}$$
$$\mathcal{E}_r^2 = \Phi_{s_2} \circ \Theta[G_r, \epsilon_2] \tag{6.24b}$$
$$\mathcal{E}_r^3 = \Theta[G_r, \epsilon_3] \tag{6.24c}$$

where $\tilde{\mathcal{G}}_r = \mathcal{G}_r \circ \mathcal{E}_r^i$ is the actual implementation of the ideal gate $\mathcal{G}_r$ for $r = 1, ..., R$ and where

$$\Lambda_s(\rho) = (1-s)\rho + s\,\mathrm{Tr}[\rho]\mathbb{I}/2 \tag{6.25a}$$
$$\Phi_s(\rho) = (1-s)\rho + sZ\rho Z \tag{6.25b}$$
$$\Theta[U, \epsilon](\rho) = \begin{cases} \rho & U \text{ is some z-rotation} \\ U^\epsilon \rho (U^\epsilon)^\dagger & \text{else} \end{cases} \tag{6.25c}$$

are the depolarizing, dephasing, and transverse overrotation channels, respectively. Therefore $\mathcal{E}_r^1$ is a gate independent depolarizing channel, $\mathcal{E}_r^2$ is gate independent dephasing combined with a gate dependent overrotation by amount ϵ_2, and $\mathcal{E}_r^3$ is purely gate dependent overrotation by amount ϵ_3. Constants were chosen by trial and error so that all three noise models result in exactly the same RB decay base $p = 0.9998$, ultimately achieved with the choices $s_1 = 0.0002$, $s_2 = 0.000028954$, $\epsilon_2 = 0.01$, and $\epsilon_3 = 0.11132$. A formula for computing p given a gate dependent noise model is provided in Ref. [178].

Data was simulated under each noise model with the initial state $\rho = |0\rangle\langle 0|$ and the measurement $M = 0.99\,|0\rangle\langle 0|$ at each of the sequence lengths $\mathbb{M} = \{1, 100, 200, 500, 1000, 2000, 5000, 10000, 20000, 50000\}$. At each sequence length, $I = 20$ random sequences were drawn and $N = 30$ repetitions were used for each. To produce histograms of the survival distributions, however, thousands of simulations were done per sequence length.

This dataset was processed in a few different ways. Posterior results using the CDPBM-survival-distribution model Equation 6.15 are summarized in Figure 6.3. The slightly simpler Beta-survival-distribution model Equation 6.13 was also used, which is compared to the CDPBM model in Figure 6.4, along with weighted least squares fitting, and a non-parametric bootstrap with 2000 samples. Additionally, estimates of the shapes of some survival distributions are seen in Figure 6.5. The prior distribution on the tying parameters was chosen to be $\pi(p, A, B) = \mathrm{Unif}([0,1]^3)$ in all cases.

6.7.2 Low Data Regime

One advantage of using the full likelihood model is that it transitions seamlessly to low data regimes where normal approximations fail and the usual sample moments are ill-defined.

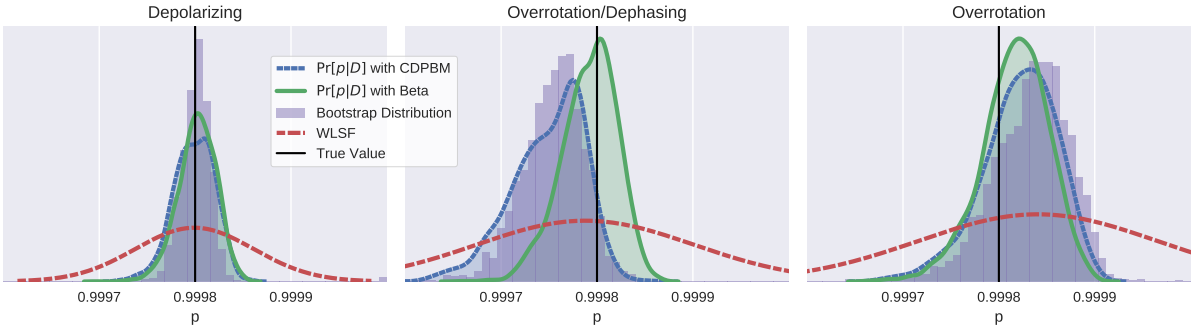

Figure 6.4: For each of the three noise models defined in Section 6.7.1, four types of data processing are performed to compare their estimates of (and uncertainties in) the parameter p from the standard RB protocol. Each dataset consists of $I = 20$ random sequences with $N = 30$ repetitions each at each of the sequence lengths 1, 100, 200, 500, 1000, 2000, 5000, 10000, 20000 and 50000. The first two methods show the posterior marginal of p under the models from Equation 6.15 and Equation 6.13, respectively. The next two methods are non-parametric bootstrapping and weighted least squares fitting, as described in Section 6.5.3.

At a given sequence length, if we only pick a handful of sequences I with a handful of shots N each, then there is a good chance that $Q_{M,e,i}$ will be equal for all $i = 1, ..., I$. This is especially true near the boundaries 0 and 1. In this event, it is difficult to use a weighted least-squares fit.

To illustrate our Bayesian model in this regime, we consider simulated data from standard RB using the gate dependent overrotation model from Equation 6.24c. We choose this model because it has very wide survival distributions, as seen in Figure 6.5.

We wish to demonstrate that posterior distributions in the low-data regime meaningfully report the parameter of interest, p. The worst thing an inference method can do in this example is predict that the RB parameter p is larger than it actually is. Therefore instead of summarizing a posterior in terms of its mean value (Bayes' estimate), it is more helpful to summarize it in terms of the the value at a one sided credibility level α,

$$p_\alpha(D) = [p_0 \text{ such that } \Pr(p > p_0|D) = \alpha]\,. \tag{6.26}$$

Here, $\Pr(p|D)$ is the posterior of p under the beta model Equation 6.13 with the same prior as in Section 6.7.1 given the RB dataset D. For example, according a given posterior, with 95% probability, $p_{0.95}(D)$ should be a lower bound for the true the value of p. Fixing the

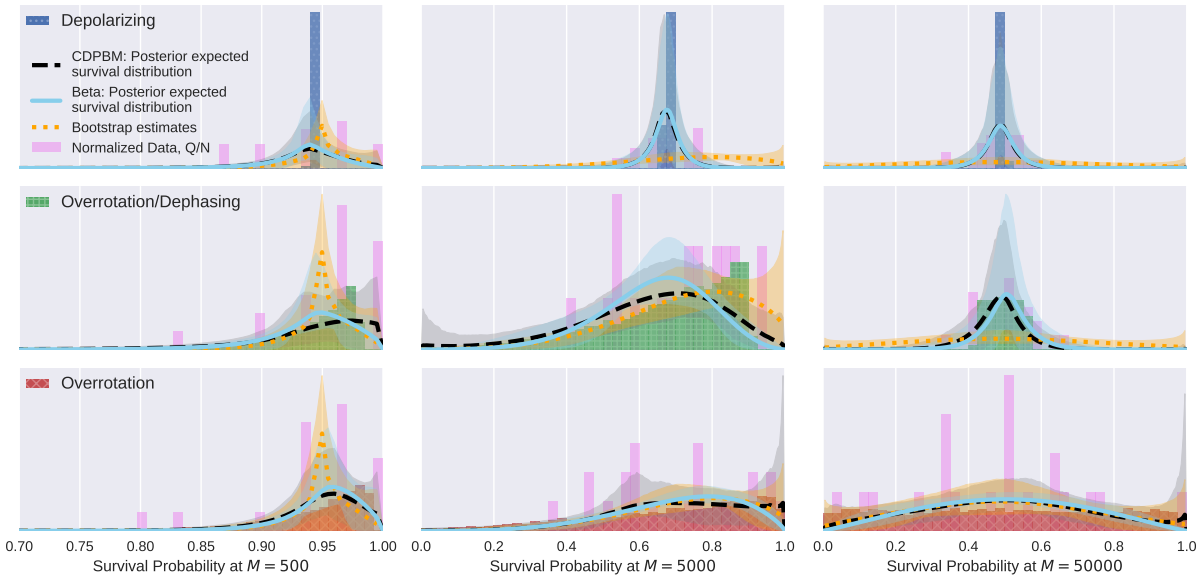

Figure 6.5: Although the survival distributions at each sequence length are considered to be nuisance parameters of the model, their posteriors are nonetheless interesting and provide a diagnostic check. Here, the three rows correspond to the noise models described in Section 6.7.1, and each column is a different sequence length. In each plot, Bayes' estimate of the survival distribution is shown for both models Equation 6.15 and Equation 6.13 along with their pointwise 95% credible envelopes. Similar 95% confidence envelopes are shown for the bootstrap method. These are overlaid on top of histograms sampled from the true survival distributions, as well as the (normalized) data that were actually used in the inference.

model and the prior, the quantity $p_\alpha(D)$ is itself a random variable as it depends on D. What we desire in our numerical test is that consistency condition

$$\Pr(p_\alpha(D) < p_{\text{true}}) \geq \alpha \tag{6.27}$$

is satisfied for any level α that we care about.

To evaluate this criterion we compute $p_\alpha(D)$ for many simulated datasets D. Each dataset uses the sequence lengths

$$\begin{aligned}\mathbb{M} = \{&1, 100, 200, 500, 1000, 2000,\\ &5000, 10000, 20000, 50000\}\end{aligned}$$

and the repetition number $N = 5$. Three-hundred data sets were considered at each of the values $I = 1, 3, 5, 10, 20, 30, 50, 80, 100$. Figure 6.6 shows both a selection of posteriors, as well as a summary of the distribution of $p_{0.95}(D)$ at each value of I. Note that the sharp elbow displayed in Figure 6.6(b) could be used in practice to decide on an appropriate amount of data to take: in this example, there is a huge advantage in moving from $I = 5$ to $I = 10$, but not much of an advantage in moving from $I = 10$ to $I = 15$.

The bootstrapped confidence bounds discussed in Section 6.5.3 are also sensibly defined in the low data regime. In Figure 6.6(d), however, we see in both parametric and non-parametric bootstrapping that the MLE has a tendency to exaggerate confidence. All bootstrap distributions contain 600 samples.

6.7.3 A pathological model: pushing the Dirichlet process to its limits

To demonstrate that CDPBM based models are capable of handling strange underlying survival distributions, we use a highly pathological error model, constructed to have multiple distinct peaks. The model has gate-independent qubit noise defined as the convex mixture of a channel that resets to a fixed pure state, a channel that resets to identity, and the identity channel, or explicitly

$$\mathcal{E}(\rho) = \text{Tr}(\rho)\left(p_1 \left|\psi_r\right\rangle\!\left\langle\psi_r\right| + p_2\frac{\mathbb{I}}{2}\right) + (1 - p_1 - p_2)\rho. \tag{6.28a}$$

We used the parameters $p_1 = 0.9$, $p_2 = 0.001$, and $\left|\psi_r\right\rangle = \mathrm{e}^{-\mathrm{i}0.05(X+Y)}\left|0\right\rangle$ in our simulations. This noise model results in an average gate fidelity of 0.5495, or a decay base of $p =$

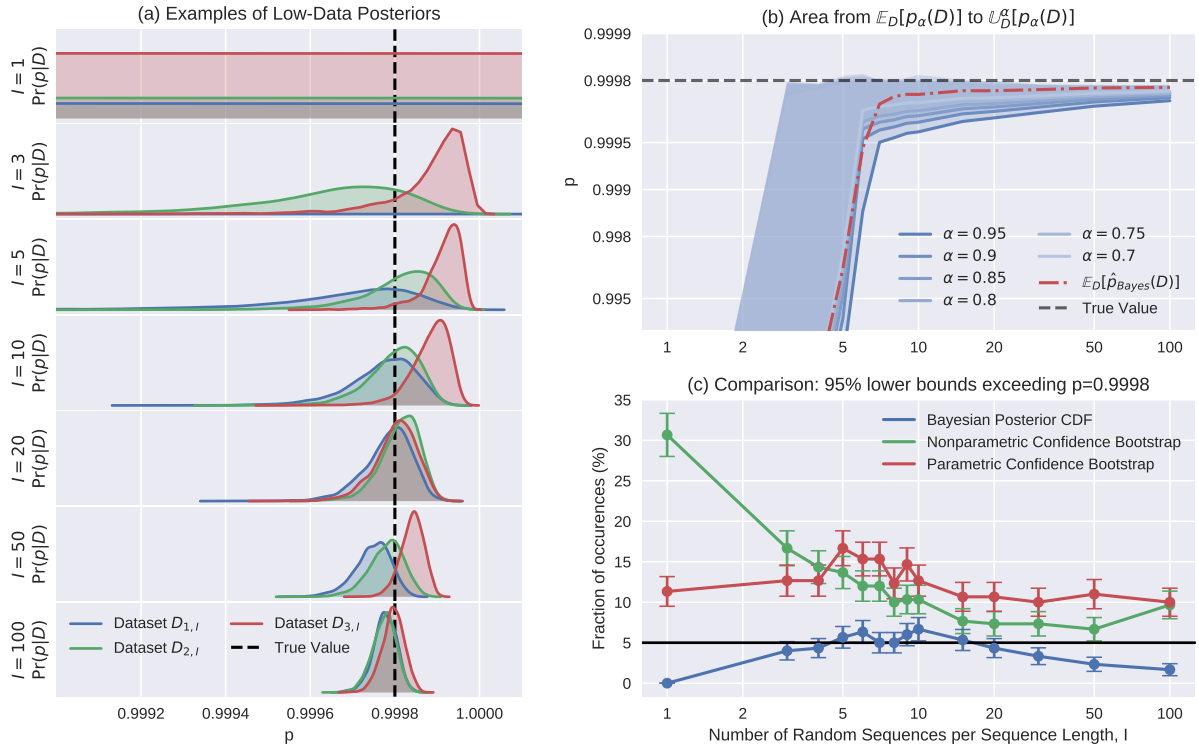

Figure 6.6: Data from the overrotation model Equation 6.24c was simulated 300 times for several values of I, the number of random sequences per sequence length. In all cases $|\mathbb{M}| = 10$ sequences were used with $N = 5$ repetitions of each random sequence. Posteriors were computed for every dataset, p-marginals for three of which are shown in (a) for several values of I. (b) The area between the upper $(100 \cdot \alpha)\%$ quantile and the mean value of $p_\alpha(D)$ is shown for several values of α, demonstrating the posterior's ability to reliably report credible lower bounds for p. (The average value of Bayes' estimate is shown for comparison.) (c) Finally, we isolate the $\alpha = 0.95$ case and display it along with bootstrapped lower 95% confidence bounds, which do not stay under the desired line. These fractions were computed by running the bootstrap method on the same collections of 300 data sets. Error bars are single standard deviations of simple binomial statistics.

0.099. Due to the high value of p_1, this error channel is so bad that running RB as a characterization tool is not a great choice in the first place, and therefore looking at the posterior distribution of (p, A, B) is of little direct use, although similarly bad channels can arise when using interleaved RB to extract tomographic information [100]. In any case, we provide certain marginals at the top of Figure 6.7 anyway. However, our point is to look at the posterior of the survival parameters, $\vec{x}_S$, which are summarized in the bottom section of Figure 6.7. This posterior was computed using the sequence lengths $\mathbb{M} = \{1, 2, 5, 20, 50, 100\}$ with $I = 30$ random sequences per sequence length, and $N = 50$ repetitions each. The same gateset as Section 6.7.1 was used, with the same initial state and measurement operators.

6.7.4 Complicated Tying Function: Leakage RB (LRB)

There are a few protocols which measure leakage of information into and/or out of the qubit subspace [189, 179, 25, 24]. Here we provide an example using our framework with the LRB protocol that is described in Ref. [189] with an experimental implementation reported as a part of Ref. [124]. We have chosen this protocol because it has one of the most complicated tying functions of existing protocols; for a single qubit there are at least seven tying parameters, three of which are not nuisances. Moreover, it is not quite a SPAM-free protocol—some of the information that is necessary to decouple the three parameters of interest from each other is contained in the constant offset term as well as the coefficients of the exponential terms.

We consider a system with a Hilbert space $\mathcal{X} = \mathcal{X}_1 \oplus \mathcal{X}_2$, where $\dim \mathcal{X}_1 = d_1 = 2$ and $\dim \mathcal{X}_2 = d_2 = 1$, and $\mathcal{X}_1$ is the computational subspace. Our noise model is gate independent, equal to the depolarizing leakage extension (DLE) [189] of $\mathcal{E}_{\text{dephasing}} \circ \mathcal{E}_{\text{rot}}$ where

$$\mathcal{E}_{\text{dephasing}}(\rho) = (1 - s)\rho + sZ\rho Z \tag{6.29a}$$

$$\mathcal{E}_{\text{rot}}(\rho) = \mathrm{e}^{-\mathrm{i}\alpha Z/2}\rho \mathrm{e}^{\mathrm{i}\alpha Z/2}, \tag{6.29b}$$

and where we denote the resulting DLE as $\mathcal{E}$. The parameters L_1 and L_2 are called the leakage and seepage respectively, and are given by

$$L_1 = 1 - \operatorname{Tr} \mathbb{I}_1 \mathcal{E}(\mathbb{I}_1/d_1) \tag{6.30a}$$

$$L_2 = \operatorname{Tr} \mathbb{I}_1 \mathcal{E}(\mathbb{I}_2/d_2) \tag{6.30b}$$

where $\mathbb{I}_1$ and $\mathbb{I}_2$ are the projectors onto $\mathcal{X}_1$ and $\mathcal{X}_2$. We see that the leakage quantifies how much population from $\mathcal{X}_1$ leaks out of $\mathcal{X}_1$, and the seepage quantifies how much population

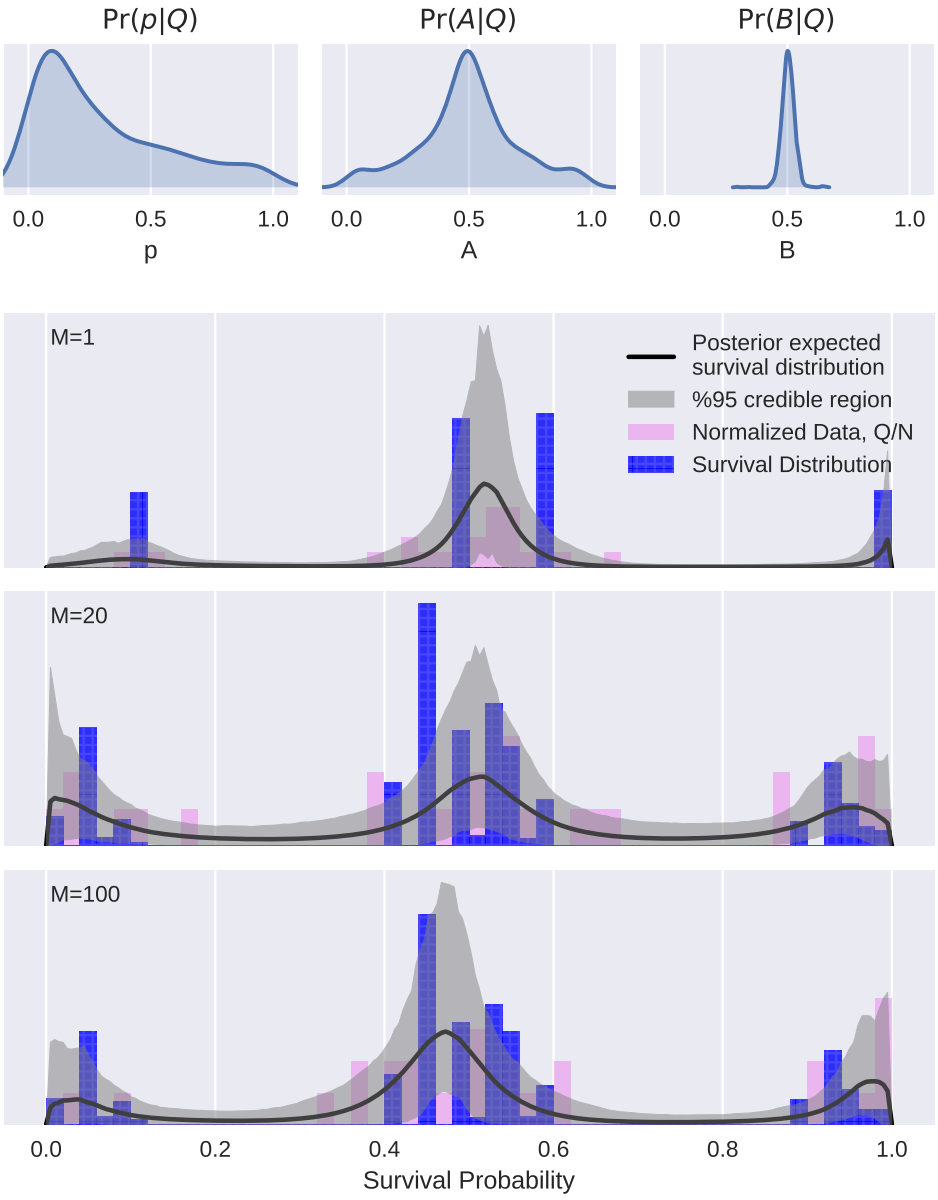

Figure 6.7: The top row of three plots show marginal posterior distributions of the standard RB protocol tying parameters with data simulated according to the pathological noise model defined in Equation 6.28. The bottom column of three plots show posterior summaries of the survival distribution at the sequence lengths $M = 1, 20$ and 100, respectively.

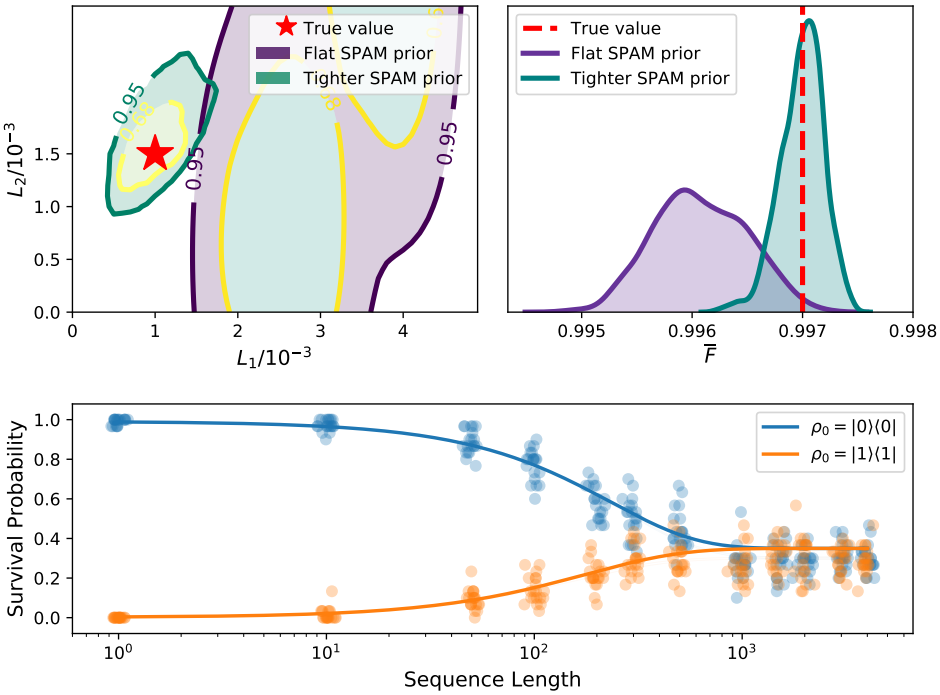

Figure 6.8: Posterior summaries for the LRB protocol under the model from Equation 6.13 and two different prior distributions. Simulated data was sampled at $|\mathbb{M}| = 12$ sequence lengths, each with $I = 15$ random sequences and $N = 30$ repetitions per sequence. The joint posterior marginals of the leakage and seepage parameters is shown (top left), as well as the posterior marginals of the average gate fidelity (top right). The LRB tying functions are plotted using parameters randomly drawn from the posterior tying distribution (bottom). Superimposed are the normalized data, where each dot comes from a unique random sequence.

seeps into $\mathcal{X}_1$ from $\mathcal{X}_2$. We have assumed that our initial states are prepared in $\mathcal{X}_1$ for simplicity in this demonstration. We use the values $s = 0.003$, $\alpha = 0.1°$, $L_1 = 0.001$, and $L_2 = 0.0015$. The average gate fidelity of $\mathcal{E}$ averaged over states in $\mathcal{X}_1$ comes out as $\overline{F} = 0.997001$ with these numbers.

One feature of the fitting method proposed along with the LRB protocol is that it implicitly asserts that certain SPAM parameters sum to unity, and certain other SPAM parameters sum to zero (respectively A and B in our appendix). Though this may be valid for some systems, it depends on the methods of state preparation and measurement for the given device. We have highlighted our ability to loosen this assertion by comparing the posterior distributions due to two priors. In the first, all SPAM parameters have flat non-informative priors, and in the second, prior information is introduced that causes the two sums in question to have support of roughly $\sim 1(\pm 0.05)$ and $\lesssim 0.05$, respectively. Explicit details of this prior, along with the LRB protocol and how we slightly modified its parameterization can be found in Section C.5. Posterior results are summarized in Figure 6.8.

6.8 Conclusions drawn from our treatment of RB+ data processing

In this chapter we have presented a Bayesian approach to analyzing data from RB+ experiments. We used a formal framework to describe such protocols to emphasize that RB and its derivative protocols, from the perspective of statistical inference, are all quite similar. Specifically, they all admit noise model dependent survival distributions which are tied together parametrically by a combination of quantities of interest and nuisance (SPAM) parameters. A handful of examples are summarized in Table 6.1.

We proposed a hierarchical Bayesian model that was constructed to be agnostic to the nature of these survival distributions, and hence to the noise model. This was achieved by modeling them non-parametrically through Dirichlet process priors. We also considered modeling them parametrically through the Beta distribution family. For physically reasonable noise models we found that this simpler family worked well. Therefore we suggest using the non-parametric model in, for example, first runs where the system is not well understood, possibly switching to the parametric model when the system is better characterized and RB+ is being used for tune-ups.

Under either model, however, one ends up with a marginal posterior distribution of the RB+ parameters, from which figures of merit can be computed. We found qualitative

similarity between the nonparametric MLE bootstrap distribution and the posterior distribution of the Bayesian nonparametric model when using a diffuse prior, which merits further study.

We tested our Bayesian models under various noise types, data regimes, and protocols. Our posterior distributions were computed numerically by drawing posterior samples with MCMC methods. As well as fitting well to survival distributions from standard error models (Figure 6.3), we were also able to fit to pathological multi-modal survival distributions (Figure 6.7). Due to our choice of parameterization, estimating probabilities very close to the boundaries $[0, 1]$ is stable. Of particular importance, we found no systematic tendency to over-report gate qualities. Specifically, a numerical study of standard RB in the low data regime showed that posteriors of our model accurately report uncertainty—for example, a 95% credible lower bound on the fidelity is indeed a lower bound to the true value at least 95% of the time (Figure 6.6). This is in contrast to the frequentist bootstrapping techniques we compared to, which do not always pass this sanity test in the low-data regime.

We assumed throughout this work that the model being used for a given dataset was correct. In practice, features like non-Markovian noise may necessitate corrections to a model. A useful direction of research would therefore be to explore Bayesian model selection and cross validation.

Chapter 7

Validating Control: Pulse Design and Benchmarking the Nitrogen Vacancy Center[1]

7.1 Introduction

So far in this thesis we have discussed dynamics and measurements of the NV defect in diamond, characterization of Hamiltonian parameters, numerical methods for generating quantum gates, and techniques for analyzing data from quantum benchmarking experiments. This chapter is an applied synthesis of these ideas. Our goal is to design high quality gates for the NV system and test their performance. We are interested in designing arbitrary unitary gates for the qutrit manifold, as well as for one of the qubit sub-manifolds.

Due to the drift Hamiltonian present when microwave controls are off, pulses implementing gates for this system cannot easily be constructed by intuition, especially in the full qutrit manifold. Therefore this platform is an ideal testbed for numerical optimal control techniques. The main confounding term in the Hamiltonian with respect to pulse

[1]This chapter is based on work to-be-published, with the benchmarking experiments done in collaboration with Chris Granade and Joel Wallman. Chris Granade helped with finding appropriate groups to benchmark, and wrote much of the code to do so. Joel Wallman provided consultation on which experiments to perform, and also on certain aspects of data analysis. This work represents the only qutrit benchmarking experiments performed to date, while also including more benchmarking protocols than are usually performed. The method of using the second rotating frame to find initial pulses for the NV qutrit system is also novel.

finding, we will find, is the coupling to the adjacent 14-nitrogen. All pulses must have high fidelity regardless of the state of the nitrogen atom, which is effectively a 3-level magnetic fluctuator with timescales much longer than experiments.

Characterization of gates and gatesets will be performed with randomized benchmarking and related protocols (RB+). In particular, we will perform the standard RB, interleaved RB, unitarity, and loss protocols on three different gateset implementations.

This work represents the first benchmarking experiments on a qutrit of any sort known to the author, though as with most solid state systems, numerical optimal control methods have been considered for NV centers (see [193] for one example), and standard RB has been performed on an NV qubit manifold and a NV(qubit)-nucleus(qubit) system [150]. The distinguishing feature of this work is that we elect to work at low field with the full three level system and a mixed state nitrogen—it is common, for example, to work in a qubit manifold at $B_z = 512\,\mathrm{G}$ where the nitrogen naturally ends up polarized. This choice is made because we are most interested in testing methods under strenuous circumstances. For example, we elect to work in the compression regime of our amplifier to introduce non-linearities in the distortion function, though it would be simpler to use a higher power amplifier, or have longer pulses.

7.2 Designing Robust Spin-1 NV Gates with GRAPE

In Chapter 5 the GRAPE algorithm was defined, including the extensions of optimizing over a static distribution of parameters, and the effects of classical circuit distortions. In this section we provide a set of steps detailing how this can be applied in practice to an NV center. The steps are as follows, with each step detailed in a subsequent subsection.

1. Find robust pulse in the second secular rotating frame with a smooth amplitude envelope and exponential distortion.
2. Convert pulse back into first rotating frame and re-optimize.
3. Introduce circuit distortion and re-optimize in units specified by Rabi experiments.
4. Tune-up pulse shape with a classical feedback loop.

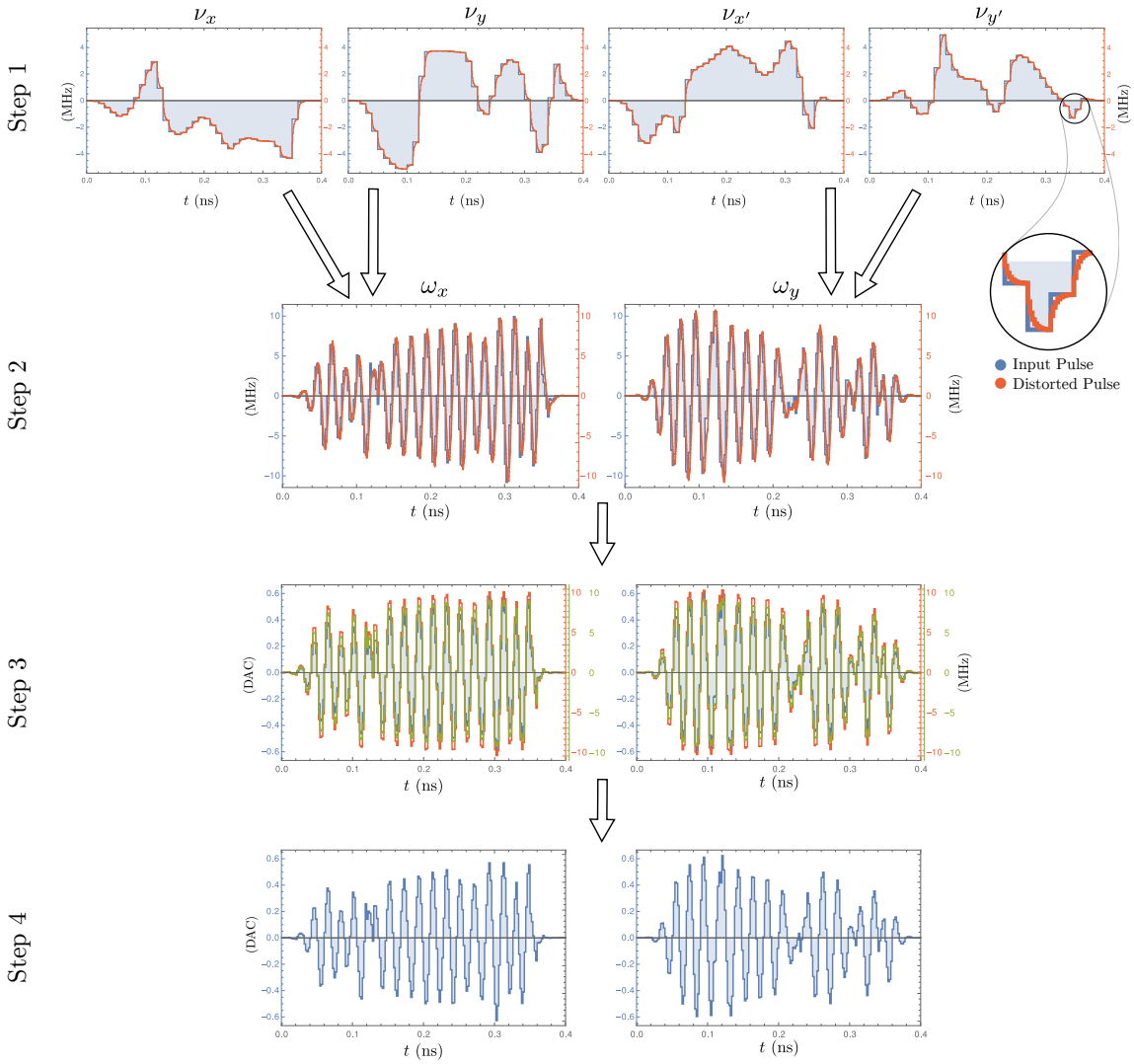

Figure 7.1: Illustration of pulse design steps for qutrit gates. (Step 1) The pulse is initially optimized with four control amplitudes for the Hamiltonians $\mathbf{S}_\mathrm{x}$, $\mathbf{S}_\mathrm{y}$, $\mathbf{S}'_\mathrm{x}$, and $\mathbf{S}'_\mathrm{y}$ in the second rotating frame. Blue indicates the input pulse, and red indicates distorted pulse. (Step 2) Channel pairs are mixed together in quadrature at the Zeeman energy, $\omega_e = 50\,\mathrm{MHz}$, and this new pulse is re-optimized in the first rotating frame. (Step 3) The pulse is re-optimized in units of the DACs, and including distortion due to amplifier compression. The red and green curves are amplitudes seen by the $|0\rangle \leftrightarrow |-1\rangle$ and $|1\rangle \leftrightarrow |+1\rangle$ transitions respectively. (Step 4) Pulses are tuned-up with a feedback loop involving a fast oscilloscope.

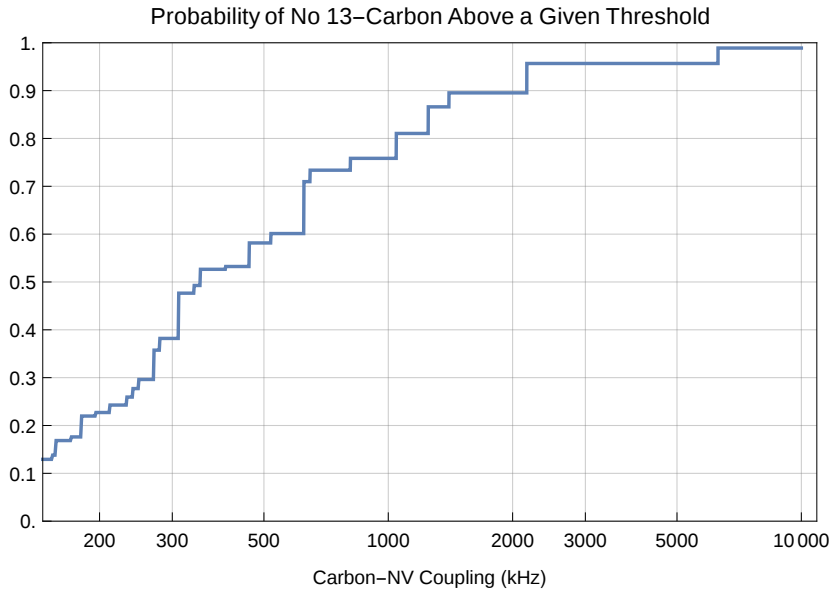

Figure 7.2: The chances of getting an NV with a clean 13-carbon environment. We see that there is roughly a 40% chance of having no 13-carbon with a coupling strength greater than 300 kHz. This figure was generated by computing the dipolar coupling strength between an electron at the coordinate origin and a 13-carbon at every diamond cubic lattice position in a 5 nm radius.

7.2.1 Step 0: Preliminary calibrations

Before describing Step 1, we discuss some preliminary calibrations of the system.

Choosing an NV center

The diamond used in our experiment is roughly 2 mm × 2 mm × 2 mm with ⟨100⟩ crystal orientation, unannealed, and has natural abundance (1.1%) carbon. The concentration of negatively charged NV centers is roughly 10-20 per 50 um × 50 um × 50 um volume. We searched several NV centres to find one with a linewidth no greater than 300 kHz. This happens with a probability of roughly 40%, as shown in Figure 7.2. We work roughly 90 um below the surface of the diamond to avoid background counts from bright features on and above the surface of the diamond.

Static Magnetic Field Calibration

Our static magnetic field is generated by a single cylindrical rare-earth magnetic with a 7.5mm diameter mounted on its own 3-axis stage. We simulated a 3-D model of the magnetic field by integrating the equation of a magnetic dipole over the cylindrical volume, assuming uniform magnetization which can be roughly inferred from the magnet's datasheet. Therefore, given a coordinate (x, y, z) relative to the center of the cylinder we have the field prediction $\vec{B}_{\text{predict}}(x, y, z)$.

To predict the field seen at the focus of the confocal microscope where the NV under study is located, we require a precise relative coordinate and orientiation between the magnet and the NV center. To learn this transformation, the magnet was moved to each point of a 3-D grid of motor stage positions, a few dozen positions in total, $\{(x_i, y_i, z_i)\}_{i=1,i=I}$, and at each point, a CW experiment was performed on the NV center of interest. Each of these data sets was fit individually to estimate the projection of the magnetic field along the z axis of the PA, $\hat{B}_{z,i}$ corresponding to (x_i, y_i, z_i) for each $1 \leq i \leq I$. Performing the least squares fit

$$\underset{x_0,y_0,z_0,\theta_0,\phi_0,B_0,\kappa}{\operatorname{argmin}} \sum_{i=1}^{I} (\hat{B}_{z,i} - (\eta(\theta_0, \phi_0) \cdot \kappa \vec{B}_{\text{predict}}(x_i - x_0, y_i - y_0, z_i - z_0) + B_0))^2 \tag{7.1}$$

where $\eta(\theta, \phi) = (\cos\phi \sin\theta, \sin\phi \sin\theta, \cos\theta)$ yields the desired values. Here,

$$(x_0, y_0, z_0, \phi_0, \theta_0)$$

comprises the coordinate transform, where (x_0, y_0, z_0) are interpretted as the magnet stage positions that would situate the magnet concentric with the microscope's focus (though such an arrangement is not physically possible due to clearances and stage limits), and the angles (ϕ_0, θ_0) include both the NV orientation and slight tilt and rotation of the diamond itself relative to the stage. The parameter $\kappa \approx 1$ is a correction to the magnetization value of the simulation, and B_0 allows for a background field independent of the permanent magnet.

Note that these data also make clear which of the four possible orientations our NV center has; only one of the four orientations will come close to giving a good least squares fit given rough bounds on the fit parameters.

Having a value for $(x_0, y_0, z_0, \theta_0, \phi_0, B_0, \kappa)$, we can optimize the inverse problem and determine the stage position that most closely produces the magetic field of interest, according to the simulated model. In our experience, this method yields the desired field to

an accuracy of about 1 MHz or, 0.3 G. When attempting to find a magnetic field aligned to the PAS, small deviations in the x and y directions are generally acceptable, as they appear in correction terms to the Hamiltonian in the rotating frame that enter with strengths of order $\gamma_e B_x/\Delta$, which is further justified by the robustness plots in Figure 7.5 discussed later. This natural robustness can be used to our advantage to generate accurate fields along the z axis. For example, suppose we were interested in creating a Zeeman splitting of 100 MHz with a field aligned to the PAS, but experimentally the splitting is found to be 100.5MHz with pulsed-ESR; an experiment is described below. We then query the field simulator for the motor position which results in an aligned splitting of 99.5MHz, which, experimentally, will yield a splitting close to 100MHz. This process generally converges in two to three rounds to within the linewidth.

Ideally, following the prescription of Chapter 4, one would use a fully adaptive statistical inference to measure the Zeeman field. Lacking this infrastructure, we may use a single quantum Ramsey experiment with time-proportional-phase (TPPI) increment on both the $m_S = 0 \leftrightarrow -1$ and the $m_S = 0 \leftrightarrow +1$ transitions. A TPPI Ramsey sequence is defined by

$$\left(\frac{\pi}{2}\right)_{\phi=0} \text{———} \tau \text{———} \left(\frac{\pi}{2}\right)_{\phi=\tau\cdot\omega_{\text{TPPI}}} \tag{7.2}$$

where ϕ is the phase of the pulse relative to x, $\pi/2$ is the single-quantum nutation angle of the pulse, τ is the wait time, and ω_{TPPI} is the TPPI frequency. A Ramsey experiment works by accumulating a relative phase between two states—in our case using one of $|0\rangle + |\pm 1\rangle$—due to a term along the z access for a varying time τ, converting this phase into population, so that the averaged output signal as a function of τ, $s(\tau)$, contains the strength of the z field as a frequency. The purpose of TPPI is to shift the Fourier transform of $s(\tau)$, denoted $\tilde{s}(\omega)$, left or right by an amount ω_{TPPI} without the pulses being off-resonance of the spins, which would reduce the amplitude of the signal.

This sequence is performed at both of the frequencies $\omega_{\text{center}} \pm \omega_{\text{offset}}$, where ω_{center} is the current best guess of the ZFS Δ, and ω_{offset} is the current best guess of $\omega_e := \gamma_e B_z$. Denoting the frequency of the central peak of $\tilde{s}(\omega)$ as ω_+ and ω_-, respectively for the cases $\omega_{\text{center}} + \omega_{\text{offset}}$ and $\omega_{\text{center}} - \omega_{\text{offset}}$, we get the improved estimates

$$\hat{\omega}_e = \frac{\omega_+ - \omega_-}{2} + \omega_{\text{offset}} \tag{7.3}$$

$$\hat{\Delta} = \omega_{\text{TPPI}} + \omega_{\text{center}} - \frac{\omega_+ + \omega_-}{2} \tag{7.4}$$

for Δ and ω_e.

7.2.2 Step 1: Optimize the second rotating frame

Having measured set the static field to be nearly aligned with the PAS with a value $\omega_e = 50\,\text{MHz}$, we proceed with pulse-finding for a particular target unitary U_{target}. In the case of finding qutrit pulses, we use the standard utility function of Definition 5.3, while for qubit pulses, we restrict our attention to the subspace spanned by $|0\rangle$ and $|+1\rangle$, using the generalized objective function of Equation 5.31 with

$$A_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad B_1 = \begin{pmatrix} [U_{\text{target}}] \\ 0 \quad 0 \end{pmatrix}. \tag{7.5}$$

However, qubit pulses skip Step 1, beginning directly with Step 2.

We first enter the rotating frame $\Delta\mathbf{S}_{\text{z}}^2$ and subsequently the rotating frame $\omega_e\mathbf{S}_{\text{z}}$, as discussed in Section 2.4. We use the second secular rotating frame, provided in generality in Equation 2.22, given by

$$H = 2\pi A\mathbf{S}_{\text{z}} + 2\pi\frac{1}{2}(1+\gamma)\left[\nu_x(t)\mathbf{S}_{\text{x}} + \nu_y(t)\mathbf{S}_{\text{y}} - \nu_{x'}(t)\mathbf{S}'_{\text{x}} + \nu_{y'}(t)\mathbf{S}'_{\text{y}}\right]. \tag{7.6}$$

Here, A is the hyperfine coupling strength to the adjacent 14-nitrogen, γ is a unitless parameter specifying over-rotation error, and we have four independent (but power-bandwidth limited) control Hamiltonians with envelopes labeled by $\{\nu_x, \nu_y, \nu_{x'}, \nu_{y'}\}$.

As in Section 3.6.1, we are treating the nitrogen as being in one of its eigenstates with equal probability, which is justified by being at room temperature in low field. Therefore, given a desired pulse U_{target}, we require that a set of controls generate a high quality pulse for all three values of the nitrogen spin, $m_{\mathbb{I}} \in \{-1, 0, +1\}$. This is encoded in our robustness distribution (see Section 5.4) as the mixture

$$\Pr(A) = \frac{1}{3}\sum_{m\in\{-1,0,+1\}} \text{Normal}\left(mA_0, \sigma_A\right) \tag{7.7}$$

where $A_0 \approx 2.14\,\text{MHz}$ is the measured hyperfine splitting of the nitrogen from the calibration step. We use an independent robustness distribution for γ,

$$\Pr(\gamma) = \text{Normal}\left(0, \sigma_\gamma\right). \tag{7.8}$$

The second rotating frame operator, $U(t) = \mathrm{e}^{\mathrm{i}2\pi t\omega_e\mathbf{S}_z}$, coincides with the trivial frame change, $\mathbb{I}$, with a period of $1/\omega_e = 20\,\text{ns}$. For convenience (though not necessary if taken

into account), we choose gate lengths which are multiples of this period to avoid frame changes. Our qutrit gates have length 400 ns. Denote the length of a pulse by T. Large $dt = 10$ ns input time-steps are used in this step. Output time-steps are smaller, $\delta t = 1$ ns, to capture the dynamics of the 2 ns exponential rise time distortion, see Section 5.7.5, which is used to help control the pulse's bandwidth.

All four control channels are constrained to lie within an envelope with gaussian rise and fall, $|\nu_\cdot(t)| <= E(t)$ for all $t \in [0, T]$, where the gaussian envelope is defined by

$$E(t) = \Omega_0 \begin{cases} e^{((r+d/2)T+t)^2/(2.5rT)^2} & t < (r + d/2)T \\ 1 & (r + d/2)T \leq t \leq (1 - r - d/2)T \\ e^{((1-r-d/2)T+t)^2/(2.5rT)^2} & t > (1 - r - d/2)T \end{cases} \tag{7.9}$$

Here, $r \in [0, 1]$ and $d \in [0, 1]$ are respectively the risetime and deadtime, each expressed as a fraction of T. This envelope is enforced by clipping the pulse after each step in the gradient descent algorithm. The purpose of the deadtime is to add small delays of no microwave energy between pulses, so that the ringdown of one pulse is unable to seep into the next pulse. However, since the pulse is optimized on the full interval $[0, T]$, deadtime is accounted for. We set $dT = 10$ ns and $r = T/7$ for our pulses. The coefficient $\Omega_0 = 6$ MHz controls the maximum amplitude of the pulse, and is deliberately set 10% below the maximum value allowed by the control circuit so that there is room for pulse adjustment in later steps.

Pulses at this stage are optimized to about $\Phi = 0.999$, picking the best result out of $\sim 20 - 50$ initial guesses. Initial guesses are formed by choosing random values for the pulses every 10 steps and sampling from a spline between them.

7.2.3 Step 2: Re-optimize in the first rotating frame

The rotating frame used in the previous section was used to smooth the landscape of control space, reducing ripples at the required modulation frequency of ω_e. In this step, pulses are converted back into the first rotating frame by cosine-mixing up at frequency ω_e, and sampling at $dt = 2$ ns. We found the second rotating frame allows most initial guesses to succeed with point-like parameter robustness distributions $\sigma_A = \sigma_\gamma$, whereas in the first rotating frame, random initial guesses rarely succeed at all with the same distribution.

The mixing operation is derived by inverting the definitions of $\{\nu_x, \nu_y, \nu_{x'}, \nu_{y'}\}$ from

Section 2.4. This results in coefficients for the control Hamiltonians $\mathbf{S}_\mathrm{x}$ and $\mathbf{S}'_\mathrm{y}$ given by

$$\begin{aligned}\omega_x(t) &= \nu_1(t)\cos(2\pi\omega_e t + \phi_1(t)) \quad \text{and} \\ \omega_y(t) &= \nu_2(t)\cos(2\pi\omega_e t + \phi_2(t))\end{aligned} \tag{7.10}$$

respectively, where

$$\begin{aligned}\nu_1(t) &= \sqrt{\nu_x(t)^2 + \nu_y(t)^2} \quad \text{and} \quad \phi_1(t) = \arctan_2(\nu_x(t), \nu_y(t)) \\ \nu_2(t) &= \sqrt{\nu_{x'}(t)^2 + \nu_{y'}(t)^2} \quad \text{and} \quad \phi_2(t) = \arctan_2(\nu_{x'}(t), \nu_{y'}(t))\end{aligned} \tag{7.11}$$

with $\arctan_2(x, y)$ the quadrant-preserving arctan function.

Because of clipping with the envelope function[2], the non-commutation of the exponential distortion operator through the mixing process, and other discretization artifacts, the utility function of the pulse drops significantly $(10 - 20\%)$. However, this loss is very quickly recovered by re-optimizing the pulse in the new frame, using it as the initial guess in the GRAPE algorithm.

Pulses for qubit gates, having skipped Step 1, use a random initial guess for $\omega_x(t)$ and $\omega_y(t)$ here instead of using the cosine-mixed output of the previous step. However, we use the rotating frame $(\Delta - \omega_e)\mathbf{S}_\mathrm{z}^2$ instead of $\Delta\mathbf{S}_\mathrm{z}^2$ so that all power is targeted at the relevant transition. Because of the reduced demands of the utility function, pulses are shorter, at $T = 100\,\mathrm{ns}$, though the envelope function, distortions, and parameter distributions are the same.

7.2.4 Step 3: Re-optimize in DAC units including power-imbalance

Pulses resulting from Step 2 have control amplitudes which are in units of the Hamiltonian, which, in our case, are MHz. To be used in practice they need to sent from computer memory to the digital-to-analog (DAC) converter of the AWG, which uses unitless numbers between -1 and 1, corresponding to $4.5V_{pp}$ at $50\,\Omega$ in our setup. While this is just a rescaling, the circuit the signal then passes through includes IQ imbalance, filters, and compression from the amplifier. In this project, we purposely work into the 3 dB compression zone of the amplifier to test our methods more thoroughly.

We characterize the amplifier compression by measuring the Rabi frequency of the NV as a function of DAC amplitude. This is done for both the $|0\rangle \leftrightarrow |-1\rangle$ transition at $\Delta - \omega_e \approx$

[2]The pulse envelope is defined independently over four channels, and the mixing process therefore allows the combined two-channel pulse to exceed new envelopes.

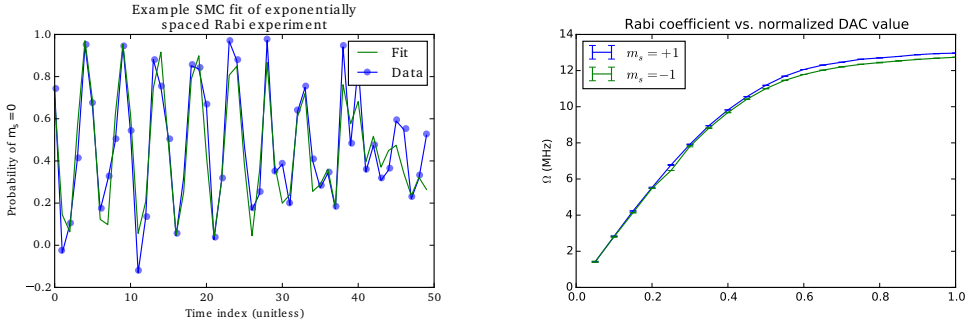

Figure 7.3: (left) Example of a single Rabi experiment, along with a fit produced by simulating the Hamiltonian with the estimates calculated using SMC. Note that the x-axis indexes arbitrary but incrementing Rabi pulse durations; the first half of the experiments are linearly spaced, whereas the latter half are exponentially spaced. (right) Power compression due to the amplifier. The output voltage of the AWG is swept and the Rabi frequency estimated at each value, for both the $m_S = -1$ transition at $IF = -50$MHz and the $m_S = +1$ transition at $IF = +50$MHz.

2820 MHz and the $|1\rangle \leftrightarrow |+1\rangle$ transition at 2920 MHz. Results of these measurements are shown in Figure 7.3. Given the observed discrepancy between the asymptotes of the curve, we wish to model the different transition rates in our pulse-finding. We label a power fit to these compression curves $c_p(v)$ and $c_m(v)$ for the positive and negative transition respectively, and use them to define the distortion operator $g_2 : \mathbb{R}^N \otimes \mathbb{R}^2 \to \mathbb{R}^M \otimes \mathbb{R}^4$, $g((v_x(t), v_y(t))) = (\omega_{p,x}(t), \omega_{p,y}(t), \omega_{m,x}(t), \omega_{m,y}(t))$ as

$$\begin{aligned} \omega_{p,x}(t) &= c_p(v_x(t)), & \omega_{p,y}(t) &= c_p(v_y(t)) \\ \omega_{m,x}(t) &= c_m(v_x(t)), & \omega_{m,y}(t) &= c_m(v_y(t)). \end{aligned} \tag{7.12}$$

These four amplitudes are used as coefficients of the single transition control Hamiltonians $\mathbf{S}_x^+$, $\mathbf{S}_y^+$, $\mathbf{S}_x^-$, and $\mathbf{S}_y^-$ defined in Table 2.2. The total distortion operator used in this step is given by $g = g_2 \circ g_1$, where g_1 is an exponential rise-time distortion matching the rise-time of the circuit.

The pulse resulting from this step is generated by inverting[3] the pulse from Step 3, and then using it as the initial guess in a re-optimization this step using the distortion operator g. The envelope function $E(t)$ is still used, with $\Omega_0 = 1$.

[3] This distortion is not technically invertible because it takes two control and outputs four controls, but we can approximate the inversion by using the inverse of the average compression $(c_p + c_m)/2$.

7.2.5 Step 4: Tune-up shape with in-hardware classical feedback loop

We found that our distortion model was not sufficiently accurate to achieve the desired pulse shape at the NV site. This was found by monitoring the pulse shape through a directional coupler with a fast oscilliscope, as shown in the circuit diagram of Figure 2.6. To make final corrections, we used a simple iterative feedback loop. At step $k = 0$, we set $\vec{q}_0 = \vec{q}$, the pulse we wish the quantum system to see. We then iterate by correcting each time step by a fraction of the measured residual,

$$\vec{q}_{k+1} \leftarrow \vec{q}_k + \epsilon_k(\vec{q} - m(\vec{q}_k)), \tag{7.13}$$

where $m(\cdot)$ represents a pulse sent through the circuit, as measured by the oscilloscope. We used the coefficients $\epsilon_k = 0.1(6-k)$ for $k = 1, 2, 3, 4, 5$. This measurement $m(\cdot)$ is performed by sampling the analog signal at a rate $10\times$ that of the AWG (5 GHz), performing a digital mix-down at the center frequency of the control circuit, and finally decimating the result by a factor of 10. The amplitude is corrected to have the same units as the input pulse by sending a constant voltage through the AWG and measuring the amplitude at the oscilloscope. Results of this loop are shown in Figure 7.4 for an example pulse.

The utility of this correction is validated by running randomized benchmarking and related protocols before and after each pulse in the gateset undergoes this correction. Results of these experiments are analyzed in section Section 7.4 and summarized in Table 7.6.

7.2.6 Robustness Plots

Although pulses are only optimized over a joint distribution of A and γ, in this section we additionally simulate the final pulse for several other parameters. In Figure 7.5 we plot the utility function $\Phi_{g,\vec{a}}[\vec{p}]$ (for the final input pulse $\vec{p}$ obtained from Step 4) as functions of parameter values—such plots are known as *robustness plots*. We also consider the effects that a single nearby carbon-13 has on the pulse, which is seen in Figure 7.6. Though we only show robustness plots for one pulse ($U_{\text{target}} = F$ in this case, see Section 7.3), such plots are studied for all generated pulses before they are used experimentally.

7.3 Selecting Unitary 2-designs for Benchmarking

In Chapter 6 we found that all RB+ protocols require the ability to physically implement any member of a unitary 2-design, or 1-design.

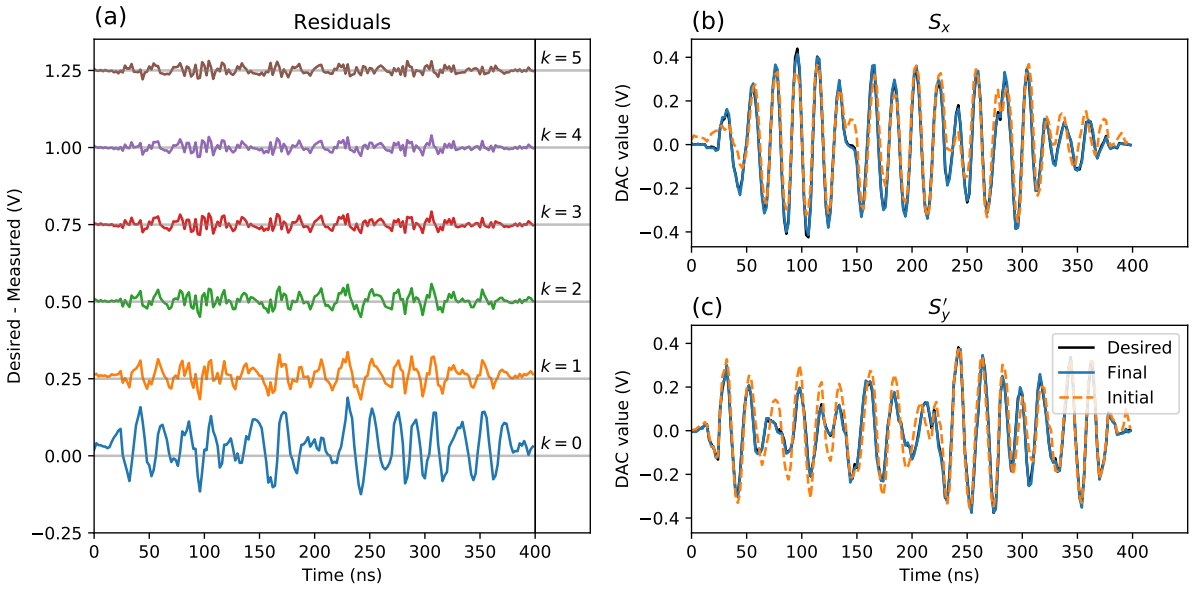

Figure 7.4: (a) Residuals of measured pulse envelopes (with an oscilliscope) to desired pulse envelopes, after k steps of correction. Residuals are offset in the y-axis by $k/4$ for visual clarity. It is seen that this method does not provide much correction past the step $k = 3$, and moreover, the similarity of $k = 3, 4, 5$ demonstrates that we are above the measurement noise threshhold. (b-c) Pulse envelopes for $\mathbf{S}_\mathrm{x}$ and $\mathbf{S}'_\mathrm{y}$ control Hamiltonians as desired (solid black), as measured after no correction ($k = 0$, dashed), and as measured after five steps of correction ($k = 5$, solid blue). Note that the 'Desired' lines are almost completely obstructed by the 'Final' lines.

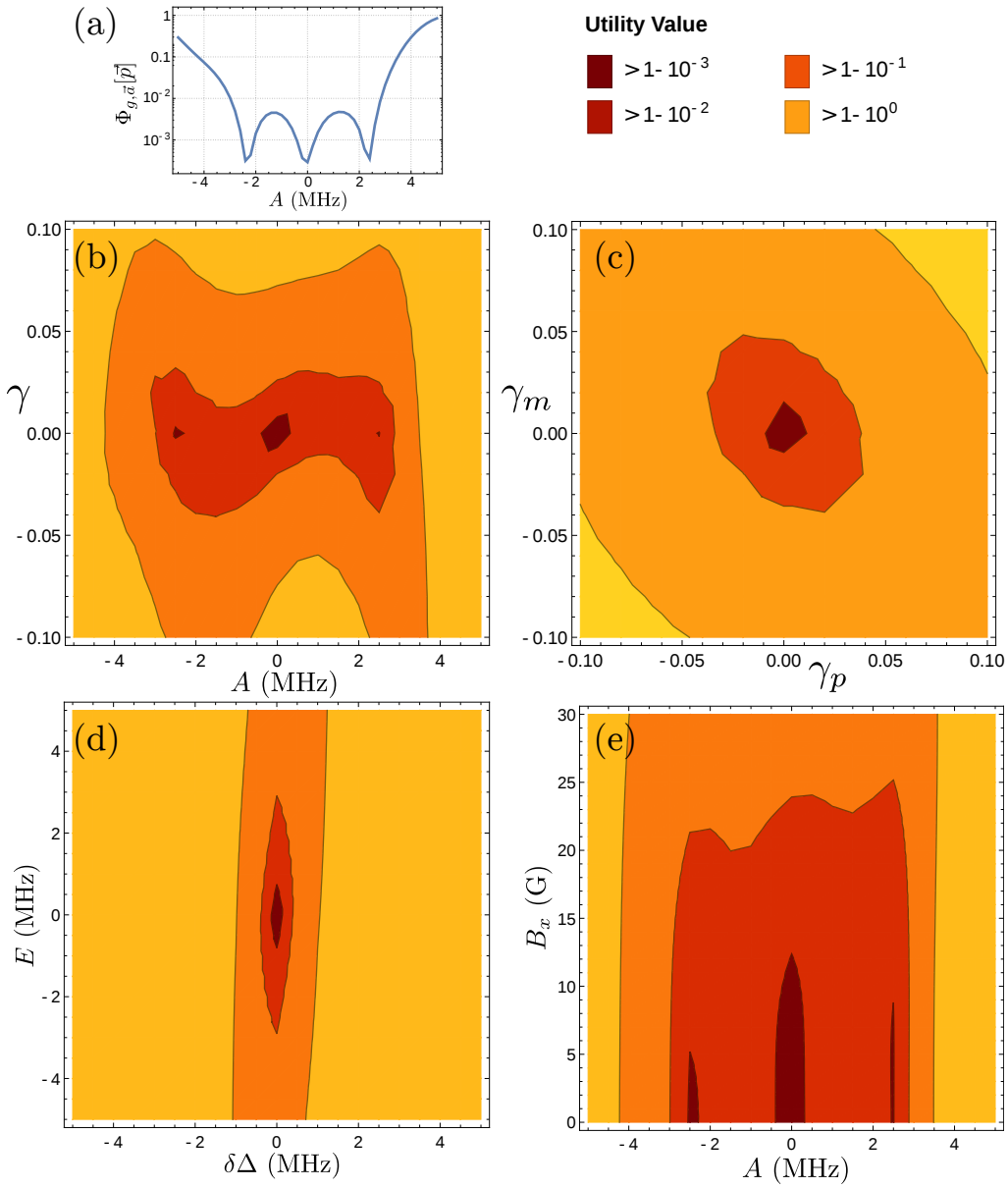

Figure 7.5: Robustness plots of the qutrit pulse F, see Section 7.3. Certain parameters of the Hamiltonian are swept and the utility function is plotting at each value. Slices include the hyperfine value A, the global power error γ, the transition-specific power adjustments γ_p and γ_m, deviations from the true ZFS value $\delta\Delta$, the crystal strain E, and the off-axis field B_x.

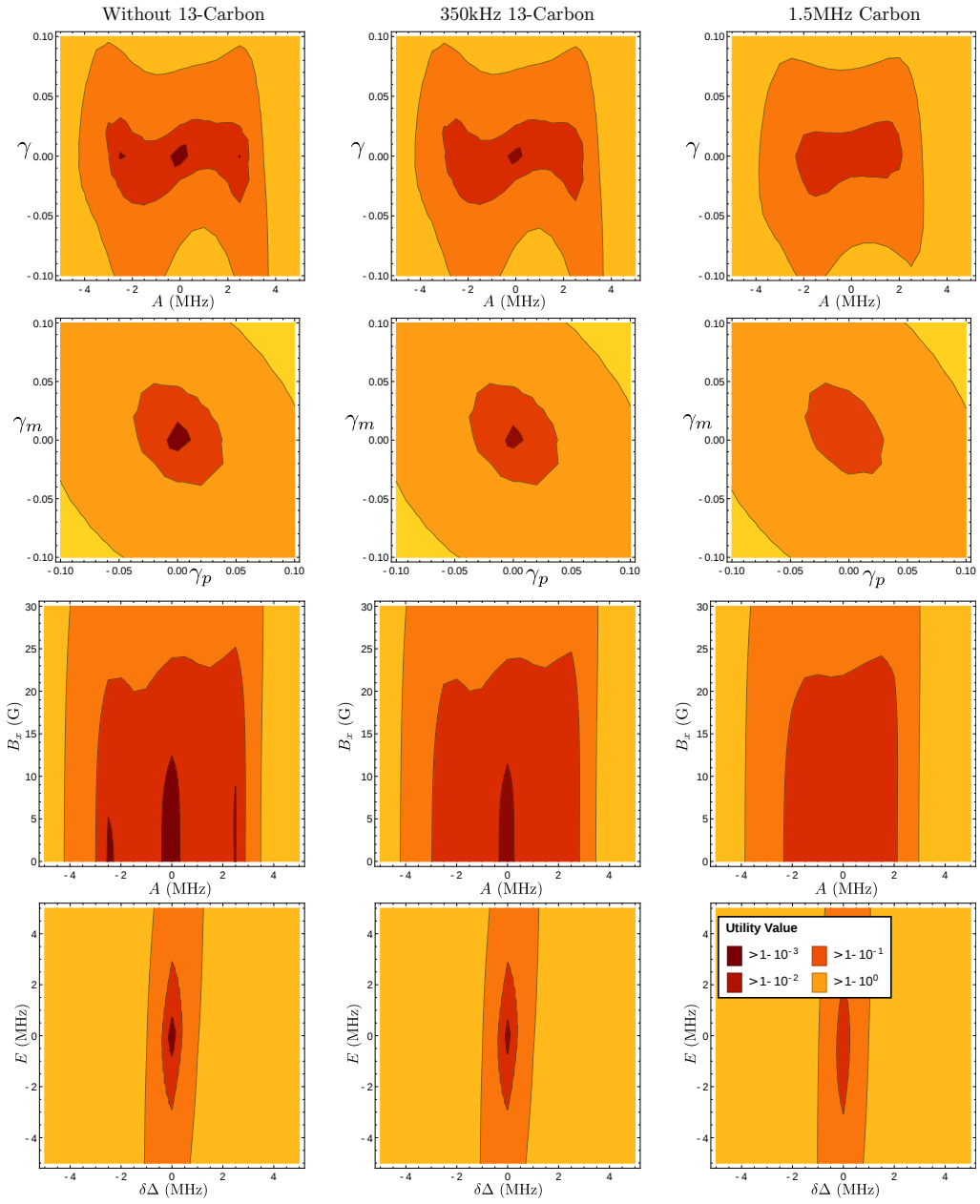

Figure 7.6: Robustness plots of the qutrit pulse F, see Section 7.3, with no nearby 13-carbon (left column), a 13-carbon with 350 kHz dipolar coupling, a 13-carbon with 1.5 kHz dipolar coupling. Parameters on the axes are described in Figure 7.5.

Definition 7.1. A collection $\{U_1, U_2, \ldots, U_L\} \subseteq \mathcal{U}(d)$ of unitary matrices is called a *unitary t-design* if for any complex homogeneous polynomial f of degree (s, s) less than or equal to (t, t) in the entries of U and $\overline{U}$, it holds that $\frac{1}{L}\sum_{i=1}^{L} f(U) = \int_{\mathcal{U}(d)} f(U)dU$. Here, integration takes place over the normalized Haar measure on $\mathcal{U}(d)$, and $f(U)$ denotes f applied to the $2d^2$ complex entries $U_{1,1}, U_{1,2}, \ldots, U_{d,d}; \overline{U}_{1,1}, \overline{U}_{1,2}, \ldots, \overline{U}_{d,d}$.

Informally, a unitary t-design is a finite set of unitaries with symmetry sufficient to approximate continuous Haar integrals with more tractable finite sums, at least for t-polynomials. Many RB+ protocols require a unitary 2-design because twirling a channel $\Lambda \in \mathcal{C}(d)$ by a unitary 2-design, i.e. the channel which acts on ρ as $\frac{1}{L}\sum_{i=1}^{L} U_i^\dagger \Lambda(U_i \rho U_i^\dagger) U_i$, results in a unitarily-invarianct depolarizing channel of one parameter whose average gate fidelity is equal to that of the original channel Λ. We see that entries of this twirl sum are homogeneous polynomials of degree $(2, 2)$ in the entries of U_i and $\overline{U}_i$.

Since every unitary 2-design is also a unitary 1-design, in this chapter we focus on the former, so that we only have one gate set to worry about. Moreover, the standard randomized benchmarking protocol also requires that the 2-design be a group. Since the Clifford group in dimension d is a 2-design, it is a useful place to start.

The qubit and qutrit Clifford groups, $\mathcal{C}(2)$ and $\mathcal{C}(3)$, are generated by S and H [27], where for qubits,

$$S = \begin{pmatrix} 1 & 0 \\ 0 & \mathrm{i} \end{pmatrix} \quad \text{and} \quad H = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \tag{7.14}$$

and for qutrits,

$$S = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \mathrm{e}^{2\pi\mathrm{i}/3} & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad H = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 1 & 1 \\ 1 & \mathrm{e}^{2\pi\mathrm{i}/3} & \mathrm{e}^{-2\pi\mathrm{i}/3} \\ 1 & \mathrm{e}^{-2\pi\mathrm{i}/3} & \mathrm{e}^{2\pi\mathrm{i}/3} \end{pmatrix}. \tag{7.15}$$

Since we are working in low dimensions, it is easiest to generate these groups by brute force, systematically multiplying generators together, discarding duplicates, until no new elements arise in an entire stage of multiplication. In this way, every member of the group $\mathcal{C}(2)$ and $\mathcal{C}(3)$ can be described as a word in the letters S and H. To reduce the length of these words, we choose to add two more generators, X and Z, which are the usual Paulis for $d = 2$, and are given by the Weyl operators

$$X = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \quad \text{and} \quad Z = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \mathrm{e}^{2\pi\mathrm{i}/3} & 0 \\ 0 & 0 & \mathrm{e}^{-2\pi\mathrm{i}/3} \end{pmatrix}. \tag{7.16}$$

for $d = 3$. Under this scheme, the groups $\mathcal{C}(2)$ and $\mathcal{C}(3)$ are presented in Table 7.1 and Table 7.2, respectively.

It is known that for any subgroup of the Clifford group $\mathcal{G} \leq \mathcal{C}(d)$ to be a 2-design, it must have an order of at least $|\mathcal{G}| \geq d^2(d^2-1)$ [69]. This bound is equal to 12 for qubits and 72 for qutrits. We found subgroups of both $\mathcal{C}(2)$ and $\mathcal{C}(3)$ saturating these bounds by systematically choosing two elements of the Clifford group, computing the generated group, and counting its elements—an inelegant brute-force approach. Indeed, defining

$$F := SHSS = \begin{cases} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ \mathrm{i} & \mathrm{i} \end{pmatrix} & d = 2 \\ \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & \mathrm{e}^{-2\pi\mathrm{i}/3} & 1 \\ \mathrm{e}^{2\pi\mathrm{i}/3} & \mathrm{e}^{2\pi\mathrm{i}/3} & 1 \\ 1 & \mathrm{e}^{2\pi\mathrm{i}/3} & \mathrm{e}^{2\pi\mathrm{i}/3} \end{pmatrix} & d = 3 \end{cases} \tag{7.17}$$

for both qubits and qutrits, we found that $\mathcal{G}_3 := \langle F, H, X, Z \rangle$ generates an order 72 subgroup of $\mathcal{C}(3)$, and $\mathcal{G}_2 := \langle F, X, Z \rangle$ generates an order 12 subgroup of $\mathcal{C}(2)$—presentations are shown in Table 7.3 and Table 7.4. We can verify that these subgroups are still 2-designs by computing the frame potential

$$\sum_{U,W \in \mathcal{G}} \frac{\left|\operatorname{Tr} U^\dagger W\right|^4}{|\mathcal{G}|^2} \tag{7.18}$$

which is equal to 2 if and only if the finite set $\mathcal{G} \subseteq \mathcal{U}(\mathcal{H})$ is a unitary 2-design [69].

If gates are to be physically implemented by composing gates that implement the group generators, then the average word count of the members of the group is highly relevant as it will affect the quality of the total gateset—the identity gate will have unit fidelity (if implemented as a length-0 pulse), and the worst gate, in our presentation of $\mathcal{C}(3)$, is composed of 7 elements. At the time that the data was collected, the effects of highly gate dependent noise had been considered, but was not yet a well-studied problem, and indeed, open problems still remain. Recently, however, it has been proven that the randomized benchmarking decay curve is exponential under arbitrary amounts of gate dependent (but non-markovian) noise [178, 141], and that the decay rate of this exponential is a simple function of a metric known as the gateset circuit fidelity (proven for a qubit, conjectured for all dimensions) [20]. Not having known these results, and more importantly being, primarily motivated by the underlying question of *'how well can we design shaped pulses for arbitrary unitaries?'*, we mitigated against such enormous gate dependent errors by choosing to separately design a pulse for every member of both $\mathcal{G}_2$ and $\mathcal{G}_3$, even the idenity elements, each of the same duration and with the same power constraints.

$\mathbb{I}$, S, H, X, Z, SH, SX, SZ, HS, HX, HZ, XS, XZ, SHS, SHX, SHZ, HSH, HSX, HSZ, HXS, HXZ, XSH, $SHSZ$, $SHXS$

Table 7.1: Presentation of $\mathcal{C}(2)$ in terms of the generators $\langle H, S, X, Z\rangle$. There are 24 elements and the average word length is ~ 2.29.

$\mathbb{I}$, H, S, X, Z, HH, HS, HX, HZ, SH, SS, SX, SZ, XH, XS, XX, XZ, ZZ, HHH, HHS, HHX, HHZ, HSH, HSS, HSX, HSZ, HXH, HXS, HXX, HXZ, SHH, SHS, SHX, SHZ, SSH, SSX, SSZ, SXH, SXS, SXX, SXZ, SZZ, XHH, XHS, XHX, XSH, XSS, XSX, XXS, XXZ, XZZ, $HHHS$, $HHHX$, $HHHZ$, $HHSH$, $HHSS$, $HHSX$, $HHSZ$, $HHXH$, $HHXS$, $HHXZ$, $HSHH$, $HSHS$, $HSHX$, $HSHZ$, $HSSH$, $HSSX$, $HSSZ$, $HSXH$, $HSXS$, $HSXX$, $HSXZ$, $HXHH$, $HXHX$, $HXSH$, $HXSS$, $HXSX$, $HXXS$, $HXXZ$, $SHHH$, $SHHS$, $SHSH$, $SHSS$, $SHSX$, $SHSZ$, $SHXS$, $SHXX$, $SHXZ$, $SSHH$, $SSHS$, $SSHX$, $SSHZ$, $SSXH$, $SSXX$, $SSZZ$, $SXHH$, $SXHS$, $SXHX$, $SXSH$, $SXSX$, $SXXS$, $XHHS$, $XHHZ$, $XHSH$, $XHSS$, $XHXH$, $XHXX$, $XSHH$, $XSSH$, $XSXH$, $XXZZ$, $HHHSH$, $HHHSX$, $HHHSZ$, $HHHXS$, $HHHXZ$, $HHSHS$, $HHSHX$, $HHSHZ$, $HHSSH$, $HHSSX$, $HHSXH$, $HHSXS$, $HHSXZ$, $HHXHX$, $HHXSH$, $HHXSS$, $HSHHH$, $HSHHS$, $HSHSS$, $HSHSX$, $HSHSZ$, $HSHXS$, $HSHXZ$, $HSSHH$, $HSSHS$, $HSSHZ$, $HSSXH$, $HSSXX$, $HSXHH$, $HSXHS$, $HSXHX$, $HSXSX$, $HSXXS$, $HXHHS$, $HXHHZ$, $HXHXH$, $HXSHH$, $HXSXH$, $SHHHS$, $SHHHZ$, $SHSHH$, $SHSHX$, $SHSSH$, $SHSSX$, $SHSSZ$, $SHSXS$, $SHSXZ$, $SHXSH$, $SHXSS$, $SSHSH$, $SSHSS$, $SSHSX$, $SSHSZ$, $SSHXS$, $SSHXX$, $SSHXZ$, $SSXHH$, $SSXHS$, $SXHHS$, $SXHSH$, $SXHSS$, $SXSHH$, $SXSXH$, $XHHSH$, $XHSHS$, $XSSHS$, $HHHSHX$, $HHHSHZ$, $HHHSXH$, $HHHSXS$, $HHHSXZ$, $HHHXSS$, $HHSHSS$, $HHSHSX$, $HHSHSZ$, $HHSHXS$, $HHSSHS$, $HHSSXH$, $HHSXHS$, $HHSXHX$, $HSHHHZ$, $HSHSSX$, $HSHSSZ$, $HSHSXS$, $HSHSXZ$, $HSHXSS$, $HSSHSZ$, $HSSXHH$, $HSSXHS$, $HSXHSS$, $HXHHSH$, $SHHHSZ$, $SHSHHS$, $SHSSHH$, $SHSSHZ$, $SHSSXH$, $SHXSHH$, $SSHSHX$, $SSHXSH$, $SSXHSH$, $XHSHSS$, $XSSHSH$, $HHHSXHX$, $HHSHSSX$, $HHSSXHS$

Table 7.2: Presentation of $\mathcal{C}(3)$ in terms of the generators $\langle H, S, X, Z\rangle$. There are 216 elements and the average word length is ~ 4.33.

$\mathbb{I}$, F, X, Z, FF, FX, FZ, XF, XZ, FFX, FFZ, FXF

Table 7.3: Presentation of $\mathcal{G}_2$ in terms of the generators $\langle F, H, X, Z\rangle$. There are 12 elements and the average word length is ~ 1.83.

$\mathbb{I}$, F, H, X, Z, FF, FH, FX, FZ, HF, HH, HX, HZ, XF, XH, XX, XZ, ZF, ZZ, FFF, FFH, FFX, FHH, FHX, FHZ, FXF, FXH, FXX, FZF, FZZ, HFF, HFX, HFZ, HHF, HHH, HHZ, HXF, HXX, HXZ, HZF, XFF, XFH, XFZ, XHF, XHX, XXF, XXZ, XZZ, ZFH, $FFFH$, $FFFX$, $FFHZ$, $FFXF$, $FFXH$, $FHHH$, $FHHZ$, $FHXF$, $FHXX$, $FXFF$, $FXFH$, $FXFZ$, $FXXF$, $HFFX$, $HFXX$, $HFZF$, $HHFZ$, $HHHF$, $HHZF$, $HXXZ$, $XFHX$, $XXZZ$, $FXFHX$

Table 7.4: Presentation of $\mathcal{G}_3$ in terms of the generators $\langle F, H, X, Z\rangle$. There are 72 elements and the average word length is ~ 2.98.

7.4 Analyzing Results

Using the steps outlined in Section 7.2, we designed qutrit pulses for all 72 elements of $\mathcal{G}_3$, each 400 ns long. We also designed qubit pulses for all 12 elements of $\mathcal{G}_2$, each 100 ns long. In subsequent subsections we analyze RB+ data that were taken on the experimental system with these two gatesets.

7.4.1 Data Sets Collected

Table 6.1 summarizes many of the popular randomized benchmarking protocols. Of these, we perform (standard) RB, Interleaved RB (IRB), and Unitarity on both of the qubit and qutrit gatesets defined in the previous section. The unitarity protocol happens to give information about leakage, too. For the qutrit gateset $\mathcal{G}_3$, we interleave the gate XF in IRB. In the qubit gateset $\mathcal{G}_2$, we interleave both of the gates $\mathbb{I}$ and F.

Importantly, these protocols are performed on the qubit gateset both before and after the tune-up sequence of Step 4, described in Section 7.2.5. The qutrit gateset, however, was only performed before the tune-up sequence because the NV under study was lost before the experiment was able to be performed. Raw datasets are shown in Figure 7.7, where they has been summed over random sequences at each value of sequence length, including both the dark and bright NV references. Table 7.5 specifies the number of random sequences used per sequence length, and which gates were interleaved in IRB.

7.4.2 Standard and Interleaved Randomized Benchmarking

Recall that standard RB consists of choosing m gates independently and uniformly at random from the 2-design group $\mathcal{G}$, and setting the $(m+1)^{\text{th}}$ gate to be the inverse of their

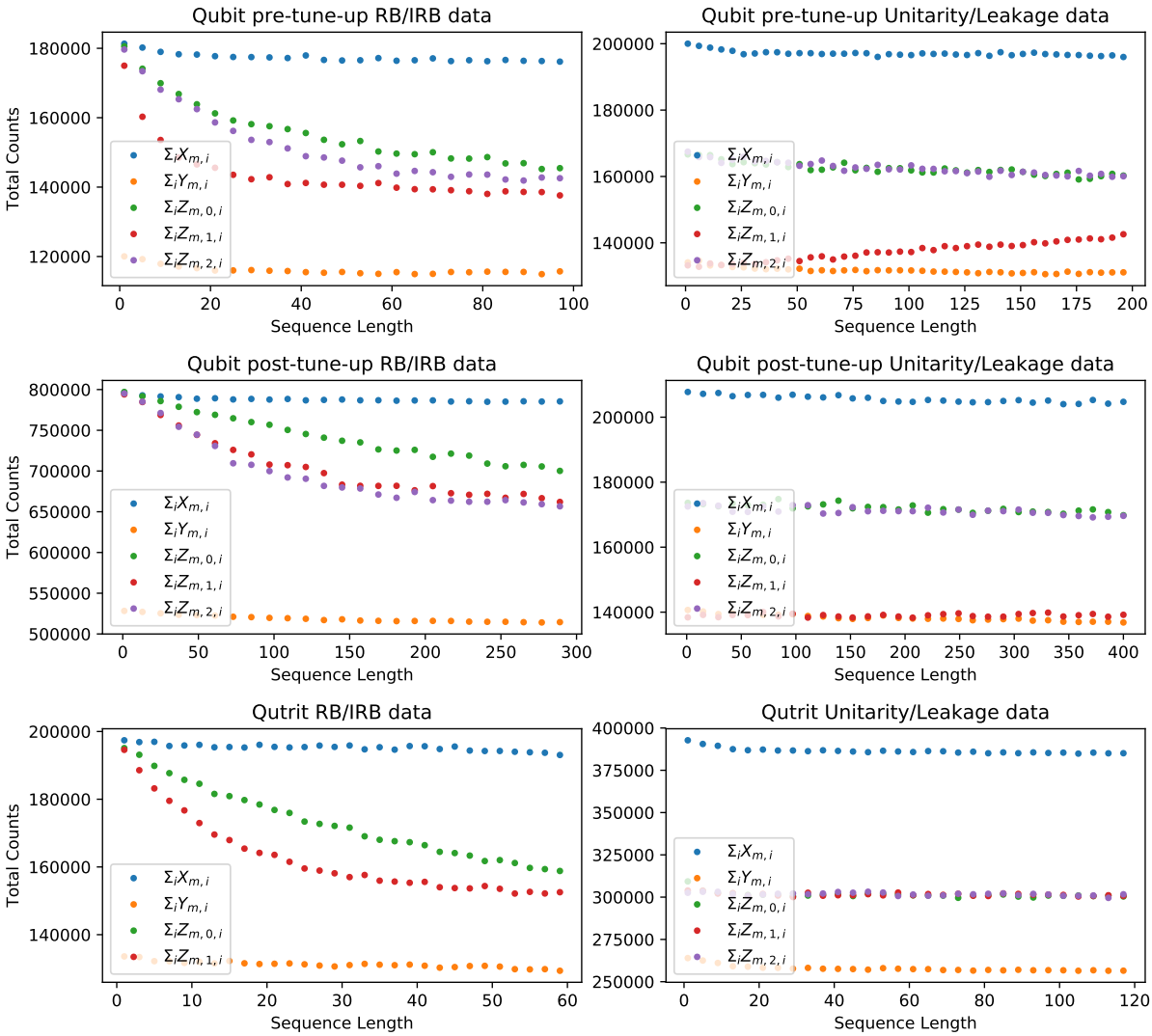

Figure 7.7: Raw summed data from RB+ experiments. Each of the three columns contain a different pulse sequences to implement a their respective gate sets, described in the main text. The top row shows data for combined RB/IRB protocols; one exponential curve corresponds to RB, and the others interleave various gates. The bottom row shows data for unitarity/leakage, where the non-reference curves correspond to measurements of $|0\rangle$, $|+1\rangle$, and $|-1\rangle$.

Name	Gateset	RB #	IRB Gates	Unitarity #
Qutrit	$\mathcal{G}_3$	998	XF	1998
Qubit pre-tune-up	$\mathcal{G}_2$	1500	I,F	998
Qubit post-tune-up	$\mathcal{G}_2$	1500	I,F	998

Table 7.5: Numbers of random sequences used for the presented datasets. IRB data were collected in tandem with RB data, and therefore use the same number of sequences.

product. This ideally implements the identity gate $\mathbb{I}$. This is done I times at many values of m, and the resulting average decay, as a function of m, is an exponential of the form

$$\mu_m = (A - B)p^m + B. \tag{7.19}$$

If the noise channel $\Lambda \in \mathcal{C}(\mathcal{H})$ is gate independent then we have

$$p = \frac{dF - 1}{d - 1} \tag{7.20}$$

with F the average gate fidelity of Λ and $d = \dim \mathcal{H}$, and where

$$\begin{aligned} A &= \mathrm{Tr}(E\Lambda(\rho_0)) \approx 1, \\ B &= \mathrm{Tr}(E\Lambda(\mathbb{I}/d)) \approx 1/d \end{aligned} \tag{7.21}$$

with $E \approx |0\rangle\langle 0|$ the measurement POVM element and $\rho_0 \approx |0\rangle\langle 0|$ the initial state. Recall further that interleaved RB modifies this protocol by interleaving a fixed gate $U_l \in \mathcal{G}$, creating sequences of length $2m + 1$ which ideally implement the identity gate, resulting in a decay of the form

$$\mu_{m,l} = (A - B)(p \cdot p_l)^m + B \tag{7.22}$$

with A, B, p as above, and where p_l corresponds to the average gate fidelity of the implementation of U_l, using the same transformation as Equation 7.20.

Following Chapter 6, making modifications to Equation 6.13 or Equation 6.15 to fit our NV measurement scheme, we can write down a hierarchical joint Bayesian model for

standard RB and interleaved RB as follows:

$$\begin{aligned} A, B, p_l &\sim \text{Unif}\left([0,1]\right) \\ \mu_{m,0} &= (A-B)p^m + B \\ \mu_{m,l} &= (A-B)(p\cdot p_l)^m + B, \text{ for } p \geq 1 \\ G_{m,l}|A,B,p_l &\sim \pi(G_{m,l}) \text{ such that } \mathbb{E}_{G_{m,l}}[p] = \mu_{m,l} \\ \alpha_{m,i}, \beta_{m,i} &\sim \text{Normal}\left((\overline{\alpha}, \overline{\beta}), \Sigma_{\alpha,\beta}\right) \end{aligned} \tag{7.23}$$

$$\begin{aligned} q_{m,l,i}|G_{m,l} &\sim G_{m,l} \\ (X_{m,i}, Y_{m,i})|\alpha,\beta &\sim \text{Pois}(\alpha_{m,i}) \times \text{Pois}(\beta_{m,i}) \\ Z_{m,l,i}|q_{m,l,i},\alpha_{m,i},\beta_{m,i} &\sim \text{Pois}(\beta_{m,i} + (\alpha_{m,i}-\beta_{m,i})q_{m,l,i}) \end{aligned} \tag{7.24}$$

Here, $m \in \mathbb{M}$ are the chosen sequence lengths, $1 \leq i \leq I$ indexes over random sequences at each sequence length, and $0 \leq l \leq L$ indexes over which gate in interleaved, where p_0 corresponds to SRB. The symbol $G_{p,l}$ denotes a random distribution describing the survival distribution at a given index (p,l) and $\pi(G_{m,l})$ its prior. This can be, for example, the family of beta distributions, or the CBMDP family of Section 6.5.2. As always, α and β are the bright and dark references of the NV system.

It is computationally helpful to integrate the $q's$ out of this distribution. Doing so in the form of a normal moment approximation results in the modification to the above program given by

$$\sigma_{m,l} \sim \pi(\sigma)$$

$$\begin{aligned} \overline{z}_{m,l,i} &= \beta_{m,i} + (\alpha_{m,i}-\beta_{m,i})\mu_{m,l} \\ \sigma^2_{z;m,l,i} &= \overline{z}_{m,i} + (\alpha_{m,i}-\beta_{m,i})^2\sigma^2_{m,l} \\ Z_{m,l,i}|\alpha_{m,i},\beta_{m,i},\mu_{m,l},\sigma_{m,l} &\sim \text{Normal}\left(\overline{z}_{m,l,i}, \sigma^2_{z;m,l,i}\right). \end{aligned} \tag{7.25}$$

Here, $\text{Var}\, G = \sigma^2$, so that a draw of G has been replaced by a draw of (some parameterization) of G's variance, greatly reducing computational overhead. This is valid in the regime where either the survival distributions are unimodal, and/or in the regime where α, β are large enough so that Z is approximately normally distributed, but their contrast is not so great that they can be easily distinguished in a single draw.

The model is used to fit the RB/IRB data-sets, seen in Figure 7.8, Figure 7.9, and Figure 7.10. We see that the qutrit average gateset fidelity is measured at 98.5%, and the

interleaved XF gate has fidelity 97.8%. The qubit gateset, before tune-up, has an average gateset fidelity of 99.1%, with average gate fidelities of I and F measured at 95.5% and 99.4% respectively. After tuneup, the same three metrics improve to 99.85%, 99.7%, and 99.55%, respectively. Fidelities are also summarized in the conclusion of this chapter, in Table 7.6.

Although we have demonstrated that a Bayesian analysis was feasible for these datasets, it was not strictly necessary. This is due both to the quantity of data at hand, and because Poisson distributions are virtually indistinguishable from normal at the rates we are working at. This is demonstrated in Figure 7.11 where a subset of the pre-tune-up RB/IRB qubit dataset (100 random sequences at each sequence length) is is analyzed with the Bayesian model, as well as with a weighted least-squares fit, with weights calculated using moment expansions. The following subsection therefore uses LSFs exclusively.

7.4.3 Unitarity and Leakage Randomized Benchmarking

The *purity* of a quantum state $\rho \in \mathcal{D}(\mathcal{H})$ is defined as $\mathrm{Tr}(\rho^2) = \langle \rho, \rho \rangle$. Pure states have a purity of 1, and the completely mixed state has a purity of $1/d$. Note that a unitary channel preserves the purity of any density matrix by the cyclic property of the trace. The *unitarity* of a superoperator $\Lambda \in \mathcal{T}(\mathcal{H})$ is the average purity of the channel's output averaged over all pure inputs [177],

$$u(\Lambda) := \frac{1}{d-1} \int \mathrm{d}\psi \, \mathrm{Tr}[\Lambda'(|\psi\rangle\langle\psi|)\Lambda'(|\psi\rangle\langle\psi|)], \tag{7.26}$$

where to account for channels which are trace decreasing (a property that this metric wishes to be independent of) the integral is over the identity-subtracted super operator Λ', defined by

$$\Lambda'(X) = \Lambda(X) - [\mathrm{Tr}\,\Lambda(X)/\sqrt{d}]\mathbb{I}. \tag{7.27}$$

The unitarity lies between 0 and 1, equal to 1 only for unitary channels, and equal to 0 for the completely depolarizing channel.

Similarly, the leakage can be defined as the average extent to which a channel decreases trace,

$$\ell(\Lambda) := 1 - \int \mathrm{d}\psi \, \mathrm{Tr}\,\Lambda(|\psi\rangle\langle\psi|), \tag{7.28}$$

equal to 0 for trace preserving maps. This is the sum of the leakage into and out of the subsystem of interest, which were defined separately in Equation 6.30.

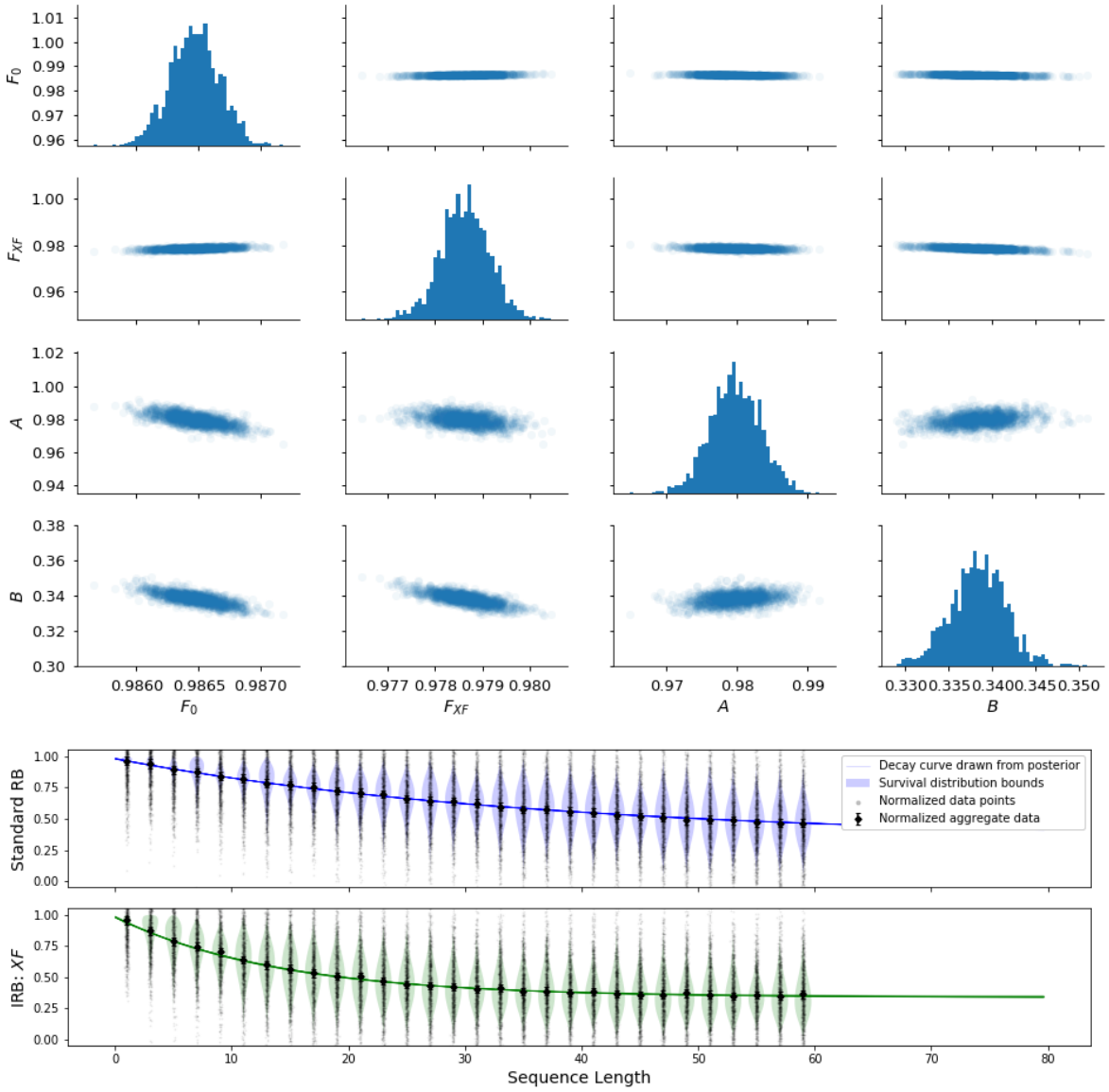

Figure 7.8: Analysis of RB/IRB data for the qutrit gateset. The top grid displays marginal posteriors over the average gate fidelity of the gateset, F_0, and of the interleaved gate XF, F_{XF}. The constants A and B are also included. The lower figures show 30 decay curves are generated by sampling from the posterior, and are plotted on top of the normalized data. 90% credibile bounds bounds on the survival distribution densities are shaded at each sequence length.

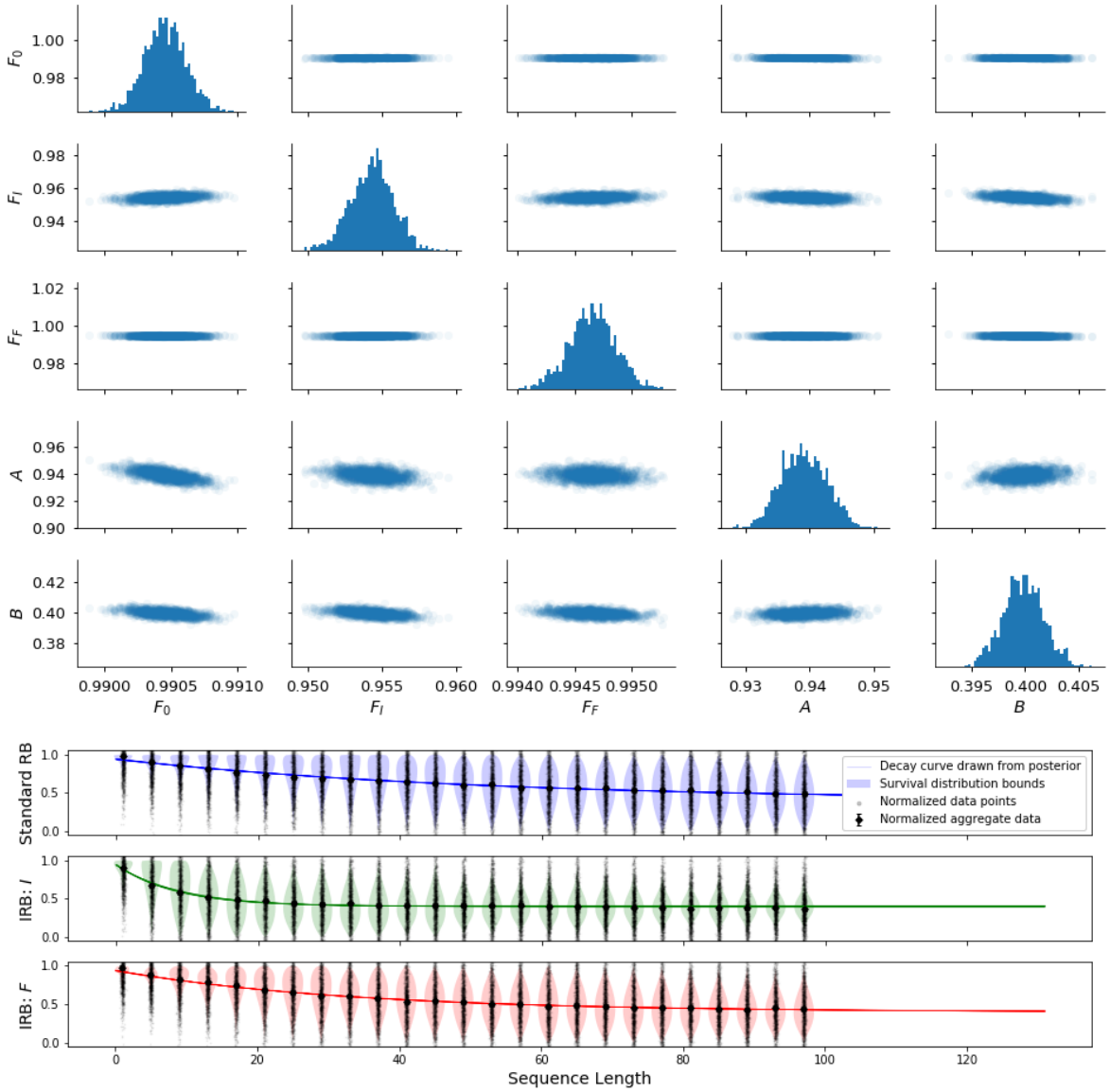

Figure 7.9: Analysis of RB/IRB data for the qubit gateset, pre-tune-up. The top grid displays marginal posteriors over the average gate fidelity of the gateset, F_0, and of the interleaved gate XF, F_{XF}. The constants A and B are also included. The lower figures show 30 decay curves are generated by sampling from the posterior, and are plotted on top of the normalized data. 90% credible bounds on the survival distribution densities are shaded at each sequence length.

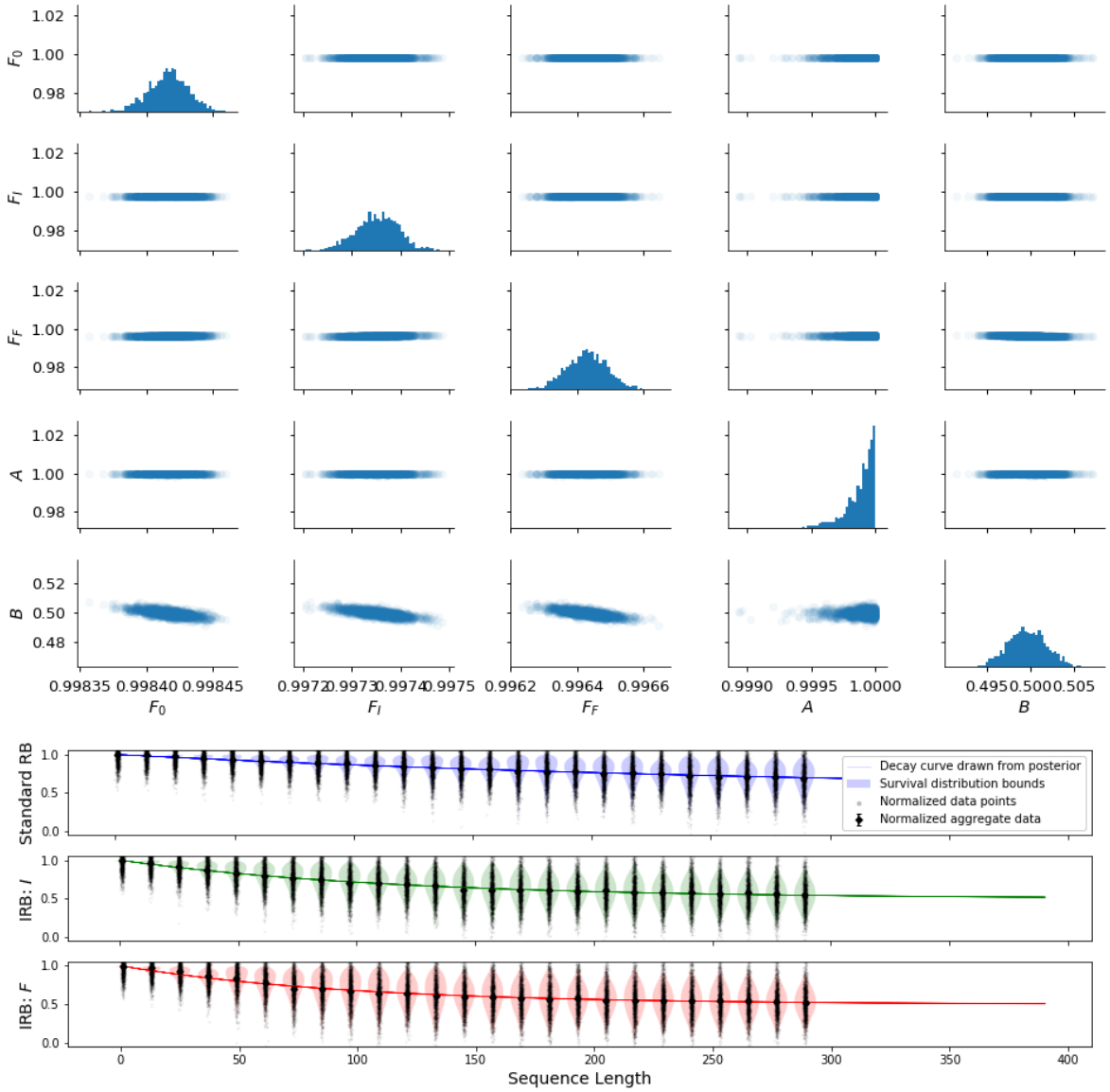

Figure 7.10: Analysis of RB/IRB data for the qubit gateset, post-tune-up. The top grid displays marginal posteriors over the average gate fidelity of the gateset, F_0, and of the interleaved gates I, F_I, and F, F_F. The constants A and B are also included. 30 decay curves are generated by sampling from the posterior, and are plotted on top of the normalized data. 90% credible bounds on the survival distribution densities are shaded at each sequence length.

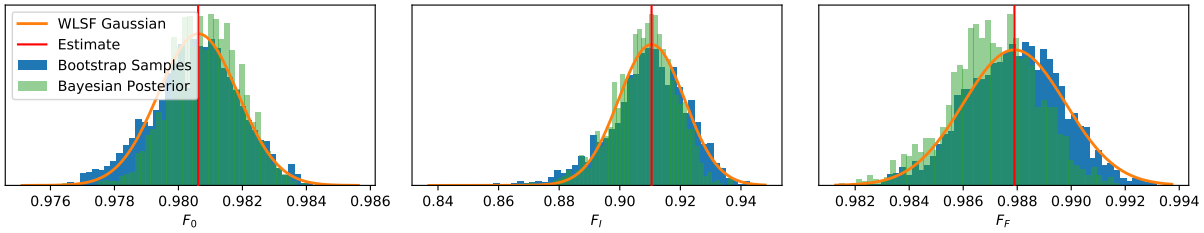

Figure 7.11: A comparison, for a subset of the RB/IRB qubit pre-tune-up dataset, between the marginal posterior distributions, the Gaussian error distribution reported by a weighted least squares fit (WLSF), and 5000 estimates based on bootstrapping the data.

Unitarity Estimator

Implementing the unitarity protocol[177] is the same as implementing the standard RB protocol, except that a final inverting gate is not used. Optionally, adding more measurement operators reduces the signal to noise, so long as the same random sequence is used for each measurement type. In our case, we measured all three operators $M_0 = |0\rangle\langle 0|$, $M_1 = |-1\rangle\langle -1|$, and $M_2 = |+1\rangle\langle +1|$ by appending an appropriate gate (selected from the gateset under study) just prior to measurement. The unitarity protocol asserts an exponential decay in the second moment of observations, rather than the first, given by

$$\nu_m := \mathbb{E}[|q|^2] = A + Bu^{m-1} \tag{7.29}$$

where the expectation is over random sequences of length m, and q is the expectation value of the measurement operator Q conditioned on some particular random gate sequence[4]. In this decay, A and B are SPAM constants, and u is equal to the unitarity of the noise channel in the case of gate-independent noise. Its meaning in the case of gate-dependent noise is an unsolved problem.

Our data structure at fixed sequence length m has the form

$$(X_i, Y_i, Z_{i,0}, \ldots, Z_{i,d-1})|\alpha_i, \beta_i, p_{i,0}, \ldots, p_{i,d-1} \tag{7.30}$$

where $1 \leq i \leq I$ index unique random sequences, X_i and Y_i are the reference counts, and $Z_{i,j}$ are the counts due to the survival as measured by M_j, that is, $p_{i,j} = \operatorname{Tr} M_j \rho_i$ where ρ_i

[4]The original paper only treats hermitian measurements Q. However the same analysis works for possible non-hermitian measurements by replacing q^2 with $|q|^2$, Q^2 with $QQ^\dagger$, and $Q \otimes Q$ with $Q \otimes Q^\dagger$

is the final state resulting from the i^{th} sequence. Therefore the likelihood of the conditional distribution is

$$\begin{aligned} X_i|\alpha_i &\sim \text{Poisson}\left(n_{\text{ref}}\alpha_i\right) \\ Y_i|\beta_i &\sim \text{Poisson}\left(n_{\text{ref}}\beta_i\right) \\ Z_{i,j}|\alpha_i,\beta_i,p_{i,j} &\sim \text{Poisson}\left(\beta_i + p_{i,j}(\alpha_i-\beta_j)\right) \end{aligned} \tag{7.31}$$

where $n_{\text{ref}} \geq 1$ references are taken in each shot. To effectively use the data from different measurements, which constitute the diagonal elements of the density matrix in our case, we synthesize the measurement of the diagonal Weyl operator $Q := Z$[5]. For a dimension d, we have $Q = \text{diag}(1, \omega, \ldots, \omega^{d-1})$. We define $Z_i = \sum_{j=0}^{d-1} \omega^j Z_{i,j}$, which is complex for $d = 3$, to synthesize a measurement of Q, with values denoted as $q_i = \operatorname{Tr} Q\rho_i = \sum_{j=0}^{d-1} \operatorname{Tr} M_j\rho_i$. We wish to derive an estimator for $\mathbb{E}[|q|^2]$, therefore we study the second moment of Z_i. We have

$$\begin{aligned} \mathbb{E}\left[|Z_i|^2 \,|\alpha_i,\beta_i,\{p_{i,j}\}_j\right] &= \text{Var}[Z_i|\alpha_i,\beta_i,\{p_{i,j}\}_j] + \left|\mathbb{E}[Z_i|\alpha_i,\beta_i,\{p_{i,j}\}_j]\right|^2 \\ &= \sum_j \left|\omega^j\right|^2 (\beta_i + p_{i,j}(\alpha_i-\beta_i)) + \left|\sum_j \omega^j(\beta_i + p_{i,j}(\alpha_i-\beta_i))\right|^2 \\ &= \sum_j (\beta_i + p_{i,j}(\alpha_i-\beta_i)) + |q_i|^2 (\alpha_i-\beta_i)^2 \end{aligned} \tag{7.32}$$

for the conditional second moment. Marginalizing over survivals gives

$$\mathbb{E}\left[|Z_i|^2 \,|\alpha_i,\beta_i\right] = \sum_j \left(\beta_i + \overline{p}_j(\alpha_i-\beta_i)\right) + \nu_m(\alpha_i-\beta_i)^2, \tag{7.33}$$

and subsequently marginalizing over references gives

$$\mathbb{E}[|Z_i|^2] = \sum_j \left(\overline{\beta} + \overline{p}_j(\overline{\alpha}-\overline{\beta})\right) + \nu_m \mathbb{E}(\alpha_i-\beta_i)^2. \tag{7.34}$$

Rearranging for the desired quantity ν_m while substituting in quantities with readily available moment estimates produces

$$\nu_m = \frac{\mathbb{E}[|Z_i|^2] - \sum_{j=0}^{d-1} \mathbb{E}[Z_{i,j}]}{\mathbb{E}[(X_i-Y_i)^2] + \mathbb{E}[X_i+Y_i]/n_{\text{ref}}^2}. \tag{7.35}$$

Equation 7.35 is valid for both $d = 2$ and $d = 3$ and can be used to fit to Equation 7.29.

[5]The Weyl operator and the measurement photon counts, unfortunately, share the same symbol 'Z'—they can be distinguished here by the presence or lack of a subscript.

Leakage Estimator

The leakage protocol is the same as the unitarity protocol (though it only requires a unitary 1-design), except the first moment of the data is analyzed [180, 179, 177]. The decay curve is given as

$$\eta_m = C\ell^{m-1} \tag{7.36}$$

where C is a SPAM constant, *ell* is equal to the leakage of the noise channel in the case of gate-independent noise, and η_m is the expected value (over sequences) of the measurement operator R. Having measured $M_0 = |0\rangle\langle 0|$,$M_1 = |-1\rangle\langle -1|$, and $M_2 = |+1\rangle\langle +1|$ for the unitarity protocol, we set $R = \sum_{j=0}^{d-1} M_j$ in order to re-use data. Deriving an estimator for η_m is straight-forward compared to the estimator we derived for ν_m above—we need only look at first moments, which give

$$\eta_m = \frac{\sum_{j=0}^{d-1} \mathbb{E}[Z_{i,j}] - d\mathbb{E}[Y_i]/n_{\text{ref}}}{\mathbb{E}[X_i - Y_i]/n_{\text{ref}}.} \tag{7.37}$$

Equation 7.37 is valid for both $d = 2$ and $d = 3$ and can be used to fit to Equation 7.36.

Results

Unitarity fits to datasets from all three gatesets are shown in Figure 7.12. Typically leakage happens on timescales much longer than gate errors, and the sequence length collected in our datasets do not extend far enough to meaningfully estimate ℓ for the qutrit or post-tune-up qubit gatesets. However, as seen in the raw data of Figure 7.7, there is considerable leakage in the pre-tune-up gateset into the third level. A simple fit to Equation 7.36 gives an estimate $\hat{\ell} = 0.0080(2)$. This is to be expected as our synthesizer frequency was set such that the lower side-band corresponds exactly to the unwanted transition $|0\rangle \leftrightarrow |-1\rangle$, which it never would be in practical qubit applications. However, it serves to demonstrate the dramatic effect of tune-up, as the slightly modified pulse profiles show negligible signs of leakage at $m = 400$. See Table 7.6 for a comparison of F, u, and ℓ for the three gatesets.

Finally, we would like to consider the extent to which our gate noise is unitary. For any channel $\Lambda \in \mathcal{C}(H)$, it holds that

$$p = \frac{dF(\Lambda) - 1}{d - 1} \leq \sqrt{u(\Lambda)} \leq 1 \tag{7.38}$$

where $F(\Lambda)$ is the average gate fidelity of Λ [177]. The left hand side is tight when Λ is the depolarizing channel of strength p, $\Lambda(X) = pX + (1-p)\mathbb{I}/d$, and the right hand side is tight when Λ is a unitary channel. Rearranging this expression gives

$$0 \leq f_u := \frac{(d-1)(1-\sqrt{u(\Lambda)})/d}{1-F} \leq 1 \tag{7.39}$$

where unitary channels saturate the upper bound and depolarizing channels saturate the lower bound. Infidelity being in the denominator, the value of f_u may be understood as the fraction of the infidelity which is attributable to unitary error. Bootstrap histograms of f_u are plotted in Figure 7.13. We see that, pre-tune-up, the qutrit pulses suffer less unitarity error than the qubit pulses, perhaps because they are four times longer and thus more subject to decoherence mechanisms. We also see that the process of tune-up decreases the fractional unitarity f_u, consistent with the intuition that the gates improve in tune-up because we are fixing unitary errors due to bad pulse shapes.

7.4.4 Time dependence of gate quality

The collected datasets represent a long span of wall-clock time. In particular, the post-tune-up RB/IRB dataset comprises nearly two weeks of continuous experiments. We can analyze subsets of this data to construct a time-trace of fidelities. There is a tradeoff between the error-bars on these subsets and the resolution of the time trace—we elect to use subsets consisting of 50 random sequences per sequence length, representing about 10.5 hrs of collection time.

This trace is plotted in Figure 7.14. We see that fidelities are roughly flat, though there is certainly structure present when compared to the same data but shuffled. There appear to be no strong daily trends.

7.5 Application: Partial pseudo-pure state preparation

Some quantum systems use initial states which are 'pseudo-pure' rather than pure. As an example, take the initial state of a proton in a large, external magnetic field, where the initial state is produced by waiting for relaxation. There is a slight preference for spin states to be aligned with this field, an effect which gives rise to the thermal state

$$\rho_{\text{thermal}} \approx (\mathbb{I} + \epsilon\sigma_z)/2 \tag{7.40}$$

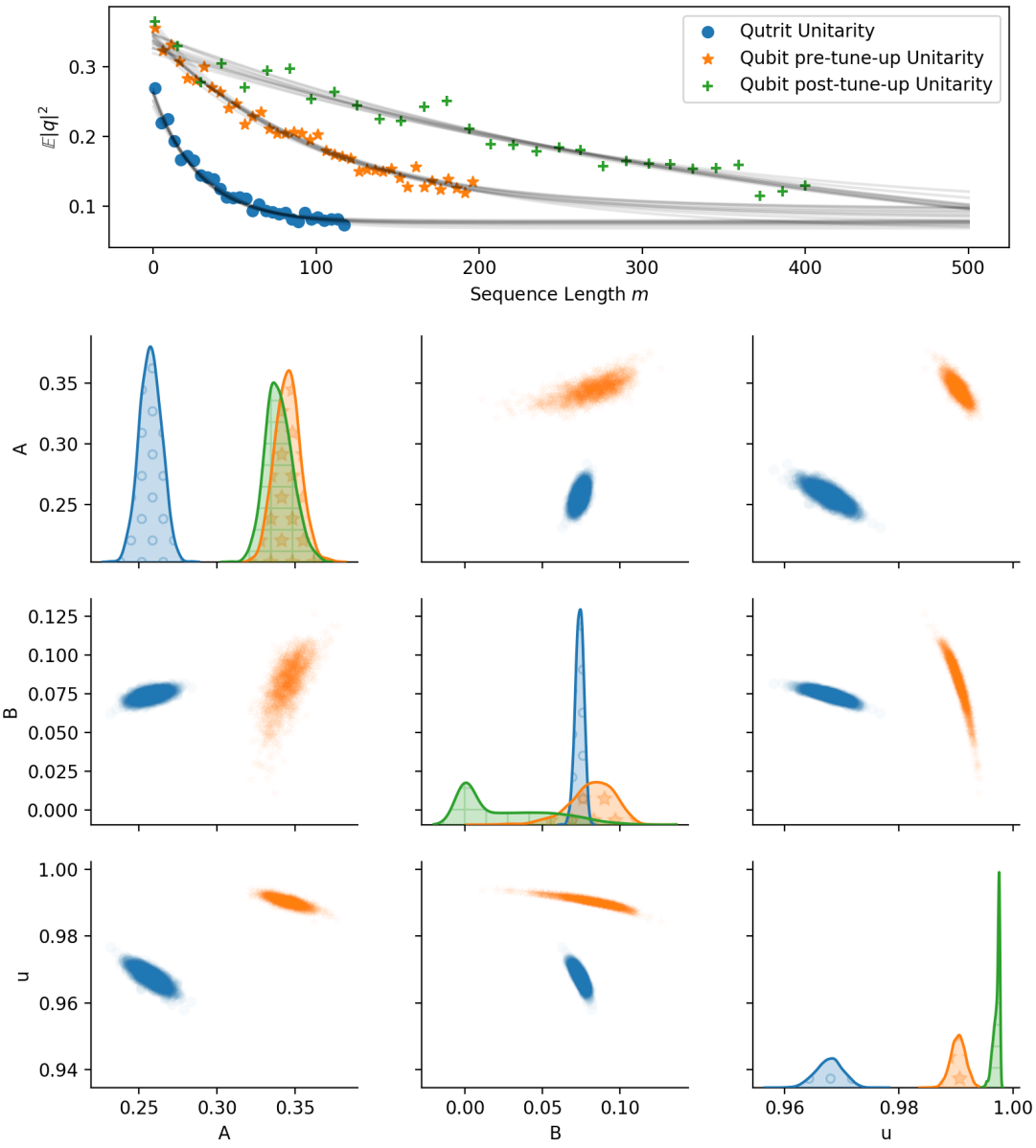

Figure 7.12: (Top) Fits to unitarity datasets from all three gatesets and (Bottom Grid) the bootstrap distribution over the parameters A, B, and u. Large correlation between u and B are due to lack of data at high sequence lengths, especially in the case of post-tuneup.

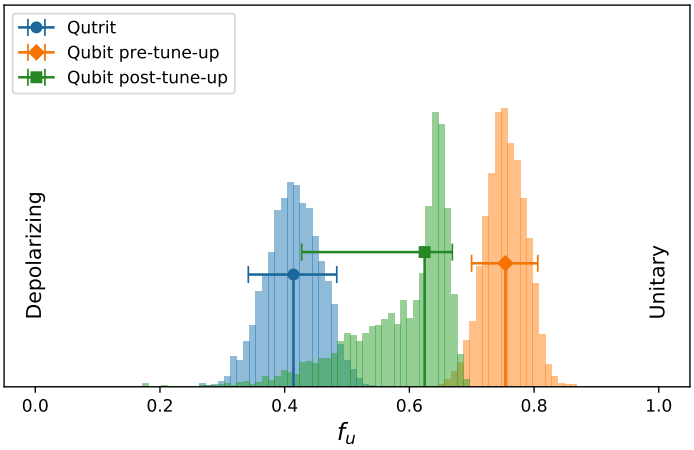

Figure 7.13: The fraction of infidelity attributable to infidelity. Distributions, left to right, are for the qutrit, qubit pre-tune-up, and qubit post-tune-up.

where ϵ depends on the temperature and the field strength as derived through Boltzmann statistics. This state is called pseudo-pure because when rewritten as

$$\rho_{\text{thermal}} = \frac{1-\epsilon}{2}\mathbb{I} + \epsilon\left|\uparrow\right\rangle\!\left\langle\uparrow\right|, \tag{7.41}$$

where $\left|\uparrow\right\rangle = \left(\begin{smallmatrix}1\\0\end{smallmatrix}\right)$, it becomes clear that any unital operation Λ (which includes all unitary channels) will affect only the latter term. This produces a final state

$$\frac{1-\epsilon}{2}\mathbb{I} + \epsilon\Lambda\left(\left|\uparrow\right\rangle\!\left\langle\uparrow\right|\right) \tag{7.42}$$

causing dynamics equivalent to having started with the pure initial state $\left|\uparrow\right\rangle$, except that we are using a copy of the Bloch sphere with radius ϵ instead of 1.

In Section 2.8.2 we concluded that the NV center's intialization procedure produces a pseudo-pure state $q\left|0\right\rangle\!\left\langle 0\right| + (1-q)\mathbb{I}/3$ for a value of q near unity. In this section we discuss pseudo-purifying the nitrogen atom, which has a fully mixed initial state in the operating conditions of our experiments: room temperature and low field. The T_1 relaxation time of the 14-nitrogen (the characteristic time scale at which its eigenstates randomly flip) has been measured to be hundreds of microseconds at our magnetic field levels [135], which is much long than a single shot including references and measurements (totaling tens of microseconds), but much shorter than an entire experiment (minutes or hours).

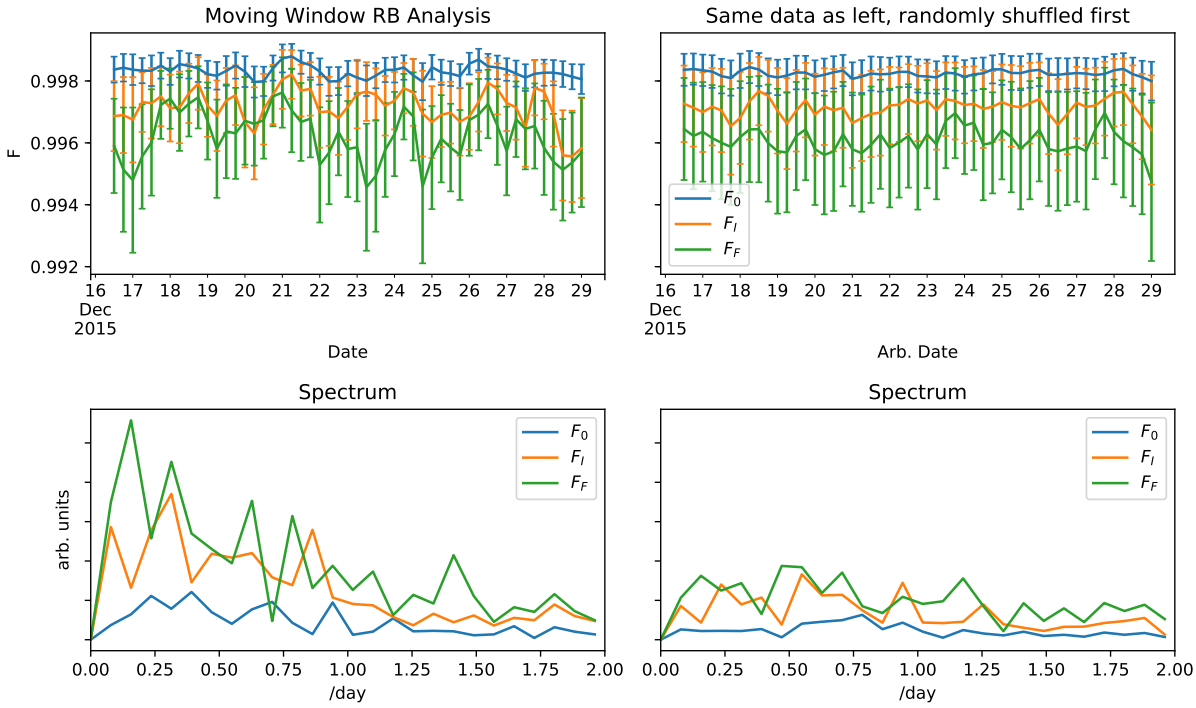

Figure 7.14: The RB/IRB post tune-up dataset is analyzed on a moving window of subsets. Each subset contains 50 random sequence per sequence length. Fidelities F_0, F_I, and F_F are plotted against timestamps of the data, with an errorbar every six hours. The right column is provided for the sake of comparision, where the data have been shuffled temporally prior to analysis. The second row shows the spectrum of the respective time traces in the first row.

Our goal is to prepare the NV-nitrogen system in a state $\rho_0 = s\,|00\rangle\langle 00| + (1-s)\sigma_0$ such that

$$(\Lambda \otimes \mathbb{I})(\rho_0) = s\Lambda(|0\rangle\langle 0|) \otimes |0\rangle\langle 0| + (1-s)\sigma_0 \tag{7.43}$$

where Λ in any unital channel acting on the NV alone, and σ_0 is a density matrix on the joint system. This is a slight modification of the usual pseudo-pure state definition because we do not require σ_0 to be proportional to $\mathbb{I}$, and our operation of interest Λ acts on only part of the system—hence the term 'partial pseudo-pure' in the title of this section. It is useful because it allows for one to consider only the possibility that the nitrogen is in the state $|0\rangle$ when designing the operation or protocol Λ, and also when analyzing the resulting data. Indeed, we saw in our Hamiltonian learning experiments that the presence of nitrogen is a nuisance to simulation and learning, and in our pulse design that it is an impediment to short, simple pulse profiles. This partial pseudo-pure state comes at the cost of an effective drop in contrast between references.

We begin by supposing access to a unitary operation W acting on the joint system such that

$$\begin{aligned} W\,|0,0\rangle &= \mathrm{e}^{\mathrm{i}\phi_1}\,|0,0\rangle \\ W\,|0,+1\rangle &= \frac{\mathrm{e}^{\mathrm{i}\phi_2}}{\sqrt{3}}(|-1\rangle + |0\rangle + |+1\rangle) \otimes |+1\rangle \\ W\,|0,-1\rangle &= \frac{\mathrm{e}^{\mathrm{i}\phi_3}}{\sqrt{3}}(|-1\rangle + |0\rangle + |+1\rangle) \otimes |-1\rangle \end{aligned} \tag{7.44}$$

for some phases $\phi_1, \phi_2, \phi_3 \in \mathbb{R}$. Composing such a unitary with a subsequent complete dephasing channel (which is implemented simply by waiting several multiples of the $T_{2,e}^*$ time, but much less than the $T_{1,e}$ time—a feat easily arranged in the NV system) produces a channel Φ_0 such that

$$\begin{aligned} \Phi_0(|0,0\rangle\langle 0,0|) &= |0,0\rangle\langle 0,0| \\ \Phi_0(|0,+1\rangle\langle 0,+1|) &= \mathbb{I} \otimes |+1\rangle\langle +1|\,/3 \\ \Phi_0(|0,-1\rangle\langle 0,-1|) &= \mathbb{I} \otimes |-1\rangle\langle -1|\,/3. \end{aligned} \tag{7.45}$$

Therefore, applying Φ_0 to the standard preparation $(q\,|0\rangle\langle 0| + (1-q)\mathbb{I}/3) \otimes \mathbb{I}/3$ gives our desired form of ρ_0 where

$$\begin{aligned} s &= q/3 \\ \sigma_0 &= \frac{\mathbb{I}}{3} \otimes \frac{q(|-1\rangle\langle -1| + |+1\rangle\langle +1|) + (1-q)\mathbb{I}}{3}. \end{aligned} \tag{7.46}$$

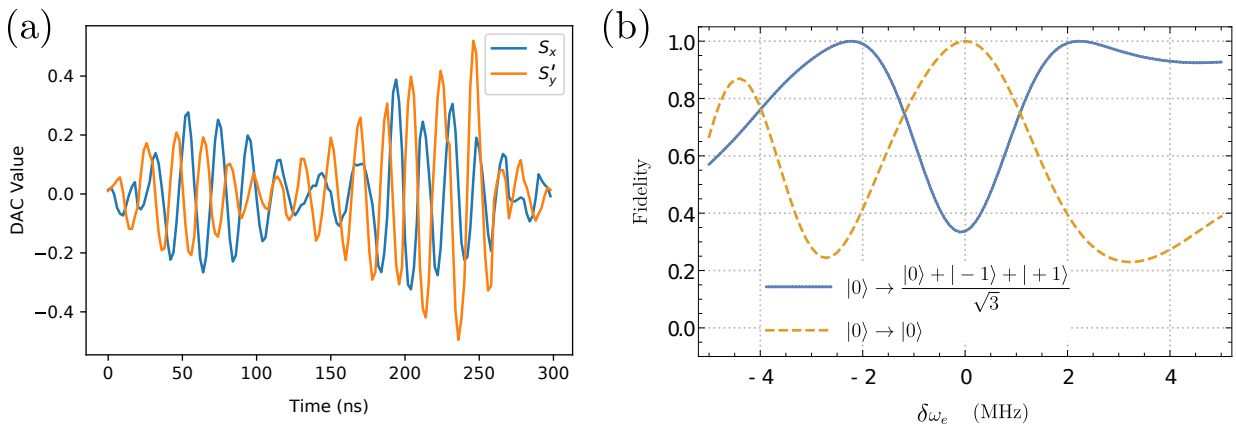

Figure 7.15: (a) The pulse profile for the control Hamiltonians $\mathbf{S}_\mathrm{x}$ and $\mathbf{S}'_\mathrm{y}$ of a pulse that implements Equation 7.44. (b) The state-to-state fidelity of this pulse from $|0\rangle$ to two different output states as a function of deviation, $\delta\omega_e$, from the nominal $\omega_e = 2\pi \cdot 50\,\mathrm{MHz}$. Note that the transfer to $|0\rangle$ takes place at $\delta\omega_e = 0\,\mathrm{MHz}$ and that the transfer to $\frac{|-1\rangle+|0\rangle+|+1\rangle}{\sqrt{3}}$ takes place at both $\delta\omega_e = \pm 2.189\,\mathrm{MHz}$.

Given that the effect of q is—by its very definition—unobservable, it is most sensible to ignore it by setting $q = 1$ in these formulas. In this case the effective reference levels are

$$\alpha' = \beta + \frac{5}{9}(\alpha - \beta) \qquad \text{and} \qquad \beta' = \beta + \frac{2}{9}(\alpha - \beta) \tag{7.47}$$

where α and β are as defined in Equation 2.71.

We may find the unitary operation W using the same methods that were used to find gatesets for randomized benchmarking experiments. However, we use a state-to-state objective function in each of the three nitrogen manifolds instead of a full unitary objective function. A pulse profile and a corresponding robustness plot are shown in Figure 7.15. In Reference [130], which is the inspiration for this example, an operation similar to W was constructed (and for the same purpose) by modifying a Dante pulse sequence. The effectiveness of this partial pseudo-pure state preparation is demonstrated in Figure 7.16, where we see that we can eliminate the nitrogen isocromats $|+1\rangle$ and $|-1\rangle$ from a Ramsey signal. This experiment was performed with the same NV and in the same environment as the randomized benchmarking experiments. Note that with an actual purification of the nitrogen atom, we would not lose contrast as we do in this experiment.

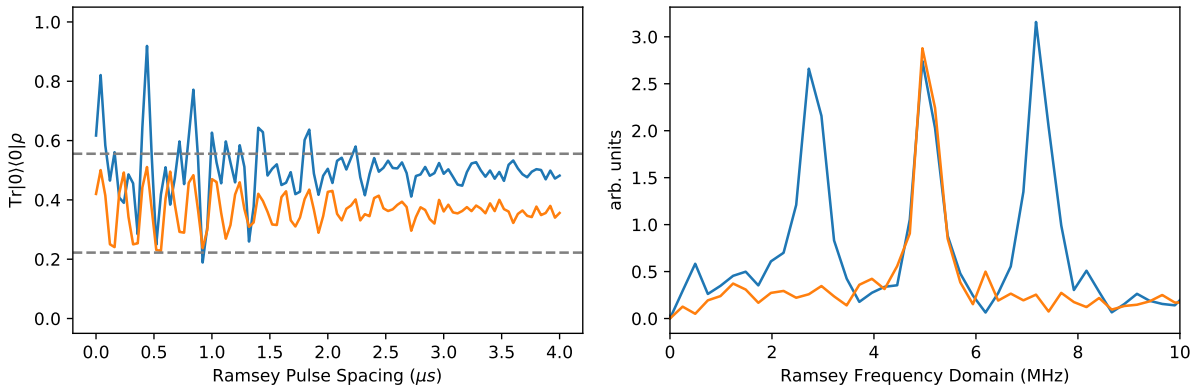

Figure 7.16: A single-quantum Ramsey experiment (with TPPI=5 MHz) preceded by standard preparation (blue) and with partial pseudo-pure state preparation (orange), plotted in both the time and frequency domains. In both cases, normalization of data to the interval $[0, 1]$ is done using the standard references α and β, not the reduced references α' and β' from Equation 7.47. In the time domain plot, dashed horizontal lines are placed at $y = 2/9, 5/9$.

7.6 Summary

The goal of this chapter has been to test our ability to generate high quality pulses that implement arbitrary unitary gates for the NV center using numerical optimal control. We laid out the concrete steps we performed to do so. This methodology was characterized experimentally by designing pulses for each member of a 2-design group, both in $d = 2$ and $d = 3$. Such gatesets were chosen so that protocols from RB+ could be applied, in particular, we considered standard RB, interleaved RB, Unitarity and Leakage. Gate qualities inferred from these protocols are summarized in Table 7.6. Though all gates generated were Clifford gates, we have no reason to suspect that some other non-Clifford gate would perform better or worse on average using the same pulse design steps.

In interpreting these results one must be careful regarding the subtleties of gate dependent errors. For example, we considered the quantity f_u as being the fraction of infidelity attributable to unitary error, when, however, little is known about what the 'unitarity of a gateset' means when there is gate dependent noise. Very little, too, is known about the effect non-markovian noise has on RB+ protocols, except that it probably leads to deviations from exponential decays[178]. On the other hand, it is known that, at least in qubits but likely higher dimensions too, the decay constant in standard RB has an interpretation

Gateset	d	F	u	ℓ
Qutrit	3	0.9816 $[0.9812, 0.9820]$	0.968 $[0.964, 0.972]$	~ 1
Qubit pre-tune-up	2	0.9902 $[0.9899, 0.9904]$	0.990 $[0.988, 0.992]$	0.0080(2)
Qubit post-tune-up	2	0.9983 $[0.9982, 0.9983]$	0.997 $[0.996, 0.998]$	~ 1
Gate				
XF	3	0.9725 $[0.9714, 0.9735]$	-	-
I (pre-tune-up)	2	0.9541 $[0.9515, 0.9566]$	-	-
F (pre-tune-up)	2	0.9947 $[0.9943, 0.9950]$	-	-
I (post-tune-up)	2	0.9972 $[0.9971, 0.9973]$	-	-
F (post-tune-up)	2	0.9962 $[0.9960, 0.9963]$	-	-

Table 7.6: A summary of RB+ measurement values. 90% confidence intervals based on percentile-bootstrap are shown in square brackets.

in the case of gate dependent errors in terms of a metric called the gateset circuit fidelity [20].

Chapter 8

Conclusions and Outlook

This thesis has discussed aspects of characterization, control, and validation as applied to small quantum systems, ranging from quantum Hamiltonian learning with Bayesian inference, to the numerical optimization of pulse shapes for high-fidelity control, to performing randomized benchmarking experiments and analyzing their results. Many of these ideas were tested experimentally, using the NV defect in diamond as a testbed, though it is our hope that these ideas are more broadly applicable across all modalities of small quantum devices.

A central theme for us has been the tradeoff between complexity and performance. In quantum Hamiltonian learning, for instance, the infrastructure to perform risk-based adaptive experiment design added overhead to the experiment's setup, and introduced substantial (though managable) classical computing requirements, but ultimately provided significant speed-ups in our parameter learning rates. Later, we considered controlling a qubit ensemble with a superconducting strip-line resonator, which provides improved sensitivity due to high quality factors, but comes at the cost of a complicated non-linear distortion operator that requires ringdown suppression. Similarly, we found that designing high-fidelity qutrit gates for the NV system at low field system and in the presence of a mixed nitrogen is feasible.

We have emphasized rigorous application of statistical methods, especially promoting Bayesian methods for their transparency and portability. This was seen throughout, such as in our treatment of data from randomized benchmarking and related protocols which relied on both parametric and non-parametric Bayesian models. Further, we found convenience in using Bayesian methods for quantum Hamiltonian learning because it provided a natural method to propagate errors to the parameters of interest, requiring only a simulator and

prior distribution as input to generate a posterior output.

Looking forward, we can consider many interesting directions for future work. One of the numerical methods we often used, sequential Monte Carlo, lends itself particularly well to online sequential learning, and adaptive experiment design. We hope that adaptive quantum experiments will become more commonplace to improve device performance. For example, in the case of randomized benchmarking, our datasets were far too large in some cases, but contained sequences that were too short in other cases. Adaptive schemes are adept at mitigating against such shortcomings, and can do so in an automated way, which could prove useful for larger, more complex devices. Regarding gate design, it seems likely that high-performance quantum devices will need to periodically adjust their gates based on measurement, another area where adaptive methods may help improve *in situ* tune-up efficiency.

Using what we have learned, we leave off with a high-level vision for the control infrastructure of a mid-sized near-term quantum information processor. First, it is accompanied by an online Bayesian engine which maintains the state of relevant parameters that parametrically capture the current quality of the device. The engine is able to suggest characterization experiments based on its current state of knowledge to routinely update its state of knowledge. Its control electronics are such that new gate sequences can be streamed into memory while experiments are being run, and faster than they are performed, so that experiments can collect one shot per random circuit without penalty, thereby maximizing the efficiency of mean-based randomized benchmarking protocols. Pulse profiles of gate operations were pre-optimized for the hardware taking into account its limitations, however, their parameters are maintained by the state of the engine.

References

[1] V. M. Acosta, E. Bauch, M. P. Ledbetter, A. Waxman, L.-S. Bouchard, and D. Budker. Temperature Dependence of the Nitrogen-Vacancy Magnetic Resonance in Diamond. *Physical Review Letters*, 104(7):070801, February 2010.

[2] Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, June 2010.

[3] Gil Bachar, Oren Suchoi, Oleg Shtempluck, Aharon Blank, and Eyal Buks. Nonlinear induction detection of electron spin resonance. *Applied Physics Letters*, 101(2):022602, July 2012.

[4] Gopalakrishnan Balasubramanian, Philipp Neumann, Daniel Twitchen, Matthew Markham, Roman Kolesov, Norikazu Mizuochi, Junichi Isoya, Jocelyn Achard, Johannes Beck, Julia Tissler, Vincent Jacques, Philip R. Hemmer, Fedor Jelezko, and Jörg Wrachtrup. Ultralong spin coherence time in isotopically engineered diamond. *Nature Materials*, 8(5):383–387, May 2009.

[5] Harrison Ball, Thomas M. Stace, Steven T. Flammia, and Michael J. Biercuk. Effect of noise correlations on randomized benchmarking. *Physical Review A*, 93(2):022303, February 2016.

[6] Thomas M Barbara, Joel F Martin, and Jon G Wurl. Phase transients in NMR probe circuits. *Journal of Magnetic Resonance (1969)*, 93(3):497–508, July 1991.

[7] R. Barends, J. Kelly, A. Megrant, A. Veitia, D. Sank, E. Jeffrey, T. C. White, J. Mutus, A. G. Fowler, B. Campbell, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, C. Neill, P. O'Malley, P. Roushan, A. Vainsencher, J. Wenner, A. N. Korotkov, A. N. Cleland, and John M. Martinis. Superconducting quantum circuits at the surface code threshold for fault tolerance. *Nature*, 508(7497):500–503, April 2014.

[8] O.W.B. Benningshof, H.R. Mohebbi, I.A.J. Taminiau, G.X. Miao, and D.G. Cory. Superconducting microstrip resonator for pulsed ESR of thin films. *Journal of Magnetic Resonance*, 230:84–87, May 2013.

[9] D. W. Berry, B. L. Higgins, S. D. Bartlett, M. W. Mitchell, G. J. Pryde, and H. M. Wiseman. How to perform the most accurate possible phase measurements. *Physical Review A*, 80(5):052114, November 2009.

[10] D. W. Berry, H. M. Wiseman, and J. K. Breslin. Optimal input states and feedback for interferometric phase estimation. *Physical Review A*, 63(5):053804, April 2001.

[11] Alexandros Beskos, Dan Crisan, and Ajay Jasra. On the stability of sequential Monte Carlo methods in high dimensions. *The Annals of Applied Probability*, 24(4):1396–1445, August 2014.

[12] M. J. Betancourt. Generalizing the No-U-Turn Sampler to Riemannian Manifolds. *arXiv:1304.1920 [stat]*, April 2013.

[13] Michael Betancourt. A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434 [stat]*, January 2017.

[14] Robin Blume-Kohout. Robust error bars for quantum tomography. *arXiv:1202.5270 [quant-ph]*, February 2012.

[15] T. W. Borneman, C. E. Granade, and D. G. Cory. Parallel Information Transfer in a Multinode Quantum Information Processor. *Physical Review Letters*, 108(14):140502, April 2012.

[16] Troy W. Borneman and David G. Cory. Bandwidth-limited control and ringdown suppression in high-Q resonators. *Journal of Magnetic Resonance*, 225:120–129, December 2012.

[17] Troy W. Borneman, Martin D. Hürlimann, and David G. Cory. Application of optimal control to CPMG refocusing pulse design. *Journal of Magnetic Resonance*, In Press, Corrected Proof, December 2010.

[18] K. R. Brown, A. C. Wilson, Y. Colombe, C. Ospelkaus, A. M. Meier, E. Knill, D. Leibfried, and D. J. Wineland. Single-qubit-gate error below \${\mathbf{10}}{\ensuremath{-}\mathbf{4}}\$ in a trapped ion. *Physical Review A*, 84(3):030303, September 2011.

[19] P. Cappellaro, J. Emerson, N. Boulant, C. Ramanathan, S. Lloyd, and D. G. Cory. Entanglement Assisted Metrology. *Physical Review Letters*, 94(2):020502, January 2005.

[20] Arnaud Carignan-Dugas, Kristine Boone, Joel J. Wallman, and Joseph Emerson. From Randomized Benchmarking Experiments to Gateset Circuit Fidelity. *arXiv:1804.01122 [quant-ph]*, April 2018.

[21] Arnaud Carignan-Dugas, Joel J. Wallman, and Joseph Emerson. Characterizing universal gate sets via dihedral benchmarking. *Physical Review A*, 92(6):060302, December 2015.

[22] Bob Carpenter, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddel. Stan: A Probabilistic Programming Language. *Journal of Statistical Software, Articles*, 76(1), 2017.

[23] Kathryn Chaloner and Isabella Verdinelli. Bayesian Experimental Design: A Review. *Statistical Science*, 10(3):273–304, August 1995.

[24] T. Chasseur and F. K. Wilhelm. Complete randomized benchmarking protocol accounting for leakage errors. *Physical Review A*, 92(4):042333, October 2015.

[25] Zijun Chen, Julian Kelly, Chris Quintana, R. Barends, B. Campbell, Yu Chen, B. Chiaro, A. Dunsworth, A. G. Fowler, E. Lucero, E. Jeffrey, A. Megrant, J. Mutus, M. Neeley, C. Neill, P. J. J. O'Malley, P. Roushan, D. Sank, A. Vainsencher, J. Wenner, T. C. White, A. N. Korotkov, and John M. Martinis. Measuring and Suppressing Quantum State Leakage in a Superconducting Qubit. *Physical Review Letters*, 116(2):020501, January 2016.

[26] Mario A. Ciampini, Nicolò Spagnolo, Chiara Vitelli, Luca Pezzè, Augusto Smerzi, and Fabio Sciarrino. Quantum-enhanced multiparameter estimation in multiarm interferometers. *Scientific Reports*, 6:28881, July 2016.

[27] Sean Clark. Valence bond solid formalism for d -level one-way quantum computation. *Journal of Physics A: Mathematical and General*, 39(11):2701, 2006.

[28] Joshua Combes, Christopher Granade, Christopher Ferrie, and Steven T. Flammia. Logical Randomized Benchmarking. *arXiv:1702.03688 [quant-ph]*, February 2017.

[29] David G. Cory, Amr F. Fahmy, and Timothy F. Havel. Ensemble quantum computing by NMR spectroscopy. *Proceedings of the National Academy of Sciences*, 94(5):1634–1639, April 1997.

[30] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, New York, NY, USA, 2006.

[31] T. Dahm and D. J. Scalapino. Theory of intermodulation in a superconducting microstrip resonator. *Journal of Applied Physics*, 81(4):2002–2009, 1996.

[32] Domenico D'Alessandro. *Introduction to Quantum Control and Dynamics*. Chapman and Hall/CRC, 1 edition, August 2007.

[33] Christoph Dankert, Richard Cleve, Joseph Emerson, and Etera Livine. Exact and approximate unitary 2-designs and their application to fidelity estimation. *Physical Review A*, 80(1):012304, July 2009.

[34] Peter K. Day, Henry G. LeDuc, Benjamin A. Mazin, Anastasios Vayonakis, and Jonas Zmuidzinas. A broadband superconducting detector suitable for use in large arrays. *Nature*, 425(6960):817–821, October 2003.

[35] Thomas J. DiCiccio and Bradley Efron. Bootstrap confidence intervals. *Statistical science*, pages 189–212, 1996.

[36] P. a. M. Dirac. A new notation for quantum mechanics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 35(3):416–418, July 1939.

[37] M. W. Doherty, F. Dolde, H. Fedder, F. Jelezko, J. Wrachtrup, N. B. Manson, and L. C. L. Hollenberg. Theory of the ground-state spin of the NV\${}{\hat\ensuremath{-}}\$ center in diamond. *Physical Review B*, 85(20):205203, May 2012.

[38] Marcus W. Doherty, Neil B. Manson, Paul Delaney, Fedor Jelezko, Jörg Wrachtrup, and Lloyd C. L. Hollenberg. The nitrogen-vacancy colour centre in diamond. *Physics Reports*, 528(1):1–45, July 2013.

[39] F. Dolde, H. Fedder, M. W. Doherty, T. Nöbauer, F. Rempp, G. Balasubramanian, T. Wolf, F. Reinhard, L. C. L. Hollenberg, F. Jelezko, and J. Wrachtrup. Electric-field sensing using single diamond spins. *Nature Physics*, 7(6):459–463, June 2011.

[40] Arnaud Doucet and Adam M. Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, 12(656-704):3, 2009.

[41] A. Dréau, J.-R. Maze, M. Lesik, J.-F. Roch, and V. Jacques. High-resolution spectroscopy of single NV defects coupled with nearby $\hat{\{}13\}$C nuclear spins in diamond. *Physical Review B*, 85(13):134107, April 2012.

[42] Simon Duane, A. D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, September 1987.

[43] M. V. Gurudev Dutt, L. Childress, L. Jiang, E. Togan, J. Maze, F. Jelezko, A. S. Zibrov, P. R. Hemmer, and M. D. Lukin. Quantum Register Based on Individual Electronic and Nuclear Spin Qubits in Diamond. *Science*, 316(5829):1312–1316, January 2007.

[44] D. J. Egger and F. K. Wilhelm. Adaptive Hybrid Optimal Quantum Control for Imprecisely Characterized Systems. *Physical Review Letters*, 112(24):240503, June 2014.

[45] Joseph Emerson, Robert Alicki, and Karol Życzkowski. Scalable noise estimation with random unitary operators. *Journal of Optics B: Quantum and Semiclassical Optics*, 7(10):S347, 2005.

[46] Jeffrey M. Epstein, Andrew W. Cross, Easwar Magesan, and Jay M. Gambetta. Investigating the limits of randomized benchmarking protocols. *Physical Review A*, 89(6):062321, June 2014.

[47] Michael D. Escobar and Mike West. Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

[48] S. Felton, A. M. Edmonds, M. E. Newton, P. M. Martineau, D. Fisher, D. J. Twitchen, and J. M. Baker. Hyperfine interaction in the ground state of the negatively charged nitrogen vacancy center in diamond. *Physical Review B*, 79(7):075203, February 2009.

[49] Guanru Feng, Joel J. Wallman, Brandon Buonacorsi, Franklin H. Cho, Daniel K. Park, Tao Xin, Dawei Lu, Jonathan Baugh, and Raymond Laflamme. Estimating the Coherence of Noise in Quantum Control of a Solid-State Qubit. *Physical Review Letters*, 117(26):260501, December 2016.

[50] Christopher Ferrie. Self-Guided Quantum Tomography. *Physical Review Letters*, 113(19):190404, November 2014.

[51] Christopher Ferrie, Christopher E. Granade, and D. G. Cory. How to best sample a periodic probability distribution, or on the accuracy of Hamiltonian finding strategies. *Quantum Information Processing*, 12(1):611–623, January 2013.

[52] Christopher Ferrie and Osama Moussa. Robust and efficient \textit{in Situ} quantum control. *Physical Review A*, 91(5):052306, May 2015.

[53] M. A. Fogarty, M. Veldhorst, R. Harper, C. H. Yang, S. D. Bartlett, S. T. Flammia, and A. S. Dzurak. Nonexponential fidelity decay in randomized benchmarking with low-frequency noise. *Physical Review A*, 92(2):022326, August 2015.

[54] Evan M. Fortunato, Marco A. Pravia, Nicolas Boulant, Grum Teklemariam, Timothy F. Havel, and David G. Cory. Design of strongly modulating pulses to implement precise effective Hamiltonians for quantum information processing. *The Journal of Chemical Physics*, 116(17):7599–7606, May 2002.

[55] Austin G. Fowler, Ashley M. Stephens, and Peter Groszkowski. High-threshold universal quantum computation on the surface code. *Physical Review A*, 80(5):052312, November 2009.

[56] Ray Freeman, H. D. W Hill, and R Kaptein. An adaptive scheme for measuring NMR spin-lattice relaxation times. *Journal of Magnetic Resonance (1969)*, 7(1):82–98, May 1972.

[57] Adam Gali, Maria Fyta, and Efthimios Kaxiras. Ab initio supercell calculations on nitrogen-vacancy center in diamond: Electronic structure and hyperfine tensors. *Physical Review B*, 77(15):155206, April 2008.

[58] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman and Hall/CRC, Boca Raton, 3 edition edition, November 2013.

[59] Michael H. Goerz, Daniel M. Reich, and Christiane P. Koch. Optimal control theory for a unitary operation under dissipative evolution. *New Journal of Physics*, 16(5):055012, May 2014.

[60] Daniel Gottesman. An Introduction to Quantum Error Correction and Fault-Tolerant Quantum Computation. *arXiv:0904.2557*, April 2009.

[61] Christopher Granade, Joshua Combes, and D. G. Cory. Practical Bayesian tomography. *New Journal of Physics*, 18(3):033024, 2016.

[62] Christopher Granade, Christopher Ferrie, and D. G. Cory. Accelerated randomized benchmarking. *New Journal of Physics*, 17(1):013042, 2015.

[63] Christopher Granade, Christopher Ferrie, Ian Hincks, Steven Casagrande, Thomas Alexander, Jonathan Gross, Michal Kononenko, and Yuval Sanders. QInfer: Statistical inference software for quantum applications. *Quantum*, 1:5, April 2017.

[64] Christopher Granade and Nathan Wiebe. Structured Filtering. *arXiv:1612.00762 [quant-ph, stat]*, December 2016.

[65] Christopher E. Granade. *Characterization, Verification and Control for Large Quantum Systems*. PhD thesis, University of Waterloo, Waterloo, ON, 2015.

[66] Christopher E. Granade, Christopher Ferrie, Nathan Wiebe, and D. G. Cory. Robust online Hamiltonian learning. *New Journal of Physics*, 14(10):103013, 2012.

[67] C. Grezes, B. Julsgaard, Y. Kubo, M. Stern, T. Umeda, J. Isoya, H. Sumiya, H. Abe, S. Onoda, T. Ohshima, V. Jacques, J. Esteve, D. Vion, D. Esteve, K. Mølmer, and P. Bertet. Multimode Storage and Retrieval of Microwave Fields in a Spin Ensemble. *Physical Review X*, 4(2):021049, June 2014.

[68] M. S. Grinolds, S. Hong, P. Maletinsky, L. Luan, M. D. Lukin, R. L. Walsworth, and A. Yacoby. Nanoscale magnetic imaging of a single electron spin under ambient conditions. *Nature Physics*, 9(4):215–219, April 2013.

[69] D. Gross, K. Audenaert, and J. Eisert. Evenly distributed unitaries: On the structure of unitary designs. *Journal of Mathematical Physics*, 48(5):052104, May 2007.

[70] A. Gruber, A. Dräbenstedt, C. Tietz, L. Fleury, J. Wrachtrup, and C. von Borczyskowski. Scanning Confocal Optical Microscopy and Magnetic Resonance on Single Defect Centers. *Science*, 276(5321):2012–2014, June 1997.

[71] Simon Gustavsson, Olger Zwier, Jonas Bylander, Fei Yan, Fumiki Yoshihara, Yasunobu Nakamura, Terry P. Orlando, and William D. Oliver. Improving Quantum Gate Fidelities by Using a Qubit to Measure Microwave Pulse Distortions. *Physical Review Letters*, 110(4):040502, January 2013.

[72] Mauricio Gutiérrez, Lukas Svec, Alexander Vargo, and Kenneth R. Brown. Approximation of realistic errors by Clifford channels and Pauli measurements. *Physical Review A*, 87(3):030302, March 2013.

[73] U. Haeberlen and J. S. Waugh. Coherent Averaging Effects in Magnetic Resonance. *Physical Review*, 175(2):453–467, November 1968.

[74] Wolfgang Härdle and Léopold Simar. *Applied Multivariate Statistical Analysis: Third Edition*. 2012.

[75] J Harrison, M. J Sellars, and N. B Manson. Optical spin polarisation of the N-V centre in diamond. *Journal of Luminescence*, 107(1–4):245–248, May 2004.

[76] W. K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57(1):97–109, 1970.

[77] Reinier W. Heeres, Philip Reinhold, Nissim Ofek, Luigi Frunzio, Liang Jiang, Michel H. Devoret, and Robert J. Schoelkopf. Implementing a universal gate set on a logical qubit encoded in an oscillator. *Nature Communications*, 8(1):94, July 2017.

[78] Michael K. Henry, Chandrasekhar Ramanathan, Jonathan S. Hodges, Colm A. Ryan, Michael J. Ditty, Raymond Laflamme, and David G. Cory. Fidelity Enhancement by Logical Qubit Encoding. *Physical Review Letters*, 99(22):220501, November 2007.

[79] B. L. Higgins, D. W. Berry, S. D. Bartlett, M. W. Mitchell, H. M. Wiseman, and G. J. Pryde. Demonstrating Heisenberg-limited unambiguous phase estimation without adaptive measurements. *New Journal of Physics*, 11(7):073023, 2009.

[80] B. L. Higgins, D. W. Berry, S. D. Bartlett, H. M. Wiseman, and G. J. Pryde. Entanglement-free Heisenberg-limited phase estimation. *Nature*, 450(7168):393–396, November 2007.

[81] N. Higham. The Scaling and Squaring Method for the Matrix Exponential Revisited. *SIAM Journal on Matrix Analysis and Applications*, 26(4):1179–1193, January 2005.

[82] I. N. Hincks, C. E. Granade, T. W. Borneman, and D. G. Cory. Controlling Quantum Devices with Nonlinear Hardware. *Physical Review Applied*, 4(2):024012, August 2015.

[83] Ian Hincks, Thomas Alexander, Michal Kononenko, Benjamin Soloway, and David G. Cory. Hamiltonian Learning with Online Bayesian Experiment Design in Practice. *arXiv:1806.02427 [quant-ph]*, June 2018.

[84] Ian Hincks, Christopher Granade, and David G. Cory. Statistical Inference with Quantum Measurements: Methodologies for Nitrogen Vacancy Centers in Diamond. *New Journal of Physics*, 20(1):013022, January 2018.

[85] Ian Hincks, Joel J. Wallman, Chris Ferrie, Chris Granade, and David G. Cory. Bayesian Inference for Randomized Benchmarking Protocols. *arXiv:1802.00401 [quant-ph]*, February 2018.

[86] Masashi Hirose, Clarice D. Aiello, and Paola Cappellaro. Continuous dynamical decoupling magnetometry. *Physical Review A*, 86(6):062320, December 2012.

[87] David Hocker, Constantin Brif, Matthew D. Grace, Ashley Donovan, Tak-San Ho, Katharine Moore Tibbetts, Rebing Wu, and Herschel Rabitz. Characterization of control noise effects in optimal quantum unitary dynamics. *Physical Review A*, 90(6):062309, December 2014.

[88] J. S. Hodges, J. C. Yang, C. Ramanathan, and D. G. Cory. Universal control of nuclear spins via anisotropic hyperfine interactions. *Physical Review A*, 78(1):010303, July 2008.

[89] Matthew D. Hoffman and Andrew Gelman. The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[90] R.A. Horn and C.R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1 edition, 1985.

[91] D. I. Hoult. Fast recovery, high sensitivity NMR probe and preamplifier for low frequencies. *Review of Scientific Instruments*, 50(2):193–200, February 1979.

[92] F. Huszár and N. M. T. Houlsby. Adaptive Bayesian quantum tomography. *Physical Review A*, 85(5):052120, May 2012.

[93] Georg Jäger and Ulrich Hohenester. Optimal quantum control of Bose-Einstein condensates in magnetic microtraps: Consideration of filter effects. *Physical Review A*, 88(3):035601, September 2013.

[94] F. Jelezko, T. Gaebel, I. Popa, M. Domhan, A. Gruber, and J. Wrachtrup. Observation of Coherent Oscillation of a Single Nuclear Spin and Realization of a Two-Qubit Conditional Quantum Gate. *Physical Review Letters*, 93(13):130501, September 2004.

[95] F. Jelezko and J. Wrachtrup. Read-out of single spins by optical spectroscopy. *Journal of Physics: Condensed Matter*, 16(30):R1089, 2004.

[96] L. Jiang, J. S Hodges, J. R Maze, P. Maurer, J. M Taylor, D. G Cory, P. R Hemmer, R. L Walsworth, A. Yacoby, A. S Zibrov, and M. D Lukin. Repetitive Readout of a Single Electronic Spin Via Quantum Logic with Nuclear Spin Ancillae. *Science*, 326(5950):267–272, September 2009.

[97] Eric Jones, Eric Oliphant, and Pearu Peterson. SciPy: Open Source Scientific Tools for Python, 2001.

[98] Dimitris Karlis and Panagiotis Tsiamyrtzis. Exact Bayesian modeling for bivariate Poisson data and extensions. *Statistics and Computing*, 18(1):27–40, March 2008.

[99] Navin Khaneja, Timo Reiss, Cindie Kehlet, Thomas Schulte-Herbrüggen, and Steffen J. Glaser. Optimal control of coupled spin dynamics: Design of NMR pulse sequences by gradient ascent algorithms. *Journal of Magnetic Resonance*, 172(2):296–305, February 2005.

[100] Shelby Kimmel, Marcus P. da Silva, Colm A. Ryan, Blake R. Johnson, and Thomas Ohki. Robust Extraction of Tomographic Information via Randomized Benchmarking. *Physical Review X*, 4(1):011050, March 2014.

[101] E. Knill, D. Leibfried, R. Reichle, J. Britton, R. B. Blakestad, J. D. Jost, C. Langer, R. Ozeri, S. Seidelin, and D. J. Wineland. Randomized benchmarking of quantum gates. *Physical Review A*, 77(1):012307, January 2008.

[102] Kyryl Kobzar, Sebastian Ehni, Thomas E. Skinner, Steffen J. Glaser, and Burkhard Luy. Exploring the limits of broadband 90° and 180° universal rotation pulses. *Journal of Magnetic Resonance*, 225:142–160, December 2012.

[103] Kyryl Kobzar, Thomas E. Skinner, Navin Khaneja, Steffen J. Glaser, and Burkhard Luy. Exploring the limits of broadband excitation and inversion: II. Rf-power optimized pulses. *Journal of Magnetic Resonance*, 194(1):58–66, September 2008.

[104] Van D. M. Koroleva, Soumyajit Mandal, Yi-Qiao Song, and Martin D. Hürlimann. Broadband CPMG sequence with short composite refocusing pulses. *Journal of Magnetic Resonance*, 230:64–75, May 2013.

[105] Robert Kosut, Ian A. Walmsley, and Herschel Rabitz. Optimal Experiment Design for Quantum State and Process Tomography and Hamiltonian Parameter Estimation. *arXiv:quant-ph/0411093*, November 2004.

[106] K. S. Kravtsov, S. S. Straupe, I. V. Radchenko, N. M. T. Houlsby, F. Huszár, and S. P. Kulik. Experimental adaptive Bayesian tomography. *Physical Review A*, 87(6):062122, June 2013.

[107] Gershon Kurizki, Patrice Bertet, Yuimaru Kubo, Klaus Mølmer, David Petrosyan, Peter Rabl, and Jörg Schmiedmayer. Quantum technologies with hybrid systems. *Proceedings of the National Academy of Sciences*, 112(13):3866–3873, March 2015.

[108] T. D. Ladd, F. Jelezko, R. Laflamme, Y. Nakamura, C. Monroe, and J. L. O'Brien. Quantum computers. *Nature*, 464(7285):45–53, March 2010.

[109] Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer Texts in Statistics. Springer-Verlag, New York, 2 edition, 1998.

[110] Michal Leskes, P.K. Madhu, and Shimon Vega. Floquet theory in solid-state nuclear magnetic resonance. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 57(4):345–380, November 2010.

[111] Malcolm H. Levitt. *Spin Dynamics: Basics of Nuclear Magnetic Resonance*. Wiley, December 2001.

[112] G. Lindblad. On the generators of quantum dynamical semigroups. *Communications in Mathematical Physics*, 48(2):119–130, June 1976.

[113] Jane Liu and Mike West. Combined parameter and state estimation in simulation-based filtering. In *Sequential Monte Carlo Methods in Practice*, pages 197–223. Springer, 2001.

[114] Maas. *Nonlinear Microwave and RF Circuits*. Artech House Publishers, Boston, MA, 2 edition edition, February 2003.

[115] Easwar Magesan, J. M. Gambetta, and Joseph Emerson. Scalable and Robust Randomized Benchmarking of Quantum Processes. *Physical Review Letters*, 106(18):180504, May 2011.

[116] Easwar Magesan, Jay M. Gambetta, and Joseph Emerson. Characterizing quantum gates via randomized benchmarking. *Physical Review A*, 85(4):042311, April 2012.

[117] Easwar Magesan, Jay M. Gambetta, B. R. Johnson, Colm A. Ryan, Jerry M. Chow, Seth T. Merkel, Marcus P. da Silva, George A. Keefe, Mary B. Rothwell,

Thomas A. Ohki, Mark B. Ketchen, and M. Steffen. Efficient Measurement of Quantum Gate Error by Interleaved Randomized Benchmarking. *Physical Review Letters*, 109(8):080505, August 2012.

[118] P. Maletinsky, S. Hong, M. S. Grinolds, B. Hausmann, M. D. Lukin, R. L. Walsworth, M. Loncar, and A. Yacoby. A robust scanning diamond sensor for nanoscale imaging with single nitrogen-vacancy centres. *Nature Nanotechnology*, 7(5):320–324, May 2012.

[119] H. Malissa, D. I. Schuster, A. M. Tyryshkin, A. A. Houck, and S. A. Lyon. Superconducting coplanar waveguide resonators for low temperature pulsed electron spin resonance spectroscopy. *Review of Scientific Instruments*, 84(2):025116, February 2013.

[120] H. J. Mamin, M. Kim, M. H. Sherwood, C. T. Rettner, K. Ohno, D. D. Awschalom, and D. Rugar. Nanoscale Nuclear Magnetic Resonance with a Nitrogen-Vacancy Spin Sensor. *Science*, 339(6119):557–560, January 2013.

[121] Soumyajit Mandal, Troy W. Borneman, Van D. M. Koroleva, and Martin D. Hürlimann. Direct optimization of signal-to-noise ratio of CPMG-like sequences in inhomogeneous fields. *Journal of Magnetic Resonance*, 247:54–66, October 2014.

[122] Soumyajit Mandal, Van D. M. Koroleva, Troy W. Borneman, Yi-Qiao Song, and Martin D. Hürlimann. Axis-matching excitation pulses for CPMG-like sequences in inhomogeneous fields. *Journal of Magnetic Resonance*, 237:1–10, December 2013.

[123] James B. McDonald and Yexiao J. Xu. A generalization of the beta distribution with applications. *Journal of Econometrics*, 66(1):133–152, March 1995.

[124] David C. McKay, Christopher J. Wood, Sarah Sheldon, Jerry M. Chow, and Jay M. Gambetta. Efficient Z-Gates for Quantum Computing. *arXiv:1612.00858 [quant-ph]*, December 2016.

[125] Michael Mehring. *High Resolution NMR in Solids*. Springer, 1976.

[126] H. R. Mohebbi, O. W. B. Benningshof, I. a. J. Taminiau, G. X. Miao, and D. G. Cory. Composite arrays of superconducting microstrip line resonators. *Journal of Applied Physics*, 115(9):094502, March 2014.

[127] Katharine Moore, Michael Hsieh, and Herschel Rabitz. On the relationship between quantum control landscape structure and optimization complexity. *The Journal of Chemical Physics*, 128(15):154117, April 2008.

[128] F. Motzoi, J. M. Gambetta, S. T. Merkel, and F. K. Wilhelm. Optimal control methods for rapidly time-varying Hamiltonians. *Physical Review A*, 84(2):022307, August 2011.

[129] Osama Moussa, Marcus P. da Silva, Colm A. Ryan, and Raymond Laflamme. Practical Experimental Certification of Computational Quantum Gates Using a Twirling Procedure. *Physical Review Letters*, 109(7):070504, August 2012.

[130] Osama Moussa, Ian Hincks, and David G. Cory. Preparing and preserving the double quantum coherence in NV- centers in Diamond at low fields. *Journal of Magnetic Resonance*, 249:24–31, December 2014.

[131] J. T. Muhonen, A. Laucht, S. Simmons, J. P. Dehollain, R. Kalra, F. E. Hudson, S. Freer, K. M. Itoh, D. N. Jamieson, J. C. McCallum, A. S. Dzurak, and A. Morello. Quantifying the quantum gate fidelity of single-atom spin qubits in silicon by randomized benchmarking. *Journal of Physics: Condensed Matter*, 27(15):154205, 2015.

[132] M. M. Müller, D. M. Reich, M. Murphy, H. Yuan, J. Vala, K. B. Whaley, T. Calarco, and C. P. Koch. Optimizing entangling quantum gates for physical systems. *Physical Review A*, 84(4):042315, October 2011.

[133] Radford M. Neal. MCMC using Hamiltonian dynamics. *arXiv:1206.1901 [physics, stat]*, June 2012.

[134] J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, January 1965.

[135] Philipp Neumann, Johannes Beck, Matthias Steiner, Florian Rempp, Helmut Fedder, Philip R. Hemmer, Jörg Wrachtrup, and Fedor Jelezko. Single-Shot Readout of a Single Nuclear Spin. *Science*, 329(5991):542–544, July 2010.

[136] Manfred Opper. On-line Learning in Neural Networks. chapter A Bayesian Approach to On-line Learning, pages 363–378. Cambridge University Press, New York, NY, USA, 1998.

[137] José P. Palao, Ronnie Kosloff, and Christiane P. Koch. Protecting coherence in optimal control theory: State-dependent constraint approach. *Physical Review A*, 77(6):063412, June 2008.

[138] Om Patange. *On an Instrument for the Coherent Investigation of Nitrogen-Vacancy Centres in Diamond*. Master's Thesis, University of Waterloo, 2013.

[139] Lev Semenovich Pontryagin. *Mathematical Theory of Optimal Processes*. CRC Press, 1987.

[140] Marco A. Pravia, Nicolas Boulant, Joseph Emerson, Amro Farid, Evan M. Fortunato, Timothy F. Havel, R. Martinez, and David G. Cory. Robust control of quantum information. *The Journal of Chemical Physics*, 119(19):9993–10001, November 2003.

[141] Timothy Proctor, Kenneth Rudinger, Kevin Young, Mohan Sarovar, and Robin Blume-Kohout. What randomized benchmarking actually measures. *arXiv:1702.01853 [quant-ph]*, February 2017.

[142] S. Putz, D. O. Krimer, R. Amsüss, A. Valookaran, T. Nöbauer, J. Schmiedmayer, S. Rotter, and J. Majer. Protecting a spin ensemble against decoherence in the strong-coupling regime of cavity QED. *Nature Physics*, 10(10):720–724, October 2014.

[143] Bo Qi, Zhibo Hou, Yuanlong Wang, Daoyi Dong, Han-Sen Zhong, Li Li, Guo-Yong Xiang, Howard M. Wiseman, Chuan-Feng Li, and Guang-Can Guo. Adaptive quantum state tomography via linear regression estimation: Theory and two-qubit experiment. *npj Quantum Information*, 3(1):19, April 2017.

[144] Herschel A. Rabitz, Michael M. Hsieh, and Carey M. Rosenthal. Quantum Optimally Controlled Transition Landscapes. *Science*, 303(5666):1998–2001, March 2004.

[145] Douglas D. Rife and John Vanderkooy. Transfer-Function Measurement with Maximum-Length Sequences. *Journal of the Audio Engineering Society*, 37(6):419–444, June 1989.

[146] Lucio Robledo, Hannes Bernien, Toeno van der Sar, and Ronald Hanson. Spin dynamics in the optical cycle of single nitrogen-vacancy centres in diamond. *New Journal of Physics*, 13(2):025013, February 2011.

[147] M. A. Rol, C. C. Bultink, T. E. O'Brien, S. R. de Jong, L. S. Theis, X. Fu, F. Luthi, R. F. L. Vermeulen, J. C. de Sterke, A. Bruno, D. Deurloo, R. N. Schouten, F. K. Wilhelm, and L. DiCarlo. Restless Tuneup of High-Fidelity Qubit Gates. *Physical Review Applied*, 7(4):041001, April 2017.

[148] L. Rondin, J.-P. Tetienne, T. Hingant, J.-F. Roch, P. Maletinsky, and V. Jacques. Magnetometry with nitrogen-vacancy defects in diamond. *Reports on Progress in Physics*, 77(5):056503, May 2014.

[149] L. Rondin, J.-P. Tetienne, S. Rohart, A. Thiaville, T. Hingant, P. Spinicelli, J.-F. Roch, and V. Jacques. Stray-field imaging of magnetic vortices with a single diamond spin. *Nature Communications*, 4:2279, July 2013.

[150] Xing Rong, Jianpei Geng, Fazhan Shi, Ying Liu, Kebiao Xu, Wenchao Ma, Fei Kong, Zhen Jiang, Yang Wu, and Jiangfeng Du. Experimental fault-tolerant universal quantum gates with solid-state spins under ambient conditions. *Nature Communications*, 6:8748, November 2015.

[151] Benjamin Rowland and Jonathan A. Jones. Implementing quantum logic gates with gradient ascent pulse engineering: Principles and practicalities. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 370(1976):4636–4650, October 2012.

[152] Colm A. Ryan, Blake R. Johnson, Diego Ristè, Brian Donovan, and Thomas A. Ohki. Hardware for dynamic quantum computing. *Review of Scientific Instruments*, 88(10):104703, October 2017.

[153] R. J. Schoelkopf and S. M. Girvin. Wiring up quantum systems. *Nature*, 451(7179):664–669, February 2008.

[154] T. Schulte-Herbrüggen, A. Spörl, N. Khaneja, and S. J. Glaser. Optimal control for generating quantum gates in open dissipative systems. *Journal of Physics B: Atomic, Molecular and Optical Physics*, 44(15):154013, August 2011.

[155] Alexandr Sergeevich, Anushya Chandran, Joshua Combes, Stephen D. Bartlett, and Howard M. Wiseman. Characterization of a qubit Hamiltonian using adaptive measurements in a fixed basis. *Physical Review A*, 84(5):052315, November 2011.

[156] Jayaram Sethuraman. A CONSTRUCTIVE DEFINITION OF DIRICHLET PRIORS. *Statistica Sinica*, 4(2):639–650, 1994.

[157] Sarah Sheldon, Lev S. Bishop, Easwar Magesan, Stefan Filipp, Jerry M. Chow, and Jay M. Gambetta. Characterizing errors on qubit operations via iterative randomized benchmarking. *Physical Review A*, 93(1):012301, January 2016.

[158] B. J. Shields, Q. P. Unterreithmeier, N. P. de Leon, H. Park, and M. D. Lukin. Efficient readout of a single spin state in diamond via spin-to-charge conversion. *Physical Review Letters*, 114(13), March 2015.

[159] A. J. Sigillito, H. Malissa, A. M. Tyryshkin, H. Riemann, N. V. Abrosimov, P. Becker, H.-J. Pohl, M. L. W. Thewalt, K. M. Itoh, J. J. L. Morton, A. A. Houck, D. I. Schuster, and S. A. Lyon. Fast, low-power manipulation of spin ensembles in superconducting microresonators. *Applied Physics Letters*, 104(22):222407, June 2014.

[160] Thomas E. Skinner, Timo O. Reiss, Burkhard Luy, Navin Khaneja, and Steffen J. Glaser. Reducing the duration of broadband excitation pulses using optimal control with limited RF amplitude. *Journal of Magnetic Resonance*, 167(1):68–74, March 2004.

[161] Benjamin Smeltzer, Lilian Childress, and Adam Gali. 13C hyperfine interactions in the nitrogen-vacancy centre in diamond. *New Journal of Physics*, 13(2):025021, February 2011.

[162] Philipp E. Spindler, Yun Zhang, Burkhard Endeward, Naum Gershernzon, Thomas E. Skinner, Steffen J. Glaser, and Thomas F. Prisner. Shaped optimal control pulses for increased excitation bandwidth in EPR. *Journal of Magnetic Resonance*, 218:49–58, May 2012.

[163] Daniel Steel. Bayesian Confirmation Theory and The Likelihood Principle. *Synthese*, 156(1):53–77, May 2007.

[164] Markku P. V. Stenberg, Oliver Köhn, and Frank K. Wilhelm. Characterization of decohering quantum systems: Machine learning approach. *Physical Review A*, 93(1):012122, January 2016.

[165] Markku P. V. Stenberg, Kevin Pack, and Frank K. Wilhelm. Adaptive identification of coherent states. *Physical Review A*, 92(6):063852, December 2015.

[166] Markku P. V. Stenberg and Frank K. Wilhelm. Simultaneous model selection and parameter estimation: A superconducting qubit coupled to a bath of incoherent two-level systems. *Physical Review A*, 94(5):052119, November 2016.

[167] G. I. Struchalin, I. A. Pogorelov, S. S. Straupe, K. S. Kravtsov, I. V. Radchenko, and S. P. Kulik. Experimental adaptive quantum tomography of two-qubit states. *Physical Review A*, 93(1):012103, January 2016.

[168] Dieter Suter. *The Physics of Laser-Atom Interactions*, volume 19. Cambridge University Press, 1997.

[169] H. Taitelbaum, J. A. Ferretti, R. G. S. Spencer, and G. H. Weiss. Two-Stage Inversion-Recovery Experiments for Measurements of T1. *Journal of Magnetic Resonance, Series A*, 105(1):59–64, October 1993.

[170] J. M. Taylor, P. Cappellaro, L. Childress, L. Jiang, D. Budker, P. R. Hemmer, A. Yacoby, R. Walsworth, and M. D. Lukin. High-sensitivity diamond magnetometer with nanoscale resolution. *Nat Phys*, 4(10):810–816, October 2008.

[171] Yee Whye Teh. Dirichlet Process. In Claude Sammut and Geoffrey I. Webb, editors, *Encyclopedia of Machine Learning*, pages 280–287. Springer US, 2011.

[172] Michael Tinkham. *Introduction to Superconductivity: Second Edition*. Dover Publications, Mineola, N.Y., second edition edition edition, June 2004.

[173] R. Tycko. Broadband Population Inversion. *Physical Review Letters*, 51(9):775–777, August 1983.

[174] Christoph W. Ueberhuber. *Numerical Computation 2: Methods, Software, and Analysis*. Springer-Verlag, Berlin Heidelberg, 1997.

[175] Harry L. Van Trees. *Detection, Estimation, and Modulation Theory*. John Wiley & Sons, 2004.

[176] M. Veldhorst, J. C. C. Hwang, C. H. Yang, A. W. Leenstra, B. de Ronde, J. P. Dehollain, J. T. Muhonen, F. E. Hudson, K. M. Itoh, A. Morello, and A. S. Dzurak. An addressable quantum dot qubit with fault-tolerant control-fidelity. *Nature Nanotechnology*, 9(12):981–985, December 2014.

[177] Joel Wallman, Christopher Granade, Robin Harper, and Steven T. Flammia. Estimating the Coherence of Noise. *arXiv:1503.07865 [quant-ph]*, March 2015.

[178] Joel J. Wallman. Randomized benchmarking with gate-dependent noise. *arXiv:1703.09835 [quant-ph]*, March 2017.

[179] Joel J. Wallman, Marie Barnhill, and Joseph Emerson. Robust Characterization of Loss Rates. *Physical Review Letters*, 115(6):060501, August 2015.

[180] Joel J. Wallman, Marie Barnhill, and Joseph Emerson. Robust characterization of leakage errors. *New Journal of Physics*, 18(4):043021, 2016.

[181] Joel J. Wallman and Steven T. Flammia. Randomized benchmarking with confidence. *New Journal of Physics*, 16(10):103032, October 2014.

[182] Jianwei Wang, Stefano Paesani, Raffaele Santagati, Sebastian Knauer, Antonio A. Gentile, Nathan Wiebe, Maurangelo Petruzzella, Jeremy L. O'Brien, John G. Rarity, Anthony Laing, and Mark G. Thompson. Experimental Quantum Hamiltonian Learning. *arXiv:1703.05402 [quant-ph]*, March 2017.

[183] John Watrous. *The Theory of Quantum Information*. Cambridge University Press, Cambridge, United Kingdom, April 2018.

[184] Nathan Wiebe, Christopher Granade, Christopher Ferrie, and D. G. Cory. Hamiltonian Learning and Certification Using Quantum Resources. *Physical Review Letters*, 112(19):190501, May 2014.

[185] Nathan Wiebe, Christopher Granade, Christopher Ferrie, and David Cory. Quantum Hamiltonian learning using imperfect quantum resources. *Physical Review A*, 89(4):042314, April 2014.

[186] S. Wimperis. Broadband, Narrowband, and Passband Composite Pulses for Use in Advanced NMR Experiments. *Journal of Magnetic Resonance, Series A*, 109(2):221–231, August 1994.

[187] null Wiseman. Adaptive phase measurements of optical modes: Going beyond the marginal Q distribution. *Physical Review Letters*, 75(25):4587–4590, December 1995.

[188] Christopher Wood, Ian Hincks, and Christopher Granade. QuantumUtils for Mathematica, 2014.

[189] Christopher J. Wood and Jay M. Gambetta. Quantification and Characterization of Leakage Errors. *arXiv:1704.03081 [quant-ph]*, April 2017.

[190] T. Xia, M. Lichtman, K. Maller, A. W. Carr, M. J. Piotrowicz, L. Isenhower, and M. Saffman. Randomized Benchmarking of Single-Qubit Gates in a 2D Array of Neutral-Atom Qubits. *Physical Review Letters*, 114(10):100503, March 2015.

[191] G. Y. Xiang, B. L. Higgins, D. W. Berry, H. M. Wiseman, and G. J. Pryde. Entanglement-enhanced measurement of a completely unknown optical phase. *Nature Photonics*, 5(1):43–47, January 2011.

[192] Ze-Liang Xiang, Sahel Ashhab, J. Q. You, and Franco Nori. Hybrid quantum circuits: Superconducting circuits interacting with other quantum systems. *Reviews of Modern Physics*, 85(2):623–653, April 2013.

[193] Kebiao Xu, Tianyu Xie, Zhaokai Li, Xiangkun Xu, Mengqi Wang, Xiangyu Ye, Fei Kong, Jianpei Geng, Changkui Duan, Fazhan Shi, and Jiangfeng Du. Experimental Adiabatic Quantum Factorization under Ambient Conditions Based on a Solid-State Single Spin System. *Physical Review Letters*, 118(13):130504, March 2017.

[194] Mingan Yang, David B. Dunson, and Donna Baird. Semiparametric Bayes hierarchical models with mean and variance constraints. *Computational Statistics & Data Analysis*, 54(9):2172–2186, September 2010.

[195] Hidehiro Yonezawa, Daisuke Nakane, Trevor A. Wheatley, Kohjiro Iwasawa, Shuntaro Takeda, Hajime Arao, Kentaro Ohki, Koji Tsumura, Dominic W. Berry, Timothy C. Ralph, Howard M. Wiseman, Elanor H. Huntington, and Akira Furusawa. Quantum-Enhanced Optical-Phase Tracking. *Science*, 337(6101):1514–1517, September 2012.

APPENDICES

Appendix A

Code for Numerics

Parts of this thesis relied heavily on numerics, and indeed, most of the effort of some projects was writing the code. In this appendix we gather together as much of it as is publicly available.

Libraries

- **QInfer -** https://github.com/QInfer/python-qinfer A Python library that performs Bayesian inference using sequential Monte Carlo, targeted for quantum systems[63].
- **Stan -** http://mc-stan.org/ A general purpose Bayesian inference algorithm whose main algorithm is HMC with NUTS [22].
- **QuantumUtils -** https://github.com/QuantumUtils/quantum-utils-mathematica A Mathematica library that contains tools for quantum control, GRAPE, simulation, channel representation conversion, and perturbations.
- **nvham -** https://github.com/ihincks/nvham Mathematica Library to help construct Nitrogen Vacancy Hamiltonians.

Project Repositories

- Chapter 3: https://github.com/ihincks/nv-meas-code
- Chapter 4: https://github.com/ihincks/nv-adaptive
- Chapter 6: https://github.com/ihincks/birb

Appendix B

Certain Details of NV Measurement

B.1 Stochastic Moments

Mathematica version 10.0.2.0 was used to execute the code snippets in Listing B.1 and Listing B.2.

Listing B.1: Mathematica code for computing the moments of the stochastic model defined in Equation Equation 3.6 conditional on the value of α_1.

```
p[expr_] := TransformedProcess[
    expr,
    {
        ν ∼ OrnsteinUhlenbeckProcess[0, σν, θν, α0 − Γ],
        κ ∼ OrnsteinUhlenbeckProcess[κ0, σκ, θκ]
    },
    t
];

αmean1 = p[Γ+ν[t]][t]//Mean
βmean1 = p[Γ+κ[t]ν[t]][t]//Mean
αvar1 = p[Γ+ν[t]][t]//Variance
βvar1 = p[Γ+κ[t]ν[t]][t]//Variance//Expand
(* We need to compute the covariance manually
    using Cov[a,b]=Mean[a*b]-Mean[a]Mean[b] *)
αβproduct =(Γ+ν[t])(Γ+κ[t]ν[t])//Expand;
αβcov1 = Sum[Mean[p[expr][t]],{expr,List@@αβproduct}]
    − αmean*βmean//Expand
```

Listing B.2: Mathematica code for using the laws of total expectation, variance, and covariance to derive the moments of the stochastic model defined in Equation Equation 3.6. The results are shown in Equation Equation 3.7.

```
totalExp[mean_] :=
    Mean[TransformedDistribution[
        mean,
        α1∼NormalDistribution[α0,σα]
    ]]

totalV\]ar[mean_, var_] := Variance[TransformedDistribution[
        mean,
        α1∼NormalDistribution[α0,σα]]
    ], + Mean[TransformedDistribution[
        var,
        α1∼NormalDistribution[α0,σα]]
    ]
totalCov[mean1_,mean2_,cov_] :=
    Covariance[TransformedDistribution[
        {mean1,mean2},
        α1∼NormalDistribution[α0,σα]],1,2
    ] + Mean[TransformedDistribution[
        cov,
        α1∼NormalDistribution[α0,σα]
    ]]

αmean = totalExp[αmean1]
βmean = totalExp[βmean1]
αvar = totalVar[αmean1,αvar1]
βvar = totalVar[βmean1,βvar1]//Expand
αβcov = totalCov[αmean1,βmean1,αβcov1]//Expand
```

B.2 Cumulants of the Conditional Likelihood Function

We wish to compute

$$I(p,\alpha,\beta) = -\mathbb{E}_{x,y,z}\left[\left(\frac{\partial}{\partial\theta_i}\frac{\partial \log\mathcal{L}}{\partial\theta_j}\right)_{i,j=1}^{3}\right] = \sum_{x=0}^{\infty}\sum_{y=0}^{\infty}\sum_{z=0}^{\infty}\mathcal{L}(x,y,z|p,\alpha,\beta)\left(\frac{\partial}{\partial\theta_i}\frac{\partial \log\mathcal{L}}{\partial\theta_j}\right)_{i,j=1}^{3}, \tag{B.1}$$

$$K_{i,j,k} = \mathbb{E}_{x,y,z}\left[\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}\frac{\partial \log\mathcal{L}}{\partial\theta_k}\right], \quad \text{and} \tag{B.2}$$

$$J_{j;i,k} = \mathbb{E}_{x,y,z}\left[\frac{\partial \log\mathcal{L}}{\partial\theta_j}\cdot\frac{\partial}{\partial\theta_i}\frac{\partial \log\mathcal{L}}{\partial\theta_k}\right] \tag{B.3}$$

where $\theta_1 = p$, $\theta_2 = \alpha$, $\theta_3 = \beta$, and from Equation Equation 3.13 we have

$$\mathcal{L}(x,y,z|p,\alpha,\beta) = \frac{\alpha^x \mathrm{e}^{-\alpha}}{x!}\cdot\frac{\beta^y \mathrm{e}^{-\beta}}{y!}\cdot\frac{(p\alpha+(1-p)\beta)^z \mathrm{e}^{-(p\alpha+(1-p)\beta)}}{z!}. \tag{B.4}$$

Listing B.3. The results of this code for the Fisher information matrix are

$$\begin{aligned} I(p,\alpha,\beta) &= -\mathbb{E}_{x,y,z}\left[\left(\frac{\partial}{\partial\theta_i}\frac{\partial \log\mathcal{L}}{\partial\theta_j}\right)_{i,j=1}^{3}\right] \\ &= \begin{pmatrix} \frac{(\alpha-\beta)^2}{p(\alpha-\beta)+\beta} & \frac{p(\alpha-\beta)}{p(\alpha-\beta)+\beta} & \frac{\alpha}{\beta+\alpha p-\beta p}-1 \\ \frac{p(\alpha-\beta)}{p(\alpha-\beta)+\beta} & \frac{p^2}{p\alpha-p\beta+\beta}+\frac{1}{\alpha} & -\frac{(p-1)p}{p(\alpha-\beta)+\beta} \\ \frac{\alpha}{p\alpha-p\beta+\beta}-1 & -\frac{(p-1)p}{p(\alpha-\beta)+\beta} & \frac{p\alpha+(p-2)(p-1)\beta}{\beta(p(\alpha-\beta)+\beta)} \end{pmatrix} \end{aligned} \tag{B.5}$$

with inverse

$$I(p,\alpha,\beta)^{-1} = \begin{pmatrix} \frac{p(p+1)\alpha+(p-2)(p-1)\beta}{(\alpha-\beta)^2} & \frac{p\alpha}{\beta-\alpha} & \frac{(p-1)\beta}{\alpha-\beta} \\ \frac{p\alpha}{\beta-\alpha} & \alpha & 0 \\ \frac{(p-1)\beta}{\alpha-\beta} & 0 & \beta \end{pmatrix}. \tag{B.6}$$

Further, the higher order cumulants turn out to be

$$
\begin{aligned}
K_{1,\cdot,\cdot} &= \begin{pmatrix} \frac{2(\alpha-\beta)^3}{(p(\alpha-\beta)+\beta)^2} & \frac{2\beta(\beta-\alpha)}{(p(\alpha-\beta)+\beta)^2} & \frac{2\alpha(\alpha-\beta)}{(p(\alpha-\beta)+\beta)^2} \\ \frac{2\beta(\beta-\alpha)}{(p(\alpha-\beta)+\beta)^2} & -\frac{2p\beta}{(p(\alpha-\beta)+\beta)^2} & \frac{p(\alpha+\beta)-\beta}{(p(\alpha-\beta)+\beta)^2} \\ \frac{2\alpha(\alpha-\beta)}{(p(\alpha-\beta)+\beta)^2} & \frac{p(\alpha+\beta)-\beta}{(p(\alpha-\beta)+\beta)^2} & -\frac{2(p-1)\alpha}{(p(\alpha-\beta)+\beta)^2} \end{pmatrix} \\
K_{2,\cdot,\cdot} &= \begin{pmatrix} \frac{2\beta(\beta-\alpha)}{(p(\alpha-\beta)+\beta)^2} & -\frac{2p\beta}{(p(\alpha-\beta)+\beta)^2} & \frac{p(\alpha+\beta)-\beta}{(p(\alpha-\beta)+\beta)^2} \\ -\frac{2p\beta}{(p(\alpha-\beta)+\beta)^2} & 2\left(\frac{p^3}{(p(\alpha-\beta)+\beta)^2}+\frac{1}{\alpha^2}\right) & -\frac{2(p-1)p^2}{(p(\alpha-\beta)+\beta)^2} \\ \frac{p(\alpha+\beta)-\beta}{(p(\alpha-\beta)+\beta)^2} & -\frac{2(p-1)p^2}{(p(\alpha-\beta)+\beta)^2} & \frac{2(p-1)^2p}{(p(\alpha-\beta)+\beta)^2} \end{pmatrix} \\
K_{3,\cdot,\cdot} &= \begin{pmatrix} \frac{2\alpha(\alpha-\beta)}{(p(\alpha-\beta)+\beta)^2} & \frac{p(\alpha+\beta)-\beta}{(p(\alpha-\beta)+\beta)^2} & -\frac{2(p-1)\alpha}{(p(\alpha-\beta)+\beta)^2} \\ \frac{p(\alpha+\beta)-\beta}{(p(\alpha-\beta)+\beta)^2} & -\frac{2(p-1)p^2}{(p(\alpha-\beta)+\beta)^2} & \frac{2(p-1)^2p}{(p(\alpha-\beta)+\beta)^2} \\ -\frac{2(p-1)\alpha}{(p(\alpha-\beta)+\beta)^2} & \frac{2(p-1)^2p}{(p(\alpha-\beta)+\beta)^2} & \frac{2}{\beta^2}-\frac{2(p-1)^3}{(p(\alpha-\beta)+\beta)^2} \end{pmatrix}
\end{aligned} \tag{B.7}
$$

and

$$
\begin{aligned}
J_{1;\cdot,\cdot} &= \begin{pmatrix} \frac{(\beta-\alpha)^3}{(p(\alpha-\beta)+\beta)^2} & \frac{(\alpha-\beta)\beta}{(p(\alpha-\beta)+\beta)^2} & \frac{\alpha(\beta-\alpha)}{(p(\alpha-\beta)+\beta)^2} \\ \frac{(\alpha-\beta)\beta}{(p(\alpha-\beta)+\beta)^2} & \frac{p^2(\beta-\alpha)}{(p(\alpha-\beta)+\beta)^2} & \frac{(p-1)p(\alpha-\beta)}{(p(\alpha-\beta)+\beta)^2} \\ \frac{\alpha(\beta-\alpha)}{(p(\alpha-\beta)+\beta)^2} & \frac{(p-1)p(\alpha-\beta)}{(p(\alpha-\beta)+\beta)^2} & -\frac{(p-1)^2(\alpha-\beta)}{(p(\alpha-\beta)+\beta)^2} \end{pmatrix} \\
J_{2;\cdot,\cdot} &= \begin{pmatrix} -\frac{p(\alpha-\beta)^2}{(p(\alpha-\beta)+\beta)^2} & \frac{p\beta}{(p(\alpha-\beta)+\beta)^2} & -\frac{p\alpha}{(p(\alpha-\beta)+\beta)^2} \\ \frac{p\beta}{(p(\alpha-\beta)+\beta)^2} & -\frac{p^3}{(p(\alpha-\beta)+\beta)^2}-\frac{1}{\alpha^2} & \frac{(p-1)p^2}{(p(\alpha-\beta)+\beta)^2} \\ -\frac{p\alpha}{(p(\alpha-\beta)+\beta)^2} & \frac{(p-1)p^2}{(p(\alpha-\beta)+\beta)^2} & -\frac{(p-1)^2p}{(p(\alpha-\beta)+\beta)^2} \end{pmatrix} \\
J_{3;\cdot,\cdot} &= \begin{pmatrix} \frac{(p-1)(\alpha-\beta)^2}{(p(\alpha-\beta)+\beta)^2} & -\frac{(p-1)\beta}{(p(\alpha-\beta)+\beta)^2} & \frac{(p-1)\alpha}{(p(\alpha-\beta)+\beta)^2} \\ -\frac{(p-1)\beta}{(p(\alpha-\beta)+\beta)^2} & \frac{(p-1)p^2}{(p(\alpha-\beta)+\beta)^2} & -\frac{(p-1)^2p}{(p(\alpha-\beta)+\beta)^2} \\ \frac{(p-1)\alpha}{(p(\alpha-\beta)+\beta)^2} & -\frac{(p-1)^2p}{(p(\alpha-\beta)+\beta)^2} & \frac{(p-1)^3}{(p(\alpha-\beta)+\beta)^2}-\frac{1}{\beta^2} \end{pmatrix}.
\end{aligned} \tag{B.8}
$$

Listing B.3: Mathematica code to find the fisher information matrix, as well as two higher order cumulants.

```
(* Define assumptions on variables *)
$Assumptions = 0<β<γ<α && 0<=p<=1 && x>0 && y>0 && z>0;
(* Define the log-likelihood *)
L = (α^x Exp[-α])/Factorial[x]
    * (β^y Exp[-β])/Factorial[y]
    * ((p α+(1-p)β)^z Exp[-(p α+(1-p)β)])/Factorial[z];
LL = Log[L] // FullSimplify;
```

```
(* Verify that L is normalized as a sanity check *)
Sum[L,{x,0,∞},{y,0,∞},{z,0,∞}] // Simplify

(* Compute Fisher matrix and its inverse *)
Ifisher = -Sum[Evaluate[Simplify[
            Outer[D[LL,#1,#2]&,{p,α,β},{p,α,β}]L
        ]],
        {x,0,∞}, {y,0,∞}, {z,0,∞}
    ] // FullSimplify;
Iinv = Ifisher // Inverse // FullSimplify;

(* Compute K and J (takes a while, especially J) *)
K = With[
        {summand=FullSimplify[L*Outer[
            D[LL,#1,#2,#3]&,
            {p,α,β},
            {p,α,β},
            {p,α,β}]]},
        Sum[summand, {x,0,∞}, {y,0,∞}, {z,0,∞}]
    ] // FullSimp\]lify;
J = With[
        {summand=L*FullSimplify[Outer[
            D[LL,#2,#3]*D[LL,#1]&,
            {p,α,β},
            {p,α,β},
            {p,α,β}]]},
        Sum[summand, {x,0,∞}, {y,0,∞}, {z,0,∞}]
    ] // FullSimplify;
```

B.3 Maximum Likelihood Estimator

B.3.1 Derivation

The log-likelihood of the basic conditional model is given by

$$\begin{aligned}\log L = (x\log\alpha - \alpha - \log x!) + (y\log\beta - \beta - \log y!) \\ + (z\log(p\alpha + (1-p)\beta) - (p\alpha + (1-p)\beta) - \log z!).\end{aligned} \tag{B.9}$$

Having fixed a some particular values of x, y, and z, the goal is to maximize the function $\log L(p,\alpha,\beta)$. The value which maximizes this function is then the MLE. The easiest

method is to consider the equivalent problem maximizing the function

$$(x \log \alpha - \alpha - \log x!) + (y \log \beta - \beta - \log y!) + (z \log \gamma - (p\alpha + (1-p)\beta) - \log z! \quad \text{(B.10)}$$

subject to the constraint $\gamma = p\alpha + (1-p)\beta$. Using the Lagrange multiplier λ, this is encoded in a Lagrangian as

$$\begin{aligned}\Phi = &[(x \log \alpha - \alpha - \log x!) + (y \log \beta - \beta - \log y!) + (z \log \gamma - (p\alpha + (1-p)\beta) - \log z!)] \\ &- \lambda[\gamma - (p\alpha + (1-p)\beta)] \end{aligned} \quad \text{(B.11)}$$

yielding a simple set of equations

$$\{\frac{\partial \Phi}{\partial \alpha} = 0, \frac{\partial \Phi}{\partial \beta} = 0, \frac{\partial \Phi}{\partial \gamma} = 0, \frac{\partial \Phi}{\partial p} = 0, \frac{\partial \Phi}{\partial \lambda} = 0\} \quad \text{(B.12)}$$

with no more logarithms. These equations can be solved for α, β, and p as a function of x, y, and z. This was done in Mathematica 10.0.2.0 using the snippet found in Listing B.4.

This calculation can also be done directly by taking partial derivatives of $\log L$ and setting them to zero, although the algebra is significantly more demanding, and the order in which the equations are solved for affects the difficulty of subsequent steps.

Listing B.4: Mathematica code to solve the Lagrange problem stated in Equation Equation B.11

```
(* Define assumptions on variables *)
$Assumptions = 0<β<γ<α&&0<=p<=1&&x>0&&y>0&& z>0&&λ∈Reals;
(* Define Lagrangian *)

Φ = Log[\]α]/Factorial[x]
      *β^y Exp[-β]/Factorial[y]
      *(p α+(1-p)β)^z Exp[-γ]/Factorial[z]
   ] - λ (γ-(p α+(1-p)β));
(* Take partial derivatives to get system of equations *)
lagrangeEquations = D[Φ,#]==0 & /@ {α,β,γ,λ,p}
   // FullSimplify;
(* Remove λ from the set of equations*)
lagrangeEquations = Rest[lagrangeEquations]
   /. First@Solve[First@lagrangeEquations,λ] // FullSimplify;
(* Remove γ from the set of equations*)
lagrangeEquations = lagrangeEquations[[{1,2,4}]]
   /. First@Solve[lagrangeEquations[[3]],γ];
(* Solve for the remaning variables to conclude *)
soln = Solve[lagrangeEquations, {α, β, p}] // FullSimplify
```

B.4 Details of Effective Strong Measurements

Given a quantum state ρ, information is accessed through the Born's probability $p = \mathrm{Tr}(|0\rangle\langle 0|\,\rho)$. In the hypothetical case of strong measurement, in the language of statistics, we would be able to draw from the Bernoulli distribution $\mathrm{Bernoulli}\,(p)$, or more generally, with n repeated preparations and strong measurements, from the binomial distribution $\mathrm{Binom}\,(n, p)$.

Standard room temperature NV setups do not allow strong measurements. Instead, access to the quantity p is obstructed by three Poisson rates, such that conditional on some values $0 < \beta < \alpha$, we can draw from the random variables

$$\begin{aligned} X|\alpha &\sim \mathrm{Poisson}\,(\alpha) \\ Y|\beta &\sim \mathrm{Poisson}\,(\beta) \\ Z|\alpha,\beta,p &\sim \mathrm{Poisson}\,(\alpha + (1-p)\beta)\,. \end{aligned} \tag{B.13}$$

The quantities α and β are known as the bright reference and the dark reference, respectively. They are defined as the expected number of photons collected (and summed over N repetitions of the experiment) using the initial NV states $|0\rangle\langle 0|$ and $|1\rangle\langle 1|$, respectively[1].

The information content about p of this referenced Poisson model is not immediately obvious, and depends both on the magnitude of $\alpha + \beta$, as well as the contrast between α and β. This is different than the strong measurement case mentioned above, where n strong measurements has a clear intuitive and operational interpretation. The goal of this section is to reduce information about the references α and β into a single number with the same interpretation as n. This will allow one, for example, to quantitatively compare two experimental setups or NVs and decide which one is better at providing information about p.

It has been shown[84] that the Fisher information matrix of this referenced Poisson model is given by

$$J(p,\alpha,\beta) = \begin{pmatrix} \frac{(\alpha-\beta)^2}{p(\alpha-\beta)+\beta} & \frac{p(\alpha-\beta)}{p(\alpha-\beta)+\beta} & \frac{\alpha}{\beta+\alpha p-\beta p} - 1 \\ \frac{p(\alpha-\beta)}{p(\alpha-\beta)+\beta} & \frac{p^2}{p\alpha-p\beta+\beta} + \frac{1}{\alpha} & -\frac{(p-1)p}{p(\alpha-\beta)+\beta} \\ \frac{\alpha}{p\alpha-p\beta+\beta} - 1 & -\frac{(p-1)p}{p(\alpha-\beta)+\beta} & \frac{p\alpha+(p-2)(p-1)\beta}{\beta(p(\alpha-\beta)+\beta)} \end{pmatrix}, \tag{B.14}$$

[1] They are more accurately defined in terms of the pseudo-pure states that are actually created in the NV initialization procedure [84].

with inverse matrix

$$J(p,\alpha,\beta)^{-1} = \begin{pmatrix} \frac{p(p+1)\alpha+(p-2)(p-1)\beta}{(\alpha-\beta)^2} & \frac{p\alpha}{\beta-\alpha} & \frac{(p-1)\beta}{\alpha-\beta} \\ \frac{p\alpha}{\beta-\alpha} & \alpha & 0 \\ \frac{(p-1)\beta}{\alpha-\beta} & 0 & \beta \end{pmatrix}. \tag{B.15}$$

Using the Cramer-Rao bound, these matrices let us estimate the information content of p in the referenced Poisson model. Specifically, they give us an estimate in each of the following extreme cases. First, the (p,p) element of J^{-1}, $(J^{-1})_{p,p} = \frac{p(p+1)\alpha+(p-2)(p-1)\beta}{(\alpha-\beta)^2}$, is a lower bound on the variance of any (unbiased) estimate of p given that a single measurement of the triple (X,Y,Z) has been made, with no prior information about p, α, or β given. Second, the inverse of the (p,p) element of J, $(J_{p,p})^{-1} = \frac{p(\alpha-\beta)+\beta}{(\alpha-\beta)^2}$, is a lower bound on the variance of any (unbiased) estimate of p given that a single measurement of Z has been made, assuming perfect knowledge of both α and β.

It will be useful for us to also be able to interpolate between these two extremes, where some, but not all, prior information about α and β is available. There are a few tacks that one might consider to achieve this, including the Bayesian Cramer-Rao bound, or looking directly at the risk of some estimator. Instead, we choose a slightly ad-hoc method as it actually produces a tractable calculation—statistics of the referenced Poisson model usually involve a triple infinite sum, and many calculations are simply not possible without numerics. To this end, let σ_α^2 and σ_β^2 represent our prior variances of α and β, respectively, before taking a measurement of $Z|\alpha,\beta,p$. We can now ask the question: how many times, M, we must measure $X|\alpha$ and $Y|\beta$ to produce these variances in the first place? We must allow M to depend on α or β in each case. The distribution Poisson $(M(\lambda)\lambda)$ has Fisher information given by $\frac{(M(\lambda)+\lambda M'(\lambda))^2}{\lambda M(\lambda}$. Equating this to $1/\sigma^2$ and soliving the differential equation at $M(0)=0$ gives $M = \lambda/4\sigma^2$. Therefore consider the distribution

$$\text{Poisson}\left(\frac{\alpha^2}{4\sigma_\alpha^2}\right) \times \text{Poisson}\left(\frac{\beta^2}{4\sigma_\beta^2}\right) \times \text{Poisson}\left(p\alpha + (1-p)\beta\right) \tag{B.16}$$

which effectively results in our desired information about α and β. Solving for the (p,p) element of the inverse Fisher information matrix of this distribution results in the formula

$$K = \frac{\beta + p\left(\alpha - \beta + p\sigma_\alpha^2 + (p-2)\sigma_\beta^2\right) + \sigma_\beta^2}{(\alpha-\beta)^2}. \tag{B.17}$$

This formula correctly interpolates between the case of perfect prior information, and prior information as collected by a single sample of $(X, Y)|\alpha, \beta$, namely,

$$\lim_{\sigma_\alpha^2, \sigma_\beta^2 \to 0} K = (J_{p,p})^{-1} \qquad \text{and} \qquad \lim_{\sigma_\alpha^2 \to \alpha, \sigma_\beta^2 \to \beta} K = (J^{-1})_{p,p}. \tag{B.18}$$

The inverse Fisher information of the binomial model $\text{Binom}\,(n, p)$ is given by $\frac{p(1-p)}{n}$, which when integrated uniformly over $[0, 1]$, produces $\frac{1}{6n}$. Our definition for the number of effective strong measurements (ESM) of a referenced Poisson model with parameters $(\alpha, \beta, \sigma_\alpha, \sigma_\beta)$ is defined by equating $\int_0^1 K \mathrm{d}p = \frac{1}{6n}$ and solving for n, which results in

$$\text{ESM} = \frac{(\alpha - \beta)^2}{3(\alpha + \beta) + 2\left(\sigma_\alpha^2 + \sigma_\beta^2\right)}. \tag{B.19}$$

This shows, for example, that having perfect information about α and β before measuring $Z|\alpha, \beta, p$ is roughly equivalent—in terms of information learned about p—to $5/3 \approx 1.67$ times more effective strong measurements than the case where the triple $(X, Y, Z)|\alpha, \beta, p$ is measured, but with no prior information.

Finally, in Figure B.1, we use some numerics to show that the ESM quantity accurately relates the mean-squared error of the Bayes estimator for the referenced Poisson model and a binomial model with $n = \text{ESM}$.

B.5 Brute-force Numerical Evaluation of Bayes Risk

The methods in this section were developed in collaboration with Thomas Alexander. See Appendix A 'nv-adaptive' for a numerical implementation.

Evaluating the full Bayes risk for continuous outcome probability distributions is not possible analytically apart from special cases such as linear models with a normal likelihood function. For finite outcome probability distributions the problem is more tractable, however as the number of possible outcomes grows to be large, or even infinite—such as the Poisson distributions considered by us—the evaluation once again becomes intractable.

The difficulty of evaluation is a result of the expectation taken in Equation 4.7. For both infinite-discrete and continuous outcome probability distributions the expectation is intractable, however for finite discrete distributions, the expectation is a bounded discrete sum and straight-forward to evaluate numerically. We, therefore, aim to evaluate the Bayes risk by approximating the possible outcomes with a finite, discrete set of outcomes—note

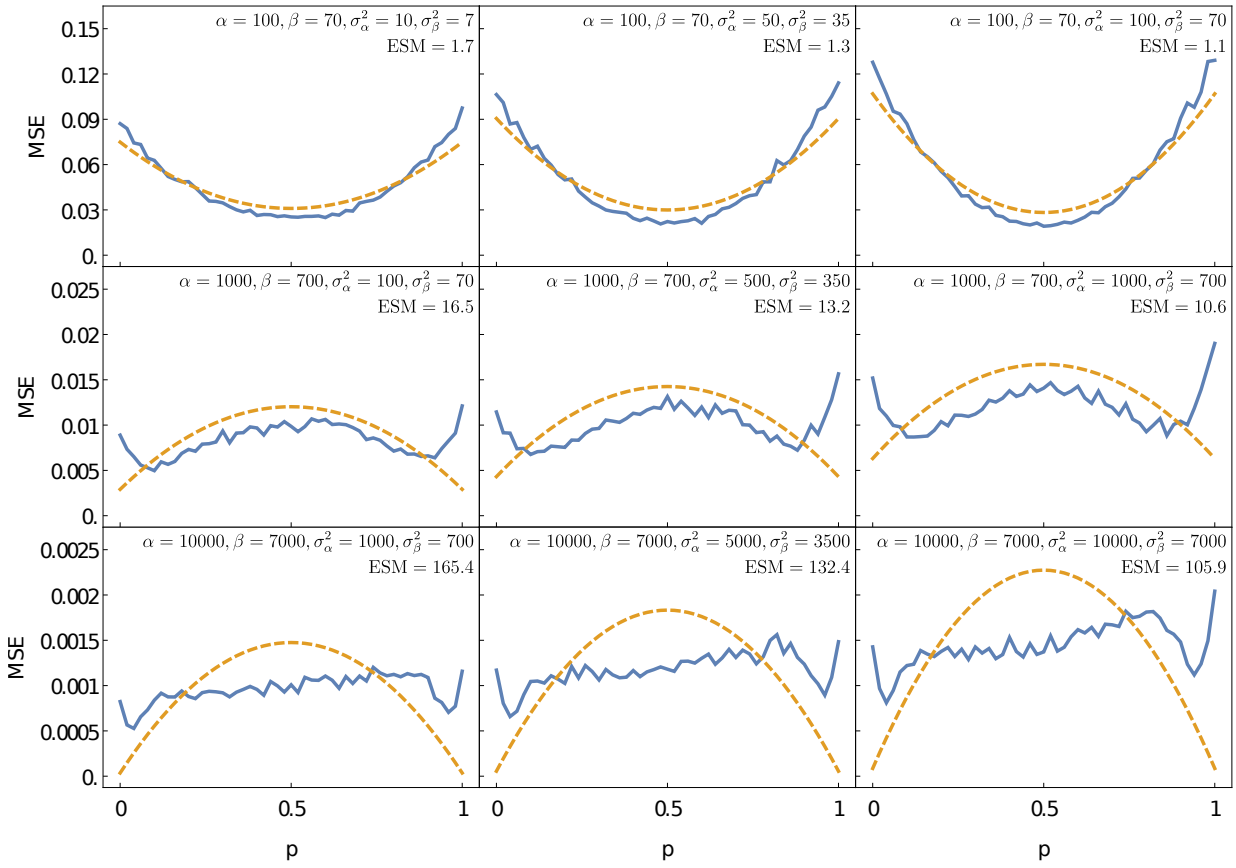

Figure B.1: The mean-squared-error of the Bayes estimator is computed as a function of p for both the referenced Poisson model (blue, solid) and for a binomial model (orange, dashed) where $n = \text{ESM}$. The prior distribution on p is uniform. This is done in nine regimes, corresponding to the nine subplots of the figure. Each row uses a different magnitude of bright reference, α, and each column uses a different amount of prior reference knowledge. The left column uses sub-Poisson error bars on α and β, and the right column uses regular Poisson error bars.

that this technique may also be used when the set of possible outcomes is finite but large enough to be computationally intractable. Typically, outcome domains are large with outcome probability mass concentrated to a small portion of the outcome domain. By fixing particle locations and sampling outcomes from these particles, we may evaluate the risk for only the outcomes that "matter" within the regions of outcome probability mass concentration.

We consider the case of evaluating the Bayes risk for the next experiment, e. We assume throughout this discussion, that this hypothetical experiment e was preceded by n experiments $e_{1:n}$ with corresponding data $d_{1:n}$. In several places, for brevity of notation, we will omit conditioning on this prior information, for example, we have $\Pr(d|e) = \Pr(d|e, d_{1:n}, e_{1:n})$. We begin by re-approximating the particle filter prior distribution with a uniformly weighted particle distribution by sampling K' particles from the prior π_n,

$$x_j \sim \pi_n(x), \tag{B.20}$$

which approximates the prior π_n as

$$\pi_n(x) \approx \frac{1}{K'} \sum_i^{K'} \delta\left(x - x_i\right). \tag{B.21}$$

For each particle we now sample a datum from the likelihood function,

$$d^{(j)} \sim \mathcal{L}\left(x_j; d, e\right) \qquad \forall\ x_j \in x_{1:K'}. \tag{B.22}$$

The set of sampled data is an approximation to the joint outcome, particle distribution

$$\Pr(d, x|e) \approx \frac{1}{K'} \sum_i^{K'} \delta\left(d - d^{(i)}\right) \delta\left(x - x_i\right). \tag{B.23}$$

The average utility—Equation 4.4—may be expanded in conjunction with Equation 4.3 as

$$\begin{aligned} U_n(e) &= \int\int \Pr(d|x, e)\tilde{\pi}_{n,d,e}(x) U_n(x, d, e) \mathrm{d}x\mathrm{d}d \\ &= \int\int \Pr(d, x|e) U_n(x, d, e) \mathrm{d}x\mathrm{d}d. \end{aligned} \tag{B.24}$$

The approximate particle, datum joint distribution, Equation B.23 may be substituted into Equation B.24 and the integrals are thus replaced by a sum,

$$U_n(e) \approx \frac{1}{K'} \sum_i^{K'} U_n(x, d, e), \tag{B.25}$$

which is the average utility of the joint sampled particle, datum distribution. When the utility is the negative mean-squared error the Bayes risk has the approximate form

$$r_{n,Q}(e) \approx \frac{1}{K'} \sum_{i}^{K'} \operatorname{Tr} Q(x_i - \hat{x}_{n,d^{(i)},e})^{\mathrm{T}}(x_i - \hat{x}_{n,d^{(i)},e}), \tag{B.26}$$

where $\hat{x}_{n,d^{(i)},e}$ is the posterior mean given the the approximate prior Equation B.21,

$$\hat{x}_{n,d^{(j)},e} = \sum_{i}^{K'} \frac{\mathcal{L}\left(x_i; d^{(j)}, e\right)}{K'} x_i. \tag{B.27}$$

The evaluation of the Bayes risk requires on the order of $O(K'^2)$ likelihood evaluations. However, typically a large number of outcome samples will be required to effectively sample the outcome domain of each particle and the total number of outcome samples will roughly be $O(Kn_d)$, where n_d is roughly the average number of outcome datum desired per particle. This may be prohibitively large when sampling is expensive.

Provided the outcome domain does not depend on the experiment—as is the case for our experiments—we may perform maximum importance sampling (MIS) to sample outcomes from an alternative distribution—chiefly the marginalized outcome distribution $\Pr(d|e)$—and properly re-weight the resultant utility functions [174]. The sampled outcome distribution $\Pr(d|e)$ is obtained from the sampled $\Pr(d, x|e)$ by neglecting the associated model parameter,

$$\Pr(d|e) \approx \frac{1}{K'} \sum_{i}^{K'} \delta\left(d - d^{(i)}\right). \tag{B.28}$$

The MIS utility is given as

$$\begin{aligned} U_n(e) &= \int \int \Pr(d|x, e)\pi_n(x) \frac{\Pr(d|e)}{\Pr(d|e)} U_n(x, d, e) \mathrm{d}x\mathrm{d}d \\ &\approx \frac{1}{K'} \sum_{i}^{K'} \sum_{j}^{K} \frac{\Pr(d^{(i)}|x_j, e)\omega_{n,j}}{\Pr(d^{(i)}|e)} U_n(x_j, d^{(i)}, e) \\ &= \frac{1}{K'} \sum_{i}^{K'} \sum_{j}^{K} \omega_{n+1|d^{(i)},j} U_n(x_j, d^{(i)}, e), \end{aligned} \tag{B.29}$$

where the judicious choice of the sampling distribution has allowed the utility to be written as the average of the posterior utility expectation over the marginalized outcome distribu-

tion. For the case of the Bayes risk this may be further simplified to

$$\begin{aligned} r_{n,Q}(e) &\approx \frac{1}{K'} \sum_{i}^{K'} \operatorname{Tr}\left[Q \operatorname{Cov}_{\tilde{\pi}}[x|d^{(i)}, e]\right] \\ &= \frac{1}{K'} \sum_{i}^{K'} \operatorname{Tr}\left[Q(\widehat{x^T x_i} - \hat{x}_i^T \hat{x}_i)\right], \end{aligned} \tag{B.30}$$

where $\widehat{x^T x_i} = \sum_{j}^{K} \omega_{n+1|d^{(i)},j} x_j^T x_j$, and $\hat{x}_i = \sum_{j}^{K} \omega_{n+1|d^{(i)},j} x_j$.

In general the the initial prior distribution may be down-sampled to some number of particles K, such that we now have two parameters that may be tuned, the number of outcome samples K', and the number of model parameter particles K. With the MIS Bayes risk, the number of likelihood function calls is now $O(KK')$, with only $O(K')$ outcome samples required. We utilize the MIS Bayes risk for experiment design within this chapter.

In practice a trade-off between accuracy and computational cost/time is necessary when selecting the number of outcomes and particle samples, K' and K respectively for the evaluation of the MIS Bayes risk. A comparison of various sampling numbers is displayed in the heatmaps of Figure B.2, which were evaluated with the wide prior of Equation 4.20 and experiments for the uniformly weighted Bayes risk experiment design heuristic given in Table 4.1. The aquistion of 4000 ESM takes roughly 10 seconds, as the full particle filter update of 30000 particles takes roughly 2 seconds, there are 8 seconds remaining in which to compute the Bayes risk and select the optimal experiment. We use $K' = 512$ outcome samples and $K = 1024$ particle filter samples, as this strikes a balance between accuracy while keeping the evaluation time below our threshold on our computational hardware. As this problem is massively parallel, if desired it is simple to use additional computational resources to refine to the evaluation accuracy.

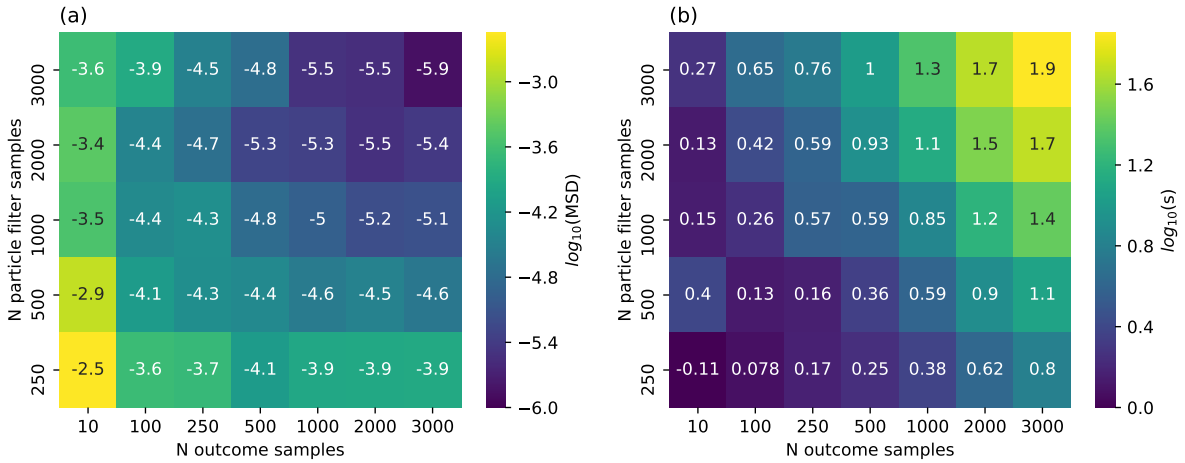

Figure B.2: Comparison of various outcomes and particle sampling accuracies and times when evaluating the MIS Bayes risk. The prior distribution over model parameters is given by Equation 4.20, and the experiments which the Bayes' risk is computed for is the uniformly weighted experiment design risk heuristic found in Table 4.1. (a) Log mean squared difference for all experiments computed with respect to a 4000 outcome, 4000 particle reference evaluation. (b) Log evaluation time(s) of Bayes risk over all experiments for a given number of outcome and particle samples. These calculations were done on an i9-7980XE CPU.

Appendix C

Supplementary Information for Bayesian RB+ Analysis

C.1 Sampling Strategies

C.1.1 Posterior Sampler

Analytic formulae for the posteriors of our models are intractable—we must instead choose a numerical inference algorithm to sample points from the posterior. A sufficient number of these points can be used to compute any quantity of interest related to the posterior. We used the Hybrid Monte Carlo (HMC) sampler using the No-U-Turns (NUTS) heuristic [42, 89]. This is a Markov chain Monte Carlo (MCMC) sampling strategy which has gained widespread use due to its lack of tuning parameters, fast mixing rate, and ability to handle large numbers of parameters. Specifically, we used the PyStan interface to the Stan library [22]. Probabilistic programs such as those written in Equation 6.13 and Equation 6.15 can entered nearly verbatim as input to this library (or other similar libraries), and samples from the posterior are returned. We suspect that sampling algorithms customized to our models could significantly outperform these generic tools, but it is hard to turn down the convenience of modern probabilistic programming languages and automatic differentiation.

It warrants mention why we have not used sequential Monte Carlo (SMC), which has emerged as a popular inference engine for quantum information processing tasks [63], including for RB and IRB with $N = 1$ [62]. Our main reason is that we wish to leave open the option of sampling from exact posterior distributions, especially while still in the proof-of-principle stage. SMC operates by storing the distribution over parameters as a weighted

mixture of delta functions. Data is entered sequentially and the prior is gradually transformed into the posterior with the inclusion of each subsequent individual datum. While SMC uses exact likelihood functions to sequentially update the weights with Bayes' rule, it also occasionally requires a resampling operation that moves the positions of the delta functions to where they are most needed. This resampling step usually only considers the first two moments of the distribution, and tends to distort the distribution toward being multivariate normal—see Appendix B of Reference [63]. Therefore, in SMC, posterior distributions are convolved with normal approximations to the true posterior distribution. However, SMC has an important advantage in that it can naturally be used with adaptive experiments, where the next experiment (sequence length, measurement type, etc.) is chosen based on the current state of knowledge. Also, SMC is often less computationally expensive and always highly parallelizable.

C.1.2 Reparameterizations

MCMC samplers benefit from using an optimal parameterization of the model — simply reparameterizing a model can make huge differences to the convergence, mixing rates, and stability. Ideally, posterior parameters are decorrelated, centered at the origin, and have a variance of order unity. Doing a perfect job at this would require knowing the posterior in advance of sampling from it, so we must instead rely on other heuristics.

For example, samplers have trouble near hard cutoffs, requiring special boundary specifications, and time can be wasted proposing random-walk values outside of the allowed region. This is relevant to our models where it is common to be inferring values that are physically restricted to the interval $[0, 1]$, and that are ideally very close to the boundary, such as the average gate fidelity of a gate-set. Modern Bayesian inference libraries, such as Stan, will automatically remove hard cutoffs by reparameterizing the model through a logit function for interval constraints, or through the logarithm for one-sided constraints. (It also multiplies the pdf by the change of variables Jacobian so that the prior is not distorted.)

We can do slightly better than this if we have prior expectations about some parameter values. For instance, if we expect a decay parameter p to be on the order of $p_0 = 0.9999$, then instead of using the sampling variable $\tilde{p} = \text{logit}(p)$ as would automatically be done by Stan, we can use the variable p_{err} where $p = \text{logit}(p_0 + p_{\text{err}})$. This is distinct from the role of a prior distribution in the sense that the correct distribution is sampled from even if we have set p_0 far from the value p that we are attempting to estimate; rather, we arrive at our samples less efficiently in that case. For reference, $\text{logit}(0.9999) \approx 9$, and we will have,

with very little effort, prevented the sampler from making an initial random walk from 0 to 9 while also keeping track of 200 other variables.

If we additionally have expectations about the standard deviation, say we expect $\delta p = p_0 \pm \delta p$ in the posterior, then we can use the changed variable p_{err} where $p = \text{logit}(p_0 + \delta\tilde{p} \cdot p_{\text{err}})$. If we let $\delta\tilde{p} = \delta p \cdot p_0^{-1}(1-p_0)^{-1}$, then $p_{\text{err}} = 0 \pm 1$ will translate to $p = p_0 \pm \delta p$. This trick does not affect the posterior in any way, it only improves sampling performance. We have had success using least-squared fits to estimate p_0 and/or δp (along with other parameters), using these values in the parameter transformation. If there is not enough data to meaningfully estimate δp with, for example, a weighted least squares fit, then $\delta\tilde{p} = 0.5$ is a fine choice.

The above heuristic should apply well to most probability parameters. There is a notable exception that comes up in low data regimes, which we will now illustrate in the case of standard RB for concreteness. Here, the tying function is $(A-B)p^M + B$, and for high quality devices, and at very low values of M, the survival probability is roughly equal to $A \lesssim 1$. Moreover, low values of M are exactly where we learn the most about A, allowing us to decorrelate its value from p and B. Suppose, however, that we are in the low data regime defined by $1/(1-A) \gg N \cdot I$, so that at the lowest values of M it's very likely that every single shot of the experiment will return 1. In this case any estimation technique will only be capable of producing a lower bound on the value of A; any value of A arbitrarily close to 1 will be consistent with the data. This is a problem for the logit rescaling discussed above because an estimate of A arbitrarily close to 1 implies a sampling parameter A_{err} that is arbitrarily large no matter the choices of A_0 and δA. There are a few potential paths forward. One is to switch sampling strategies to something like Riemannian Manifold HMC that fairs better with varying curvature in parameter space [12]. Another is to reparameterize in a different way, for example through an exponential distribution. Perhaps the easiest, however, is to recall that the lowest values of M are dubious in the case of gate dependent noise, and no data should be taken there anyway. We can just take a new definition of the initial state to be our old initial state acted on by a fixed number of random gates, effectively lowering the value of A.

C.2 Nonparametric Families

C.2.1 Dirichlet Processes

Dirichlet processes (DP) can be introduced in many ways. Given how they are used in the main body of this chapter, we will introduce them as a natural extension to beta

and Dirichlet priors as follows. Much more comprehensive introductions can be found elsewhere, for example, see this article of Teh [171].

First, consider of coin with an unknown bias p that we wish to infer. If we flip it N times and sum the resulting number of heads we get the random variable $X \sim \text{Binom}(N, p)$. In a Bayesian setting, we start by assigning a prior $\pi(p)$ to the unknown quantity p. Having collected the variate x of X, our posterior is proportional to $\Pr(p|x) \propto \int \binom{N}{x} p^x (1-p)^{N-x} d\pi_p(p)$. An important property the beta distribution is that when it is used as the prior in this example, say $\pi(p) = \text{Beta}(a, b)$ for some choices $a > 0$ and $b > 0$, then this integral has a nice closed form solution,

$$\Pr(p|x) = \text{Beta}(a + x, b + N - x). \tag{C.1}$$

This is one of the reasons the beta distribution family is the canonical family of distributions with support on $[0, 1]$. Moreover, from this formula, an operational interpretation of the prior parameters a and b is apparent: a can be thought of as the number of prior 'heads' observations, and b as the number of prior 'tails' observations. For example, $\pi(p) = \text{Beta}(1, 1)$ is asserting that one's prior knowledge of p is equivalent to having already flipped the coin twice, with each a heads and a tails landing once.

Let us generalize one step further before mentioning Dirichlet processes. Suppose we are interested in inferring the weights $p = (p_1, ..., p_K)$ of a K-sided die, where p is a finite probability distribution, so that $\sum_k p_k = 1$ and $p_k \geq 0$. (The coin example above is the case $K = 2$.) Rolling this die N times and binning the number of times each side lands face up results in the random variable $X = (X_1, ..., X_K) \sim \text{Multinomial}(N, p)$. The Dirichlet distribution family is the natural extension to the beta distribution family for $K > 2$. Namely, if we set the prior $\pi(p) = \text{Dir}(a_1, ..., a_K)$, then the posterior distribution is given by

$$\Pr(p|x) = \text{Dir}(a_1 + x_1, ..., a_K + x_K) \tag{C.2}$$

where $x = (x_1, ..., x_K)$ is the data. As before, this provides an operational interpretation of the prior parameters $a_1, ..., a_K$ — the value a_k can be interpreted as the number of prior observations of side k out a total of $\sum_k a_k$ prior observations.

Dirichlet processes can be thought of as the next logical step in this progression. We move from probability distributions with two sides, to K sides, and now to a continuum of sides; Dirichlet processes are natural priors for probability density functions. Suppose that f is an unknown probability density function on the sample space Ω that we wish to infer. Therefore $\int_\Omega f(x) d\mu(x) = 1$ where μ is some measure on Ω. Data is collected from this unknown distribution through the random variable $X \sim f$. We wish to set

our prior on f to be a Dirichlet process, which is a distribution of distributions on the sample space Ω. First, we need to define what a Dirichlet process is: given a distribution G_0 defined on Ω and a positive real number $\alpha > 0$, we say that the random distribution G is Dirichlet process distributed with base distribution G_0 and concentration parameter α, writing $G \sim \text{DP}(\alpha, G_0)$ to denote this, if for any finite disjoint measurable partition $\cup_{k=1}^{K} B_k = \Omega$, it holds that

$$(G(B_1), ..., G(B_K)) \sim \text{Dir}(\alpha G_0(B_1), ..., \alpha G_0(B_K)). \tag{C.3}$$

Note that for $B \subset \Omega$, all we mean by $G(B)$ is the probability of an event in B under distribution G. This means that $\alpha G_0(B)$ has the interpretation of being the number of prior observations in the region $B \subset \Omega$, and that to be Dirichlet process distributed means to be a distribution which obeys this condition for every possible partition of Ω into regions. In our previous example with the K-sided die, we could have reparameterized the Dirichlet prior as $\text{Dir}(\alpha, g)$ where $\alpha := \sum_k \alpha_k$ and $g := (\alpha_1/\alpha, ..., \alpha_K/\alpha)$ to be more notationally analogous to the present example.

If we let $\pi(f) = \text{DP}(\alpha, G_0)$ be the prior distribution of f, and suppose we make N iid measurements $X \sim f$, then the posterior is also Dirichlet process distributed, with

$$\Pr(f|x) = \text{DP}\left(\alpha + N, \frac{\alpha}{\alpha + N} G_0 + \frac{N}{\alpha + N} \frac{\sum_{k=1}^{N} \delta_{x_k}}{N}\right) \tag{C.4}$$

where δ_{x_k} is the delta distribution centered at the datum $x_k \in \Omega$. We see that the base distribution of the posterior of f is a mixture of the prior's base distribution and the empirical distribution of the data. It also makes it clear that α still has the interpretation as the total number of prior observations. Despite these nice interpretations, so far it might seem like Dirichlet processes are too abstractly defined to make them practical; at the end of this section, we will see that they have a alternate and procedural description which is not too hard to work with.

Dirichlet processes can be used as a generic stand-in for parametric priors in Bayesian models. For example, suppose we have samples $X_k|\lambda \sim \text{Pois}(T_k\lambda)$ for some rate of events λ measured for durations of time T_k. Moreover, suppose that there is not just one underlying rate of emission, but that there truly is a distribution of rates taking place, and we would like to infer what this distribution looks like. A parametric Bayesian approach might be through the model

$$\begin{aligned} X_k|\lambda &\sim \text{Pois}(T_k\lambda) \\ \lambda &\sim \text{Gam}(a, b) \\ a, b &\sim \pi(a, b) \end{aligned} \tag{C.5}$$

where $\pi(a, b)$ is some prior on a and b. We have parameterized the unknown distribution over rates with a gamma distribution with hyperparameters a and b; if we infer a and b, we can plot an estimate of the distribution of λ. Hovever, if we are unable to confidently assert that the distribution over rates λ must be gamma distributed, then we might choose a *nonparametric* Bayesian approach with the model

$$\begin{aligned} X_k|\lambda &\sim \text{Pois}(T_k\lambda) \\ \lambda &\sim G \\ G &\sim \text{DP}(\alpha, G_0). \end{aligned} \tag{C.6}$$

Just as we looked at the posterior of the parameters a and b above, here we can look at the posterior of G.

One caveat to random distributions drawn from $\text{DP}(\alpha, G_0)$ is that they are almost surely discrete in nature, even when G_0 is a continuous distribution. With probability one, $G \sim \text{DP}(\alpha, G_0)$ will be of the form

$$G(\cdot) = \sum_{k=1}^{\infty} p_k \delta_{\theta_k}(\cdot) \tag{C.7}$$

where the p_k are probabilities summing to unity and θ_k are members of Ω. However, this is not a big deal in practice for two reasons. The first is that any continous function can be approximated with arbitrary accuracy in L_1 distance using distributions of the form Equation C.7. Secondly, we always have the option of convolving G with some smooth distribution to end up with a smooth distribution. This is called a Dirichlet process *mixture model* [47].

Sethuraman found a way to construct instances of $\text{DP}(\alpha, G_0)$ in the form of Equation C.7 using a *stick breaking process* [156]. A random variate $G \sim \text{DP}(\alpha, G_0)$ can be construction as follows. The points θ_k are simply drawn from G_0 independently and identically. Their weights p_k, however, are derived from the following process. A stick of unit length is broken in two at the random location $V_1 \sim \text{Beta}(1, \alpha)$. The first piece is kept and its length is assigned to the first weight, $p_1 = V_1$. The remaining piece has length $1 - V_1$ and is broken again at a random fraction $V_2 \sim \text{Beta}(1, \alpha)$ of its length. The first piece is kept and its length is assigned to the second weight, $p_2 = V_2(1 - V_1)$. This process is repeated until the stick has been broken up a countably infinite number of times, giving

$p_k = V_k \prod_{l<k}(1 - V_l)$. We therefore have the representation

$$G = \sum_{k=1}^{\infty} \left[V_k \prod_{l=1}^{k-1} (1 - V_l) \right] \delta_{\theta_k}$$
$$V_k \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha), \quad \theta_k \overset{\text{iid}}{\sim} G_0 \tag{C.8}$$

which is equivalent to $G \sim \text{DP}(\alpha, G_0)$.

Finally, we remark that it is standard practice to assign a distribution to the parameter α, acknowledging one doesn't know *a priori* how good the base distribution G_0 is. We can see this in the stick breaking process, where low values of α lead to few important modes, and high values of α lead to many modes. In practice, Dirichlet processes are parameterized by their weights and locations, and the number of possible modes is truncated. One can verify that a certain truncation is sufficient by making sure the last weights (which must decrease in size) are negligibly small.

C.2.2 Constrained Dirichlet Process Beta Mixtures

We wish to modify the Dirichlet Process, defined in the previous section, so as to make it a suitable prior for survival distributions. In order for such a prior to work well with state-of-the-art MCMC samplers, which depend on gradients, we require a sample space of smooth distributions. This is easily done, as in the previous section, by convolving variates of the Dirichlet process with smooth distributions. It is natural for us to convolve with Beta distributions rather than the typically used normal distributions because survival distributions have support only within the interval $[0, 1]$.

The main difficulty of our construction lies in our second demand, which is the ability to constrain certain moments of these random distributions to specific values. We draw inspiration from Yang et al. [194] who propose a method to specify either or both of the first two moments of Dirichlet process variates. This method consists simply of shifting and scaling the delta locations θ_k (see Equation C.8) so that the mean and variance of G are as desired. We cannot use this approach directly because our domain is $[0, 1]$; for example, we might need to shift some of our locations θ_k to be outside of this interval to obtain the correct mean, which is not allowed. To overcome this, we use the logit function and its inverse to constrain and unconstrain variables between $\mathbb{R}$ and $(0, 1)$, as follows.

Using a sample space $\Omega = \mathbb{R} \times (0, 1)$, we begin by drawing a standard Dirichlet process distributed variate $\sum_{k=1}^{K} w_k \delta_{(\nu_k^*, r_k)}$. Here, r_k is the scaled variance parameter (Section C.3)

and ν_k^* represents an unconstrained beta mean. We then constrain each of these latter values to $(0, 1)$ by using the inverse logit function, $\nu_k = \text{logit}^{-1}(\nu_k^* + h) = 1/(1+\text{e}^{-\nu_k^* - h})$. The value of h is chosen as the unique real number which enforces the condition $\sum_{k=1}^{K} w_k \nu_k = \mu_1$. This in turn guarantees that $\mathbb{E}[G] = \mu_1$. There is no analytic formula for h, but it can be found efficiently with numerical optimization. In particular, Newton's method with an initial guess $h = \text{logit}\left(\mu_1 - \sum_{k=1}^{K} w_k \nu_k^*\right)$ has quadratic convergence. A code sample is shown in Listing C.1. This procedure produces a variate from what we call the mean-constrained Dirichlet process beta mixture mean-CDPBM distribution, which is summarized in Equation 6.14 of the main body.

Draws from the second-moment-constrained version, $\text{CDPBM}_K(\alpha, G_0, \mu_2)$, and the first-second-moment-constrained version, $\text{CDPBM}_K(\alpha, G_0, (\mu_1, \mu_2))$, are similar, except that a transform of the form $\text{logit}^{-1}(h_1 \nu^* + h_2)$ is necessary to constrain the variance as well as the mean.

If a protocol were to tie together the first two moments, following Equation 6.15, we would have the probabilistic program

$$\vec{x}_T \sim \pi(\vec{x}_T) \tag{C.9a}$$
$$\mu_{1,M,e} | \vec{x}_T = T(1, M, e, \vec{x}_T) \tag{C.9b}$$
$$\mu_{2,M,e} | \vec{x}_T = T(2, M, e, \vec{x}_T) \tag{C.9c}$$
$$\alpha_{M,e} \overset{\text{iid}}{\sim} \text{Gam}(1, 1) \tag{C.9d}$$
$$G_{M,e} | \alpha_{M,e}, \mu_{1,M,e}, \mu_{2,M,e} \overset{\text{ind}}{\sim} \text{CDPBM}_K\left(\alpha_{M,e}, G_0, (\mu_{1,M,e}, \mu_{2,M,e})\right) \tag{C.9e}$$

$$q_{M,e,i} | G_{M,e} \overset{\text{ind}}{\sim} G_{M,e} \tag{C.9f}$$
$$Q_{M,e,i} | q_{M,e,i} \overset{\text{ind}}{\sim} \text{Binom}(N, q_{m,i}). \tag{C.9g}$$

Listing C.1: Stan function (similar to C) to transform input weighted locations (ν_k^*) into output locations (ν_k) whose weighted mean is equal to μ.

```
// nu_star is a vector of input locations
// w is a length-K vector of weights
// mu is the desired mean value
vector compute_nu(vector nu_star, vector w, real mu) {
```

```
6
    real h;
    vector[size[w]] nu;

    // initial guess for h is exact when var(nu_star)=0
11  h = logit(mu) - dot_product(w, nu_star);

    // fixed descent of five steps
    for (newton_loops in 1:5) {
      nu = inv_logit(h + nu_star);
16    h = h - (dot_product(w, nu) - mu) /
          (dot_product(w, nu .* (1-nu)));
    }

    nu = inv_logit(h + nu_star);
21
    return nu;
  }
```

C.3 Beta Reparameterizations

In this section we provide some useful reparameterizations of the beta distribution, along with their inverses. As noted in Equation 6.11 of the main body, a beta distribution $\text{Beta}(\alpha, \beta)$ has a density function given by

$$\text{pdf}_{\text{Beta}}(q) = \frac{q^{\alpha-1}(1-q)^{\beta-1}}{\text{B}(\alpha, \beta)} \tag{C.10}$$

for any $q \in [0, 1]$. The normalization constant $\text{B}(\alpha, \beta)$ is the beta function, which is defined in terms of the gamma function, $\text{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. The parameters α and β must both be positive. Its mean and variance are given by

$$\mu = \frac{\alpha}{\alpha + \beta} \tag{C.11a}$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{C.11b}$$

respectively. The conditions $\alpha, \beta > 0$ are exactly equivalent to the conditions $0 < \mu < 1$ and $0 < \sigma^2 < \mu(1-\mu)$.

Parameters	Bounds	Transform	Inverse Transform	Variance
(μ, σ^2)	$\mu \in (0,1)$ $\sigma^2 \in (0, \mu(1-\mu))$	$\mu = \alpha/(\alpha+\beta)$ $\sigma^2 = \alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$	$\alpha = \mu^2(1-\mu)/\sigma^2 - \mu$ $\beta = \mu(1-\mu)^2/\sigma^2 - (1-\mu)$	σ^2
(μ, μ_2)	$\mu \in (0,1)$ $\mu_2 \in (\mu^2, \mu)$	$\mu = \alpha/(\alpha+\beta)$ $\mu_2 = \alpha(1+\alpha)/(\alpha+\beta)(1+\alpha+\beta)$	$\alpha = \mu(\mu-\mu_2)/(\mu_2-\mu^2)$ $\beta = (1-\mu)(\mu-\mu_2)/(\mu_2-\mu^2)$	$\sigma^2 = \mu_2 - \mu^2$
(μ, t)	$\mu \in (0,1)$ $t \in (0,1)$	$\mu = \alpha/(\alpha+\beta)$ $t = 1/(1+\alpha+\beta)$	$\alpha = \mu(1/t-1)$ $\beta = (1-\mu)(1-t)/t$	$\sigma^2 = t\mu(1-\mu)$
(μ, r)	$\mu \in (0,1)$ $r \in (0,1)$	$\mu = \alpha/(\alpha+\beta)$ $r = (\alpha+\beta)^2/(\alpha\beta(1+\alpha+\beta))$	$\alpha = 1/(r-r\mu) - \mu$ $\beta = 1/(r\mu) + \mu - 1$	$\sigma^2 = r\mu^2(1-\mu)^2$
(μ, s)	$\mu \in (0,1)$ $s \in (0,\infty)$	$\mu = \alpha/(\alpha+\beta)$ $s = \alpha+\beta$	$\alpha = s\mu$ $\beta = s(1-\mu)$	$\sigma^2 = \mu(1-\mu)/(s+1)$

Table C.1: Five reparameterizations of the beta distribution Beta(α, β). The first two, (μ, σ^2) and (μ, μ_2), simply reparameterize into mean and variance (or second moment), which yields non-rectangular bounds. The other four parameterizations have rectangular bounds. In the parameterization (μ, t), t represents the fraction of the maximum possible variance given a mean μ. Conversely, in the parameterization (μ, s), we have $t = 1/(s+1)$ so that large s corresponds to small variance. The parameterization (μ, r) does not allow the full range of varinace—the maximum possible variance (assuming $0 < r < 1$) is $\mu^2(1-\mu)^2$.

The parameters α and β have operational interpretations in terms of 'prior observations'; α is the number of prior observations of heads, and β is the number of prior observations of tails. Therefore, for example, the uniform prior Beta$(1,1)$ asserts that two prior observations have been made: one of heads, and one of tails. not entirely intuitive. In Table C.1, four reparameterizations along with their inverse transformations are given.

C.4 Priors on Heavily Biased Coins

In all of the examples in the main text we used an uninformative uniform prior on probability parameters such as A, B, and p in the standard RB protocol. There may be situations, especially in low data regimes, where incorporating prior knowledge has a noticeable effect on the posterior width, thereby reducing the necessary amount of data needed to attain a desired credibility lower bound.

For this purpose, we suggest a two-parameter family of distribution with support on $[0,1]$ which we call *probably at least* (PAL). A member of this family with parameters

$0 < p_0 < 1$ and $0 < z < p_0$ has a continuous density function given by

$$\mathrm{pdf}_{\mathrm{PAL}}(x) = \frac{1-z}{1-p_0} \begin{cases} \left(\frac{x}{p_0}\right)^{\frac{p_0-z}{z(1-p_0)}} & x < p_0 \\ 1 & x \geq p_0 \end{cases}. \tag{C.12}$$

This distribution is a sort of hedged version of Unif $([p_0, 1])$ that admits a finite but decreasing probability that $x < p_0$. Indeed, it is parameterized so that the probability $0 < x < p_0$ is equal to z. Observe that $p_0 = z$ gives Unif $([0, 1])$, and $z \to 0$ approaches Unif $([p_0, 1])$.

If the discontinuity of the derivative of this prior at $x = p_0$ poses a problem for the sampler at hand, this distribution can be smoothed over as follows:

$$\begin{aligned} g(x) &= \frac{p_0^2(2-z) + 2xz + p_0(2x+z)}{zp_0(1-p_0)} \\ \mathrm{pdf}_{\mathrm{PAL}'}(x) &= \frac{1-z}{1-p_0} \begin{cases} g(x)\left(\frac{x}{p_0}\right)^{2\frac{p_0-z}{z(1-p_0)}} & x < p_0 \\ 1 & x \geq p_0 \end{cases}. \end{aligned} \tag{C.13}$$

The parameters p_0 and z have the same interpretations, but now the decaying piece moves smoothly into the constant piece, at the cost of a bit more complexity.

C.5 A Reparameterization of LRB

Consider SPAM configurations $e = (\lambda, i) \in \mathfrak{E}$ where E_λ is a measurement operator, and ρ_i is an initial state. LRB as described in [189] has a first moment tying function $T(1, M, e, \vec{x}_T)$ defined by

$$\begin{aligned} &\frac{1}{L_1 + L_2} \operatorname{Tr}\left[E_\lambda^\dagger \mathcal{E}\left(L_2 \frac{\mathbb{I}_1}{d_1} + L_1 \frac{\mathbb{I}_2}{d_2}\right)\right] \\ &\quad + \left(\frac{L_1}{L_1 + L_2} - p_i\right) \operatorname{Tr}\left[E_\lambda^\dagger \mathcal{E}\left(\frac{\mathbb{I}_1}{d_1} - \frac{I_2}{d_2}\right)\right] \lambda_1^M \\ &\quad + (1 - p_i) \operatorname{Tr}\left[E_\lambda^\dagger \mathcal{E}\left(\rho_i' - \frac{\mathbb{I}_1}{d_1}\right)\right] \lambda_2^M \end{aligned} \tag{C.14}$$

where $\mathcal{E}$ is the gate independent noise acting on $\mathcal{X}_1 \oplus \mathcal{X}_2$, with $\dim \mathcal{X}_k = d_k$ and $\mathbb{I}_k = \mathbb{I}_{\mathcal{X}_k}$ for $k = 1, 2$. Here, $p_i := \operatorname{Tr}[\mathbb{I}_1 \rho_i]$ and $\rho_i' := \mathbb{I}_1 \rho_i \mathbb{I}_1 / (1 - p_i)$. The protocol recommends choosing

$\lambda = 0, ..., d_1 - 1$ with $E_\lambda \approx |\lambda\rangle\langle\lambda|$ and $i = 0$ with $\rho_0 \approx |0\rangle\langle 0|$. Other quantities are defined as

$$L_1 = 1 - \operatorname{Tr} \mathcal{E}(\mathbb{I}_1/d_1)\mathbb{I}_1 \tag{C.15a}$$
$$L_2 = \operatorname{Tr} \mathcal{E}(\mathbb{I}_2/d_2)\mathbb{I}_1 \tag{C.15b}$$
$$\lambda_1 = 1 - L_1 - L_2 \tag{C.15c}$$
$$\lambda_2 = \mu_1(1 - L_1) = \frac{d_1\overline{F}(\mathcal{E}) - (1 - L_1)}{d_1 - 1} \tag{C.15d}$$

with $\overline{F}(\mathcal{E})$ the average gate fidelity of $\mathcal{E}$ averaged over states in $\mathcal{X}_1$. L_1, called the *leakage*, measures $\mathcal{E}$'s average loss of population from $\mathcal{X}_1$ into $\mathcal{X}_2$, and L_2, called the *seepage*, measures the reverse effect. Some easy bounds on these parameters include

$$L_1, L_2 \geq 0 \tag{C.16a}$$
$$L_1 + L_2 \leq 1. \tag{C.16b}$$

Wood and Gambetta suggest extracting the parameters of interest, $(L_1, L_2, \overline{F})$, as follows. First, the data are summed over λ and the sample mean is taken over sequences I and sequence repetitions N. Under this sum the third term of the tying function, $\sum_\lambda \operatorname{Tr}\left[E_\lambda^\dagger \mathcal{E}\left(\rho_0' - \frac{\mathbb{I}_1}{d_1}\right)\right] \approx 0$, approximately cancels out leaving a single exponential term of base λ_1. Fitting to this curve yields λ_1 and hence $L_1 + L_2$, and combining this with the constant offset of the curve, $\sum_\lambda \frac{1}{L_1+L_2} \operatorname{Tr}\left[E_\lambda^\dagger \mathcal{E}\left(L_2\frac{\mathbb{I}_1}{d_1} + L_1\frac{\mathbb{I}_2}{d_2}\right)\right] \approx \frac{L_2}{L_1+L_2}$, we can separate to get L_1 and L_2. Note that this protocol is not truly SPAM free because part of the inference relies on the constant term which contains SPAM parameters. Next we go back to the unsummed data, plug in our estimate of λ_1, and fit to λ_2 to deduce $\overline{F}$.

In our scheme, we are able to process the data all at once, instead of this two step fitting procedure. It is helpful to rewrite the tying function a bit to make it a bit more clear what all of the independent parameters are. We generalize the protocol to possibly use multiple initial states ρ_i, $i = 0, ..., d - 1$. Then if we define $A_\lambda = \operatorname{Tr}[E_\lambda^\dagger \mathcal{E}(\mathbb{I}_1/d_1)]$, $B_\lambda = \operatorname{Tr}[E_\lambda^\dagger \mathcal{E}(\mathbb{I}_2/d_2)]$, and $C_{i,\lambda} = \operatorname{Tr}[E_\lambda^\dagger \mathcal{E}(\rho_i)]$ the tying function is expressed as

$$\begin{aligned} T(1, M, (\lambda, i), \vec{x}_T) = {} & \frac{L_2 A_\lambda + L_1 B_\lambda}{L_1 + L_2} \\ & + \left(\frac{L_1}{L_1 + L_2} - p_i\right)(A_\lambda - B_\lambda)\lambda_1^M \\ & + (1 - p_i)(C_{i,\lambda} - A_\lambda)\lambda_2^M. \end{aligned} \tag{C.17}$$

There are two reasons that one might prefer to use an orthogonal basis of pure initial states with one measurement operator, rather than vice versa, as suggested in the LRB paper. The first is that it requires fewer nuisance parameters — both A_λ and B_λ depend on the measurement but not the initial state. The second is that the offset term $\frac{L_2 A_\lambda + L_1 B_\lambda}{L_1 + L_2}$ is exactly equal for all experiments (under the assumption of gate-independent noise), which means it can effectively be measured independently by including very long sequence lengths in the data collection.

In Section 6.7.4 of the main text, we used one measurement operator, $M = 0.99999\,|0\rangle\langle 0|$, and two initial states, $\rho_0 = 0.9999\,|0\rangle\langle 0|$ and $\rho_1 = 0.9995\,|1\rangle\langle 1|$. The two prior distributions used for tying parameters were

$$L_1, L_2, \cdot \sim \mathrm{Dir}(1, 1, 100) \tag{C.18a}$$
$$\mu_1, A_\lambda, B_\lambda, C_{i,\lambda} \sim \mathrm{Unif}([0, 1]). \tag{C.18b}$$

and

$$L_1, L_2, \cdot \sim \mathrm{Dir}(1, 1, 100) \tag{C.19a}$$
$$\mu_1, C_{i,\lambda} \sim \mathrm{Unif}([0, 1]) \tag{C.19b}$$
$$A_\lambda \sim \mathrm{Beta}(100, 100) \tag{C.19c}$$
$$B_\lambda \sim \mathrm{Beta}(1, 100) \tag{C.19d}$$

labeled 'Flat SPAM prior' and 'Tighter SPAM prior' in Figure 6.8, respectively. The Dirichlet distribution on L_1 and L_2 was chosen because of the additive constraint $L_1 + L_2 \leq 1$ means that the triple $(L_1, L_2, 1 - L_1 - L_2)$ is a probability vector. The variable $0 \leq \mu_1 \leq 1$ has the interpretation of the depolarizing parameter of $\mathcal{E}$ restricted to $\mathcal{X}_1$.