Statistical inference with quantum measurements: methodologies for nitrogen vacancy centers in diamond

Ian Hincks1,2, Christopher Granade3,4 and David G Cory2,5,6,7

1 Department of Applied Mathematics, University of Waterloo, Waterloo, ON, Canada
2 Institute for Quantum Computing, University of Waterloo, Waterloo, ON, Canada
3 School of Physics, University of Sydney, Sydney, NSW, Australia
4 Centre for Engineered Quantum Systems, University of Sydney, Sydney, NSW, Australia
5 Department of Chemistry, University of Waterloo, Waterloo, ON, Canada
6 Canadian Institute for Advanced Research, Toronto, ON, Canada
7 Perimeter Institute for Theoretical Physics, 31 Caroline St. N, Waterloo, Ontario, N2L 2Y5, Canada

E-mail: ihincks@uwaterloo.ca

Keywords: nitrogen vacancy centers, statistical inference, quantum measurements, quantum Hamiltonian learning

Abstract

The analysis of photon count data from the standard nitrogen vacancy (NV) measurement process is treated as a statistical inference problem. This has applications toward gaining better and more rigorous error bars for tasks such as parameter estimation (e.g. magnetometry), tomography, and randomized benchmarking. We start by providing a summary of the standard phenomenological model of the NV optical process in terms of Lindblad jump operators. This model is used to derive random variables describing emitted photons during measurement, to which finite visibility, dark counts, and imperfect state preparation are added. NV spin-state measurement is then stated as an abstract statistical inference problem consisting of an underlying biased coin obstructed by three Poisson rates. Relevant frequentist and Bayesian estimators are provided, discussed, and quantitatively compared. We show numerically that the risk of the maximum likelihood estimator is well approximated by the Cramér–Rao bound, for which we provide a simple formula. Of the estimators, we in particular promote the Bayes estimator, owing to its slightly better risk performance, and straightforward error propagation into more complex experiments. This is illustrated on experimental data, where quantum Hamiltonian learning is performed and cross-validated in a fully Bayesian setting, and compared to a more traditional weighted least squares fit.

1. Introduction

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 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 [1–4].

These considerations imply that data from quantum experiments should be analyzed on firm statistical footing. This, in particular, 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 [5–7], randomized benchmarking (RB) and derivative protocols are inherently statistical [8–11], and Hamiltonian parameter learning is often considered from a machine learning perspective [12–15]. 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-platform generality.

Specific platforms, however, really do need to think carefully about how to append their own messy details onto these well-established models. Incorrect models may lead to estimates biased in unknown ways, or, just as bad, to error-bars which are inaccurate given the actual data. This leads us to the purpose of the present paper: a detailed and explicit model of the measurement process of the negatively charged nitrogen vacancy (NV−) center in diamond, along with an analysis of relevant estimators and methodologies.

NV− centers have been studied extensively due to a number of remarkable physical properties [16]. These include long coherence times at room temperature [17], the ability to address and readout a single defect in isolation [18, 19], the ability to initialize to a (nearly) pure state on demand [20], and the ability to selectively interact with nearby nuclei [21–26]. 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 [27–29].

As such, it is of value to carefully examine precisely what one learns from the NV− center when a measurement is made. If the application is quantum sensing, then it is crucial to be able to accurately report error bars of the quantity of interest. Or, if an application requires high-fidelity control, it is important to be able to quantify one’s knowledge of the system parameters so that control sequences can be designed to be robust against uncertainty. If these same control sequences are being tested in a tomography or benchmarking protocol, then sensible estimates of figures of merit also presuppose a thorough understanding of the measurement data.

These demands all reduce to statistical inference, which is the process of inferring values of interest from a model, given a dataset of measurements. However, in order to trust an inference one must first be able to trust the sufficiency of the model. Therefore, we begin by taking the accepted physical description of the NV− measurement process and use it to derive a statistical model describing the expected distributions of experimental measurements, including annoyances such as limited visibility, dark counts, imperfect state preparation, and reference drift.

This paper is organized as follows. Section 2 introduces a phenomenological model of the optical dynamics of an NV− center in terms of Lindblad jump operators. Section 3 provides a model for the measurement process, including photon loss and dark counts. Section 4 describes the state initialization process and derives the pseudo-pure state and its contrast-reduced references. Section 5 introduces a stochastic model of the reference drift process. Section 6 uses the derived description of the measurement process to state it as a statistical inference problem—readers who are less interested in the physical derivation of the model may choose to begin with this section; though it is the culmination of the preceding sections, it has been written to be comprehensible without them. Section 7 provides a couple of estimators for this inference problem, and compares them. Section 8 gives an example of the model being used with experimental data for quantum Hamiltonian learning (QHL) [12]. All raw data and code necessary to reproduce the results of this paper are openly available online [30].

2. Optical dynamics of the NV− center in diamond

Here we briefly summarize the optical dynamics of the NV− center in diamond, setting up our notation for future sections. More complete descriptions can be found elsewhere, see, for example, the review article by Doherty et al [31] and references therein.

The 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 spacial wavefunction extends on the order of a dozen lattice sites before becoming negligible compared to nuclear dipolar interactions [26, 32, 33]. These six electrons form an effective spin-1 electron in the electronic ground state [16]. It is standard practice to pick two out of these three spin energy levels to define a qubit.

Importantly, the NV− center also has optical properties; it can absorb and emit photons. The energy difference between the optical ground state and the first excited state is 637 nm (red). 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 [18]: 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 [31], but adding them to this discussion will not change our model of the measurement process.

We decompose our Hilbert space as the direct sum

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

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

$$|g, +1\rangle = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad |g, 0\rangle = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad |g, -1\rangle = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \quad \langle e, +1\rangle = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad \langle e, 0\rangle = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \quad \langle e, -1\rangle = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \quad |s\rangle = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix},$$

where $\{+1, 0, -1\}$ are spin labels corresponding to the eigenvalues of $S_z = \text{diag}(1, 0, -1)$, and $(g, e, s)$ 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 $S_z \oplus 0 \oplus 0 = |g, +1\rangle \langle g, +1| - |g, -1\rangle \langle g, -1|.$

Given an external magnetic field applied along the z-axis with strength $\omega_0$ (in angular frequency units), the Hamiltonian of the system is given by

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

with

$$H_{\text{ground}} = \Delta_z S_z^2 + \omega_0 S_z,$$

$$H_{\text{excited}} = \Delta_e S_e^2 + \omega_0 S_z,$$

The terms $\Delta_z \approx 2.87\text{GHz}$ and $\Delta_e \approx 1.4\text{GHz}$ 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,

$$L_1 = \sqrt{\gamma_0} (|g, +1\rangle \langle e, +1| + |g, 0\rangle \langle e, 0| + |g, -1\rangle \langle e, -1|),$$

$$L_2 = \sqrt{k \cdot \gamma_0} (|e, +1\rangle \langle g, +1| + |e, 0\rangle \langle g, 0| + |e, -1\rangle \langle g, -1|),$$

where $\gamma_0$ 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 $1/\gamma_0 = 13\text{ns}$. The dimensionless parameter $k$ corresponds to the laser power 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 600–800 nm emission spectrum. This route preferentially allows the excited $|e, -1\rangle$ and $|e, +1\rangle$ states to decay to the ground states through the ISC to the state $|s\rangle$. It can be modeled using the Lindblad dissipation terms

$$L_3 = \sqrt{\gamma_0/2} |s\rangle \langle e, +1|,$$

$$L_4 = \sqrt{\gamma_0/2} |s\rangle \langle e, -1|,$$

$$L_5 = \sqrt{\gamma_0/3} |g, +1\rangle \langle s|,$$
The first two move support on the excited $|e, -1\rangle$ and $|e, +1\rangle$ 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 $|e, +1\rangle$, $|e, -1\rangle$ to $|s\rangle$ is what actually drives polarization, as measured by Robledo et al [34]. The rate of the spin-selective decay is roughly $\gamma_{01} = 30$ MHz, and comparing this to $\gamma_{es}$ shows that the excited $\pm 1$ states take the ISC path roughly 1/3rd of the time. The lifetime of the singlet $|s\rangle$ is quite long, with a decay rate of roughly $\gamma_{ge} = 3$ 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 $\gamma_{01} = 1$ MHz [31]. They can be modelled as the Lindblad operators

$$L_8 = \sqrt{\gamma_{ge}/3} \langle g, 0 | s \rangle,$$

$$L_9 = \sqrt{\gamma_{ge}/3} \langle g, -1 | s \rangle.$$

Solving the Lindblad master equation,

$$\frac{d \rho(t)}{dt} = -i[H, \rho(t)] + \sum_{i=1}^{11} (L_i \rho(t) L_i^\dagger - (\rho(t) L_i^\dagger L_i + L_i L_i^\dagger \rho(t))/2),$$

with some initial state $\rho(0)$ allows us to track the populations and coherences of the quantum system through time. 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 3.2 would not apply.

3. 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 span$|g, -1\rangle$, $|g, +1\rangle$, 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.

3.1. 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 [35]:

$$Pr(\text{photon}; [t, t + dt]) = \gamma_{01} \text{Tr}[P_e \cdot \rho(t)] dt,$$

where $P_e = |e, +1\rangle \langle e, +1 | + |e, 0 \rangle \langle e, 0 | + |e, -1\rangle \langle e, -1 |$ is the projector onto the excited subspace $\mathcal{H}_{\text{exc}}$ and $\rho(t) \in D(\mathcal{H})$ 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

$$\mu(t) = \gamma_{01} \text{Tr}[P_e \cdot \rho(t)].$$

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

$$Pr(n_e|\rho(0)) = \left(\int_0^t \mu(t) dt\right)^{n_e} e^{-\int_0^t \mu(t) dt},$$
where \( n_t \) is the number of emitted photons during the interval \([0, \Delta t]\). The expected number of emitted photons is then

\[
\mathbb{E}[n_t | \rho(0)] = \int_0^{\Delta t} \mu(t) \, dt.
\]

(17)

With our typical parameters, the expected number of emitted photons is on the order of a dozen. Optimal measurement times \( \Delta t \) will be discussed in section 6.6.

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 \( 0 \leq t \leq \Delta t \) 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.

### 3.2. 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

\[
\begin{align*}
\rho_{gg}(t) &= \text{Tr}[|g\rangle \langle g| \rho(t)], \\
\rho_{ge}(t) &= \text{Tr}[|g\rangle \langle e| \rho(t)], \\
\rho_{eg}(t) &= \text{Tr}[|e\rangle \langle g| \rho(t)], \\
\rho_{ee}(t) &= \text{Tr}[|e\rangle \langle e| \rho(t)], \\
\rho_s(t) &= \text{Tr}[|s\rangle \langle s| \rho(t)],
\end{align*}
\]

(18)

as well as the vector \( \vec{\rho}(t) = (\rho_{gg}(t), \rho_{ge}(t), \rho_{eg}(t), \rho_{ee}(t), \rho_s(t))^T \in [0, 1]^5 \). Here, \( \rho(t) \) is the solution to the Lindblad master equation (13). Notice that the components of this vector sum to unity, \( \sum_{i=1}^{5} \rho_i(t) = 1 \) because the projectors used in the definitions of its components sum to the identity.

Combining the master equation with the definitions from equation (18), we can compute the time evolution of each component of \( \vec{\rho}(t) \). For example, we have

\[
\dot{\rho}_{gg}(t) = \text{Tr}[|g\rangle \langle g| \rho(t)]
\]

\[
= \text{Tr}\left[ |g\rangle \langle g| \left( -i[H, \rho(t)] + \sum_k (L_k \rho(t)L_k^\dagger - (\rho(t)L_k^\dagger L_k + L_k^\dagger L_k \rho(t))/2) \right) \right]
\]

\[
= -k_{eg}\rho_{ge}(t) + \gamma_{eg}\rho_{eg}(t) + \frac{\gamma_{eg}}{2} \rho_s(t),
\]

(19)

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 \( DE \)

\[
\begin{bmatrix}
\dot{\rho}_{gg}(t) \\
\dot{\rho}_{ge}(t) \\
\dot{\rho}_{eg}(t) \\
\dot{\rho}_{ee}(t) \\
\dot{\rho}_s(t)
\end{bmatrix} =
\begin{bmatrix}
-k_{eg} & 0 & \gamma_{eg} & \gamma_{01}/2 & \gamma_{eg}/3 \\
0 & -k_{eg} & \gamma_{01}/2 & \gamma_{eg} & 2\gamma_{eg}/3 \\
k_{eg} & 0 & -\gamma_{eg} - \gamma_{01}/2 & 0 & 0 \\
0 & k_{eg} & 0 & -\gamma_{eg} - \gamma_{01}/2 & 0 \\
0 & 0 & 0 & \gamma_{eg} & -\gamma_{eg}
\end{bmatrix}
\begin{bmatrix}
\rho_{gg}(t) \\
\rho_{ge}(t) \\
\rho_{eg}(t) \\
\rho_{ee}(t) \\
\rho_s(t)
\end{bmatrix}.
\]

(20)

which we write in condensed notation as

\[
\dot{\vec{\rho}}(t) = R \cdot \vec{\rho}(t).
\]

(21)

Notice that the columns of \( R \) sum to 0 which ensures that \( \vec{\rho}(t) \) 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 (21) is \( \vec{\rho}(t) = e^{tR}\vec{\rho}(0) \). Thus the inhomogeneous Poisson rate from equation (15),

\[
\mu(t) = \gamma_{ea} \text{Tr}[P_e \rho(t)],
\]

can be simplified to \( \mu(t) = \gamma_{ea}(P_{gg}(t) + P_s(t)) \), or in terms of the initial state,
\[ \mu(t) = \gamma_{\text{m}} \cdot e^{i\phi} \mathcal{P}(0), \]

with \[ \mathcal{P}(0) = (\text{Tr}[|g, 0\rangle \langle g, 0| \rho(0)], \text{Tr}[|g, -1\rangle \langle g, -1| + |g, +1\rangle \langle g, +1|] \rho(0))]_0, 0, 0, 0^T, \]

where \( \mathcal{P} = (0, 0, 0, 0)^T \) 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 \( 5 \times 5 \) matrix exponential, instead of a \( 49 \times 49 \) matrix exponential in superoperator space. It also makes it simpler to derive the following relationship:

\[ \mu(t|\rho(0)) = \gamma_{\text{m}} \cdot e^{i\phi} \mathcal{P}(0), \]

\[ = \gamma_{\text{m}} \cdot e^{i\phi} \text{Tr}[|g, 0\rangle \langle g, 0| \rho(0)], \text{Tr}[|g, -1\rangle \langle g, -1| + |g, +1\rangle \langle g, +1|] \rho(0))]_0, 0, 0, 0^T \]

\[ + \text{Tr}[|g, -1\rangle \langle g, -1| + |g, +1\rangle \langle g, +1|] \rho(0)] \gamma_{\text{m}} \cdot e^{i\phi}(1, 0, 0, 0, 0^T) \]

\[ = \text{Tr}[|g, 0\rangle \langle g, 0| \rho(0)] \mu(t, |g, 0\rangle \langle g, 0|) + \text{Tr}[|g, -1\rangle \langle g, -1| + |g, +1\rangle \langle g, +1|] \rho(0)] \mu(t, |g, +1\rangle \langle g, +1|) \]

\[ + p \cdot \mu(t, |g, 0\rangle \langle g, 0|) + (1 - p) \cdot \mu(t, |g, +1\rangle \langle g, +1|), \]

(23)

where \( p = \text{Tr}[|g, 0\rangle \langle g, 0| \rho(0)] \), again assuming \( \rho(0) \) has support only on \( \mathcal{H}_{\text{ground}} \). Note that the choice of using \(|g, +1\rangle\) rather than \(|g, -1\rangle\) (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 rate of photon emission rates of the states \(|g, 0\rangle\) and \(|g, +1\rangle\), where the convex combination parameter is the overlap of \( \rho(0) \) with \(|g, 0\rangle\). 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 (17) can be simplified to

\[ \mathbb{E}[n_d|\rho(0)] = p\mathbb{E}[n_{dc}|g, 0\rangle \langle g, 0|] + (1 - p)\mathbb{E}[n_{td}|g, +1\rangle \langle g, +1|] \]

(24)

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 \(|g, 0\rangle\) and \(|g, +1\rangle\). We already know that \( n_{dc} \) is always a Poisson distribution for any pre-measurement state (section 3.1), and that a Poisson distribution is characterized completely in terms of its expected value, therefore once

\[ \mu_0 = \mathbb{E}[n_{dc}|g, 0\rangle \langle g, 0|], \mu_1 = \mathbb{E}[n_{td}|g, +1\rangle \langle g, +1|], \text{and } p = \text{Tr}[|g, 0\rangle \langle g, 0| \rho(0)] \text{ are known, everything} \]

about \( \text{Pr}(n_d|\rho(0)) \) is also known:

\[ \text{Pr}(n_d|\rho(0)) = \frac{(p\mu_0 + (1 - p)\mu_1)^n}{n!} e^{-(p\mu_0 + (1 - p)\mu_1)} \]

(25)

which is a more tractable version of equation (16).

### 3.3. 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 3.1 we derived the probability of the NV emitting some number of photons during a measurement, \( \text{Pr}(n_d) \), and in section 3.2 we provided a simpler formula for the same quantity.

In this section we introduce two new variables, \( \Gamma \) and \( \eta \). Let \( \Gamma \) be the rate of dark counts per unit time, due both to noise in the photon counter itself and to stray photons. Similarly, let \( \eta \) 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 \( n_{td} \) 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

\[ n_d = n_{dc} + n_{td}, \]

(26)

where \( n_{dc} \) 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 \( \Gamma \) over the a measurement duration \( \Delta t \) we simply have

\[ n_{dc} \sim \text{Poisson}(\Gamma \Delta t). \]

(27)

The true detections are slightly more complicated. Each photon emitted by the NV has an (independent) probability \( \eta \) of arriving at the detector and being detected. We therefore have the conditional distribution
\[ n_{id}|n_e \sim \text{Binom}(n_e, \eta), \]  
\[ \text{Pr}(n_{id} = n) = \sum_{m=0}^{\infty} \text{Pr}(n_{id} = n|n_e = m)\Pr(n_e = m) \]
\[ = \sum_{m=0}^{\infty} \binom{m}{n} \eta^n(1 - \eta)^{m-n} \frac{\mu^m e^{-\mu}}{m!} \]
\[ = \frac{(\eta \mu)^n e^{-\eta \mu}}{n!}, \]
which shows that \( n_{id} \) 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 variable of interest, \( n_{id} = n_{id} + n_{dd} \), is Poisson distributed,

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

This, in combination with (24), gives

\[ n_{id} \sim \text{Poisson} \left( \Gamma \Delta t + \eta (p \mu_0 + (1 - p) \mu_1) \right) \]
\[ = \text{Poisson} \left( p(\Gamma \Delta t + \eta \mu_0) + (1 - p)(\Gamma \Delta t + \eta \mu_1) \right), \]

where \( \mu_0 = \mathbb{E}[n_e|\{ g, 0 \}] \), \( \mu_1 = \mathbb{E}[n_e|\{ g, +1 \}] \), and \( p = \text{Tr}[\{ g, 0 \} | \{ g, 0 \} \rho(0)] \).

Therefore, just as \( \mu_0 \) and \( \mu_1 \) 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 \( \mathbb{E}[n_e] \).

### 4. State initialization

Simply illuminating the 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.

#### 4.1. The steady state solution to the rate equation

The rate equation derived in section 3.2, while much simpler than the Lindblad master equation it was derived from, is still not analytically solvable unless certain terms like \( \gamma_0 \) 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 (21) has only non-positive eigenvalues, and must have a non-trivial null space. To see this, examine the solution to the rate equation, \( \tilde{p}(t) = e^{\Delta t \tilde{R}} \tilde{p}(0) \). If \( R \) had positive or complex eigenvalues, the probability vector \( \tilde{p}(t) \) 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, \( \tilde{p}(t) \) 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 \( \tilde{p}(0) \) and decompose this into an eigenbasis \( \tilde{v}_1, \ldots, \tilde{v}_s \) for \( R \), where it holds that \( \tilde{v}_1, \ldots, \tilde{v}_s \) all have eigenvalue zero, and \( \tilde{v}_{i+1}, \ldots, \tilde{v}_s \) have strictly negative eigenvalues \( \lambda_{i+1}, \ldots, \lambda_s \). This gives \( \tilde{p}(0) = \sum_{i=1}^{s} a_i \tilde{v}_i \) for some real coefficients \( a_1, \ldots, a_s \). The evolution is now described by

\[ \tilde{p}(t) = \sum_{i=1}^{s} a_i e^{\lambda_i t} \tilde{v}_i = \sum_{i=1}^{s} a_i e^{\lambda_i t} \tilde{v}_i + \sum_{i=i_0+1}^{s} a_i e^{\lambda_i t} \tilde{v}_i \rightarrow \sum_{i=i_0}^{s} a_i \tilde{v}_i, \]

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 $(1, 0, 0, 0, 0)^T$ and $(0, 1, 0, 0, 0)^T$, which correspond to the populations of the optical ground state—this is not surprising. The three remaining eigenvalues are $-(\gamma_s + \gamma_{01}/2)$, $-(\gamma_s + \gamma_s + \gamma_{01}/2)$, and $-\gamma_{eg}$. The first two die off quickly—these correspond to population escaping from the optically excited states. The last rate, $\gamma_{eg}$, is the most important, having an eigenvector $(1/3, 2/3, 0, 1)^T$, 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 $1/\gamma_{eg}$ before the population of $|s\rangle$ 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. Firstly, it can be shown that the null space is now only one dimensional. We denote the probability vector spanning the null subspace as $\vec{p}_n = (p_0, p_0, p_0, p_0, p_0)^T$. 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 equilibrate under optical illumination, it will reach a unique steady-state.

Under the approximation that $\gamma_{01} = 0$, we have $\vec{p}_n = (1/(1 + k), 0, k/(1 + k), 0, 0)^T$ 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 equilibrate 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 we plot the components of $\vec{p}_n$ as a function of $\gamma_{01}$. As $\gamma_{01}$ 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 $\gamma_{01}$, there is always some population to feed it. Indeed, expanding $\vec{p}_n$ in a $\gamma_{01}$ power series about 0, the steady state population of the singlet is $p_s = \frac{3k\gamma_{01}}{2(k + 1)\gamma_s} + O(\gamma_{01}^2)$.

### 4.2. Imperfect preparation: the pseudo-pure state

Due to small but important spin-non-conserving terms in the master equation during continuous optical excitation (equation (12)), the steady state density matrix must have non-zero support on the states $|g, -1\rangle$, $|g, +1\rangle, |e, -1\rangle$, and $|e, +1\rangle$. This can be seen in figure 2. This means that the initialization process does not asymptote to the pure state $|g, 0\rangle$, but rather to a mixed state $\rho_0$ 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 $\rho_0$ can be written as a pseudo-pure state [36],

$$\rho_0 = q|g, 0\rangle \langle g, 0| + (1 - q)\frac{I}{2},$$

(33)

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.
so that its lack of complete purity, in practice, serves only to limit the contrast of measurement.

To see this, we begin by computing the expected number photons emitted during the measurement of our preparation procedure \( \rho_0 \) using equation (24),

\[
\mu_0' := \mathbb{E}[n_d|\rho_0] = \frac{1 + 2q}{3} \mu_0 + \frac{2 - 2q}{3} \mu_1,
\]

Next, if prior to measuring \( \rho_0 \) we perform an operation \( |g, 0\rangle \mapsto |g, +1\rangle \), which may be implemented as an adiabatic inversion with a microwave pulse, we have a pre-measurement state

\[
\rho_1 = q|g, +1\rangle \langle g, +1| + (1 - q) \frac{I}{3}
\]

which when measured emits an expected number of photons

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

Previously we had considered \( \mu_0 \) and \( \mu_1 \) as the quantities that define the reference measurements in the absence of noise. However, given that \( \rho_0 \) is the best achievable initial state using standard techniques, it is in practice more convenient to use \( \mu_0' \) and \( \mu_1' \). Indeed, if we prepare the state \( \rho_0 \) and perform any unital operation resulting in a pre-measurement state

\[
\rho = q\rho_0 + (1 - q) \frac{I}{3},
\]

then some simple algebra shows that

\[
\mathbb{E}[n_d|\rho] = p\mu_0' + (1 - p)\mu_1',
\]

where \( p = \text{Tr}[\rho_0|g, 0\rangle \langle g, 0]|. \)

Finally, if we take into consideration finite visibility \( \eta \) and dark count rate \( \Gamma \) as discussed in section 3.3, we may define

\[
\alpha := \mathbb{E}[n_d|\rho_0] = \Gamma \Delta t + \eta \mu_0',
\]

\[
\beta := \mathbb{E}[n_d|\rho_1] = \Gamma \Delta t + \eta \mu_1',
\]

\[
\gamma := \mathbb{E}[n_d|\rho],
\]

to arrive at

\[
\mathbb{E}[n_d|\rho] = \gamma = p\alpha + (1 - p)\beta
\]

which is analogous to equation (31) but for our pseudo-pure state preparation. The quantity \( \gamma \), which we will call the signal, is by definition conditioned on the pre-measurement state \( \rho \), 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 \( \rho_0 \) yields \( \rho \). We will call the quantities \( \alpha \)
and $\beta$ our references because they bound the expected values of detected photons for arbitrary states of the form in equation (38). More specifically, $\alpha$ is the bright reference and $\beta$ is the dark reference since $\alpha > \beta$.

5. Drifting references

One of the main complications of experimental NV measurement is drifting of the references $\alpha$ and $\beta$ 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, $\eta$. Other mechanisms may include fluctuation of the laser power, and quality of the confocal microscope’s alignment, which will affect $\eta$, $k$, and $\Gamma$. 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 of spatial images and using feedback to realign.

5.1. 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; variations of this ordering may require a modified (though similar) analysis to that which follows. We denote the list of experiment parameter configurations as $\bar{a}_1, \ldots, \bar{a}_G$. For example, in the case of a Ramsey experiment for magnetometry [37], the distance between two $\frac{\pi}{2}$ pulses is varied, so that $\bar{a}_i = (t_i)$ where $t_i$ is the pulse spacing. If additionally the phase of the second Ramsey pulse is varied in proportion to the spacing, we have instead $\bar{a}_i = (t_i, \omega t_i)$ for some angular frequency $\omega$. A parameter configuration $\bar{a} \in \{\bar{a}_1, \ldots, \bar{a}_G\}$ 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 $\bar{a}$. 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 $\gamma$ depends only on the current value of $\alpha$ and $\beta$, the current parameter configuration $\bar{a}$, and any noise operations acting on $\bar{a}$. We denote the true value of $\alpha$, $\beta$ and $\gamma$ at the $r$th repetition of the $s$th experiment $\bar{a}$ 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 $\gamma_{n,s,r} = p\alpha_{n,s,r} + (1 - p)\beta_{n,s,r}$ where we have used our assumption of the independence of $p$ from the drift processes to label $p$ by only $s$. This yields the random variables

$$X_{n,s,r} \sim \text{Poisson}(\alpha_{n,s,r}),$$

$$Y_{n,s,r} \sim \text{Poisson}(\beta_{n,s,r}),$$

$$Z_{n,s,r} \sim \text{Poisson}(\gamma_{n,s,r}),$$

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

---

Figure 3. The considered experimental ordering. Another popular ordering transposes the inner two levels. Given a particular parameter configuration of the experiment, $\bar{a} \in \{\bar{a}_1, \ldots, \bar{a}_G\}$ (in the above example, the parameter is the distance between the two last microwave pulses), $N$ repetitions are performed of both the experiment, $\gamma$, and the references, $\alpha$ and $\beta$. The bright reference $\alpha$ 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 $\beta$ is similar, except an inversion pulse is applied prior to measurement. The pulse sequence prior to the reference measurement $\gamma$ depends on the current parameter $\bar{a}$. 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 $R$. 

---

New J. Phys. 20 (2018) 013022 I Hincks et al
5.2. Combining experiments

Due to low visibility, the repetition number $N$ in figure 3 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 $X_{n,t} = \sum_{n=1}^{N} X_{n,t}^r$, $Y_{n,t} = \sum_{n=1}^{N} Y_{n,t}^r$, and $Z_{n,t} = \sum_{n=1}^{N} Z_{n,t}$, we get

\[
X_{n,t} \sim \text{Poisson} (\alpha_{n,t}), \\
Y_{n,t} \sim \text{Poisson} (\beta_{n,t}), \\
Z_{n,t} \sim \text{Poisson} (\gamma_{n,t}),
\]

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

In a slight abuse of notation, in section 6 and on, when we talk about unscripted $\alpha$, $\beta$, and $\gamma$, we are referring to $\alpha_{n,t}$, $\beta_{n,t}$, and $\gamma_{n,t}$ for a particular index $(s, t)$ where some $N > 1$ is implicitly understood. Furthermore, in this context $\delta$ will refer to $\sum_{n=1}^{N} \Gamma_{n,t} \Delta t$, that is, the total contribution to the dark counts. Similarly, $\nu$ will refer to $\sum_{n=1}^{N} \eta_{n,t} (\mu_{n,t})^0$, and $\kappa$ to $\sum_{n=1}^{N} \eta_{n,t} (\mu_{n,t})^0$. The fraction $0 < \kappa < 1$ represents how much dimmer the reference state $\rho_{e}$ is as compared to $\rho_{g}$. This results in the relationship

\[
\alpha = \delta + \nu, \\
\beta = \delta + \kappa \nu.
\]

5.3. Correlations of model parameters

The three quantities $\mu_{0}$, $\mu_{1}$, and $\eta$ 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, $\eta$ decreases. This could be due to, for example, temperature changes in the lab which expand or contract the components on the optical table. However, $\mu_{0}$ and $\mu_{1}$ also change with decreased optical coupling because the preparation state $\rho_{0}$ depends on the laser power which is coupling dependent. Therefore we expect correlations between $\mu_{0}$, $\mu_{1}$, and $\eta$.

Nominally $\Gamma \Delta t$ should be independent of each of the quantities $\eta$, $\mu_{0}$, $\mu_{1}$, 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 $\eta$, $\mu_{0}$, and $\mu_{1}$.

Finally, and perhaps most importantly, the quantity $p$ will normally be independent of the quantities $\eta$, $\mu_{0}$, $\mu_{1}$, and $\Gamma \Delta t$. However, it is also possible for this to fail. For example, if the process that takes the preparation state $\rho_{0}$ to the pre-measurement state $\rho_{e}$ is non-unital, we will not end up with exactly the form $q_{\rho_{e}} + (1 - q) ||/3$ (as seen in section 4.2) 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.

5.4. A stochastic model of drift

The references $\alpha$ and $\beta$ 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 4.2, conditioned on set of parameters, including $k$, $\Gamma$, $\Delta t$, $\eta$, $\gamma_{0}$, $\gamma_{1}$, $\gamma_{2}$, and $\gamma_{3}$, 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 $k$, $\Gamma$, and $\eta$) on top of the already complex conditional model stated above, writing down an analytic model for the stochastic processes $\alpha$ and $\beta$ 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 $(\alpha, \beta)$ 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 5.1 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 $\alpha_{n,t} = \alpha(t_{n,t})$ and $\beta_{n,t} = \beta(t_{n,t})$ where $t_{n,t}$ is the time of the $r$th repetition of the $r$th experiment relative to the start of the $r$th average. This gives, for example, the discrete Cox process $X_{n,t} \sim \text{Poisson} (\alpha(t_{n,t}))$.

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

\[
(\alpha(0), \beta(0)) \sim \text{Normal}(\alpha_{0}, \beta_{0}, \Sigma),
\]

where the variances in $\Sigma$ 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

\[ (\alpha(t), \beta(t)) \sim \text{Normal}(\alpha_i, \beta_i, \Sigma_t) \]  

at time \( t \geq 0 \).

As a concrete example, consider the stochastic model

\[
\begin{align*}
\alpha_1 &\sim \text{Normal}(\alpha_0, \sigma_0) \\
\nu &\sim \text{Ornstein-Uhlenbeck}(0, \sigma_\nu, \theta_\nu, \alpha_1 - \Gamma \Delta t) \\
\kappa &\sim \text{Ornstein-Uhlenbeck}(\kappa_0, \sigma_\kappa, \theta_\kappa, 0) \\
(\alpha, \beta) &= (\Gamma \Delta t + \nu, \Gamma \Delta t + \kappa \cdot \nu).
\end{align*}
\]

(47)

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 \( r_0 \), denoted \( \alpha_n \), is an Ornstein–Uhlenbeck process with long-time mean \( 0 \), volatility \( \sigma_n \), mean reversion speed \( q_n \), and initial value \( \alpha - GD_t \), where the initial value is marginalized over \( \text{Normal}(0, \sigma_n) \) representing imperfections in the tracking process. This implies that, on average, \( \alpha = \alpha_0 \) and \( \alpha(\infty) = \Gamma \Delta t \) with average decay rate \( \mu_n \), where 'shakiness' in getting there is determined by \( \sigma_n \). In order to correlate \( \alpha \) with \( \beta \), and also to help enforce the physical constraint \( \beta \leq \alpha \), we relate the two components with another Ornstein–Uhlenbeck process \( \kappa \).

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

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

(48)

These calculations can be found in appendix B. In figure 4, 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 6.1.

Figure 4. Simulation of a severe case of drift. A random instance of the process \( (\alpha(t), \beta(t)) \) defined in equation (47) is shown on top of their first moments with a shaded single standard deviation. The dotted–dashed purple line shows the square root of covariance.

The model parameters used are \( \alpha_0 = 10^{-5}, \sigma_0 = 5 \times 10^{-5}, \sigma_\nu = 5 \times 10^{-5}, \theta_\nu = 0.03, \Gamma \Delta t = 3 \times 10^{-4}, \kappa_0 = 1/3, \sigma_\kappa = 0.01, \) and \( \theta_\kappa = 0.01 \). The time units are arbitrary; scaling the x-axis is equivalent to scaling \( \theta_\nu \) and \( \theta_\kappa \).
6. Statistical models of measurement

In this section we state the measurement of an NV center as a statistical inference problem. This statement is a direct result of the derivations of the previous sections, however, it is intentionally written so that previous sections may be ignored by those presently uninterested in the physical derivation of the model.

6.1. The statistical model of measurements

Consider the inner-product space $H = \mathbb{C}^3$ with the canonical basis $|1\rangle = (1, 0, 0)^T$, $|0\rangle = (0, 1, 0)^T$, and $|-1\rangle = (0, 0, 1)^T$. 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 $\rho$. Define $p = \text{Tr}[\rho P_0]$ where $P_0$ is the projector onto $|0\rangle$. In the case of a strong quantum measurement we would have access to random variables drawn from the distribution Bernoulli $p$. Instead, however, we have access to the random triplet $(X, Y, Z)$ defined by

$$
X \sim \text{Poisson} (\alpha),
Y \sim \text{Poisson} (\beta),
Z \sim \text{Poisson} (p\alpha + (1 - p)\beta) \equiv \text{Poisson} (\gamma),
$$

(49)

where $\alpha$ is the expected number of photons collected in $N$ independent measurements of the pre-measurement state $|0\rangle$, $\beta$ is the expected number of photons collected in $N$ independent measurements of the pre-measurement state $|1\rangle$, and $\gamma$ is the expected number of photons collected in $N$ independent measurements of the pre-measurement state $\rho$. It is always true that $0 \leq p \leq 1$ and $0 \leq \beta \leq \gamma \leq \alpha$.

The references $\alpha$ and $\beta$ are in turn random variables drawn from the distribution

$$
\left(\begin{array}{c}
\alpha \\
\beta
\end{array}\right) \sim \text{Normal} \left(\begin{array}{c}
\alpha_0 \\
\beta_0
\end{array}\right), \left(\begin{array}{cc}
\sigma^2_{\alpha} & \sigma_{\alpha \beta} \\
\sigma_{\alpha \beta} & \sigma^2_{\beta}
\end{array}\right)
$$

(50)

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 5. In later sections this multinormal distribution will be replaced with a product gamma distribution, or a mixture of product gamma distributions, as discussed in appendix F. 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 $\alpha$ and $\beta$.

This creates a hierarchical model with nuisance hyperparameters $\alpha_0, \beta_0, \sigma_\alpha, \sigma_\beta, \text{and } \sigma_{\alpha \beta}$. 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 $\alpha$ and $\beta$ will change each time the set of $N$ measurements are made.

6.2. Moment calculations

We first work out the first two moments of the random variables $X, Y, \text{and } Z$, both conditional and unconditional on the hyperparameters $\alpha$ and $\beta$.

The conditional moments are trivially given by

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

(51)

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

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

(52)

---

8 For those astute readers who have read the previous sections: we are abusing language slightly to keep the statement of the model compact; $\alpha$ and $\beta$ actually refer to the expected photons collected due to the pseudo-pure states $\rho_0$ and $\rho_1$. 
showing that the variance of \( \alpha \) and \( \beta \) do not affect the mean. Similarly, the law of total variance gives

\[
\begin{align*}
\text{Var}[X] &= E_{\alpha, \beta} [\text{Var}[X|\alpha]] + \text{Var}_{\alpha, \beta} [E[X|\alpha]] = \alpha + \sigma_\alpha^2 \\
\text{Var}[Y] &= E_{\alpha, \beta} [\text{Var}[Y|\beta]] + \text{Var}_{\alpha, \beta} [E[Y|\beta]] = \beta + \sigma_\beta^2 \\
\text{Var}[Z] &= E_{\alpha, \beta} [\text{Var}[Z|p, \alpha, \beta]] + \text{Var}_{\alpha, \beta} [E[Z|p, \alpha, \beta]] \\
&= p\alpha + (1 - p)\beta + p^2\sigma_\alpha^2 + (1 - p)^2\sigma_\beta^2 + p(1 - p)\sigma_{\alpha, \beta}
\end{align*}
\] (53)

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.

6.3. Three advantages of the conditional model

If the measurement model is not stated as concretely as it was in section 6.1, 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 \( (x_0, y_1, z_1), (x_0, y_2, z_2), \ldots, (x_0, y_R, z_R) \). They go back to their separate offices and try to analyse the data. Yves calculates the sample mean and variance of \( x_0, x_1, \ldots, x_R \) as \( x_{\text{samp}} = \frac{1}{R} \sum_{r=1}^{R} x_r \) and \( \sigma_{x_{\text{samp}}}^2 = \frac{1}{R - 1} \sum_{r=1}^{R} (x_r - x_{\text{samp}})^2 \), respectively. He gets \( x_{\text{samp}} = 200 \) and \( \sigma_{x_{\text{samp}}} = 20 \). Looking at equation (52), and using the standard error of the mean, this informs his rough belief that \( \bar{\alpha} \approx x_{\text{samp}} \pm \frac{\sigma_{x_{\text{samp}}}}{\sqrt{R}} = 200 \pm 20/\sqrt{R} \).

Zoe recollects that each variate \( x_r \) was obtained by summing \( N \) independent measurements. She wonder why they bothered batching the results into \( R \) sets, and why they did not 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 \( \{\alpha_r\} \) for some specific but unknown value of \( \alpha_r \). Therefore, knowing the summative property of Poisson distributions, she correctly deduces that \( x_{\text{sum}} \) was sampled from Poisson \( \{\sum_{r=1}^{R} \alpha_r\} \).

The Poisson distribution’s standard deviation is the square-root of its mean, so she therefore adopts the rough belief that \( \bar{\alpha} \approx x_{\text{sum}} / \sqrt{x_{\text{sum}}} \), and therefore that \( \sum_{r=1}^{R} \alpha_r \approx x_{\text{sum}} / \sqrt{x_{\text{sum}}} / \sqrt{R} = 200 \pm 14/\sqrt{R} \).

Zoe and Yves get back together and compare their results, and wonder why Zoe 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 \( \tilde{\alpha} \) from the hierarchical model in section 6.1, justified by the moment calculations in section 6.2. Zoey is foregoing the hyperparameter layer of the model and making a direct inference about the sum of the particular references \( \alpha_1, \alpha_2, \ldots, \alpha_R \) they happened to draw in their measurements. Neither is wrong, they are simply estimating different but related quantities.

We saw in section 6.2 that \( \text{Var}[X|\alpha] = \tilde{\alpha} \) and \( \text{Var}[X] = \tilde{\alpha} + \sigma_\alpha^2 \). 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 \( \sigma_\alpha \).

The end goal, of course, is not to estimate the reference \( \alpha \) or its mean, but to estimate quantities related to the quantum state like \( p = \text{Tr}[\rho \rho_0] \). 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 \( \alpha, \beta \) is a good approximation to the moments of the stochastic process governing \( \alpha \) and \( \beta \) discussed in section 5.4. 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.

6.4. 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 \((X, Y, Z)\). The inference problem is, given the likelihood function

\[
L(p, \alpha, \beta | x, y, z) = \text{Pr}(X = x, Y = y, Z = z | p, \alpha, \beta)
\]

\[
= \text{pdf}_{\text{pos}}(x; \alpha) \cdot \text{pdf}_{\text{pos}}(y; \beta) \cdot \text{pdf}_{\text{pos}}(z; p\alpha + (1 - p)\beta)
\]

\[
= \frac{\alpha e^{-\alpha x}}{x!} \cdot \frac{\beta e^{-\beta y}}{y!} \cdot (p\alpha + (1 - p)\beta)^{-(p\alpha + (1 - p)\beta)}
\]

(54)

to infer the value of \(p\). Note that \(\alpha\) and \(\beta\) are nuisance parameters.

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

\[
L(p, \bar{\alpha}, \bar{\beta}, \sigma_\alpha, \sigma_\beta, \sigma_{\alpha\beta}; (x_1, y_1, z_1), (x_2, y_2, z_2), \ldots, (x_R, y_R, z_R))
\]

\[
= \prod_{r=1}^{R} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \text{pdf}_{\text{pos}}(x_r; \alpha) \cdot \text{pdf}_{\text{pos}}(y_r; \beta) \cdot \text{pdf}_{\text{pos}}(z_r; \gamma) \cdot \text{pdf}_{\text{Norm}}(\alpha, \beta; \bar{\alpha}, \bar{\beta}) \cdot \text{pdf}_{\text{Norm}}(\sigma_\alpha, \sigma_\beta, \sigma_{\alpha\beta}) \text{d} \alpha \text{d} \beta \text{d} \gamma
\]

(55)

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

### 6.5. Generalized inference problems

In the previous subsection, the inference problem was stated such that the survival probability \(p = \text{Tr}[\rho P_\|]\) 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 \(\{ U_n \ldots, U_0 \} \in \mathcal{U}(\mathcal{H})\) in such a way that \(\{ U_n | 0 \rangle \langle 0 | U_n^\dagger \}_{n=1}^{9}\) is a basis for \(\mathcal{L}(\mathcal{H})\). Our scheme is to prepare \(\rho\), implement the gate \(U_n\) for some \(n\), and measure the resulting state, so that \(p_n = \text{Tr}[U_n \rho U_n^\dagger P_\|]\) with corresponding random variables

\[
Z_n \sim \text{Poisson} (p_n \alpha + (1 - p_n)\beta) \equiv \text{Poisson} (\gamma)
\]

(56)

Then our likelihood function becomes

\[
L(\rho, \alpha, \beta | x, y, z_1, z_2, \ldots, z_9) = \text{pdf}_{\text{pos}}(x; \alpha) \cdot \text{pdf}_{\text{pos}}(y; \beta) \cdot \prod_{n=1}^{9} \text{pdf}_{\text{pos}}(z_n; p_n \alpha + (1 - p_n)\beta)
\]

(57)

and we are interested in inferring \(\rho \in \mathcal{D}(\mathcal{H})\). This will be similar for the hierarchical model. We have taken one set of reference measurements for all \(Z_1, \ldots, Z_9\). We could also choose to take one for each, resulting in the data \((x_1, y_1, z_1), \ldots, (x_9, y_9, z_9)\) with a similar likelihood function of the form \(L(\rho, \alpha_1, \beta_1, \ldots, \alpha_9, \beta_9 | x, y, z_1, z_2, \ldots, z_9)\). 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 8.

### 6.6. 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 [38].

Given a likelihood function \(L(\hat{\theta}|d)\) of parameters \(\hat{\theta}\) given data \(d\), the Fisher information is the average curvature of \(L\), more specifically, it is the negative expected Hessian matrix of the log-likelihood, \(I(\hat{\theta})_{ij} = -\mathbb{E}_{\hat{\theta}}[\partial^2 \log L(\hat{\theta} | d) / \partial \hat{\theta}_i \partial \hat{\theta}_j]\). The Cramér–Rao bound asserts that no unbiased estimator of \(\hat{\theta}\) can outperform this intrinsic curvature. Namely, if \(\hat{\theta}\) is any unbiased estimator, which takes data \(d\) and outputs estimates of the true value of \(\theta\), then its covariance is lower-bounded by the inverse Fisher information matrix,

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

(58)

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 7.1.
The Fisher information matrix of the conditional model \((X, Y, Z) \mid \alpha, \beta\) can be computed exactly as

\[
I(p, \alpha, \beta) = \begin{pmatrix}
\frac{(a - \beta)^2}{p(a - \beta) + \beta} & \frac{p(a - \beta)}{p(a - \beta) + \beta} & \frac{\alpha}{\beta + \alpha - \beta} \\
\frac{p(a - \beta)}{p(a - \beta) + \beta} & \frac{p^2}{p(a - \beta) + \beta} + \frac{1}{\alpha} & -\frac{(p - 1)p}{p(a - \beta) + \beta} \\
\frac{\alpha}{p(a - \beta) + \beta} - 1 & -\frac{(p - 1)p}{p(a - \beta) + \beta} & \frac{\alpha}{\beta(p(a - \beta) + \beta)}
\end{pmatrix}
\] (59)

with an inverse matrix given by

\[
I(p, \alpha, \beta)^{-1} = \begin{pmatrix}
\frac{p(a + (p - 2)(p - 1)\beta)}{(a - \beta)^2} & \frac{p\alpha}{\beta - \alpha} & \frac{p\alpha}{(p - 1)\beta} \\
\frac{p\alpha}{\beta - \alpha} & \alpha & 0 \\
\frac{p\alpha}{(p - 1)\beta} & 0 & \beta
\end{pmatrix}
\] (60)

See appendix C for the calculation. The Cramér–Rao bound for the top left entry states that

\[
\text{Var}[^{\hat{p}}(x, y, z)] \geq \frac{p(p + 1)\alpha + (p - 2)(p - 1)\beta}{(a - \beta)^2},
\] (61)

for any unbiased estimator \(^{\hat{p}}\) given the data triple of photon counts \((x, y, z)\). This bound is plotted in figure 5 for the slice \(\beta = \alpha/2\). In section 7.3 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 (MLE), 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 \(^{\hat{p}} \sim \text{Normal}(p, (I(p, \alpha, \beta)^{-1})^{1/2})\) with \(\alpha\) and \(\beta\) known (whereas we will generally only have estimates of them). This gives the 100(1 – \(\zeta\))% confidence interval

\[
p \pm \Delta p = p \pm c_{\zeta/2} \sqrt{\frac{p(p + 1)\alpha + (p - 2)(p - 1)\beta}{(a - \beta)^2}}
\] (62)

for \(^{\hat{p}}\) where \(c_{\zeta/2} = \sqrt{2} \text{erf}^{-1}(1 - \zeta)\) with \(\text{erf}(x) = \frac{1}{2} \int_{-\infty}^{x} e^{-t^2} dt\) the error function. Supposing a reference contrast of \(C = \frac{\alpha - \beta}{\alpha + \beta}\), we need
To make $\hat{p} \approx p \pm \Delta p$ a 100(1 − $\zeta$)% confidence interval. 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 $\alpha$ is on the order of $\sim 170,000$.

This heuristic holds for any optical transition rates and choice of measurement time $\Delta t$ in the measurement protocol detailed in section 3. 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 $\Delta t$ 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 $\Delta t$ by 3 to account for all three of the signal, bright reference, and dark reference counting windows. Writing $\alpha = N\bar{\alpha}$ and $\beta = N\bar{\beta}$, with $\bar{\alpha}$ and $\bar{\beta}$ 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

$$\Delta p \sqrt{T} = \frac{\sqrt{2(T_e + 3\Delta t)\bar{\alpha}(\Delta t)}}{\bar{\pi}(\Delta t) - \bar{\beta}(\Delta t)}. \quad (64)$$

We have written $\bar{\pi} = \bar{\pi}(\Delta t)$ and $\bar{\beta} = \bar{\beta}(\Delta t)$ to emphasize their implicit dependence on $\Delta t$. 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 minimized, visualized in figure 6. 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 Taylor et al [37] or Hirose et al [39] for two examples out of many.

---

9 We use this definition for contrast simply because it shows up naturally in (69); $\frac{\alpha - \beta}{\alpha}$ is arguably a better choice.
Of note is the steep increase of $\Delta p \sqrt{T}$ as $\Delta t \to 0$. While the slope is relatively gentle as $\Delta t$ gets larger, causing little harm even if $\Delta t$ is twice as big as the optimal value for a given $T$, there is a large penalty for choosing a measurement of $\Delta t$ which is too short. Long refocusing sequences like CPMG are especially at risk of falling into this trap.

### 6.7. 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 [40], or another that incoherently stores pre-measurement spin populations on an ancilla nuclear qubit that is subsequently read out several times successively [24], 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.

#### 7. Simple estimators

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 6.4. We also compare their relative strengths and weaknesses.

##### 7.1. 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 (51) shows that $x, y, z$ are unbiased estimates of $\alpha, \beta$, 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

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

(65)

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 MLE of the model, that is, $(\frac{z - y}{x - y}, x, y)$ is the (unique) maximum of the function $L(p, \alpha, \beta; x, y, z)$ on the domain $0 \leq p \leq 1, 0 < \beta \leq \alpha$, for any $x, y, z \geq 0$. See appendix D.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 $\alpha$ and $\beta$ 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

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

(66)

for some non-integer value of $\epsilon$. One might consider using this estimator instead of the MLE if $x \approx y$ has a significant probability. After quite a bit of work, shown in appendix D.2, it is seen that the bias of $\hat{p}_{\text{MLE}, \epsilon}$ is non-zero, a linear function of $p$, and exactly given by the integral

$$\text{Bias}[\hat{p}_{\text{MLE}, \epsilon}] = \text{re} \int_0^\pi i (-1)^\epsilon e^{-\alpha(1 + \epsilon^2) - \beta(1 + \epsilon^2)}((\beta + (\alpha - \beta)p)e^{i\phi} + \beta e^{i\phi(1 - \epsilon)})d\phi - p.$$  

(67)

Taking the limit as $\epsilon$ approaches zero gives an expression for the bias of the original estimator,

$$\text{Bias}[\hat{p}_{\text{MLE}}] = \lim_{\epsilon \to 0} \text{Bias}[\hat{p}_{\text{MLE}, \epsilon}]$$

$$= \int_0^\pi e^{-(\alpha + \beta)(1 + \cos \phi)[(\gamma + \beta \cos \phi)\sin((\alpha - \beta)\sin \phi) + (\beta \sin \phi)\cos((\alpha - \beta)\sin \phi)]}d\phi - p.$$  

(68)

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,

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

(69)

which through numerics can be shown to be valid for $\alpha + \beta \gg 300$. 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 $\frac{\alpha + \beta}{(\alpha - \beta)^2} = \mathcal{O}(\alpha + \beta)^{-1}$.

Although estimators with no bias are generally preferred, we see that this estimator has the more important property of being consistent, meaning that $\text{Bias}[\hat{p}_{\text{MLE}}] \to 0$ 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 (58) known as the van Trees inequality or the Bayesian Cramér–Rao bound [41], stating that

$$\text{Cov}[\hat{\theta}] \geq I_{\theta}(p, \alpha, \beta) \text{I}(p, \alpha, \beta)^{-1} J_{\theta}(p, \alpha, \beta),$$

(70)

where $I_{\theta}(p, \alpha, \beta)$ is the expectation value of the Jacobian matrix of the possibly biased estimator $\hat{\theta}$. Using the approximation from (69), this gives us the inequality

$$\text{Var}[\hat{\theta}_{\text{MLE}}] \geq p(p + 1)\alpha + (p - 2)(p - 1)\beta \div (\alpha - \beta)^2 + O((\alpha + \beta)^{-2})$$

(71)

for the MLE, again assuming that the contrast is fixed as $\alpha$ and $\beta$ 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 [42].

7.2. Bayes estimator

If we assume our prior knowledge of the parameters $(p, \alpha, \beta)$ 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 $(x, y, z)$:

$$\pi^*(p, \alpha, \beta) \equiv \Pr(p, \alpha, \beta, x, y) = \frac{\Pr(x, y, z|p, \alpha, \beta)p(p, \alpha, \beta)}{\int \Pr(x, y, z|p, \alpha, \beta)p(p, \alpha, \beta)dp \, d\alpha \, d\beta} = L(p, \alpha, \beta|x, y, z)p(p, \alpha, \beta) \quad \mathcal{N}.$$

(72)

Here, $L$ is the likelihood function from equation (54) 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 $X, Y, Z$. We can first update our prior distribution using the datum $(x, y)$ to get

$$\Pr(p, \alpha, \beta|x, y) = \frac{\Pr(x, y|p, \alpha, \beta)p(p, \alpha, \beta)}{\int \Pr(x, y|p, \alpha, \beta)p(p, \alpha, \beta)dp \, d\alpha \, d\beta} \equiv \pi^*(\alpha, \beta)p(p)$$

(73)

and subsequently

$$\Pr(p, \alpha, \beta|x, y, z) = \frac{\Pr(z|p, \alpha, \beta)p^*(\alpha, \beta)p(p)}{\int \Pr(z|p, \alpha, \beta)p^*(\alpha, \beta)p(p)dp \, d\alpha \, d\beta} = \pi^*(p, \alpha, \beta).$$

(74)

This sequential breakdown 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 appendix F. The full likelihood $L$, however, almost certainly does not have a conjugate prior.

Given the posterior distribution $\pi^*$, 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 $\pi^*$. For the particular parameter $p$, we have

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

(75)

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

$$(\hat{p}_{\text{Bayes}}, \hat{\alpha}_{\text{Bayes}}, \hat{\beta}_{\text{Bayes}}) = \text{arg min}_q \mathbb{E}[\mathbb{E}[(\hat{\theta}(x, y, z) - (p, \alpha, \beta))^2]].$$

(76)

7.3. 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, $p$ 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 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.

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

$$L_{\text{MSE}}(\hat{p}(x, y, z), p) = (\hat{p}(x, y, z) - p)^2,$$

(77)

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 $P_S(\alpha, \beta) = Pr(\alpha, \beta|\text{experimental setup } S)$ 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 $(p, \alpha, \beta)$ is defined in the standard way as the average value of the loss,

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

(78)

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

$$R_S(\hat{p}, p) = \mathbb{E}_{\alpha,\beta}[R(\hat{p}, p, \alpha, \beta)] = \int R(\hat{p}, p, \alpha, \beta)P_S(\alpha, \beta)\,d\alpha\,d\beta,$$

(79)

by marginalizing over our knowledge about it. Note that the units of $\sqrt{R_S(\hat{p}, p)}$ are the same as those of $p$, and that, for example, a value of 0.01 has the straightforward 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 $\alpha$ and $\beta$, 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, $\hat{p}_{\text{Bayes}}$ with itself under different priors. The prior on $\alpha$ and $\beta$ does not need to have any relationship with $P_S(\alpha, \beta)$. 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, $S_{\text{hd}}$, where $\pi = 100 000$, a mid-data regime, $S_{\text{md}}$, where $\pi = 10 000$, and a low-data regime, $S_{\text{ld}}$ where $\pi = 1000$. In these three cases the contrast is the same, $C = \frac{\pi - \beta}{\pi + \beta} = 0.6$. We additionally consider mid-data regimes of varying contrast, where $\pi = 10000$ for $C = 0.05, 0.33, 0.82$. This defines the respective low, medium, and high contrast setups $S_{\text{lc}}, S_{\text{mc}}$, and $S_{\text{nc}}$. The above setup descriptions only supply the mean values of their respective distributions $P_S$, so that, $\mathbb{E}_S[\alpha|\beta] = (\pi, \beta)$. To keep things simple, we take these distributions to all be binormal with super-Poisson standard deviations $\sigma_1 = 2\sqrt{\pi}$ and $\sigma_2 = 2\sqrt{\pi}$, and covariances defined by $\sigma_{1,2} = 1.5\beta$.

In figure 7, $\sqrt{R_S(\hat{p}, p)}$ is plotted for each of the setups described above, and for each of the estimators $\hat{p}_{\text{MLE}}, \hat{p}_{\text{BCE}},$ and $\hat{p}_{\text{Bayes}}$. Two different priors are used for the Bayes estimator on each setup, both are product gamma distributions, discussed further in appendix F.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 appendix F.2, which allows for correlations between $\alpha$ and $\beta$, should in theory be strictly better than the ones used in these calculations, but were found to be too computationally expensive for naïve implementations of risk computation. For all setups and estimators, risk is computed by Monte Carlo sampling: for each value of $p$, many pairs $(\alpha, \beta)$ are sampled from $P_S(\alpha, \beta)$, for each pair many variates $(x, y, z)$ are drawn from the likelihood distribution, and the loss $L_{\text{MSE}}$ is computed for each. The average of these loss values for this value of $p$ forms an estimate of $R_S(\hat{p}, p)$.

These plots show that the Cramér–Rao bound (61) 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.

8. Example: Hamiltonian learning with Bayesian inference

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 justifyably 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 QHL using experimental data. In a recent and related experimental work, QHL was shown to be a powerful method of characterizing quantum systems. Raw data and reproducible code for this paper can be found in our online repository.

Other advantages of the Bayesian inference algorithm we employ, though not used by us here, include the ability to adaptively choose the next experiment to maximize information gain, and the ability to treat reference drift as a time dependent stochastic process.

8.1. Hamiltonian model

We assume that our spin-1 Hamiltonian (in the optical ground state) is of the form

$$H_{m_1} = \delta \Delta S_z^2 + (\omega_z + m_1 A_N) S_z + \Omega(t) S_y,$$

where we are in a frame rotating near the ground state ZFS, 2.87 GHz, and have invoked the secular and rotating wave approximations. Here, $\delta \Delta$ is the mismatch between the applied microwave frequency and the ZFS, $\omega_z$ is the projection of the static magnetic field onto the $z$-axis, $A_N$ is the hyperfine splitting due to 14-Nitrogen, $m_1$ is the spin number of the nitrogen atom, and $\Omega$ is the microwave nutation strength. All of the parameter units are angular frequency, $2\pi \cdot$ MHz, except $m_1$ which is unitless. Our goal is to learn these parameters, as well as $T_2$.

Figure 7. 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, $\hat{p}_{MLE}$, the bias corrected estimator (see appendix E), $\hat{p}_{BCE}$, and the Bayes estimator, $\hat{p}_{Bayes}$, 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 0–1 in steps of 0.05. The risks of $\hat{p}_{MLE}$ and $\hat{p}_{BCE}$ are much bigger than 1 for the low-contrast regime due to the common occurrence of $y > x$, and are therefore not plotted.
decay time. In particular, we would like good error bars on \( \omega_t \) 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 \( |m_n| \in \{-1, 0, +1\} \). 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 \( m_n \) dependent shift in the static magnetic field.

With an initial state \( \rho_t = |g_s, 0\rangle \langle g_s, 0| \), at time \( t \) the density matrix is described by \( \rho(t) = \frac{1}{\sum_m} S(t)_{m} \rho_{t} \) where \( S_{m}(t) \) is the evolution superoperator under the Hamiltonian \( H_{m} \), along with a single dephasing Lindblad term \( L = \sqrt{T_1} S_{m} \). The superoperators \( S_{m} \) can be computed by exponentiating the superoperator 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.

8.2. Experiment choices and data

The time dependence of the nutation envelope \( \Omega \) 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 \( \Omega \) 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 \( \Omega \) 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 photons \( \beta = 0.004 \), 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 \( N = 30 \, 000 \) 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 linearly spaced points between \( t_w = 0.01 \) µs and \( t_w = 2 \) µs, with a pulse time \( t_p = 44 \) ns. Raw (summed) data is plotted in appendix H.

8.3. QHL likelihood function

As we did for the tomography model sketched in section 6.5, we now write down a model for our QHL problem. We label a given experimental configuration as \( \tilde{c} = (t_r, t_w, t_p, k) \), where the three timing parameters were defined in section 8.1 and \( k \in \{ \text{RABI, RAMSEY} \} \) selects the experiment types explained in section 8.2.

Similarly, we denote a hypothetical parameter set as a vector \( \tilde{x} = (\omega_t, \delta \Delta, \Omega, A_X, T_1^{-1}) \). A specific pair, \( (\tilde{x}, \tilde{c}) \), 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

\[
P_{\tilde{x}, \tilde{c}} = \text{Tr}(|0\rangle \langle 0| \, S_{\tilde{x}, \tilde{c}} |0\rangle \langle 0|).
\]  

Here, \( S_{\tilde{x}, \tilde{c}} \) is the solution to the Lindblad master equation under the Hamiltonian model described in section 8.1. This yields the conditional model \( X, Y, Z|\alpha, \beta, \tilde{x}, \tilde{c} \sim \text{with} \ X \sim \text{Poisson} (\alpha), Y \sim \text{Poisson} (\beta), \) and \( Z \sim \text{Poisson} (\beta + P_{\tilde{x}, \tilde{c}} (\alpha - \beta)) \). The goal of the inference problem is to deduce the true values of \( \tilde{x} \), and in particular, \( \omega_t \), given a dataset of photon counts.

8.4. Bayesian inference with sequential Monte Carlo

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

\[
\omega_t \sim \text{Unif} (0 \text{ MHz}, 10 \text{ MHz}), \\
\delta \Delta \sim \text{Unif} (-5 \text{ MHz}, 5 \text{ MHz}), \\
\Omega \sim \text{Unif} (0 \text{ MHz}, 10 \text{ MHz}), \\
A_X \sim \text{Unif} (1.5 \text{ MHz}, 3.5 \text{ MHz}),
\]  

(82) (83) (84) (85)
Additionally, we empirically choose a prior for the references $\alpha$ and $\beta$ 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 appendix F.1. We label this prior with the data from a late time point, the posterior will have many peaks because every divisor of the argument is stored as a so-called particle approximation consisting of a finite list of hypothetical values, called particles, labeled as $\{(\alpha_k, \beta_k, \vec{x}_k)\}$ with corresponding weights $\{w_i\}$, $\sum_i w_i = 1$. Typically on the order of 10 000 particles are used for numerical stability. The distribution for the references $\alpha$ and $\beta$ is reset to $\pi_0(\alpha, \beta)$ before each triple of data is used. As discussed in section 7.2, we may process the reference pair $(\alpha_1, \beta_1)$ first and subsequently process the signal count $z_i$. The reference pair is processed by replacing the reference prior with the analytically derived posterior, discussed in appendix F.1. This is done by replacing the $(\alpha, \beta)$ coordinates of each particle $(\alpha_k, \beta_k, \vec{x}_k)$ with a random variate drawn from the posterior $\pi^*(\alpha, \beta|x_k, y_k)$. The signal is processed using the sequential Monte Carlo (SMC) algorithm, as implemented by the software package QInfer.

While Bayes’ update rule is agnostic to the order in which we enter data since each data triple is independently, 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—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 $t_n$ and $t_l$, shuffling the Rabi and Ramsey data together randomly. Alternatively, an algorithm without a unimodality constraint could be considered.

One nice feature of the SMC algorithm is that it typically heralds its own failure through the effective sample size criterion. 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 7.2, and also by a technique called bridging the transition, discussed in appendix G.

8.5. 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 the appendix). Recall that our entire dataset consists of 400 averages of 30 000 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 $\alpha$ and $\beta$), wider posteriors are expected for these than for the entire data set. The main results are shown in figure 8 and table 1, with supplementary figures found in appendix H. The top two plots, figures 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.

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 the parameters

\[ \hat{\vec{x}} = (\omega, \delta \Delta, \Omega, A_N, \tau_2^{-1}) \]

of the parameters \( \vec{x} = (\omega_x, \delta \Delta, \Omega, A_N, \tau_2^{-1}) \). This was done by using the SciPy [48] function optimization.curve_fit to minimize the quantity

\[ \Phi(\vec{x}) = \sum_{\xi} \left( \frac{\hat{p}_\xi - p_{\xi,\vec{x}}}{\sigma_{\xi}} \right)^2 \]

\[ \hat{p}_\xi = \frac{z_\xi - y_\xi}{x_\xi - y_\xi}, \quad \sigma_{\xi}^2 = \frac{\hat{p}_\xi (\hat{p}_\xi + 1)x_\xi + (\hat{p}_\xi - 2)(\hat{p}_\xi - 1)y_\xi}{(x_\xi - y_\xi)^2}, \]

(87)
where the sum is taken over all experiment configurations \( \bar{z} = (t_r, t_w, t_p, k) \) that were performed, \( p_{\bar{z}} \) is the simulation of hypothesis \( \bar{z} \) under conditions \( \bar{z} \) defined in (81), \( \hat{p}_c \) is the MLE of \( p \) given the data \( (x_{\bar{z}}, y_{\bar{z}}, z_{\bar{z}}) \), and the formula for the estimated variance \( \sigma_\bar{z}^2 \) is derived from the Cramér–Rao bound (61). For simplicity, the initial guess of the WLSF function was taken to be the SMC point estimate. The WLSQ fit is shown in figures 8(e)–(f) on top of the SMC marginal posteriors. Table 1 provides a more comprehensive comparison, where WLSQ fits are also performed on each of the ten batches. 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 \( d_D \) and \( w_e \).

9. Conclusions

We started with a standard phenomenological model of NV⁰ optical dynamics and derived random variables corresponding to photon emission during the conventional measurement process. Noise, limited visibility, and imperfect state preparation were considered in detail and added to the model, which did not affect it very much; one still ends up with three conditionally independent Poisson random variables, just with lower count rates and contrast.

This material will have all been review for those familiar with this quantum system. However, we deemed it important to include it explicitly so that our subsequent statement of the NV⁰ measurement process as a formal statistical inference problem would be entirely self-contained and justified. We were able to give a straightforward argument for why summing up all of the photon counts for repeated experiments of the same type is almost always the best thing to do, regardless of data-processing techniques, and despite our discussion about the headaches involved in drifting reference counts.

### Table 1. Results of the SMC and WLSQ fits are shown for for the entire data sat, as well as for each batch. All units are kHz. Both the point estimate (\( \bar{E} \)) and the marginal standard deviation (\( \sigma \)) are displayed for each of the five fit parameters, \( (\omega_c, \delta \Delta, \Omega, A_N, T_2^\phi) \).

| Batch | SMC | WLSF |
|-------|-----|------|
|       | \( \omega_c \) | \( \delta \Delta \) | \( \Omega \) | \( A_N \) | \( T_2^\phi \) |
| All data | \( \bar{E} \) | 1432 | 0.5 | 597 | 13.9 | 5555 | 1.1 | 2171 | 0.8 | 35 | 0.5 |
|         | \( \sigma \) | 1432 | 1.4 | 158 | 54.7 | 5564 | 3.3 | 2169 | 1.8 | 43 | 1.7 |
| Batch 0 | \( \bar{E} \) | 1430 | 3.6 | 639 | 115.0 | 5554 | 6.2 | 2181 | 4.7 | 44 | 2.8 |
|         | \( \sigma \) | 1424 | 4.0 | 773 | 127.8 | 5551 | 9.5 | 2176 | 5.2 | 49 | 4.8 |
| Batch 1 | \( \bar{E} \) | 1422 | 2.8 | 463 | 211.0 | 5555 | 9.6 | 2172 | 3.1 | 40 | 3.6 |
|         | \( \sigma \) | 1425 | 3.5 | 773 | 127.8 | 5551 | 9.5 | 2176 | 4.6 | 41 | 4.0 |
| Batch 2 | \( \bar{E} \) | 1437 | 3.2 | 323 | 279.6 | 5567 | 8.9 | 2163 | 4.9 | 48 | 4.2 |
|         | \( \sigma \) | 1433 | 4.0 | 188 | 153.3 | 5566 | 8.6 | 2164 | 4.9 | 49 | 4.5 |
| Batch 3 | \( \bar{E} \) | 1434 | 2.6 | 30 | 247.2 | 5569 | 7.6 | 2166 | 3.3 | 31 | 2.0 |
|         | \( \sigma \) | 1435 | 3.3 | 153 | 140.4 | 5569 | 8.3 | 2164 | 4.3 | 36 | 3.9 |
| Batch 4 | \( \bar{E} \) | 1432 | 2.5 | 120 | 233.3 | 5560 | 8.0 | 2173 | 4.2 | 40 | 2.3 |
|         | \( \sigma \) | 1431 | 3.6 | 132 | 133.3 | 5563 | 7.8 | 2175 | 4.8 | 46 | 4.3 |
| Batch 5 | \( \bar{E} \) | 1435 | 2.9 | 125 | 147.8 | 5561 | 7.3 | 2177 | 4.2 | 40 | 2.7 |
|         | \( \sigma \) | 1435 | 3.7 | 163 | 140.1 | 5566 | 8.4 | 2178 | 4.8 | 45 | 4.2 |
| Batch 6 | \( \bar{E} \) | 1433 | 3.7 | 41 | 248.9 | 5567 | 7.1 | 2169 | 4.5 | 44 | 3.7 |
|         | \( \sigma \) | 1433 | 3.5 | 0 | 6962.6 | 5571 | 8.1 | 2166 | 4.5 | 44 | 4.0 |
| Batch 7 | \( \bar{E} \) | 1426 | 3.4 | 390 | 499.5 | 5559 | 10.4 | 2168 | 4.3 | 40 | 3.3 |
|         | \( \sigma \) | 1429 | 3.3 | 410 | 127.2 | 5562 | 7.8 | 2165 | 4.1 | 40 | 3.8 |
| Batch 8 | \( \bar{E} \) | 1438 | 4.3 | 203 | 348.1 | 5549 | 7.2 | 2168 | 3.8 | 41 | 2.9 |
|         | \( \sigma \) | 1440 | 3.7 | 511 | 144.9 | 5550 | 8.9 | 2168 | 4.6 | 43 | 4.3 |
| Batch 9 | \( \bar{E} \) | 1439 | 3.5 | 226 | 650.5 | 5554 | 7.8 | 2167 | 3.6 | 45 | 2.9 |
|         | \( \sigma \) | 1439 | 3.9 | 538 | 140.1 | 5554 | 8.8 | 2169 | 4.6 | 46 | 4.3 |
We derived both frequentist and Bayesian estimators for this inference problem. We found that the most commonly used frequentist estimator is exactly the MLE, and its MSE risk is well approximated by a simple equation derived from the Cramér–Rao bound. This fact produced a useful formula, (63), for estimating the amount of data required to reduce the error bars of $p = \text{Tr}(\rho|0\rangle\langle 0|)$ to a specified value $\Delta p$. Quantitatively comparing the estimators revealed that the Bayesian estimator for $p$ is superior to the MLE with respect to MSE risk, especially near the boundaries of $[0, 1]$.

Finally, in order to convince the reader that a Bayesian approach is tractable in practice, we perform QHL on experimental data in a fully Bayesian setting using the SMC inference algorithm. In addition to giving a good fit even in the case of a very wide prior, cross validation gives convincing evidence that the posterior distribution over Hamiltonian parameters accurately reports its confidence in their values. Credible regions reported by SMC are generally better than confidence intervals due to weighted least squares fitting, but in some cases, not by much.

Our methods and results are just a first step toward more efficient and robust NV based devices, with many directions for future research. There would be benefit to performing similar analyses for other measurement protocols with higher read-out fidelity. One benefit to the Bayesian hierarchical approach we have taken in our software implementation is that one measurement protocol can be seamlessly swapped out for another. We also expect learning rates could be improved by considering more general pulse programs, and moving to adaptive experiment design where Bayesian methods have a clearer advantage.

Acknowledgments

IH is grateful to Toeno van der Sar and Thomas Alexander for helpful conversations. This research was undertaken thanks in part to funding from: The Canada First Research Excellence Fund (CFREF) Canadian Excellence Research Chairs (CERC), the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canadian Institute for Advanced Research, the Province of Ontario, and Industry Canada.

Appendix A. Data and code for this paper

This paper used Wolfram Mathematica 11 for several calculations, both numerical and symbolic, along with the corresponding plots found in the figures. QHL was performed in Python 2.7 using a custom model written for the QInfer software package [43]. All code and data necessary to reproduce the results of this paper and its appendices are openly hosted in a GitHub repository; follow the link of [30].

Appendix B. Stochastic moments

Mathematica version 10.0.2.0 was used to execute the code snippets in listings 1 and 2.

Listing 1. Mathematica code for computing the moments of the stochastic model defined in equation (47) conditional on the value of $\alpha_1$.

```
p[expr_] := TransformedProcess[
   expr,
   {
      \nu \sim \text{OrnsteinUhlenbeckProcess}[0, \sigma_{\nu}, \theta_\nu, \alpha_0 - \Gamma],
      \kappa \sim \text{OrnsteinUhlenbeckProcess}[\sigma_0, \sigma_{\kappa}, \theta_\kappa]
   },
   t
];

\alpha_{\text{mean}1} = p[\Gamma + \nu[t]]/\text{Mean}
\beta_{\text{mean}1} = p[\Gamma + \kappa[t]]/\text{Mean}
\alpha_{\text{var}1} = p[\Gamma + \nu[t]]/\text{Variance}
\beta_{\text{var}1} = p[\Gamma + \kappa[t]]/\text{Variance}/\text{Expand}
(* We need to compute the covariance manually using Cov[a, b] = \text{Mean}[a + b] - \text{Mean}[a]\text{Mean}[b] *)
\alpha_{\text{product}} = (\Gamma + \nu[t])(\Gamma + \kappa[t])/\text{Expand};
\alpha_{\text{cov}1} = \text{Sum}[\text{Mean}[p[\text{expr}[t]], \{\text{expr}, \text{List}@\alpha_{\text{product}}\}] - \alpha_{\text{mean}1}/\text{mean}/\text{Expand} 
```
Listing 2. Mathematica code for using the laws of total expectation, variance, and covariance to derive the moments of the stochastic model defined in equation (47). The results are shown in equation (48).

```mathematica
totalExp = Mean[TransformedDistribution[{mean, α1 ~ NormalDistribution[α0, σα]}]] + Mean[TransformedDistribution[{var, α1 ~ NormalDistribution[α0, σα]}]]

Covariance[TransformedDistribution[{[mean1, mean2], α1 ~ NormalDistribution[α0, σα]}, Covariance[TransformedDistribution[{cov, α1 ~ NormalDistribution[α0, σα]}]]]

αmean = totalExp[αmean1]
βmean = totalExp[βmean1]
αvar = totalVar[αmean1, αvar1]
βvar = totalVar[βmean1, βvar1] // Expand

α Kov = totalCov[αmean1, βmean1, α Kov1] // Expand
```

Appendix C. Cumulants of the conditional likelihood function

We wish to compute

\[ I(p, \alpha, \beta) = -\mathbb{E}_{x,y,z} \left[ \frac{\partial}{\partial \theta_i} \frac{\partial}{\partial \theta_j} \log L \right]_{i,j=1}^3 \]

where \( \theta_1 = p, \theta_2 = \alpha, \theta_3 = \beta \), and from equation (54) we have

\[ L(x, y, z|p, \alpha, \beta) = \frac{\alpha^x e^{-\alpha}}{x!} \cdot \frac{\beta^y e^{-\beta}}{y!} \cdot \frac{(p\alpha + (1-p)\beta)^z}{z!} \]

Listing 3. The results of this code for the Fisher information matrix are

\[
I(p, \alpha, \beta) = -\mathbb{E}_{x,y,z} \left[ \frac{\partial}{\partial \theta_i} \frac{\partial}{\partial \theta_j} \log L \right]_{i,j=1}^3
\]

\[
= \begin{pmatrix}
\frac{(\alpha - \beta)^2}{p(\alpha - \beta) + \beta} & \frac{p(\alpha - \beta)}{p(\alpha - \beta) + \beta} & \frac{\alpha}{\beta + p\alpha - p\beta} - 1 \\
\frac{p(\alpha - \beta)}{p(\alpha - \beta) + \beta} & \frac{p(\alpha - \beta) + 1}{p(\alpha - \beta) + \beta} & -\frac{(p-1)p}{\beta(p(\alpha - \beta) + \beta)} \\
\frac{\alpha}{p\alpha - p\beta + \beta} & -\frac{(p-1)p}{\beta(p(\alpha - \beta) + \beta)} & \frac{p(\alpha - \beta)}{\alpha - \beta}
\end{pmatrix}
\]

with inverse

\[
I(p, \alpha, \beta)^{-1} = \begin{pmatrix}
\frac{p(p+1)+p-2(p-1)\beta}{\alpha - \beta} & \frac{p(\alpha - \beta)}{\alpha - \beta} & \frac{(p-1)\beta}{\alpha - \beta} \\
\frac{p(\alpha - \beta)}{\alpha - \beta} & \alpha & 0 \\
\frac{(p-1)\beta}{\alpha - \beta} & 0 & \beta
\end{pmatrix}
\]
Further, the higher order cumulants turn out to be

\[
K_{1,\cdot} = \begin{pmatrix}
\frac{2(\alpha - \beta)^2}{\beta (\beta + \alpha)^2} & \frac{2(\beta - \alpha)}{(\beta + \alpha)^2} & \frac{2(\alpha - \beta)}{\beta (\beta + \alpha)^2} \\
\frac{2(\beta - \alpha)}{(\beta + \alpha)^2} & \frac{2\beta}{(\beta + \alpha)^2} & \frac{\beta}{(\beta + \alpha)^2} \\
\frac{2(\alpha - \beta)}{\beta (\beta + \alpha)^2} & \frac{\beta}{(\beta + \alpha)^2} & \frac{\beta}{(\beta + \alpha)^2}
\end{pmatrix}
\]

\[
K_{2,\cdot} = \begin{pmatrix}
\frac{2(\beta - \alpha)}{(\beta + \alpha)^2} & \frac{2\beta}{(\beta + \alpha)^2} & \frac{p}{(\beta + \alpha)^2} \\
\frac{2\beta}{(\beta + \alpha)^2} & \frac{2(\beta - \alpha)}{(\beta + \alpha)^2} & \frac{p + 1}{(\beta + \alpha)^2} \\
\frac{p}{(\beta + \alpha)^2} & \frac{p + 1}{(\beta + \alpha)^2} & \frac{2(\beta - \alpha)}{(\beta + \alpha)^2}
\end{pmatrix}
\]

\[
K_{3,\cdot} = \begin{pmatrix}
\frac{2(\alpha - \beta)}{\beta (\beta + \alpha)^2} & \frac{p - 1}{(\beta + \alpha)^2} & \frac{p - 1}{(\beta + \alpha)^2} \\
\frac{p - 1}{(\beta + \alpha)^2} & \frac{2(\alpha - \beta)}{\beta (\beta + \alpha)^2} & \frac{p - 1}{(\beta + \alpha)^2} \\
\frac{p - 1}{(\beta + \alpha)^2} & \frac{p - 1}{(\beta + \alpha)^2} & \frac{2(\alpha - \beta)}{\beta (\beta + \alpha)^2}
\end{pmatrix}
\]

\begin{equation}
(C7)
\end{equation}

and

\[
J_{1,\cdot} = \begin{pmatrix}
\beta (\beta + \alpha) & \beta (\beta + \alpha) & p (\beta - \alpha) \\
\beta (\beta + \alpha) & \beta (\beta + \alpha) & p (\beta - \alpha) \\
\beta (\beta + \alpha) & \beta (\beta + \alpha) & p (\beta - \alpha)
\end{pmatrix}
\]

\[
J_{2,\cdot} = \begin{pmatrix}
\beta (\beta + \alpha) & \beta (\beta + \alpha) & p (\beta - \alpha) \\
\beta (\beta + \alpha) & \beta (\beta + \alpha) & p (\beta - \alpha) \\
\beta (\beta + \alpha) & \beta (\beta + \alpha) & p (\beta - \alpha)
\end{pmatrix}
\]

\[
J_{3,\cdot} = \begin{pmatrix}
\beta (\beta + \alpha) & \beta (\beta + \alpha) & p (\beta - \alpha) \\
\beta (\beta + \alpha) & \beta (\beta + \alpha) & p (\beta - \alpha) \\
\beta (\beta + \alpha) & \beta (\beta + \alpha) & p (\beta - \alpha)
\end{pmatrix}
\]

\begin{equation}
(C8)
\end{equation}

Listing 3. Mathematica code to find the fisher information matrix, as well as two higher order cumulants.

1 (* Define assumptions on variables *)
2 $Assumptions = \beta < \gamma < \alpha \&\& 0 <= p <= 1 \&\& x > 0 \&\& y > 0 \&\& z > 0; (* Define the log-likelihood *)
3 L = ((\alpha Exp[-\alpha])/Factorial[x])
4 (* ((\beta Exp[-\beta])/Factorial[y]) *)
5 (* (p \alpha + (1-p)\beta) Exp[-(p \alpha + (1-p)\beta)]/Factorial[z]; *)
6 LL = Log[L] // FullSimplify;
(* Verify that L is normalized as a sanity check *)
7 Sum[LL, {x, 0, \infty}, {y, 0, \infty}, {z, 0, \infty}]] // FullSimplify
8 (* Compute Fisher matrix and its inverse *)
9 Ifisher = -Sum[Evaluate[Simplify]
10 Outer[D[LL, \#1, \#2] &\{p, \alpha, \beta, \{p, \alpha, \beta\}][L,]
11 \{x, 0, \infty\}, \{y, 0, \infty\}, \{z, 0, \infty\}],
16 ] // FullSimplify;
17 invfisher = Inverse // FullSimplify;
(* Compute K and J (takes a while, especially J) *)
18 K = With[
19 \{summand = FullSimplify[LL * Outer[D[LL, \#1, \#2, \#3] &\{p, \alpha, \beta, \{p, \alpha, \beta\}];[L,]
20 \{x, 0, \infty\}, \{y, 0, \infty\}, \{z, 0, \infty\}]
21 \}
28
Appendix D. Maximum likelihood estimator

D.1. Derivation

The log-likelihood of the basic conditional model is given by

\[
\log L = (x \log \alpha - \alpha - \log x!) + (y \log \beta - \beta - \log y!) \\
+ (z \log (p\alpha + (1 - p)\beta) - (p\alpha + (1 - p)\beta) - \log z!).
\]  

\[\text{(D1)}\]

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

\[
(x \log \alpha - \alpha - \log x!) + (y \log \beta - \beta - \log y!) + (z \log \gamma - (p\alpha + (1 - p)\beta) - \log z! \\
= \Phi = [(x \log \alpha - \alpha - \log x!) + (y \log \beta - \beta - \log y!) + (z \log \gamma - (p\alpha + (1 - p)\beta) - \log z!)]
\]

subject to the constraint \(\gamma = p\alpha + (1 - p)\beta.\) Using the Lagrange multiplier \(\lambda,\) this is encoded in a Lagrangian as

\[
\Phi = [(x \log \alpha - \alpha - \log x!) + (y \log \beta - \beta - \log y!) + (z \log \gamma - (p\alpha + (1 - p)\beta) - \log z!)] \\
- \lambda[\gamma - (p\alpha + (1 - p)\beta)]
\]

\[\text{(D3)}\]

yielding a simple set of equations

\[
\left\{ \frac{\partial \Phi}{\partial \alpha} = 0, \ \frac{\partial \Phi}{\partial \beta} = 0, \ \frac{\partial \Phi}{\partial \gamma} = 0, \ \frac{\partial \Phi}{\partial \lambda} = 0 \right\}
\]

\[\text{(D4)}\]

with no more logarithms. These equations can be solved for \(\alpha, \beta,\) 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 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 4. Mathematica code to solve the Lagrange problem stated in equation (D3).

\[
\text{\texttt{(* Define assumptions on variables *)}} \\
\text{\texttt{Assumptions = 0 < \beta < \gamma < \alpha&&0 < = p < = 1 && x > 0 && y > 0 && z > 0 && \Lambda \in \text{Reals};}} \\
\text{\texttt{(* Define Lagrangian *)}} \\
\text{\texttt{\Phi = \log[}} \\
\text{\texttt{\alpha x \exp[-\alpha]/\text{Factorial}[x]}} \\
\text{\texttt{\beta y \exp[-\beta]/\text{Factorial}[y]}} \\
\text{\texttt{\(p\alpha + (1 - p)\beta \gamma \exp[-\gamma]/\text{Factorial}[z]}} \\
\text{\texttt{\]}} \\
\text{\texttt{\lambda[\gamma - (p\alpha + (1 - p)\beta)]}}} \\
\text{\texttt{(* Take partial derivatives to get system of equations*)}} \\
\text{\texttt{lagrangeEquations = D[\Phi, \#] \[\rightarrow\] 0 & /@ \{\alpha, \beta, \gamma, \Lambda, p\} // \text{FullSimplify};}} \\
\text{\texttt{(* Remove }\Lambda\text{from the set of equations*)}} \\
\text{\texttt{lagrangeequations = \text{Rest}[lagrangeequations] /. First@Solve[First@lagrangeEquations, \Lambda] // \text{FullSimplify};}} \\
\text{\texttt{(* Remove }\gamma\text{from the set of equations*)}} \\
\text{\texttt{lagrangeEquations = lagrangeEquations[[1, 2, 4]] /. First@Solve[lagrangeEquations[[3]], \gamma];}} \\
\text{\texttt{(* Solve for the remaining variables, giving the desired result.*)}} \\
\text{\texttt{soln = Solve[lagrangeEquations, \{\alpha, \beta, p\}] // \text{FullSimplify}}}
\]

D.2. Bias

Throughout this section, if an integral is performed or a series is summed without explanation, the reader can assume it was done using Mathematica. For a variate \((x, y, z)\) of \((X, Y, Z)\) \(|\alpha, \beta\) consider the estimator
\[ \hat{p}_{\text{MLE}, \epsilon} = \frac{z - y}{x - y + \epsilon} \]  

(D5)

for \( p = \frac{\gamma - \beta}{\alpha - \beta} \) where \( \epsilon \) is any non-integer. We are ultimately interested in the limiting case as \( \epsilon \to 0 \) since this gives the MLE. We wish to compute the bias of \( \hat{p}_{\text{MLE}, \epsilon} \), which is given by

\[ \text{Bias}[\hat{p}_{\text{MLE}, \epsilon}] = E_{x,y,z}[\hat{p}_{\text{MLE}, \epsilon} - p] \]

(D6)

\[ = \sum_{z=0}^{\infty} \sum_{y=0}^{\infty} \sum_{x=0}^{\infty} \frac{z - y}{x - y + \epsilon} \alpha^x e^{-\alpha} \beta^y e^{-\beta} \gamma^z e^{-\gamma} - p. \]

This triple sum is not straightforward to compute. Our strategy is to first sum over \( z \) and \( x \) resulting in

\[ E_{x,y,z}[\hat{p}_{\text{MLE}, \epsilon}] = \sum_{y=0}^{\infty} \left[ \frac{\gamma e^{-\beta \sum_{y=\epsilon}^{y} (\gamma - \beta)(\epsilon - \gamma)} }{\Gamma(y + 1)} \right] - \sum_{y=0}^{\infty} \left[ \frac{\beta e^{-\beta \sum_{y=\epsilon}^{y} (\gamma - \beta)(\epsilon - \gamma)} }{\Gamma(y + 1)} \right] \int_{0}^{\infty} t^2 e^{-\alpha t} dt \]

where \( \Gamma(x) = \int_{0}^{\infty} t^{x-1} e^{-t} dt \) is the gamma function, and \( \Gamma(x, y) = \int_{y}^{\infty} t^{x-1} e^{-t} dt \) is the upper incomplete gamma function. Note that the non-integer value of \( \epsilon \) 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,

\[ \Gamma(x-y, -\alpha) = \int_{0}^{\infty} t^{-\alpha} e^{-(x-y)t} dt \]

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

\[ \sum_{y=0}^{\infty} \frac{\gamma e^{-\beta \sum_{y=\epsilon}^{y} (\gamma - \beta)(\epsilon - \gamma)} }{\Gamma(y + 1)} \int_{0}^{\infty} t^2 e^{-\alpha t} dt \]

and similarly

\[ \sum_{y=0}^{\infty} \frac{\beta e^{-\beta \sum_{y=\epsilon}^{y} (\gamma - \beta)(\epsilon - \gamma)} }{\Gamma(y + 1)} \int_{0}^{\infty} t^2 e^{-\alpha t} dt \]

for the last term. Therefore

\[ E_{x,y,z}[\hat{p}_{\text{MLE}, \epsilon}] = \int_{-1}^{\infty} (-1)! e^{-\alpha (1 + t) - \beta (1 + 1/t)} (\gamma t^2 - 1 + \beta t^2) dt \]

(D11)

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 \( t = e^{\phi} \) we are left with

\[ E_{x,y,z}[\hat{p}_{\text{MLE}, \epsilon}] = \text{re} \int_{0}^{\pi} i(-1)! e^{-(\alpha (1 + e^{\phi}) - \beta (1 + e^{i\phi})) (\gamma e^{i\phi} + \beta e^{i(\gamma - 1)} d\phi \]

(D12)

from which we conclude

\[ E_{x,y,z}[\hat{p}_{\text{MLE}, \epsilon}] = \lim_{\epsilon \to 0} E_{x,y,z}[\hat{p}_{\text{MLE}, \epsilon}] \]

\[ = \text{re} \int_{0}^{\pi} i(-\alpha (1 + e^{\phi}) - \beta (1 + e^{i\phi})) (\gamma e^{i\phi} + \beta e^{i(\gamma - 1)} d\phi \]

\[ = \int_{0}^{\pi} e^{-(\alpha (1 + e^{\phi}) - \beta (1 + e^{i\phi})) (\gamma + \beta \cos \phi) \sin((\alpha - \beta) \sin \phi)) + (\beta \sin \phi) \cos((\alpha - \beta) \sin \phi)d\phi. \]

(D13)

From this expression we can see that \( E_{x,y,z}[\hat{p}_{\text{MLE}, \epsilon}] \) is exactly linear with respect to \( \gamma \), 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 \( \phi \) decreases from \( \pi \). Since it will usually hold that \( \alpha + \beta \gg 2 \), the integrand will have most of its support in a region close to \( \pi \). This justifies a low order approximation which will usually be more practical.

We see that the integrand falls off exponentially with rate \( \alpha + \beta \) as \( \phi \) decreases from \( \pi \). Since it will usually hold that \( \alpha + \beta \gg 2 \), the integrand will have most of its support in a region close to \( \pi \). This justifies a low order approximation which will usually be more practical.
series expansion of trigonometric functions about $\phi = \pi$, giving

$$
\mathbb{E}_{x,y}[\hat{P}_{\text{MLE}}] \approx \int_0^\pi e^{-(\alpha + \beta)\phi^2/2} \{ - (\gamma - \beta) \sin((\alpha - \beta)(\phi - \pi)) - (\phi - \pi) \beta \cos((\alpha - \beta)(\phi - \pi)) \} \, d\phi
$$

$$
= \int_0^\pi e^{-(\alpha + \beta)\phi^2/2} \{ - (\gamma - \beta) \sin((\alpha - \beta)\phi) - \phi \beta \cos((\alpha - \beta)\phi) \} \, d\phi
$$

$$
\approx \int_{-\infty}^0 e^{-(\alpha + \beta)\phi^2/2} \{ - (\gamma - \beta) \sin((\alpha - \beta)\phi) - \phi \beta \cos((\alpha - \beta)\phi) \} \, d\phi
$$

$$
= \frac{\gamma - \beta}{\alpha - \beta} + \left( \frac{\gamma - \beta}{\alpha - \beta} - \frac{\beta}{\alpha + \beta} \right) \left( \frac{\alpha - \beta}{\sqrt{2} \alpha + \beta} F \left( \frac{\alpha - \beta}{\sqrt{2} \alpha + \beta} \right) - 1 \right)
$$

$$
P + \left( p - \frac{\beta}{\alpha + \beta} \right) f \left( \frac{\alpha - \beta}{\sqrt{2} \alpha + \beta} \right)
$$

(D14)

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

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

(D15)

We see that the estimator is unbiased at the single point $p = \frac{\alpha}{\alpha + \beta}$ 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.

**Appendix E. Bias corrected estimator**

We can use equation (69) to attempt to derive an estimator which is less biased than the MLE by subtracting off an estimate of the bias. This gives

$$
\hat{P}_{\text{BCE}} = \hat{P}_{\text{MLE}} + \left( \hat{P}_{\text{MLE}} - \frac{\beta}{\alpha + \beta} \right) \frac{\hat{\alpha} + \hat{\beta}}{(\hat{\alpha} - \hat{\beta})^2}
$$

$$
= \frac{z((x - y)^2 - x - y) - y((x - y)^2 - 2x - 2y)}{(x - y)^3}.
$$

(E1)

However, recall from section 6.6 that the Cramér–Rao bound gives us the lower bound

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

(E2)

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 $\mathcal{O}(\alpha + \beta)^{-1}$, well contained within a single standard deviation of $\hat{P}_{\text{MLE}}$, $\mathcal{O}(\alpha + \beta)^{-2}$.

Moreover, the estimate of the bias $\left( \hat{P}_{\text{MLE}} - \frac{\beta}{\alpha + \beta} \right) \frac{\hat{\alpha} + \hat{\beta}}{(\hat{\alpha} - \hat{\beta})^2}$ 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.

**Appendix F. Conjugate priors for drift 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.

**F.1. Uncorrelated conjugate prior**

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

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

(F1)
the product of Gamma distributions will be a conjugate prior. Indeed, with the prior

\[ \pi(\alpha, \beta) = \pi(\alpha)\pi(\beta) = \text{pdf}_{\text{Gamma}}(\alpha; a_\alpha, b_\alpha) \cdot \text{pdf}_{\text{Gamma}}(\beta; a_\beta, b_\beta), \]

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

\[ \pi^*(\alpha, \beta) = \Pr[\alpha, \beta|x, y] = \frac{\Pr|x, y|\alpha, \beta, \pi(\alpha, \beta)}{\int \Pr|x, y|\alpha, \beta, \pi(\alpha, \beta) d\alpha d\beta} = \text{pdf}_{\text{Gamma}}(\alpha; a_\alpha + x, b_\alpha + 1) \cdot \text{pdf}_{\text{Gamma}}(\beta; a_\beta + y, b_\beta + 1). \]

This convenient fact means that if we describe our knowledge of the references \((a_\alpha, b_\alpha, a_\beta, b_\beta)\), 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 \(\text{Gamma}(a, b)\) 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 \((\mu_\alpha, \sigma^2_\alpha, \mu_\beta, \sigma^2_\beta)\) which give posterior hyperparameters

\[ a_\alpha^* = x + \frac{\mu_\alpha^2}{\sigma^2_\alpha}, \quad b_\alpha^* = 1 + \frac{\mu_\alpha^2}{\sigma^2_\alpha}, \quad a_\beta^* = y + \frac{\mu_\beta^2}{\sigma^2_\beta}, \quad b_\beta^* = 1 + \frac{\mu_\beta^2}{\sigma^2_\beta}, \]

or, in terms of mean and variance,

\[ \mu_\alpha^* = \frac{\mu_\alpha^2 + \sigma^2_\alpha x}{\mu_\alpha + \sigma^2_\alpha}, \quad \sigma^2_\alpha = \frac{\sigma^2_\alpha (\mu_\alpha^2 + \sigma^2_\alpha x)}{(\mu_\alpha + \sigma^2_\alpha)^2}, \quad \mu_\beta^* = \frac{\mu_\beta^2 + \sigma^2_\beta y}{\mu_\beta + \sigma^2_\beta}, \quad \sigma^2_\beta = \frac{\sigma^2_\beta (\mu_\beta^2 + \sigma^2_\beta y)}{(\mu_\beta + \sigma^2_\beta)^2}. \]

\textbf{F.2. Correlated conjugate prior}

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

To this end we consider a bivariate Poisson model inspired by equation (44). In that equation it is clear that variations in \(\rho\) will cause correlations between \(\alpha\) and \(\beta\). A bivariate Poisson random variable is defined as \((A, B) \sim BP(\theta_0, \theta_1, \theta_2)\), where \(A = C_0 + C_1\) and \(B = C_0 + C_2\), with \(C_i \sim \text{Poisson} (\theta_0), i = 0, 1, 2\). This produces the probability density

\[ \text{pdf}_{\text{BP}}(x, y; \theta_0, \theta_1, \theta_2) = \frac{e^{-(\theta_0 + \theta_1 + \theta_2)} \theta_0^x \theta_1^y \theta_2^{min(x,y)} \Gamma(x+y)}{x!y! \Gamma(x) \Gamma(y) \Gamma(\theta_0) \Gamma(\theta_1) \Gamma(\theta_2)} \cdot \frac{\theta_0^{\theta_0} \theta_1^{\theta_1} \theta_2^{\theta_2}}{\theta_0^{\theta_0} \theta_1^{\theta_1} \theta_2^{\theta_2}}. \]

This distribution has marginal distributions \(A \sim \text{Poisson} (\theta_0 + \theta_1)\) and \(B \sim \text{Poisson} (\theta_0 + \theta_2)\), and covariance \(\text{Cov}(A, B) = \theta_0\). As discovered by Karlis and Tsiamyrtzis [49], it has an exact family of conjugate priors given by the mixture distributions

\[ \sum_{j=0}^{\infty} w_j G(\theta_0; a_j, b_j) \cdot G(\theta_1; a_1 - j, b_1) \cdot G(\theta_2; a_2 - j, b_2), \]

where \(G\) is the probability density of the gamma distribution, \(a_j, b_j > 0, r \in \{0, 1, 2, \ldots\}, 0 \leq w_j \leq 1\), and \(\sum_{j=0}^{\infty} w_j = 1\).

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 \((\mu_\alpha, \sigma^2_\alpha, \mu_\beta, \sigma^2_\beta, \sigma_{\alpha\beta})\).

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

\[ \pi(\alpha, \beta, \theta_0) = G(\alpha - \theta_0; a_1, b_1) \cdot G(\beta - \theta_0; a_2, b_2) \cdot G(\theta_0; a_0, 1), \]

where

\[ a_1 = \frac{(\mu_\alpha - \sigma_{\alpha\beta})^2}{\sigma^2_\alpha - \sigma_{\alpha\beta}} \quad b_1 = \frac{\mu_\alpha - \sigma_{\alpha\beta}}{\sigma^2_\alpha - \sigma_{\alpha\beta}} \]

\[ a_2 = \frac{(\mu_\beta - \sigma_{\alpha\beta})^2}{\sigma^2_\beta - \sigma_{\alpha\beta}} \quad b_2 = \frac{\mu_\beta - \sigma_{\alpha\beta}}{\sigma^2_\beta - \sigma_{\alpha\beta}} \]

\[ a_0 = \sigma_{\alpha\beta}. \]
It then holds that, for example,

\[
\mathbb{E}[\alpha^2] = \int_0^\infty d\theta_0 \int_0^\infty d\alpha \int_0^\infty d\beta \alpha^2 \pi(\alpha, \beta, \theta_0)
\]

\[
= \int_0^\infty d\theta_0 \int_0^\infty d\theta_0 \theta_0^2 + \theta_1^2 + 2\theta_0 \theta_1) G(\theta_0; a_0, b_1) \cdot G(\theta_0; a_0, 1)
\]

\[
= \left( \frac{a_0}{\Gamma} + \frac{a_0^2}{\Gamma^2} \right) + \left( \frac{a_1}{\beta_1} + \frac{a_1^2}{\beta_1^2} \right) + \frac{2}{b_1} a_0
\]

\[
= \sigma^2_\alpha + \mu^2_\alpha
\]

so that \(\text{Var}[\alpha] = \sigma^2_\alpha\). Similarly we get \(\mathbb{E}[\beta] = \mu_\beta\), \(\text{Var}[\alpha] = \sigma^2_\alpha\), \(\text{Var}[\beta] = \sigma^2_\beta\), and \(\text{Cov}(\alpha, \beta) = \sigma_{\alpha\beta}\).

Since this prior is conjugate to the likelihood function of \((X, Y) \sim \text{BP}(\theta_0, \alpha - \theta_0, \beta - \theta_1)\), if we receive the i.i.d. data \((x_i, y_i), \ldots, (x_n, y_n)\) sampled from this distribution, the posterior will have the form of equation (F7) 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 \(a_0^* = a_0, a_1^* = a_1 + x, a_2^* = a_2 + y, b_1^* = b_1 + n\) for \(i = 0, 1, 2\), which is very similar to the uncorrelated case of the previous section. The new weights \(w_k^*\) 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 \(0 \leq k \leq S_n\) we have \(w_k = \tilde{p}_k / \sum_{k=0}^S \tilde{p}_k\) where

\[
\tilde{p}_k = \sum_{k=0}^{S_n} \sum_{r=\max(0, k-s_i)}^{s_i} c_k^{(n)} c_r^{(n-1)} \Gamma(a_0 + k + \chi) \Gamma(a_2 - k + y) \Gamma(a_0 + k + (n + b_1)(n + b_2))
\]

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

\[
c_k^{(n)} = \frac{\min(k, s_i)}{\max(k, s_i)} \Gamma(a_2^{(n)} k^{(n-1)}),
\]

where \(w_k^{(n)} = (x_{n-r})!(x_{n-r}!)^{-1} k^{(n)} = k^{(1)} , \text{ and } s_i^{(n)} = \min(s_{n+1}, s_{n-1})\). For generality, we have provided these formulas given \(n\) random samples, when in practice, because of the 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

\[
w_k^* = \frac{x!y!(1 + (x_0 + x) \sin(\pi (x_1 + x)) \sin(\pi (x_2 + y)) \Gamma(k + \alpha_0) \Gamma(x - k + \alpha_0) \Gamma(y - k + \alpha_2)}{\pi^2 k!(x - k)! (y - k)! \Gamma(\alpha_0 + 1) \Gamma(\alpha_2 + 1) \sqrt{2}} \left( 1 - x, -y, \alpha_0, \{1 - x - \alpha_0, 1 - y - \alpha_2, \frac{1}{2} (\beta_1 + 1) (\beta_2 + 1) \right) \right)
\]

\[
\times \left( \frac{(1 + b_1)(1 + b_2)}{2} \right)^k,
\]

where \(\tilde{p}_k^{e}\) is the regularized generalized hypergeometric function and \(\tilde{p}_k^{f}\) is the generalized hypergeometric function. However, the only factors in F13 which are relevant to numerical implementations are those which involve \(k\). Everything else can be implicitly calculated by demanding the normalization \(\sum_{k=0}^S w_k^{*} = 1\). An example of posterior mixture weights is shown in figure F1.

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

\[
\pi^*(\alpha, \beta, \theta_0) = \sum_{j=0}^{S_n} w_k^{*} G(\theta_0; a_0 + j, b_0 + n) \cdot G(\alpha - \theta_0; a_0 + x - j, b_1 + n)
\]

\[
\cdot G(\beta - \theta_0; a_2 + y - j, b_2 + n).
\]

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 \(w_k^{*}\) have been computed, drawing a sample is essentially the same cost as drawing a sample from three Gamma distributions.

**Appendix G. 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 in section 8.4.

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, $n_{\text{eff}} = \frac{1}{\sum_{n=1}^{N} w_n^2}$, 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 16 000 actual particles. Storing a multi-dimensional distribution (seven dimensions in our 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 160 000 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 $G_1$ 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 [50], 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:

- $s \sim \pi(s)$, (G2a)
- $\alpha, \beta \sim \pi(\alpha, \beta)$, (G2b)
- $X|\alpha \sim \text{Poisson}(\alpha)$, (G2c)
- $Y|\beta \sim \text{Poisson}(\beta)$, (G2d)
- $Z|s, \alpha, \beta \sim \text{Poisson}(p(s)\alpha + (1 - p(s))\beta)$, (G2e)

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 $p = \text{Tr} \rho P_{0s}$, 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 $\text{Pr}(s, \alpha, \beta | x, y, z)$, which, using Bayes’ rule, is proportional to

$$
\text{Pr}(s, \alpha, \beta | x, y, z) \propto \text{Pr}(x, y, z | s, \alpha, \beta) \pi(\alpha, \beta) \pi(s)
$$

As noted in section 7.2, this formula shows that our update consists of first updating $\alpha$ and $\beta$ analytically using a conjugate prior, and subsequently updating $s$, $\alpha$, $\beta$ using $\pi(\alpha, \beta) \pi(s)$ as a prior. The proportionalities allow

**Figure F1.** An example of posterior mixture weights $w_k^*$ for $k = 0, \ldots, S$. Parameters used were $\mu_x = 200$, $\mu_y = 140$, $\sigma_x = 40$, $\sigma_y = 15$, $\sigma_{x,y} = 90$. The prior was updated with a single sample $x = 220$ and $y = 120$. 

I Hincks et al
us to neglect those factors which depend only on \( x, y \), and/or \( z \) but not any of \( s, \alpha, \) or \( \beta \); 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 \( \pi_{x,y}^{\mu_0}(\alpha, \beta) \pi(s) \) and notice that

\[
Pr(z|s, \alpha, \beta) \propto (p(s)\alpha + (1 - p(s))\beta)^2 e^{-p(s)\alpha - (1-p(s))\beta} \\
= \left( \frac{p(s)}{m} + \frac{1 - p(s)}{m} \right)^m e^{-p(s)\alpha} - m(1-p(s)\beta) \\
\propto \text{pdf}_\text{Poisson}\left( \frac{z}{m}; p(s)\frac{\alpha}{m} + (1 - p(s))\frac{\beta}{m} \right)^m
\]

so that the update can be achieved instead with \( m \) Poisson updates with reduced count data \( z/m \) and reference particle locations for \( \alpha \) and \( \beta \) also reduced by \( m \). This is illustrated in figure G2. Note that our fictional data can have fractional photon counts; this is 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.
Appendix H. QHL supplemental figures

In this appendix section a few extra plots are included related to section 8 of the main body. Figure H1 shows the raw (summed) data. Figure H2 shows all two-parameter marginals of the parameter posterior distribution using the entire data-set. Figure H3 shows simulation results due to SMC fits for each of the 10 batches.
Figure H2. 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 H3. 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 (61).

References

[1] Rabitz H A, Hsieh M M and Rosenthal C M 2004 Science 303 1998
[2] Pravia M A, Boulant N, Emerson J, Farid A, Fortunato E M, Havel T F, Martinez R and Cory D G 2003 J. Chem. Phys. 119 9993
[3] Hocker D, Brif C, Grace M D, Donovan A, Ho T-S, Tabbetts K M, Wu R and Rabitz H 2014 Phys. Rev. A 90 062309
[4] Henry M K, Ramanathan C, Hodges J S, Ryan C A, Ditty M J, Lalramme R and Cory D G 2007 Phys. Rev. Lett. 99 220501
[5] Huszar F and Houlsby N M T 2012 Phys. Rev. A 85 052120
[6] Blume-Kohout R 2012 arXiv:1202.5270 [quant-ph]
[7] Granade C, Combes J and Cory D G 2016 New J. Phys. 18 033024
[8] Magesan E, Gambetta J M and Emerson J 2012 Phys. Rev. A 85 042311
[9] Epstein J M, Cross A W, Magesan E and Gambetta J M 2014 Phys. Rev. A 89 062321
[10] Granade C, Ferrie C and Cory D G 2015 New J. Phys. 17 013042
[11] Wallman J and Flammini S T 2014 New J. Phys. 16 103032
[12] Wiebe N, Granade C, Ferrie C and Cory D 2014 Phys. Rev. A 89 042314
[13] Granade C E, Ferrie C, Wiebe N and Cory D G 2012 New J. Phys. 14 103013
[14] Kosut R, Walmsley I A and Rabitz H 2004 arXiv:quant-ph/0411093
[15] Wang I et al 2017 arXiv:1703.05402 [quant-ph]
[16] Doherty M W, Dolde F, Fedder H, Jelezko F,Wrachtrup J, Manson N B and Hollenberg L C J 2012 Phys. Rev. B 85 205203
[17] Balasubramanian G et al 2009 Nat. Mater. 8 383
[18] Gruber A, Drahenstedt A, Tietz C, Fleury L, Wrachtrup J and Borczyskowski C V 1997 Science 276 2012
[19] Jelezko F and Wrachtrup J 2004 J. Phys.: Condens. Matter 16 R1089
[20] Harrison J, Sellars M J and Manson N B 2004 J. Lumine. 107 245
[21] Jelezko F, Gaebel T, Popa I, Domhan M, Gruber A and Wrachtrup J 2004 Phys. Rev. Lett. 93 130501
[22] Dutt M V G, Childress L, Jiang L, Togan E, Maze J, Jelezko F, Zibrov A S, Hemmer P R and Lukin M D 2007 Science 316 1312
[23] Felton S, Edmonds A M, Newton M E, Martineau P M, Fisher D, Twitchen D J and Baker J M 2009 Phys. Rev. B 79 075203
[24] Jiang L et al 2009 Science 326 267
[25] Neumann P, Beck J, Steiner M, Rempp F, Fedder H, Hemmer P R, Wrachtrup J and Jelezko F 2010 Science 329 542
[26] Dreau A, Maze J-R, Lesik M, Roch J-F and Jacques V 2012 Phys. Rev. B 85 134107
[27] Dolde F et al 2011 Nat. Phys. 7 459
[28] Acosta V M, Bauch E, Ledbetter M P, Waxman A, Bouchard L-S and Budker D 2010 Phys. Rev. Lett. 104 070801
[29] Rondin L, Tetienne J-P, Hingant T, Roch J-F, Maletinsky P and Jacques V 2014 Rep. Prog. Phys. 77 056503
[30] Hincks I, Granade C and Cory D G 2017 Source code and data for ‘Statistical Inference with Quantum Measurements: Methodologies for Nitrogen Vacancy Centers in Diamond’ (version a579e540829) https://github.com/ihincks/nv-meas-code
[31] Doherty M W, Manson N B, Delaney P, Jelezko F, Wrachtrup J and Hollenberg L C J 2013 The nitrogen-vacancy colour centre in diamond Phys. Rep. 528 1
[32] Smeltzer B, Childress L and Gali A 2011 New J. Phys. 13 025021
[33] Gali A, Pyta M and Kaxiras E 2008 Phys. Rev. B 77 155206
[34] Robled L, Bernien H, Sar T V D and Hanson R 2011 New J. Phys. 13 025013
[35] Suter D 1997 The Physics of Laser-Atom Interactions Vol 19 1st edn (Cambridge: Cambridge University Press)
[36] Cory D G, Fahmy A F and Havel T F 1997 Proc. Natl Acad. Sci. 94 1634
[37] Taylor J M, Cappellaro P, Childress L, Jiang L, Budker D, Hemmer P R, Yacoby A, Walsworth R and Lukin M D 2008 Nat. Phys. 4 810
[38] Cover T M and Thomas J A 2006 Elements of Information Theory (New York: Wiley)
[39] Hirose M, Aiello C D and Cappellaro P 2012 Phys. Rev. A 86 062320
[40] Shields B, Unterreithmeier Q, De Leon N, Park H and Lukin M 2015 Phys. Rev. Lett. 114 136402
[41] Van Trees H L and Bell K L 1968 Detection Estimation and Modulation Theory. Part 2 2nd edn (New York: Wiley)
[42] Opper M 1998 A Bayesian Approach to On-line Learning (New York: Cambridge University Press) pp 363–78
[43] Granade C, Ferrie C, Hincks I, Casagrande S, Alexander T, Gross J, Kononenko M and Sanders Y 2017 Quantum 1 5
[44] Liu J and West M 2001 Sequential Monte Carlo Methods in Practice ed D Freitas and N Gordon (New York: Springer)
[45] Granade C and Wiebe N 2016 arXiv:1612.00762 [quant-ph, stat]
[46] Beskos A, Crisan D and Jara A 2014 Ann. Appl. Probab. 24 1396
[47] Gelman A, Carlin J B, Stern H S, Dunson D B, Vehtari A and Rubin D B 2013 Bayesian Data Analysis 3rd edn (London: Chapman and Hall)
[48] Jones E, Oliphant E and Peterson P 2001 SciPy: Open Source Scientific Tools for Python
[49] Karlis D and Tsamarias P 2008 Stat. Comput. 18 27
[50] Andrieu C, Doucet A and Holmes N 2010 J. R. Stat. Soc. B 72 269