



HAL
open science

Maximum likelihood estimation in the context of an optical measurement

Anthony Vella, Miguel Alonso

► **To cite this version:**

Anthony Vella, Miguel Alonso. Maximum likelihood estimation in the context of an optical measurement. A Tribute to Emil Wolf, 65, pp.231-311, 2020, Progress in optics, 10.1016/bs.po.2019.11.007 . hal-02554686

HAL Id: hal-02554686

<https://hal.science/hal-02554686>

Submitted on 1 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Maximum Likelihood Estimation in the Context of an Optical Measurement

Anthony Vella¹ and Miguel A. Alonso^{1,2,*}

¹The Institute of Optics, University of Rochester
Rochester, NY 14627, USA

²Aix Marseille Univ., CNRS, Centrale Marseille, Institut Fresnel, UMR 7249, 13397
Marseille Cedex 20, France

*miguel.alonso@fresnel.fr

August 27, 2019

Abstract

Maximum likelihood estimation (MLE) is a widely-used statistical approach for estimating one or more unknown parameters of a probabilistic model based on observed data. In this tutorial, we review MLE's mathematical foundations within the context of measurements of optical intensity distributions. Here, the detection of each photon is treated as a random event, the outcome being the photon's location within a pixelized array. These detected photons accumulate to form an intensity profile. We give a straightforward derivation for the likelihood function and Fisher information matrix (FIM) associated with a measurement. An estimate for the parameter(s) of interest is then obtained by maximizing the likelihood function, while the FIM determines the estimate's uncertainty. These concepts are illustrated with several simple examples involving a small number of pixels, for one and two parameters, revealing interesting properties and practical considerations for optical measurements. Connections are drawn to weak (off-null) measurements.

Keywords: Maximum likelihood, Fisher information, optical measurement, weak value, weak measurement, off-null metrology.

Table of Contents

Foreword	3
1 Introduction	6
2 Overview of MLE: likelihood, Fisher information, and the Cramér-Rao bound	8
3 MLE formalism for an optical measurement	12
4 Comparison to Bayesian statistics	14
5 Off-null and weak measurements	16
6 One-parameter optical MLE examples	20
6.1 Linear dependence on p_1	22
6.2 Weaker linear dependence on p_1	30
6.3 Null and off-null (weak) measurements	36
6.4 Far-from-null (high intensity) measurement	44
7 Two-parameter optical MLE examples	47
7.1 Linear and sinusoidal variations (case 1)	48
7.2 Linear and sinusoidal variations (case 2)	54
7.3 Piecewise linear dependence (nonzero covariance)	60
7.4 Piecewise linear dependence (zero covariance)	65
7.5 Two-parameter off-null measurement	69
7.6 Two-parameter off-null measurement with smaller departure from null	73
8 Concluding remarks	80
A Mathematica code	82
A.1 Syntax and usage	82
A.2 Code	85
A.3 Examples	87
Acknowledgments	88
References	89

Foreword

This chapter is a tribute to the memory of Emil and Marlies Wolf. As the reader undoubtedly knows, Prof. Emil Wolf was a highly influential figure in the field of optics in general, not only for his seminal research on physical optics, but also for his dissemination of knowledge through the three excellent books he wrote or co-wrote [1–3], and for launching the highly prestigious periodical publication in which this article is appearing. Emil had a particularly strong influence on the authors of this chapter, both through his work and through personal interactions. While the topic presented here is not one that we discussed with him specifically, we see this work as a natural consequence of our interactions with Emil. Our motivation for studying maximum likelihood estimation resulted from our research of the metrology of nanostructures based on tailored polarization illumination [4], as well as our study of devices that measure polarization [5]. The polarization of light and the statistical theory of optics, which we largely learned from Emil, are therefore at the heart of our research in this area.

In the following paragraphs we would like to share some reminiscences of the impact that Emil had on each of us.

MAA: I had the great fortune of meeting Emil in the mid-nineties, when as a PhD student I took his course on Advanced Physical Optics. Towards the end of my second year in the PhD program my research advisor, Prof. Greg Forbes, moved permanently to Australia. I eventually also moved there to finish my dissertation, but I first had to complete another semester of coursework. During this semester Emil took me and my friend Ron Gordon under his wing, and even taught his course on Coherence Theory exclusively for us, since he knew it was our only opportunity to take it. Soon after the end of this course I left for Sydney, but I continued to see Emil every now and then during the OSA Annual Meetings. I remember being very excited about buying what I thought was the very first copy to be sold of his book with Leonard Mandel on Coherence and Quantum Optics, which officially went on sale during the Exhibit of the 1995 OSA Meeting in Dallas. (I later found out that three other copies had been sold prior to this meeting, so I only had the fourth copy!) After completing my PhD (where Emil served in my committee and gave me a run for my money during the defense questioning), and after spending a few years as a postdoc in Australia

and as a junior faculty member in Mexico, the opportunity arrived in 2003 for me to return to Rochester as a member of the Faculty of The Institute of Optics. Emil and Marlies were very welcoming to me and my partner Kristel. One day not long after we moved into our first (completely unfurnished) apartment, we heard the bell ring; it was Emil and Marlies, who brought along a table and chairs, some dishes and utensils, and a portable short-wave radio. Some years later, after my first car in Rochester finally succumbed to rust, Emil insisted that I should keep his old Ford Taurus, as he was buying a new car. As anyone who met Emil knows, once he decided to give you something there was no further discussion.

Being based in Rochester gave me the wonderful opportunity of collaborating with Emil and coauthoring a couple of articles with him, but more importantly, of meeting him regularly for lunch. By this time, he had retired from teaching (but most certainly not from research), so his course on coherence was no longer being offered. However, based on his course notes and on his recent research, Emil published a short book on Coherence and Polarization [3] that was perfectly suited for a one-semester graduate course. I used this book as the basis of reading courses I organized a couple of times over the years for my graduate students to learn about this topic. These courses consisted of weekly meetings where the students presented the material. When I told Emil about these courses he immediately asked if he could also participate. He ended up attending most of the meetings, peppering them with insights, clarifications, and anecdotes. Sometimes after the session we would all go for lunch together, and Emil would insist on inviting us all. My last published collaboration with him was based on a video that our colleague Wayne Knox took of his ducks jumping into a pond early in the morning, and how the waves they created illustrated the analogous behavior of light waves known as the Van Cittert-Zernike theorem. We published together a short note in *Physics Today* under the unusual title “Spatial Coherence from Ducks”.

As I write this, I have been based in Europe for the last two years. The day before I left Rochester happened to coincide with Emil’s 94th birthday. Emil’s son Bruno and his wife Laura contacted me that day to ask if I could come to their place for a small picnic to celebrate Emil with a few friends. I was in the middle of moving out of my apartment, but I managed to stop by. I am extremely happy I did, because that was the last time I ever saw Emil, and it was for a happy occasion. By then he was quiet most of the time, not communicating much. However, after an hour or so I told him that unfortunately I had to leave to finish with the move,

to which he responded: “so early?” Needless to say, I stayed until the end of the party.

AV: I had the pleasure of meeting Emil in 2014 when I was a graduate student enrolled in the coherence theory reading course mentioned above. Emil’s work had a large impact on my graduate research, and although our paths only overlapped for that one semester, it was an invaluable glimpse at his storied career and his kindness, generosity, and care for students. Throughout the reading course, the students would take turns presenting material from Emil’s book with Emil himself sitting in the audience – a tall task, for certain. On one occasion, I finished writing a particularly long equation on the whiteboard and Emil, sharp as ever, simply remarked, “I think you forgot something.” Sure enough, I had omitted a subscript “ i ” on one of the many variables in the equation!

1 Introduction

Perhaps the main use of light in modern technology is as a source and vehicle of information. This information can reveal intrinsic aspects of the optical field, or it might have been encoded by someone sending us a message. Light also carries information about the spatial, temporal, and/or chemical structure of the sources that emitted it, the objects that reflected it, or the media through which it traveled. The accurate extraction of this information from optical measurements is therefore essential, and a range of mathematical techniques have been formulated for this purpose, many of which have applications in other contexts as well. One such technique discussed here is maximum likelihood estimation (MLE).

MLE was introduced by R. A. Fisher in the early 20th century as a way to estimate the parameters associated with an observed quantity based on some statistical model [6–8]. Since then, it has been used in a wide range of applications, and several comprehensive monographs on this subject have been published [9–11], including some that focus on signal processing [12–14] and optics [15], but also others devoted to entirely different topics such as social and political sciences [16, 17]. Notably, two particularly useful tutorials have been published within the context of mathematical psychology [18, 19]. MLE is widely used within the optical sciences, where applications include fundamental research on photon statistics [20] and its influence in the formation of images [21–23]. These studies have led to image processing methods applicable to a range of imaging techniques used in astronomical observation, microscopy and optical coherence tomography [24–26]. MLE has also been used extensively in optical metrology techniques not necessarily based on image formation, such as scatterometry [4, 27], polarimetry [28–30], interferometry [31–33], wavefront sensing [34], and spectroscopy [35, 36].

This tutorial concentrates on the application of MLE to measured optical intensity distributions that encode information about a set of unknown physical parameters p_n , grouped here in a vector as $\mathbf{p} = (p_1, \dots, p_N)$. These parameters can take a continuous range of values, and in general they might each have different units. In this context, the goal of MLE is to determine the most likely value of \mathbf{p} from a measurement of the intensity $I(\mathbf{x}; \mathbf{p})$, where \mathbf{x} is typically a one- or two-dimensional variable. For the purpose of this tutorial, the variable \mathbf{x} be taken to correspond to a coordinate in a detector array, which in practice is discretized due to pixelation of the detector.

(Notice, however, that \mathbf{x} could also correspond to another physical parameter such as time or frequency, which would also be discretized in bins for practical purposes.) The treatment shown in this discussion emphasizes the information gained from the shape of I (i.e., its dependence on \mathbf{x}) without regard for the overall intensity (i.e., the total power incident on the detector). One advantage of this approach is that the accuracy of the parameter estimate is not influenced by power fluctuations of the light source, which would otherwise be especially problematic when operating under low-light conditions, as discussed further in Section 6.3. Particular attention is paid to a type of measurement, referred to in the engineering literature as “off-null” and in the context of quantum physics as “weak”, which consist on a careful choice of illumination in combination with a filtering process prior to detection, in order to enhance the dependence on the parameters p_n of the shape of the detected intensity. In this work we primarily consider real-valued parameters p_n , although they may be complex in general.

The key concepts of MLE and the related topic of Fisher information are summarized in Section 2 for the case of a discrete random variable that depends on one or more parameters p_n . This situation applies directly to most real-world optical measurements, in which the detector is divided into a discrete pixel array, implying that a measurement consisting of a finite number of photon detections has a finite number of possible outcomes. A mathematical description of this scenario is derived explicitly in Section 3. For context and further insight, the results are then compared in Section 4 to the Bayesian statistical approach employed in Ref. [30]. Section 5 gives a brief overview of off-null/weak measurements. Lastly, Sections 6 and 7 contain a number of examples, corresponding to the retrieval of one and two parameters, respectively, and where for simplicity the detector is assumed to consist only of a sequence of nine pixels. These simplified examples were selected to illustrate not only the procedure of MLE for optical measurements, but also the role of Fisher information in evaluating and optimizing the accuracy of an experiment, and the resulting advantage of certain types of measurement. The Mathematica code for these calculations is provided in the appendix.

2 Overview of MLE: likelihood, Fisher information, and the Cramér-Rao bound

Before discussing its application to an optical measurement, in this section the basic concepts of MLE are reviewed in a general context. Consider a discrete random variable Y , and let $P(y|\mathbf{p})$ denote the probability mass function (PMF) specifying the conditional probability of the outcome $Y = y$ given some vector of parameters \mathbf{p} . The PMF is normalized such that

$$\sum_{y \in \mathcal{Y}} P(y|\mathbf{p}) = 1, \quad (1)$$

where \mathcal{Y} is the set of all possible outcomes of Y . It should be emphasized that the PMF is interpreted as a function of y . That is, given a fixed value of \mathbf{p} , the function $P(y|\mathbf{p})$ provides the probability of each possible outcome y . In a typical measurement, however, we require just the opposite: given an observed value of y , we wish to determine the value of \mathbf{p} that is most likely to have produced the measured outcome. This inverse problem is solved by introducing the likelihood function, defined as¹ $L(\mathbf{p}|y) = P(y|\mathbf{p})$. Although the likelihood function and the PMF appear to be mathematically identical (and indeed they are in their unevaluated symbolic forms), they actually have quite different meanings. In contrast to the PMF, the likelihood function is regarded as a continuous function of \mathbf{p} for some fixed value of y . It is not subject to any normalization condition over \mathbf{p} . Given an observation $Y = y$, $L(\mathbf{p}|y)$ represents the likelihood (relative probability) of a vector \mathbf{p} of candidate parameter values. Accordingly, the maximum likelihood estimate (also abbreviated as MLE) for the unknown parameter values is obtained by determining the value of \mathbf{p} that maximizes $L(\mathbf{p}|y)$. For computational convenience, the log-likelihood function $\ell(\mathbf{p}|y) = \ln L(\mathbf{p}|y)$ is often equivalently maximized instead.

Next, consider the related problems of (1) evaluating the uncertainty of a maximum likelihood estimate and (2) designing an experiment for optimal sensitivity. These problems both pertain to the Fisher information, which quantifies the amount

¹Often, the likelihood is used to describe of a set of measurements $\mathcal{S} = (y_1, y_2, \dots)$, in which case it could be denoted as $L(\mathcal{S}|\mathbf{p})$. In this discussion, the notation $L(\mathbf{p}|y)$ is used with the understanding that y could represent either a single measurement or an ensemble of measurements, e.g., an optical intensity distribution, which is a collection of many individual photon detection events.

of information about \mathbf{p} that is contained within a measurement of Y . For the case of N parameters, the Fisher information matrix (FIM) $\mathbb{J}(\mathbf{p})$ is defined as the $N \times N$ symmetric, positive semi-definite matrix with elements

$$[\mathbb{J}(\mathbf{p})]_{mn} = \mathbb{E} \left[\left(\frac{\partial}{\partial p_m} \ell(\mathbf{p}|y) \right) \left(\frac{\partial}{\partial p_n} \ell(\mathbf{p}|y) \right) \right] \quad (2a)$$

$$= \sum_{y \in \mathcal{Y}} \left(\frac{\partial}{\partial p_m} \ell(\mathbf{p}|y) \right) \left(\frac{\partial}{\partial p_n} \ell(\mathbf{p}|y) \right) L(\mathbf{p}|y), \quad (2b)$$

where \mathbb{E} denotes the expectation value over \mathcal{Y} . Under mild regularity conditions [37], the FIM is equivalently defined as²

$$[\mathbb{J}(\mathbf{p})]_{mn} = -\mathbb{E} \left[\frac{\partial^2}{\partial p_m \partial p_n} \ell(\mathbf{p}|y) \right] \quad (3a)$$

$$= -\sum_{y \in \mathcal{Y}} \left(\frac{\partial^2}{\partial p_m \partial p_n} \ell(\mathbf{p}|y) \right) L(\mathbf{p}|y). \quad (3b)$$

Since $\mathbb{J}(\mathbf{p})$ represents the information contained in a single observation of the random variable Y , it is sometimes called the *unit* Fisher information. If the measurement is repeated for T independent trials, it can be shown that the total information obtained is $T\mathbb{J}(\mathbf{p})$. Note that while the Fisher information is a function of the true parameter values \mathbf{p} , it is independent of y . This indicates that $\mathbb{J}(\mathbf{p})$ is not a property of an individual measurement, but rather of the measurement scheme (and its expected outcome). For this reason, $\mathbb{J}(\mathbf{p})$ is often referred to as the *expected* Fisher information. Some texts also define the *observed* Fisher information $\mathbb{J}^{(\text{obs})}(\mathbf{p}; y)$ associated with a particular measured outcome y by dropping the expectation values from Eqs. (2a) and (3a) and evaluating at the maximum likelihood estimate for \mathbf{p} . There has been debate regarding the conditions under which it is more appropriate to use the observed or expected Fisher information [38, 39]. In the asymptotic limit of a large number of observations, it can be shown that the two definitions are equivalent [40].

The statistical significance of the FIM is that its inverse $\mathbb{J}^{-1}(\mathbf{p})$ places a lower

²To prove this result, one can expand the derivatives in Eq. (3b) using the chain rule and product rule. This produces the RHS of Eq. (2b) plus an additional term $-\sum_{y \in \mathcal{Y}} \frac{\partial^2}{\partial p_m \partial p_n} L(\mathbf{p}|y) = -\frac{\partial^2}{\partial p_m \partial p_n} \sum_{y \in \mathcal{Y}} L(\mathbf{p}|y)$. By Eq. (1), the sum over $L(\mathbf{p}|y)$ is equal to 1, so its derivative is zero. The “regularity conditions” for this proof essentially require that $L(\mathbf{p}|y)$ is twice differentiable and that the order of summation and differentiation can be swapped. In practice, these conditions are met in all but the most pathological cases.

limit on the covariance matrix $\mathbb{C}(\mathbf{p})$ for a maximum likelihood estimate of \mathbf{p} . More precisely, for any unbiased estimator³, the Cramér-Rao bound [11] states that the matrix $\mathbb{C} - \mathbb{J}^{-1}$ must be positive semi-definite, i.e., for any vector \mathbf{p} ,

$$\mathbf{p}^T \mathbb{C} \mathbf{p} \geq \mathbf{p}^T \mathbb{J}^{-1} \mathbf{p}. \quad (4)$$

The diagonal elements $[\mathbb{J}^{-1}]_{nn}$ provide the minimum variance of each parameter p_n , while the off-diagonal elements $[\mathbb{J}^{-1}]_{mn}$ (where $m \neq n$) represent the expected covariances between parameters p_m and p_n . The uncertainty of the measurement can be visualized as an ellipsoid in N -dimensional parameter space (centered at the MLE) representing the standard deviation confidence interval. The principal axis orientations of the ellipsoid are given by the eigenvectors of \mathbb{J}^{-1} , and the semi-axis lengths are the square roots of the corresponding eigenvalues [42]. Four examples are illustrated in Table 1 for the case of a two-parameter measurement in which the true parameter values for p_1 and p_2 are both zero. Since \mathbb{J}^{-1} is a function of \mathbf{p} , in general the size and shape of the error ellipsoid also varies over the parameter space. This dependence can be visualized for the two-parameter case (or a 2D slice of a higher-dimensional parameter space) by plotting a grid of ellipses over a selection of parameter values, as seen in Section 7 and in Ref. [4].

In summary, the Cramér-Rao lower bound can be used to assess the minimum expected error of a maximum likelihood estimate based on the inverse of the expected Fisher information matrix for the measurement. In a similar manner, the FIM can be used to predict and optimize the accuracy of an experiment before any measurements are taken. This is done by minimizing a suitable merit function (chosen based on the desired relative accuracies of each parameter) over the range of interest of \mathbf{p} . It is often convenient to reparametrize \mathbf{p} to be unitless, such that the intervals $-1 \leq p_n \leq 1$ (for $n = 1, \dots, N$) correspond to each physical parameter's range of interest.⁴ Then one reasonable choice for the merit function would be the product of the eigenvalues of \mathbb{J} , which is inversely proportional to the square root of the area (for two parameters) or volume/hypervolume (for three or more parameters) of the error ellipsoid. Another option is the root mean square (RMS) of the eigenvalues of \mathbb{J}^{-1} , which is half of the

³In general, the MLE can be biased. However, it is asymptotically unbiased for a sufficiently large sample size [41]. The form of the Cramér-Rao bound given in Eq. (4) only applies when the MLE is unbiased.

⁴One of the advantages of MLE is that it is invariant to the choice of parametrization [11].

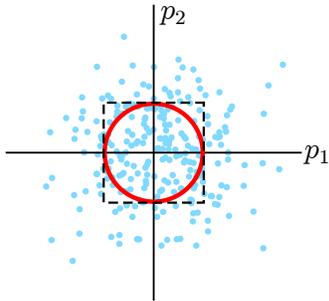
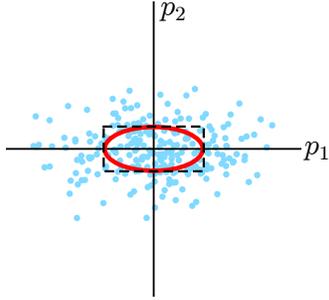
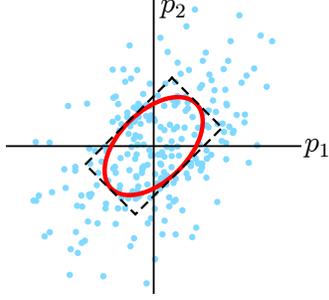
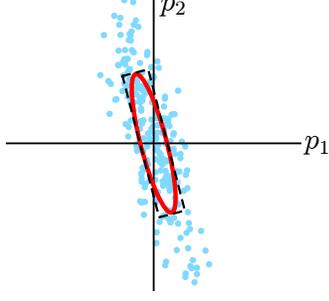
\mathbb{J}^{-1}	Eigenvalues	Eigenvectors	Error ellipse
$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	1, 1	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}$	
$\begin{bmatrix} 1 & 0 \\ 0 & 0.2 \end{bmatrix}$	1, 0.2	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}$	
$\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$	1.5, 0.5	$\begin{bmatrix} 0.71 \\ 0.71 \end{bmatrix}, \begin{bmatrix} 0.71 \\ -0.71 \end{bmatrix}$	
$\begin{bmatrix} 0.2 & -0.5 \\ -0.5 & 2 \end{bmatrix}$	2.13, 0.07	$\begin{bmatrix} -0.25 \\ 0.97 \end{bmatrix}, \begin{bmatrix} 0.97 \\ 0.25 \end{bmatrix}$	

Table 1: Plots of the error ellipses associated with four different 2×2 Fisher information matrices. The square roots of the eigenvalues of \mathbb{J}^{-1} determine the semi-axis lengths of the ellipse, i.e., the dimensions of the bounding rectangle, while the eigenvectors determine the orientation. The blue points in each plot represent the estimated parameters from 250 observations of the random variable Y (assuming a bivariate normal distribution) given true parameter values $p_1 = p_2 = 0$. In these examples, p_1 and p_2 are taken to be unitless, and they are plotted over the range $-3 \leq p_1, p_2 \leq 3$.

diagonal length of the rectangle/box containing the ellipse/ellipsoid.

3 MLE formalism for an optical measurement

The MLE formalism is now applied to the optical measurement described previously, in which one or more parameters \mathbf{p} are to be estimated from a measurement of an intensity distribution $I(\mathbf{x}; \mathbf{p})$. The functional form of $I(\mathbf{x}; \mathbf{p})$ (not to be confused with the measured intensity $\tilde{\mathbf{I}}$ defined below) is generally obtained from either a theoretical model, simulated data, experimental calibration data, or some combination thereof. Suppose that the detector is discretized into a finite number of pixels $i = 1, 2, \dots$ centered at coordinates \mathbf{x}_i , and assume the pixels are sufficiently small so that $I(\mathbf{x}; \mathbf{p})$ is nearly constant over the area of one pixel. Then, given some vector of true parameter values \mathbf{p} , the probability that a single incident photon will hit the detector at pixel i is prescribed by the normalized intensity distribution:

$$P(i|\mathbf{p}) = \frac{I(\mathbf{x}_i; \mathbf{p})}{\sum_i I(\mathbf{x}_i; \mathbf{p})}, \quad (5)$$

where the sum is taken over all pixels.⁵ This equation represents the PMF for a single detected photon. Notice that in this context, the outcome of a measurement (denoted as y in the previous section) is the pixel i where a photon is detected. For a classical measurement, each photon detection can be considered as an independent event, so the probability of M photons hitting pixels i_1, \dots, i_M is given by the product

$$P(i_1 \cap \dots \cap i_M | \mathbf{p}) = \prod_{m=1}^M P(i_m | \mathbf{p}). \quad (6)$$

Now consider a measured intensity $\tilde{\mathbf{I}} = (\tilde{I}_1, \tilde{I}_2, \dots)$, where \tilde{I}_i is the number of photons detected at pixel i . Since the detector is indifferent to the order in which photons arrive (i.e., photons are indistinguishable), the probability of obtaining this

⁵This approximation for small pixels is acceptable for most applications involving sensors with dense pixel arrays. For large pixels, however, one should instead use the exact expression $P(i|\mathbf{p}) = \langle I \rangle_i / \sum_i \langle I \rangle_i$, where $\langle I \rangle_i$ is the integral of $I(\mathbf{x}; \mathbf{p})$ over the area of pixel i . For experiments in which the expected intensity distribution is obtained from a set of calibration images (which themselves are discretized), Eq. (5) is an exact result.

distribution is

$$P(\tilde{\mathbf{I}}|\mathbf{p}) = P_0 \prod_i P(i|\mathbf{p})^{\tilde{I}_i}, \quad (7)$$

where the leading factor $P_0 = (\sum_i \tilde{I}_i)! / \prod_i \tilde{I}_i!$ accounts for all possible permutations. When regarded as a function of \mathbf{p} , the right-hand side of Eq. (7) represents the likelihood function $L(\mathbf{p}|\tilde{\mathbf{I}})$. The log-likelihood is therefore given by

$$\ell(\mathbf{p}|\tilde{\mathbf{I}}) = \ln P_0 + \sum_i \tilde{I}_i \ln P(i|\mathbf{p}). \quad (8)$$

Since P_0 is a constant, the maximum likelihood estimate for \mathbf{p} is obtained by maximizing the sum in the second term of this expression. As described in Section 2, the inverse of the Fisher information matrix places a lower bound on the covariance matrix for this estimate. The expected FIM for a single photon can be calculated using Eq. (2) or (3), with y replaced by the pixel index i specifying the photon's location. For a measurement of \mathcal{N} photons, the total information is⁶

$$[\mathcal{N}\mathbb{J}(\mathbf{p})]_{mn} = \mathcal{N} \sum_i P(i|\mathbf{p}) \left(\frac{\partial}{\partial p_m} \ln P(i|\mathbf{p}) \right) \left(\frac{\partial}{\partial p_n} \ln P(i|\mathbf{p}) \right) \quad (9a)$$

$$= -\mathcal{N} \sum_i P(i|\mathbf{p}) \left(\frac{\partial^2}{\partial p_n \partial p_m} \ln P(i|\mathbf{p}) \right). \quad (9b)$$

On the other hand, the observed FIM associated with a particular measurement $\tilde{\mathbf{I}}$ is obtained by summing the derivatives of $\ell(\mathbf{p}|\tilde{\mathbf{I}})$ over all detected photons:

$$[\mathbb{J}^{(\text{obs})}(\mathbf{p}; \tilde{\mathbf{I}})]_{mn} = \sum_i \tilde{I}_i \left(\frac{\partial}{\partial p_m} \ln P(i|\mathbf{p}) \right) \left(\frac{\partial}{\partial p_n} \ln P(i|\mathbf{p}) \right) \quad (10a)$$

$$= -\sum_i \tilde{I}_i \left(\frac{\partial^2}{\partial p_n \partial p_m} \ln P(i|\mathbf{p}) \right). \quad (10b)$$

Since $\tilde{I}_i \approx \mathcal{N}P(i|\mathbf{p})$ when a large number of photons are measured, the expected and observed information converge in the limit as $\mathcal{N} \rightarrow \infty$, in agreement with the claim

⁶Here the FIM is written in terms of the PMF $P(i|\mathbf{p})$ to emphasize the dependence on the normalized intensity distribution. However, the likelihood function $L(\mathbf{p}|i)$ associated with pixel i , which has the same functional form, could also be used. Also, note that in this analysis \mathcal{N} is taken as an integer representing the actual number of measured photons (i.e., the number of photoelectrons registered by the detector), as opposed to the mean or expected number of photons over a particular time interval.

made in the previous section. In practice, they should yield nearly identical results in most applications, with the exception of extreme low-light measurements using single-photon detectors.

In the above analysis, it has been implicitly assumed that the detector is capable of measuring any arbitrary number of photons incident on a pixel, i.e., that it can resolve individual photons. However, most real detectors have a finite bit depth, meaning that they can only resolve some finite number of distinct intensity levels. For example, in an 8-bit sensor, each pixel has an integer readout value between 0 and 255. This discretization of pixel values is analogous to the discreteness of photons; therefore, in this situation, Eqs. (7) through (10) can be used with \tilde{I}_i interpreted as the readout value of pixel i . In the absence of thermal noise or other sources of error, the equivalent “photon count” of the signal from a sensor with finite bit depth must be less than or equal to \mathcal{N} , the actual number of photons incident on the detector. As needed, the effective bit depth of the sensor can be increased by averaging the output signal over multiple exposures. This time-averaging has the added benefit of reducing the impact of electronic shot noise.

4 Comparison to Bayesian statistics

The method of MLE is considered a “frequentist” approach in the sense that it does not assign a probability distribution to the unknown parameter \mathbf{p} , but rather it estimates the value of \mathbf{p} that is most consistent with the observed data. A popular alternative is the Bayesian approach, which is predicated on the calculation of a posterior probability density function (PDF) $P(\mathbf{p}|\tilde{\mathbf{I}})$ describing the probability of every possible value of \mathbf{p} given an observed intensity $\tilde{\mathbf{I}}$. In general, $P(\mathbf{p}|\tilde{\mathbf{I}})$ depends on a prior distribution $P(\mathbf{p})$ as well as the observed intensity. The prior distribution $P(\mathbf{p})$ may be uniformly distributed (i.e., constant), or it may be used to introduce known (or assumed) information about \mathbf{p} before the measurement takes place. For example, in polarimetry measurements in which $\mathbf{p} = (p_1, p_2, p_3)$ represent the normalized Stokes parameters [30], $P(\mathbf{p})$ could be used to incorporate prior knowledge. One such case is that of a measurement where coherent monochromatic illumination is used and the illuminated object does not introduce significant randomness, so the measured light is known to be fully polarized and therefore $P(\mathbf{p}) = 0$ except when $|\mathbf{p}| = 1$.

Using Bayes' theorem, the posterior PDF can be written as

$$P(\mathbf{p}|\tilde{\mathbf{I}}) = \frac{P(\mathbf{p})}{P(\tilde{\mathbf{I}})}P(\tilde{\mathbf{I}}|\mathbf{p}), \quad (11)$$

where the constant term in the denominator, given by

$$P(\tilde{\mathbf{I}}) = \int P(\mathbf{p})P(\tilde{\mathbf{I}}|\mathbf{p})d^N p, \quad (12)$$

ensures the normalization condition $\int P(\mathbf{p}|\tilde{\mathbf{I}})d^N p = 1$. Substituting Eq. (7) into Eq. (11), one obtains

$$P(\mathbf{p}|\tilde{\mathbf{I}}) = \frac{P(\mathbf{p})}{P(\tilde{\mathbf{I}})}P_0 \prod_i P(i|\mathbf{p})^{\tilde{I}_i} \quad (13a)$$

$$= \frac{P(\mathbf{p})}{P(\tilde{\mathbf{I}})}P_0 \exp\left(\sum_i \tilde{I}_i \ln P(i|\mathbf{p})\right). \quad (13b)$$

Notice that $P(\mathbf{p}|\tilde{\mathbf{I}})$ is proportional to the prior distribution times the likelihood. If no prior information is assumed about \mathbf{p} (as is the case for all examples discussed throughout this tutorial), then $P(\mathbf{p})$ is constant and the peak of $P(\mathbf{p}|\tilde{\mathbf{I}})$ coincides with the maximum likelihood estimate for \mathbf{p} . More generally, if $P(\mathbf{p})$ is nonuniform, the two values converge in the limit as $\mathcal{N} \rightarrow \infty$, assuming that $P(\mathbf{p})$ is smooth and nonzero near the true value of \mathbf{p} .

As discussed in Ref. [30], if the measurement is limited by photon noise (as opposed to other noise mechanisms or systematic errors) and \mathcal{N} is large, then $P(\mathbf{p}|\tilde{\mathbf{I}})$ is approximately a narrow, generally anisotropic Gaussian distribution that is maximized by the true parameter values \mathbf{p}_0 :

$$P(\mathbf{p}|\tilde{\mathbf{I}}) \propto \exp\left[-\frac{1}{2}(\mathbf{p} - \mathbf{p}_0)^T \Sigma^{-1}(\mathbf{p} - \mathbf{p}_0)\right]. \quad (14)$$

Here the covariance matrix Σ determines the shape and width of the distribution, and its inverse Σ^{-1} is the Hessian matrix of second derivatives of $\ln P(\tilde{\mathbf{I}}|\mathbf{p})$ evaluated at \mathbf{p}_0 . Recalling the results of the previous sections, one can see that if $P(\mathbf{p})$ is constant, then Σ^{-1} is equal to the observed FIM $\mathbb{J}^{(\text{obs})}(\mathbf{p}_0; \tilde{\mathbf{I}})$, and its expected value (taken over all possible outcomes for $\tilde{\mathbf{I}}$) is the expected FIM $\mathbb{J}(\mathbf{p}_0)$. Intuitively, a measurement with high information content, for which the FIM is large and nearly diagonal, will result in

a narrow posterior distribution $P(\mathbf{p}|\tilde{\mathbf{I}})$, enabling a precise estimate of \mathbf{p} . Thus, even in a Bayesian framework, the maximum likelihood estimate and the Fisher information matrix can both be shown to have clear statistical meanings.

The ideas presented so far will be illustrated in Sections 6 and 7 using a series of simple examples, chosen to highlight specific aspects of measurements with different characteristics. Before this, however, we introduce in the following section the concept of off-null and weak measurements.

5 Off-null and weak measurements

For several applications, it is advantageous to design the measurement setup so that the illumination and filtering are tailored to increase parameter sensitivity, even if this implies a significantly reduced light level being detected. One example is off-null ellipsometry, in which polarization elements before and after the sample are configured to produce a high extinction ratio over the range of interest of the parameter(s) under test [43]. Other common examples include imaging techniques such as phase contrast microscopy [44, 45] and differential interference contrast microscopy [46].

Related ideas were introduced in 1988 by Aharonov, Albert, and Vaidman [47–49] within the context of the quantum theory of measurements. These authors introduced the concepts of *weak value* and *weak measurement* as an alternative to the standard quantum measurement formalism based on *expected values*. The expected value of a standard measurement of an operator \hat{A} is the normalized inner product $\langle\psi|\hat{A}|\psi\rangle/\langle\psi|\psi\rangle$, where the quantum state vector $|\psi\rangle$ is typically normalized so that $\langle\psi|\psi\rangle = 1$. For an observable operator (which must be Hermitian), the eigenvalues and expected value of \hat{A} are real, and the expected value is limited to the range spanned by the eigenvalues. In contrast, in a weak measurement, the system is prepared in an initial state $|i\rangle$ and measured in a final state $|f\rangle$, which are referred to as *preselected* and *postselected* states, respectively. The expected outcome of the measurement, known as the *weak value*, is then $\langle f|\hat{A}|i\rangle/\langle f|i\rangle$. In general, the weak value may be complex, even for a Hermitian operator; furthermore, it is not bounded by the eigenvalues of \hat{A} since its denominator can be made arbitrarily small by appropriate preselection and postselection. This makes it possible to measure a strong signal in response to a very small deviation in a physical parameter — for example, the angular and/or spatial shift of a beam of light upon reflection [50–55].

While framed in notation often associated with quantum measurements, the weak value formalism also applies to classical measurements; in fact, some of its most successful applications can be understood classically [50, 51], and pre-existing techniques such as off-null ellipsometry [43, 56], phase contrast microscopy [44, 45], and differential interference contrast microscopy [46] mentioned earlier, can be interpreted in terms of weak values. In this context, the preselected state describes a specifically tailored illumination and postselection is achieved by a filtering process of the resulting light prior to measuring the signal, where both the illumination tailoring and the detection filtering can exploit the spatial, directional, temporal, spectral, or polarization degrees of freedom of light.

Traditionally, the weak value formalism treats the measurement of a single parameter. However, many interesting problems require the simultaneous measurement of multiple parameters. The focused beam scatterometry experiment in Ref. [4] presented a description of how the weak value formalism can be used for the measurement several parameters \mathbf{p} simultaneously. We now give a brief review of this formalism. While we consider specifically the case of classical light, we use the notation of operators and bra-kets to stress the analogy with quantum measurements. Let $|i\rangle$ denote the carefully chosen field used for illumination, and let $\hat{T}(\mathbf{p})$ describe the passage of light through a system that includes the sample whose parameters \mathbf{p} are being measured. The state of the light field after emerging from the system is then $\hat{T}(\mathbf{p})|i\rangle$, but before being detected this light is made to pass through a filter described by $|f\rangle$. The measured intensity is then

$$I(\mathbf{x}; \mathbf{p}) = \left| \langle f | \hat{T}(\mathbf{p}) | i \rangle \right|^2, \quad (15)$$

where the dependence on the detector variable \mathbf{x} can be due to the illumination, filtering, and/or the operator.

The basic assumption of this formalism is that, within the range of interest for \mathbf{p} , the dependence of $\hat{T}(\mathbf{p})$ on these parameters can be approximated as linear, namely, $\hat{T}(\mathbf{p}) \approx \hat{T}_0 + \mathbf{p} \cdot \hat{\mathbf{T}}$, where without loss of generality we assume that the parameters are defined so that their nominal value is zero. The \mathbf{p} -independent scalar and vector operators \hat{T}_0 and $\hat{\mathbf{T}}$ can be calculated via a Taylor expansion, namely $\hat{T}_0 = \hat{T}(\mathbf{0})$ and $\hat{\mathbf{T}} = \nabla_{\mathbf{p}} \hat{T}(\mathbf{0})$, or they can be estimated by fitting a series of reference measurements.

The substitution of this linear approximation in Eq. (15) gives

$$I(\mathbf{x}; \mathbf{p}) \approx |\langle f | \hat{T}_0 | i \rangle + \mathbf{p} \cdot \langle f | \hat{\mathbf{T}} | i \rangle|^2. \quad (16)$$

Notice that the first term inside the absolute values is independent of the parameters being measured. It is then tempting to think that the sensitivity to the parameters is maximized if we choose the illumination and filtering such that $\langle f | \hat{T}_0 | i \rangle = 0$, since then any measured photon would indicate a departure from zero of the estimated parameters. This would constitute a *null* measurement scheme. The problems with this approach, however, is that the dependence on \mathbf{p} is quadratic, meaning that the signal would be very small for moderate values of these parameters and, perhaps more importantly, that there would be a global sign uncertainty of the parameter estimates. Also, as will be shown later, the values of the parameters in this case cannot be retrieved solely from the shape of the intensity distribution.

To understand the off-null/weak-value approach, let us consider first the case of a single parameter p_1 . Note that the literature on weak values [47–49] typically considers operators of the form $\hat{T}(p_1) = \exp(ip_1 \hat{A})$ so that if a Taylor expansion is used we get $\hat{T}_0 = 1$ and $\mathbf{p} \cdot \hat{\mathbf{T}} = ip_1 \hat{A}$. Equation (16) can then be written as

$$I(\mathbf{x}; p_1) \approx |\langle f | i \rangle + ip_1 \langle f | \hat{A} | i \rangle|^2 = |\langle f | i \rangle|^2 \left| 1 + ip_1 \frac{\langle f | \hat{A} | i \rangle}{\langle f | i \rangle} \right|^2, \quad (17)$$

where we recognize the weak value of \hat{A} in the second term inside absolute values. Recall that this weak value is a function of \mathbf{x} . While in the limit of a null measurement this second term dominates, for a weak measurement this term is often assumed to be considerably smaller than unity, so that one more approximation can be performed:

$$I(\mathbf{x}; p_1) \approx |\langle f | i \rangle|^2 \left| \exp \left(ip_1 \frac{\langle f | \hat{A} | i \rangle}{\langle f | i \rangle} \right) \right|^2. \quad (18)$$

This expression shows that, in order for the detected intensity to provide information about the parameter p_1 , the illumination and filtering must be chosen such that the weak value is imaginary or at least contains an imaginary part. For example, in phase contrast microscopy [44, 45] where the measured parameter p_1 corresponds to the optical thickness of a sample at a given point, the weak value is made imaginary

by placing at a Fourier-conjugate plane of the object/image a transparent mask with a small semitransparent spot at its center, whose thickness is chosen so that the spatial DC component of the field is dephased by $\pi/2$ and attenuated by a factor τ . This small obstacle has a negligible effect on $\langle f|\hat{A}|i\rangle$ but makes $\langle f|i\rangle$ acquire a factor of $i\tau$.

Let us return to the general case in which the system can depend on several parameters and this dependence is not necessarily exponential. It can be seen from Eq. (16) that the dependence of the measured intensity on the parameters can be made strong and unambiguous by ensuring that the two terms inside the absolute values are in phase and of comparable magnitudes over the range of interest for the parameters. To achieve this, it is convenient to write the first term inside absolute values in Eq. (16) in the following form [4]:

$$\langle f|\hat{T}_0|i\rangle = -\bar{\mathbf{p}}(\mathbf{x}) \cdot \langle f|\hat{\mathbf{T}}|i\rangle, \quad (19)$$

where the vector $\bar{\mathbf{p}}(\mathbf{x})$ has as elements functions $\bar{p}_n(\mathbf{x})$ that are real. This way the measured intensity can be written as

$$I(\mathbf{x}; \mathbf{p}) \approx |[\mathbf{p} - \bar{\mathbf{p}}(\mathbf{x})] \cdot \mathbf{T}(\mathbf{x})|^2, \quad (20)$$

where we defined $\mathbf{T}(\mathbf{x}) = [T_1(\mathbf{x}), \dots, T_N(\mathbf{x})] = \langle f|\hat{\mathbf{T}}|i\rangle$, which is in general complex. Note that for the case of only one parameter, the form in Eq. (19) with real \bar{p}_1 can be achieved only through the appropriate choice of the illumination and filtering. For the case of two parameters, on the other hand, it is always possible to find real functions \bar{p}_1 and \bar{p}_2 for any illumination and filtering distributions, according to

$$\bar{p}_{1,2}(\mathbf{x}) = \frac{\text{Im}[T_{2,1}^*(\mathbf{x})T_0(\mathbf{x})]}{\text{Im}[T_{1,2}^*(\mathbf{x})T_{2,1}(\mathbf{x})]}, \quad (21)$$

where $T_0(\mathbf{x}) = \langle f|\hat{T}_0|i\rangle$. However, a good measurement is achieved by tailoring $|i\rangle$ and $|f\rangle$ such that the functions $\bar{p}_n(\mathbf{x})$ are mutually linearly independent (preferably as different from each other as possible), and where each takes values that span the complete range of variations of interest of the corresponding parameters p_n . (Note that it is useful to ensure that T_1 and T_2 are not in phase so that the denominator in Eq. (21) does not vanish; this fact will be used in the examples in Sections 7.5 and 7.6.) This approach was used in Ref. [4] to simultaneously measure the orientation

and duty cycle of a subwavelength grating.

When more than two parameters are measured, real functions $\bar{p}_n(\mathbf{x})$ can always be found that satisfy Eq. (20) for any illumination and filtering, which is now an underdetermined system of equations. However, for any number of parameters, the measurement is optimized by tailoring illumination and filtering distributions so that the standard deviations found from the Fisher information are minimized (according to some metric) over the range of interest of the parameters being measured.

6 One-parameter optical MLE examples

This section contains a series of four simple thought experiments involving one-dimensional intensity distributions $I_j(x; p_1)$ (where the subindex $j = 1, 2, 3, 4$ labels each of these examples) that depend on a single parameter p_1 . Without loss of generality, let us assume that p_1 is unitless and that its range of interest is $-1 \leq p_1 \leq 1$. (As noted on page 10, any physical parameter can be reparametrized in this way without affecting the MLE.) The one-dimensional coordinate x is also taken to be unitless. In the examples that follow, the function

$$\Pi(x) = \begin{cases} I_0, & -1 \leq x \leq 1, \\ 0 & \text{otherwise,} \end{cases} \quad (22)$$

where I_0 represents some reference intensity level, is used as a normalization factor that also serves to limit each intensity distribution to the spatial extent of the sensor (e.g., a beam being truncated by a hard aperture). Each intensity distribution is normalized such that it reaches a maximum value of I_0 over the range of interest of p_1 . Note, however, that this does not preclude the possibility of intensities greater than I_0 when $|p_1| > 1$.

For simplicity, in all examples we consider a the detector consisting of a one-dimensional array of 9 pixels, with pixel i centered at coordinate $x_i = (i - 5)/4$, so that

$$(x_1, \dots, x_9) = (-1, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1). \quad (23)$$

According to Eq. (5), the probability of an incident photon hitting pixel i is

$$P_j(i|p_1) = \frac{I_j(x_i; p_1)}{\sum_i I_j(x_i; p_1)}. \quad (24)$$

As mentioned earlier, for such a sparse array of pixels, this is a relatively poor approximation since the intensity may vary significantly over the width of each pixel. However, since the approximation is reasonable for most real applications, it is used here for instructive purposes. If desired, the exact expression for $P_j(i|p_1)$ (which is provided in footnote 5 following Eq. (5)) could be substituted into the analysis with minimal modifications required. Similarly, while the concepts of Fisher information and the Cramér-Rao bound are usually applied to measurements consisting of many observations (photons), the calculations below are demonstrated for measurements of just a few photons and then extended to larger sample sizes. Also note that while the following examples all involve intensity distributions over a 1D spatial coordinate, the more general 2D case can be treated in the same manner by rearranging the numerical output of the detector’s 2D pixel array into a 1D array during signal processing.

The intensity distributions considered in each of the following sections are summarized in Table 2. In Section 6.1, an in-depth analysis is performed for a simple intensity distribution that depends linearly on p_1 . In Section 6.2, the results are compared to a similar intensity distribution with a weaker linear dependence on p_1 . Next, the commonly-used experimental configurations of null and off-null measurements are explored in Section 6.3. Finally, Section 6.4 examines the case of an intensity that may be far from perfect nulling conditions, and the results are compared to the near-null case.

Section	Intensity distribution
6.1	$I_1(x; p_1) = \Pi(x)(0.5 + 0.5 p_1 x)$
6.2	$I_2(x; p_1) = \Pi(x)(0.9 + 0.1 p_1 x)$
6.3	$I_3(x; p_1) = \Pi(x) \frac{1}{(c + 1)^2} (p_1 - cx)^2$, where $c = \text{constant}$
6.4	$I_4(x; p_1) = \Pi(x) \frac{1}{(d + 2)^2} (p - x - d)^2$, where $d = \text{constant}$

Table 2: Intensity distributions for each example considered in Section 6.

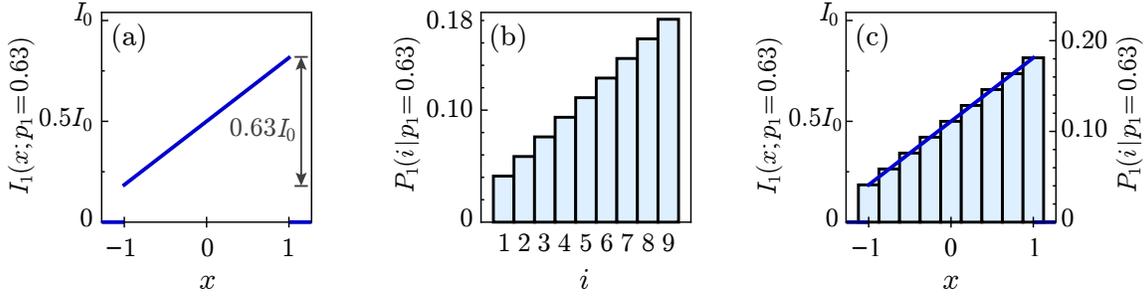


Figure 1: (a) Linear intensity distribution $I_1(x; p_1)$ and (b) the corresponding PMF for each pixel i , both shown for the case that $p_1 = 0.63$. The two plots are shown together in part (c). For practical reasons, the axis labels for i are excluded from the combined plot. In all subsequent figures, the vertical axis labels are also omitted to reduce clutter.

6.1 Linear dependence on p_1

For the first example, consider the intensity distribution

$$I_1(x; p_1) = \Pi(x)(0.5 + 0.5 p_1 x). \quad (25)$$

The distribution is only valid when $-1 \leq p_1 \leq 1$ since larger parameter values would result in negative intensity values, which are not allowed. This is an extreme case of a common real-world scenario in which an approximation is made for the intensity that is only valid over some range of parameter values. In practice, for reliable parameter estimation, the range of interest of \mathbf{p} should be smaller than the region where the approximation is valid (within some prescribed accuracy).

Using Eq. (24), it is straightforward to calculate the PMF for a detected photon:

$$P_1(i|p_1) = \frac{1}{9} \left(1 + \frac{i-5}{4} p_1 \right). \quad (26)$$

The continuous intensity distribution $I_1(x; p_1)$ and discrete PMF $P_1(i; p_1)$ are plotted in Figs. 1(a) and 1(b) for the case that $p_1 = 0.63$. To visualize the relationship between the intensity and PMF, it is useful to combine the two plots with appropriately chosen scales, as seen in Fig. 1(c). The dependence of each quantity on p_1 is illustrated in Fig. 2, which contains plots of $I_1(x; p_1)$ and $P_1(i|p_1)$ for five different parameter values over the range of interest.

As discussed previously, the likelihood function $L_1(p_1|i)$ has the same algebraic

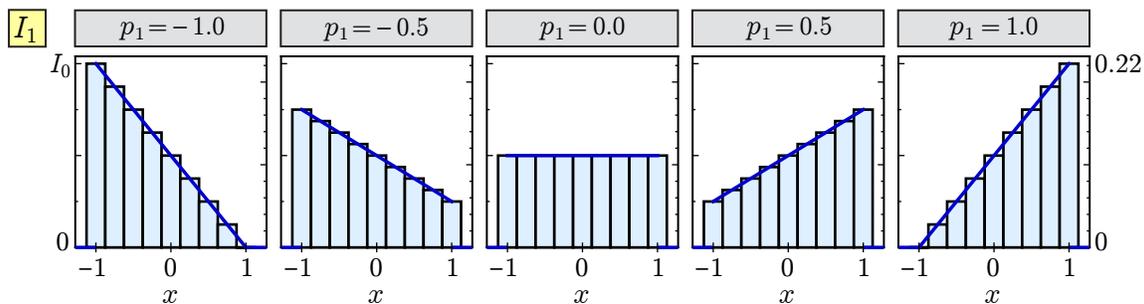


Figure 2: Plots of $I_1(x; p_1)$ (left axis) and $P_1(i|p_1)$ (right axis) for several values of p_1 .

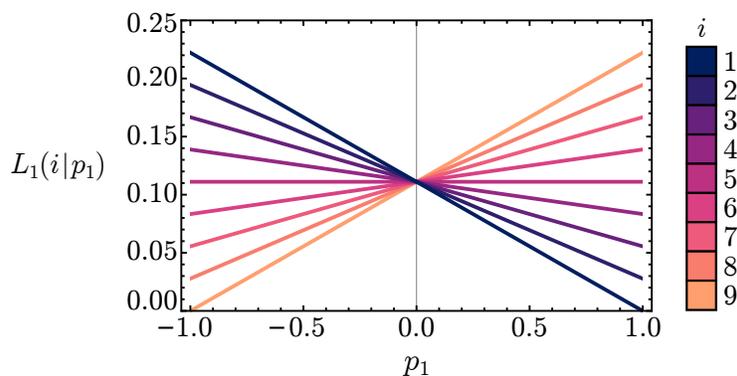


Figure 3: Likelihood functions $L_1(i|p_1)$ associated with each pixel i in a measurement with theoretical intensity distribution $I_1(p_1)$.

form as $P_1(i|p_1)$, but it is regarded as a continuous function of p_1 . The likelihood functions associated with individual photons detected at each pixel $i = 1, \dots, 9$ are plotted in Fig. 3. To illustrate the procedure of calculating the MLE from the likelihood function, let us now consider a simulated measurement of the intensity for which the true parameter value is $p_1 = 0.63$. The simulated intensity $\tilde{\mathbf{I}}$ is constructed by randomly selecting individual photons according to the probability distribution $P(i|p_1 = 0.63)$ that was shown previously in Fig. 1(b). For demonstrative purposes, suppose that the sensor is capable of detecting individual photons, even though this is typically not the case in real experiments where many photons accumulate within the sensor's exposure time. This will allow us to examine the influence of each photon on the likelihood and the MLE, as well as the evolution of the MLE as photons accumulate.

Suppose that the first simulated photon hits the detector at pixel 1. From Eq. (26), the likelihood of this event is found to be $L_1(p_1|i = 1) = \frac{1}{9}(1 - p_1)$. The MLE based

on this single photon is obtained by maximizing the likelihood with respect to p_1 . This example illustrates the fact that the MLE is not guaranteed to exist in general, since $L_1(p_1|i = 1)$ would be unbounded if p_1 were allowed to take any real value. A sufficient condition for the existence of an MLE is that the parameter space is compact [57, 58], such as the closed interval $p_1 \in [-1, 1]$. Within this interval, the likelihood function is maximized by $p_1 = -1$.⁷ Notice from Fig. 3 that a single photon detected at pixel 2, 3, or 4 also would have produced the same MLE, albeit with lower confidence.

Now suppose that a second photon is detected at pixel 7, so that the measured intensity becomes $\tilde{\mathbf{I}} = (1, 0, 0, 0, 0, 0, 1, 0, 0)$. The likelihood function associated with this second photon is $L_1(p_1|i = 7) = \frac{1}{9}(1 - \frac{1}{2}p_1)$. Using Eq. (7) (and remembering that the probability and likelihood are algebraically equivalent), the likelihood of measuring this two-photon intensity distribution is

$$L_1(p_1|\tilde{\mathbf{I}}) = \frac{2!}{1! \times 1!} L_1(p_1|i = 1) L_1(p_1|i = 7) = \frac{1}{81}(-p_1^2 - p_1 + 2). \quad (27)$$

It is easy to show that this function is maximized when $p_1 = -0.5$, which becomes the new MLE. Similarly, suppose that a third photon is detected, also at pixel 7, so that the measured intensity becomes $\tilde{\mathbf{I}} = (1, 0, 0, 0, 0, 0, 2, 0, 0)$. The likelihood of measuring this intensity distribution is

$$L_1(p_1|\tilde{\mathbf{I}}) = \frac{3!}{1! \times 2!} L_1(p_1|i = 1) L_1(p_1|i = 7)^2 = \frac{1}{972}(-p_1^3 - 3p_1^2 + 4), \quad (28)$$

which is maximized when $p_1 = 0$.

The likelihood functions for individual photons at pixels 1 and 7 are plotted in Fig. 4(a), as well as the likelihoods of the two- and three-photon intensity distributions from above. The latter two functions are also plotted separately in Fig. 4(b,c). From these plots one can see the effect of each photon: as photons are detected at pixel 1, then pixel 7, then pixel 7 again, the peak of the likelihood function shifts from $p_1 = -1$ to $p_1 = -0.5$ to $p_1 = 0$. Additionally, the distribution becomes more sharply peaked with each accumulated photon, reducing the uncertainty in the MLE.

⁷Note that the condition of compactness is sufficient but not necessary. In fact, in the present example, the restriction quickly becomes unnecessary as soon as multiple photons are detected at different pixels. Another example is that of polarimetry, in which the Stokes parameters are restricted to the interval $[-1, 1]$ by definition, guaranteeing the existence of an MLE.

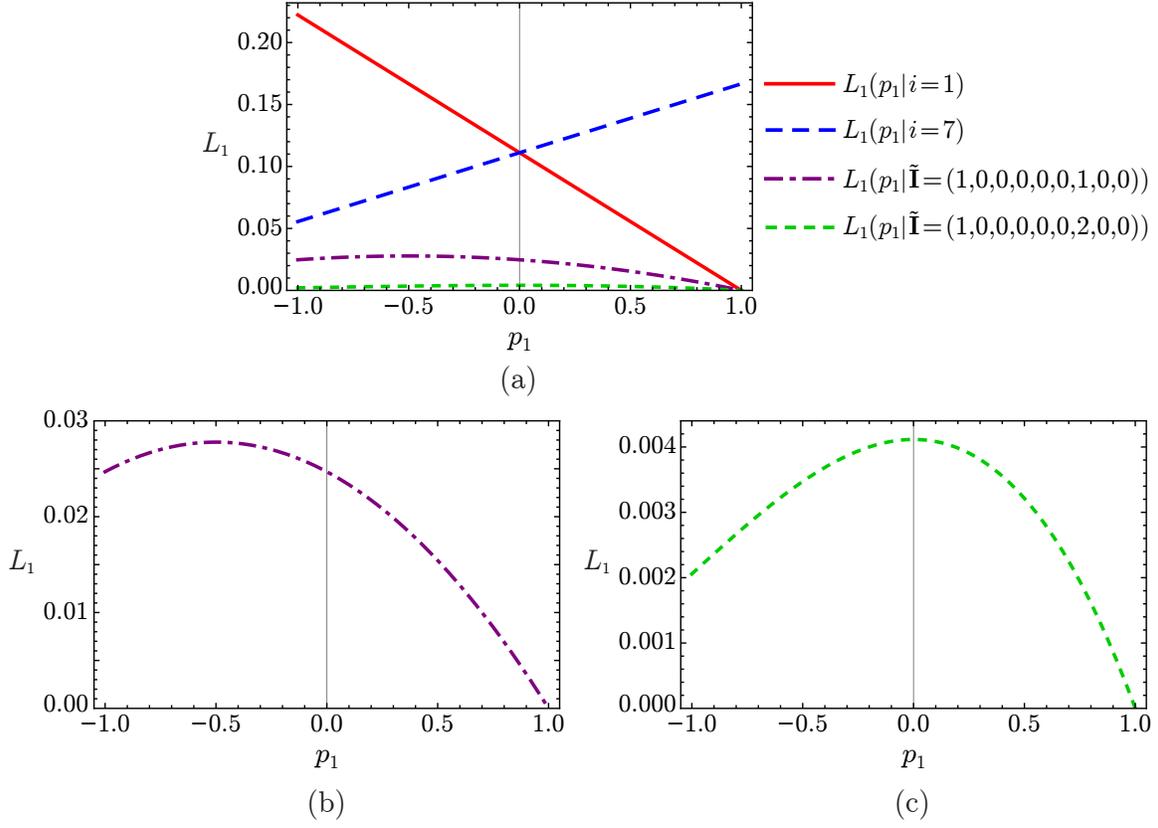


Figure 4: (a) Likelihood functions (based on intensity distribution I_1) for detected photons at pixels $i = 1$ and $i = 7$ and for intensity measurements consisting of one photon at pixel 1 and one or two photons at pixel 7. The two- and three-photon likelihoods are also shown on independent scales in plots (b) and (c).

This uncertainty can be quantified by using Eq. (10) to calculate the observed Fisher information, which is a 1×1 “matrix” (i.e., a scalar) in the one-parameter case. For example, for the three-photon measurement $\tilde{\mathbf{I}} = (1, 0, 0, 0, 0, 0, 2, 0, 0)$, Eq. (10a) yields

$$\begin{aligned}
 J_1^{(\text{obs})}(p_1; \tilde{\mathbf{I}}) &= \sum_i \tilde{I}_i \left(\frac{\partial}{\partial p_1} \ln P(i|p_1) \right)^2 \\
 &= \left(\frac{i-5}{4+(i-5)p_1} \Big|_{i=1} \right)^2 + 2 \left(\frac{i-5}{4+(i-5)p_1} \Big|_{i=7} \right)^2 \\
 &= \frac{1}{(p_1-1)^2} + \frac{2}{(p_1+2)^2}, \tag{29}
 \end{aligned}$$

which produces $J_1^{(\text{obs})} = 1.5$ when evaluated at the MLE $p_1 = 0$. In the one-parameter

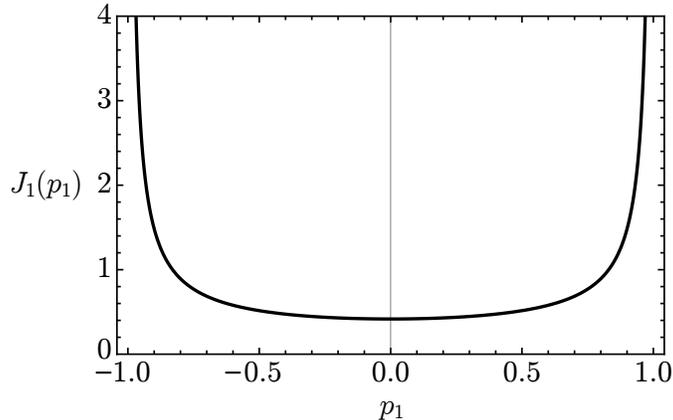


Figure 5: Expected unit Fisher information for a measurement of $I_1(x; p_1)$.

case, the eigenvalue of the “matrix” $J_1^{(\text{obs})}$ is just the value of $J_1^{(\text{obs})}$ itself. Therefore, the minimum expected standard deviation uncertainty of the measurement is $1/\sqrt{1.5} = 0.816$. Considering the fact that only three photons were detected, this large uncertainty (relative to the range of interest) is not surprising.

Alternatively, using Eq. (9a), the minimum error for a measurement of \mathcal{N} photons (independent of the specific outcome of the measurement) can be quantified by calculating the expected Fisher information

$$\mathcal{N}J_1(p_1) = \frac{\mathcal{N}}{36} \sum_{i=1}^9 \frac{(i-5)^2}{4 + (i-5)p_1}. \quad (30)$$

For example, for a three-photon measurement with MLE $p_1 = 0$, the expected standard deviation error is $[3J_1(0)]^{-1/2} = 0.894$. Keep in mind, however, that the expected Fisher information is not necessarily appropriate for a measurement containing very few photons. As seen in Fig. 5, $J_1(p_1)$ grows infinitely large in the limit that $|p_1| \rightarrow 1$, implying that the uncertainty approaches zero. Although this is a meaningful limit for the case of large \mathcal{N} , it would clearly be nonsensical to suggest that a single photon could produce an MLE with zero uncertainty!

To observe these concepts on a larger scale, suppose that the simulation continues until 100,000 photons have accumulated. For a single random trial of the experiment, Table 3 contains the measured intensities and corresponding MLEs obtained throughout the simulation for several values of \mathcal{N} . Notice that the MLE approaches the true parameter value ($p_1 = 0.63$) as \mathcal{N} increases. As seen in Fig. 6, the log-

\mathcal{N}	MLE (p_1)	$\tilde{\mathbf{I}} = (\tilde{I}_1, \dots, \tilde{I}_9)$
1	-1.0000	(1, 0, 0, 0, 0, 0, 0, 0, 0)
2	-0.5000	(1, 0, 0, 0, 0, 0, 1, 0, 0)
3	0.0000	(1, 0, 0, 0, 0, 0, 2, 0, 0)
4	0.3187	(1, 0, 0, 0, 0, 0, 2, 1, 0)
5	0.5024	(1, 0, 0, 0, 0, 0, 2, 1, 1)
6	0.5429	(1, 0, 0, 0, 0, 1, 2, 1, 1)
7	0.6187	(1, 0, 0, 0, 0, 1, 2, 2, 1)
8	0.6727	(1, 0, 0, 0, 0, 1, 2, 3, 1)
9	0.6916	(1, 0, 0, 0, 0, 2, 2, 3, 1)
10	0.6646	(1, 0, 0, 1, 0, 2, 2, 3, 1)
100	0.7114	(6, 1, 8, 9, 8, 9, 15, 19, 25)
1000	0.6656	(41, 56, 64, 91, 112, 121, 166, 160, 189)
10000	0.6243	(413, 583, 784, 956, 1112, 1262, 1446, 1615, 1829)
100000	0.6329	(4009, 5847, 7696, 9460, 11151, 12839, 14588, 16160, 18250)

Table 3: Evolution of the MLE for p_1 and the measured intensity distribution $\tilde{\mathbf{I}}$ as individual photons accumulate for a simulated measurement of $I_1(x; p_1)$ with true parameter value $p_1 = 0.63$.

likelihood function $\ell_1(p_1|\tilde{\mathbf{I}})$ becomes increasingly narrow as photons accumulate, and its shape becomes approximately parabolic; therefore, the likelihood $L_1(p_1|\tilde{\mathbf{I}})$ approaches a Gaussian distribution, i.e., an exponentiated concave-downward quadratic function. Furthermore, as observed above, the location of the peak likelihood (which by definition determines the MLE) approaches the true parameter value. The MLE is plotted against \mathcal{N} in Fig. 7, with shaded regions representing the standard deviation confidence intervals based on the expected and observed Fisher information. Notice that as \mathcal{N} increases, not only does the MLE approach the true value of p_1 with increasing confidence, but the expected and observed information rapidly converge.

Although the above simulation is a representative example of the behavior of the MLE, it is merely a single observation of a random process. To gain a broader view of the statistical behavior of $I_1(x; p_1)$, a Monte Carlo simulation of 50,000 trials of a 100-photon intensity measurement was performed, first for a true parameter value of $p_1 = 0$ and then for $p_1 = 0.63$. The results of the simulations are plotted in Figs. 8(a) and 8(b), which contain histograms showing the distribution of the MLE over all trials. As seen in the upper left corner of each plot, the mean MLE over all trials differs from

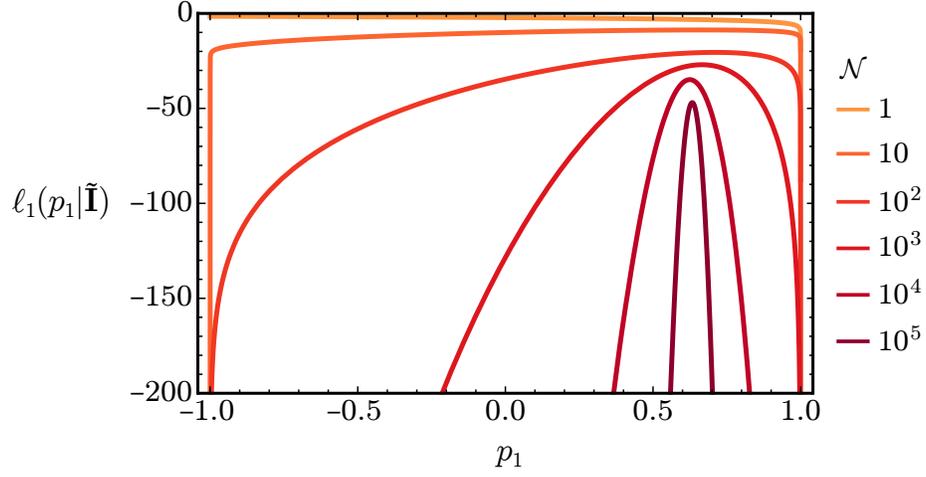


Figure 6: Log-likelihood functions associated with the simulated intensities listed in Table 3.

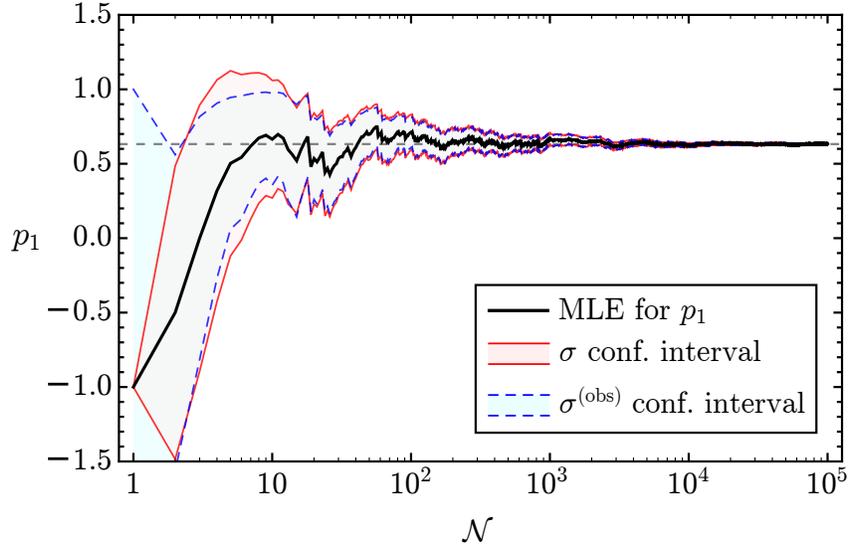


Figure 7: Evolution of the maximum likelihood estimate and standard deviation confidence interval for p_1 as 100,000 photons accumulate for a simulated measurement of $I_1(x; p_1)$ with true parameter value $p_1 = 0.63$. The solid red and dashed blue regions represent the confidence intervals based on the expected and observed Fisher information, respectively.

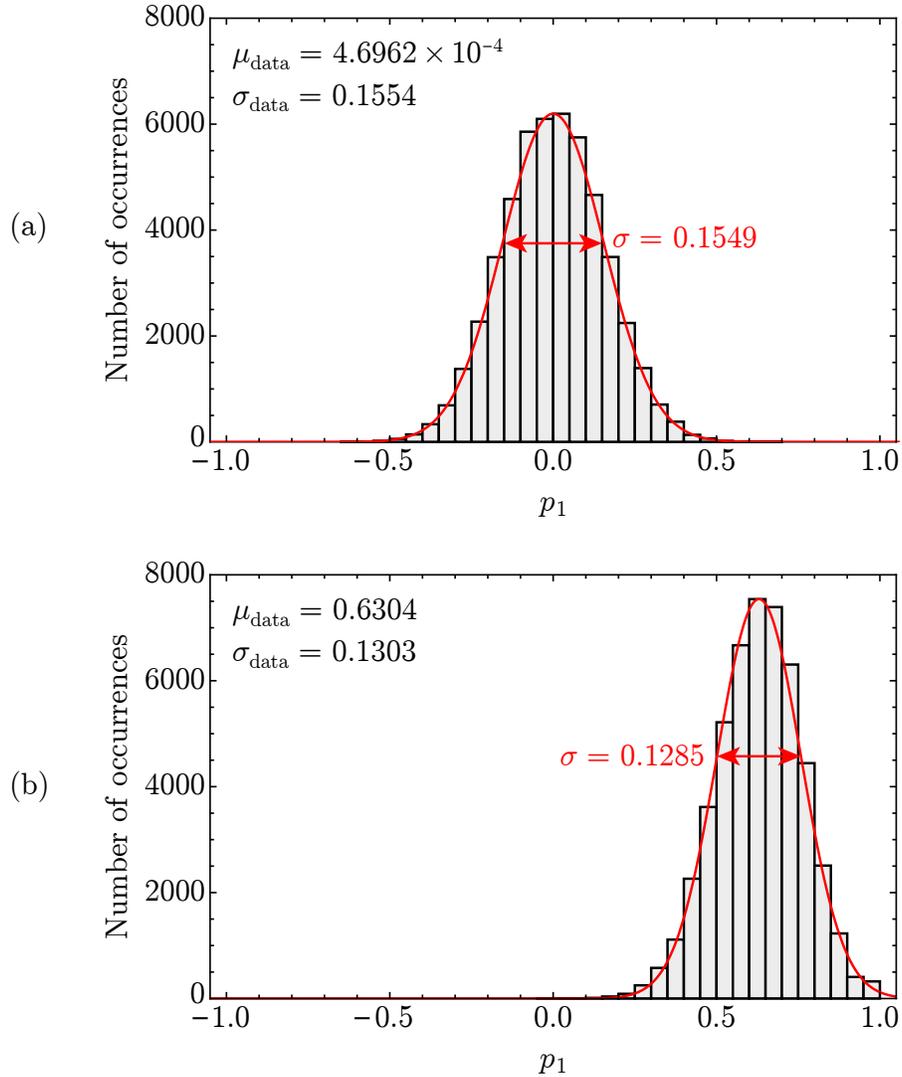


Figure 8: Histograms of the maximum likelihood estimates obtained from 50,000 trials of a 100-photon simulation of $I_1(x; p_1)$ with true parameter values (a) $p_1 = 0$ and (b) $p_1 = 0.63$. The mean (μ_{data}) and standard deviation (σ_{data}) of each distribution are indicated in the upper left corner of the plot. For comparison, a normal distribution with mean p_1 and standard deviation $\sigma = [100J_1(p_1)]^{-1/2}$ is overlaid in red; the value of σ is indicated alongside each curve.

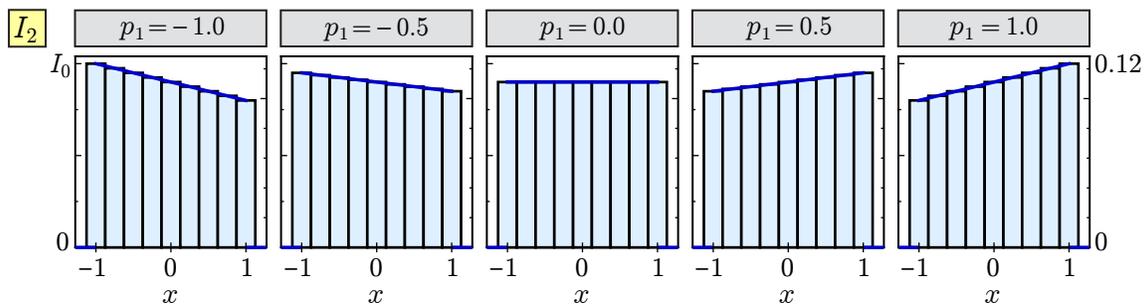


Figure 9: Plots of $I_2(x; p_1)$ (left axis) and $P_2(i|p_1)$ (right axis) for several values of p_1 .

the true parameter value by less than 0.001. The standard deviations of the MLEs obtained for the $p_1 = 0$ and $p_1 = 0.63$ cases are 0.1554 and 0.1303, respectively. In comparison, using Eq. (30), the expected Fisher information for the $p_1 = 0$ case is $100J_1(0) = 41.67$, corresponding to a standard deviation error of 0.1549. Similarly, the expected error for the $p_1 = 0.63$ case is found to be 0.1285. These values closely agree with the results of the simulation. To help visualize this, a normal distribution with the expected standard deviation is overlaid in red on top of each histogram in Fig. 8; notice that each curve almost exactly matches the distribution of MLEs over 50,000 trials.

6.2 Weaker linear dependence on p_1

For the next example, consider the intensity distribution

$$I_2(x; p_1) = \Pi(x)(0.9 + 0.1 p_1 x), \quad (31)$$

which is valid when $-9 \leq p_1 \leq 9$. (However, the range of interest is still $-1 \leq p_1 \leq 1$.) Using Eq. (24), the PMF for a single photon is

$$P_2(i|p_1) = \frac{1}{9} \left(1 + \frac{i-5}{36} p_1 \right). \quad (32)$$

This distribution is nearly the same as the first example except that the linear p_1 term is 9 times smaller. As a result, the variations in intensity, PMF, and likelihood with respect to p_1 have much lower contrast over the range of interest, as seen in Figs. 9 and 10. Analogously to Section 6.1, suppose that we simulate a measurement of

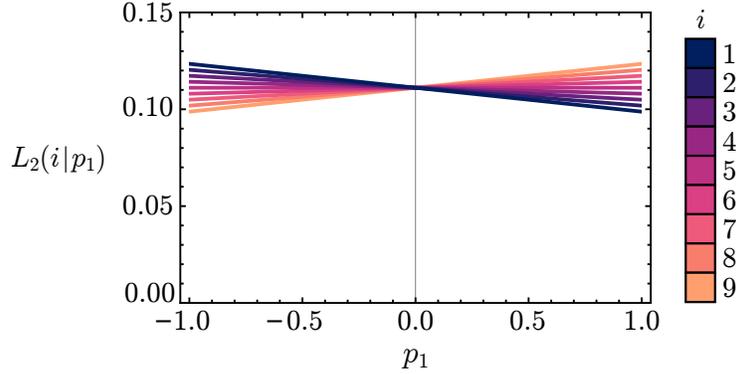


Figure 10: Likelihood functions $L_2(i|p_1)$ associated with each pixel i in a measurement with theoretical intensity distribution $I_2(p_1)$.

$I_2(x; p_1)$ and that the first three photons are again detected at pixels 1, 7, and 7. Following the same procedure as in the previous example, it can be shown that the maximum likelihood estimates after each photon detection are $p_1 = -9, -4.5,$ and 0 . The corresponding likelihood functions, shown in Fig. 11, are nearly flat, which is a sign that the MLE has a large uncertainty. Indeed, for $\tilde{\mathbf{I}} = (1, 0, 0, 0, 0, 0, 2, 0, 0)$, the observed Fisher information is found to be

$$J_2^{(\text{obs})}(p_1; \tilde{\mathbf{I}}) = \frac{1}{(p_1 - 9)^2} + \frac{2}{(p_1 + 18)^2}, \quad (33)$$

which yields $J_1^{(\text{obs})} = 0.0185$ when evaluated at the MLE $p_1 = 0$, corresponding to a standard deviation uncertainty of $1/\sqrt{0.0185} = 7.35$. Similarly, the expected Fisher information

$$\mathcal{N}J_2(p_1) = \frac{\mathcal{N}}{324} \sum_{i=1}^9 \frac{(i - 5)^2}{36 + (i - 5)p_1} \quad (34)$$

for an \mathcal{N} -photon measurement of I_2 is significantly smaller than the information contained in a measurement of I_1 , as shown in Fig. 12. For example, the expected standard deviation error for a three-photon measurement, given by $[3J_2(0)]^{-1/2} = 8.05$, is nine times larger than it was in the previous example. The discrepancy grows even larger as $|p_1|$ increases.

Similarly to the previous section, a 100,000 photon simulation of $I_2(x; p_1)$ was performed, and the results were monitored along the way as photons accumulated. The intensities and corresponding MLEs obtained at several steps throughout the simulation are listed in Table 4, and the MLE and standard deviation confidence

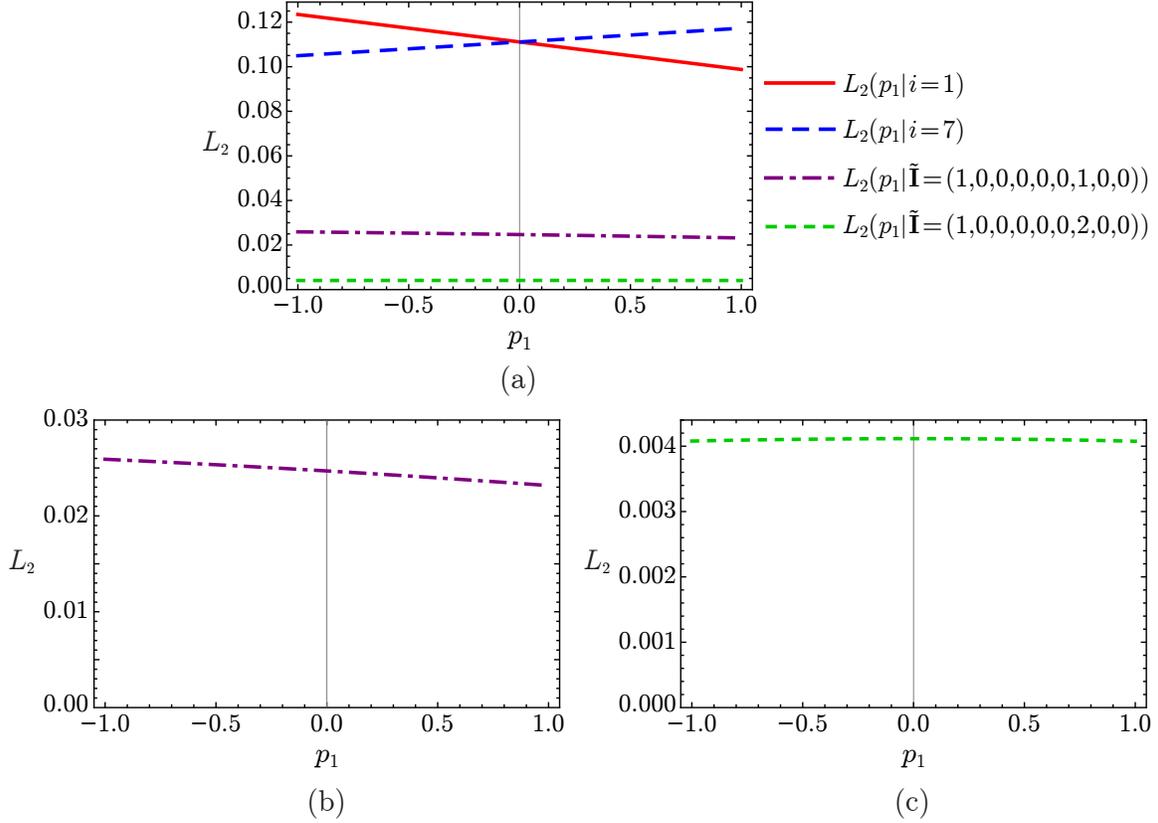


Figure 11: (a) Likelihood functions (based on intensity distribution I_2) for detected photons at pixels $i = 1$ and $i = 7$ and for intensity measurements consisting of one photon at pixel 1 and one or two photons at pixel 7. The two- and three-photon likelihoods are also plotted on independent scales in plots (b) and (c).

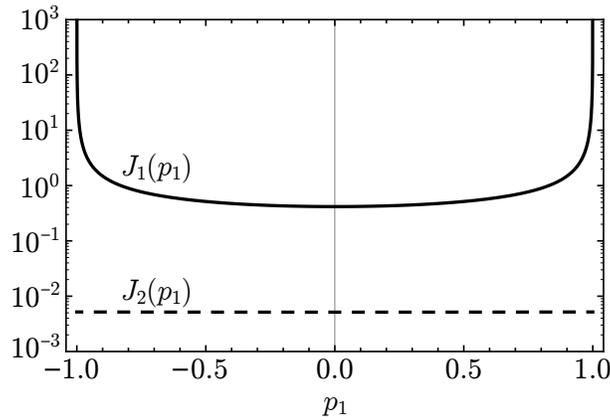


Figure 12: Expected unit Fisher information $J_1(p_1)$ and $J_2(p_1)$ for measurements of $I_1(x; p_1)$ and $I_2(x; p_1)$, respectively, plotted on a logarithmic scale.

\mathcal{N}	MLE (p_1)	$\tilde{\mathbf{I}} = (\tilde{I}_1, \dots, \tilde{I}_9)$
1	-9.0000	(1, 0, 0, 0, 0, 0, 0, 0, 0)
2	-4.5000	(1, 0, 0, 0, 0, 0, 1, 0, 0)
3	0.0000	(1, 0, 0, 0, 0, 0, 2, 0, 0)
4	-3.8285	(1, 1, 0, 0, 0, 0, 2, 0, 0)
5	-3.8285	(1, 1, 0, 0, 1, 0, 2, 0, 0)
6	-2.3629	(1, 1, 0, 0, 1, 1, 2, 0, 0)
7	-5.1192	(1, 2, 0, 0, 1, 1, 2, 0, 0)
8	-5.1192	(1, 2, 0, 0, 2, 1, 2, 0, 0)
9	-6.0605	(1, 2, 0, 1, 2, 1, 2, 0, 0)
10	-4.8152	(1, 2, 0, 1, 2, 2, 2, 0, 0)
100	2.3159	(6, 6, 12, 17, 13, 11, 9, 13, 13)
1000	1.8366	(91, 98, 89, 105, 113, 108, 145, 120, 131)
10000	0.7542	(1000, 1044, 1101, 1077, 1117, 1088, 1204, 1168, 1201)
100000	0.6331	(10278, 10541, 10629, 11026, 11138, 11377, 11438, 11843, 11730)

Table 4: Evolution of the MLE for p_1 and the measured intensity distribution $\tilde{\mathbf{I}}$ as individual photons accumulate for a simulated measurement of $I_2(x; p_1)$ with true parameter value $p = 0.63$.

interval are plotted as a function of \mathcal{N} in Fig. 14. From these results, one can see that the MLE approaches the true parameter value more slowly than in the previous example, with a much larger uncertainty. (Take note of the increased scale of the plot compared to Fig. 7.)

Finally, to complete the comparison to Section 6.1, a Monte Carlo simulation was performed for 50,000 trials of a 1000-photon measurement of $I_2(x; p_1)$. For true parameter values $p_1 = 0$ and $p_1 = 0.63$, the expected standard deviation errors are 0.4409 and 0.4401, respectively. Histograms of the results of each simulation for 50,000 trials are shown in Fig. 15; as indicated on the plots, the standard deviations of the MLEs obtained for each case are 0.4413 and 0.4394, closely matching expectations. Notice that the errors are larger than they were in the previous example (0.1554 and 0.1303) despite the fact that the measured intensity contains ten times as many photons. This is noteworthy because for any value of p_1 , the total power incident on the detector (given by the sum of the intensity over all pixels) is 1.8 times larger for I_2 than it is for I_1 , indicating that on average nearly twice as many photons will be measured within a given exposure time. Even so, based on the above results, we can

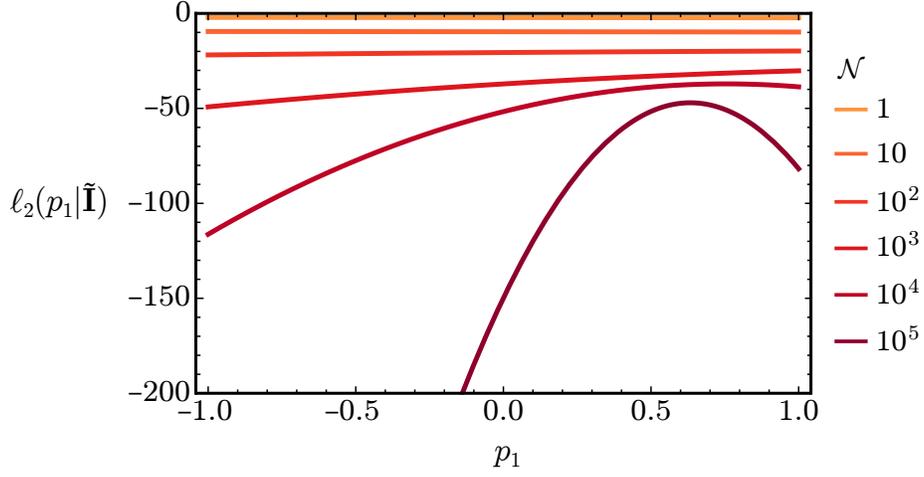


Figure 13: Log-likelihood functions associated with the simulated intensities listed in Table 4.

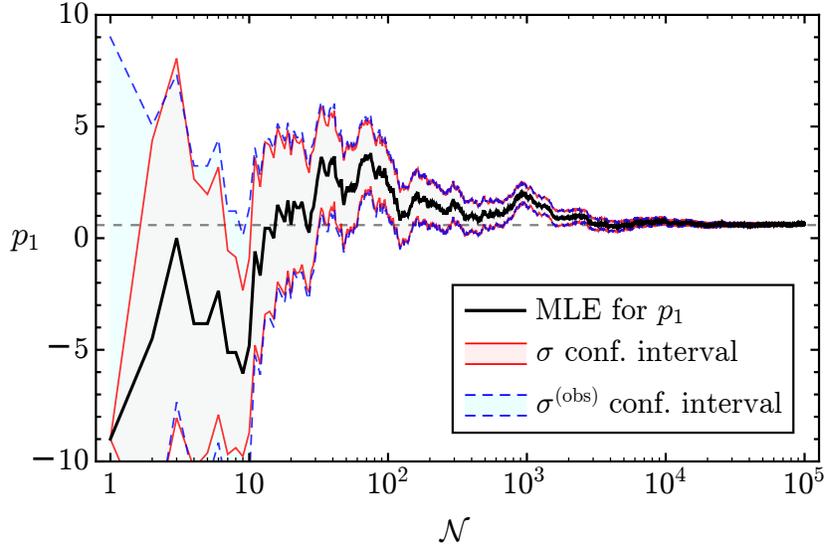


Figure 14: Evolution of the maximum likelihood estimate and standard deviation confidence interval for p_1 as 100,000 photons accumulate for a simulated measurement of $I_2(x; p_1)$ with true parameter value $p_1 = 0.63$. The solid red and dashed blue regions represent the confidence intervals based on the expected and observed Fisher information, respectively.

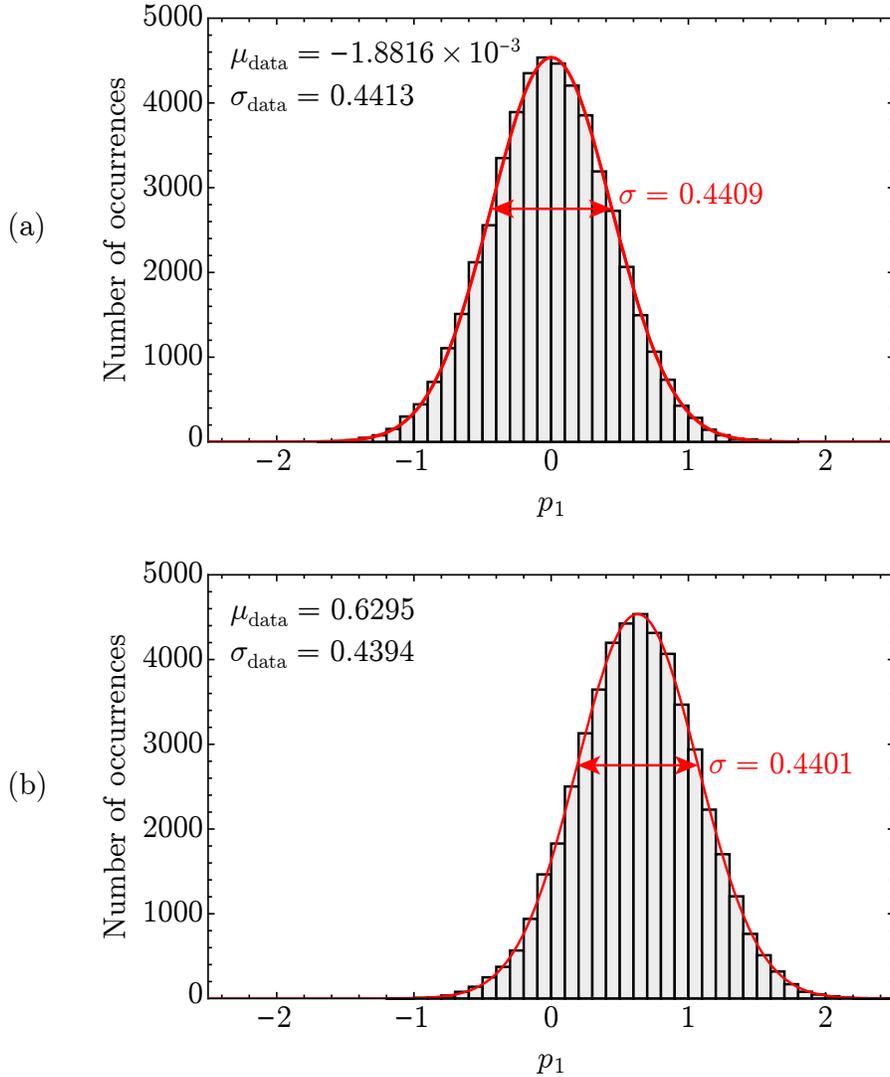


Figure 15: Histograms of the maximum likelihood estimates obtained from 50,000 trials of a 1000-photon simulation of $I_2(x; p_1)$ with true parameter values (a) $p_1 = 0$ and (b) $p_2 = 0.63$. The mean (μ_{data}) and standard deviation (σ_{data}) of each distribution are indicated in the upper left corner of the plot. For comparison, a normal distribution with mean p_1 and standard deviation $\sigma = [1000J_1(p_1)]^{-1/2}$ is overlaid in red; the value of σ is indicated alongside each curve.

conclude that if measurements of I_1 and I_2 were conducted with identical exposure times, then the measurement of I_1 (for which the output signal would contain fewer photons) would be expected to produce a more accurate parameter estimate. This is an important lesson to keep in mind when designing an experiment: the most informative measurement is not always the one with the strongest signal. On the contrary, it can be beneficial to filter out a large fraction of the light before it reaches the detector (e.g., via polarization selection) in such a way that the measured signal contains only the photons emitted from the source that provide the most information about p_1 .⁸ This idea is explored further in the next example.

6.3 Null and off-null (weak) measurements

Let us now consider a simple example representative of null and off-null measurements, for which the measured intensity has the form in Eq. (20) with $\bar{p}_1(x)$ linear in x and $T_1(x)$ constant:

$$I_3(x; p_1) = \Pi(x) \frac{1}{(|c| + 1)^2} (p_1 - cx)^2, \quad (35)$$

where c is a real constant. For $c = 0$, this represents a null measurement for which the (spatially uniform) intensity vanishes when $p_1 = 0$ and increases quadratically with p_1 . For $c \neq 0$, the value of p_1 for zero intensity (i.e., the departure from perfect nulling) varies linearly with the coordinate x . Using Eq. (24), the PMF for a detected photon is found to be

$$P_3(i|p_1) = \frac{(4p_1 - (i - 5)c)^2}{144p_1^2 + 60c^2}. \quad (36)$$

Let us begin by examining the case of perfect nulling ($c = 0$), for which the intensity $I_3(x; p_1) = \Pi(x)p_1^2$ and PMF $P_3(i|p_1) = 1/9$ are plotted in Fig. 16. In contrast to the previous two examples, these plots illustrate that for a given coordinate x_i , the ratio between the measured intensities at two different parameter values need not be the same as the ratio between the corresponding PMF values. In fact, in this example the PMF is the same for all values of p_1 with the exception of $p_1 = 0$, for which it is undefined (due to the fact that no photons are detected). Consequently, the likelihood function is completely flat and the Fisher information is zero, implying that

⁸When possible, it would be preferential to encode information by rearranging the light rather than filtering it out. However, sometimes this is not possible, e.g., when measuring the coupling induced by a scattering process between a pair of specific input and output polarization states.

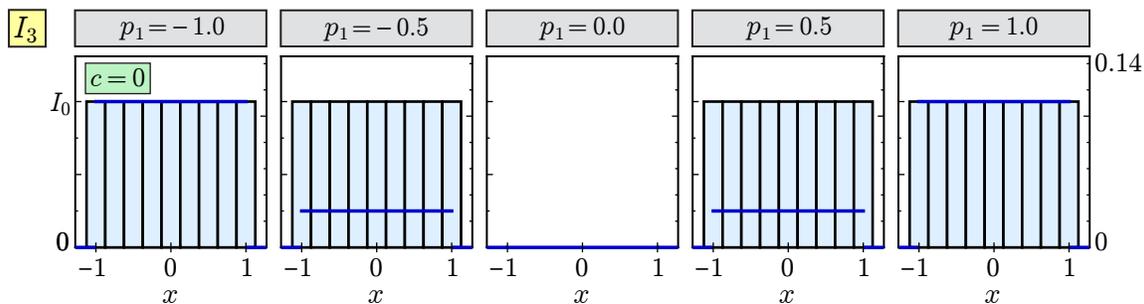


Figure 16: Plots of $I_3(x; p_1)$ (left axis) and $P_3(i|p_1)$ (right axis) for several values of p_1 for the case of perfect nulling.

it is impossible to determine p_1 from the shape of the measured intensity distribution.⁹ (Of course, this is also obvious from the simple fact that the PMF is independent of p_1 .) In this situation, it would only be possible to deduce the value of p_1 from the total optical power incident on the detector, which is beyond the scope of the current statistical approach. Even then, it would only be possible to determine the magnitude of p_1 but not its sign (since I_3 is an even function of p_1), and the measurement would be susceptible to temporal fluctuation errors unless the illumination source power were very stable.

The aforementioned shortcomings of a null measurement can be avoided by designing the experiment to operate under an off-null condition, which corresponds to the choice of some constant $c \neq 0$ in the present example. The intensity and PMF are plotted in Fig. 17 for several positive values of c ; symmetric behavior is observed when c is negative. Notice in each plot that the null in intensity (when one exists within the range of interest) is located at $x = p_1/c$. When $|c| = 1$, the null shifts across the entire width of the sensor as p_1 varies from -1 to 1 , causing the shape of $P_3(i|p_1)$ to vary substantially over the entire parameter range. When $|c| \gg 1$, the null is confined to a narrow region near the center of the sensor, resulting in very little variation in $P_3(i|p_1)$ with respect to p_1 . On the other hand, when $|c| \ll 1$, the null shifts away from the origin very quickly when p_1 is nonzero. This results in dramatic variations in $P_3(i|p_1)$ (and very low intensity levels) when $|p_1|$ is small, but much smaller changes near the edge of the parameter range.

This behavior can also be visualized by plotting the likelihood functions $L_3(i|p_1)$ for each pixel, which are shown in Fig. 18. From the definition of the Fisher in-

⁹In this case, the MLE exists but it is not unique, since all values of p_1 within the range of interest maximize the likelihood function.

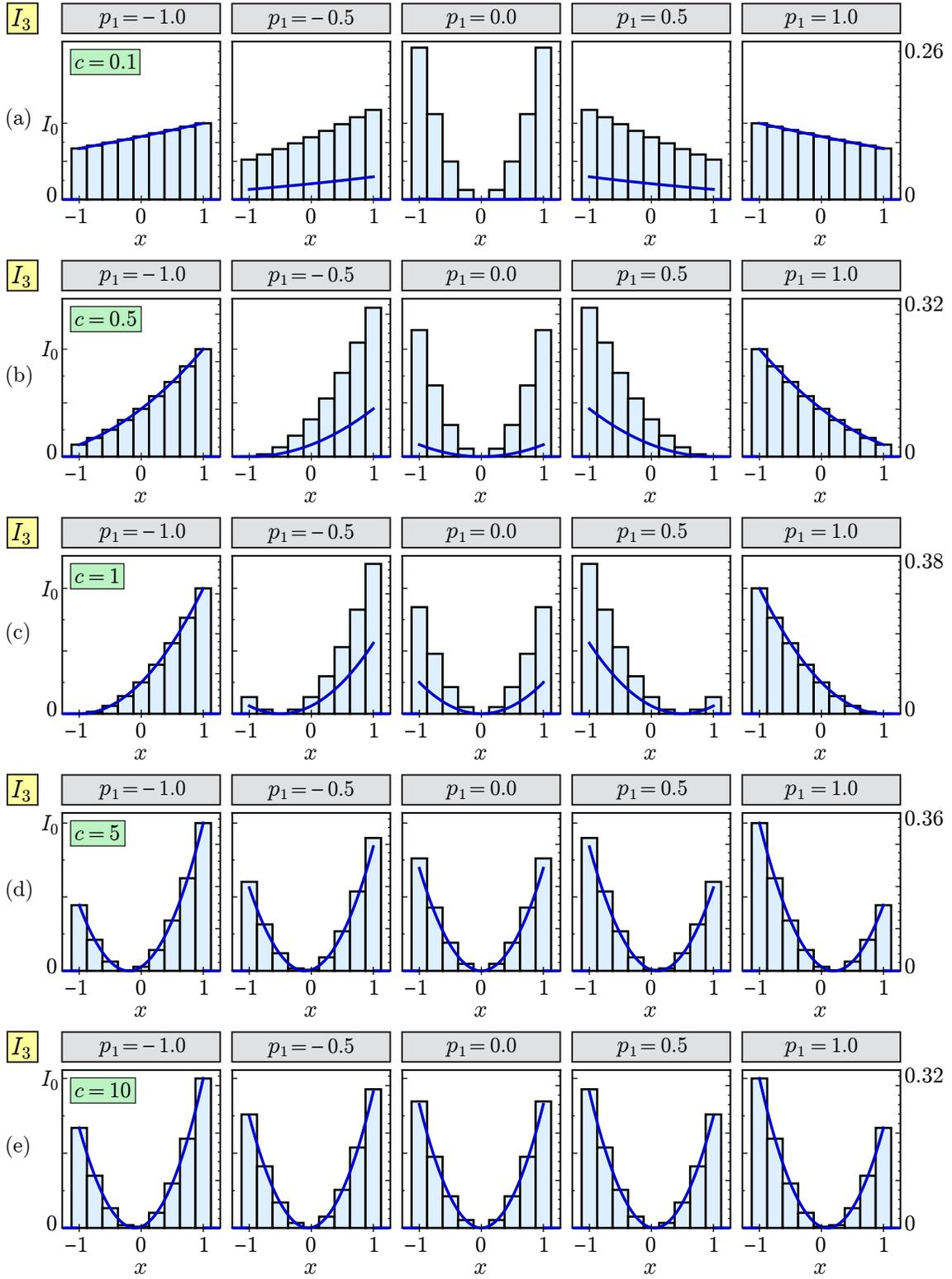


Figure 17: Plots of $I_3(x; p_1)$ (left axes) and $P_3(i|p_1)$ (right axes) for several values of p_1 . Each row of plots corresponds to a different value of c , as indicated in the leftmost plot.

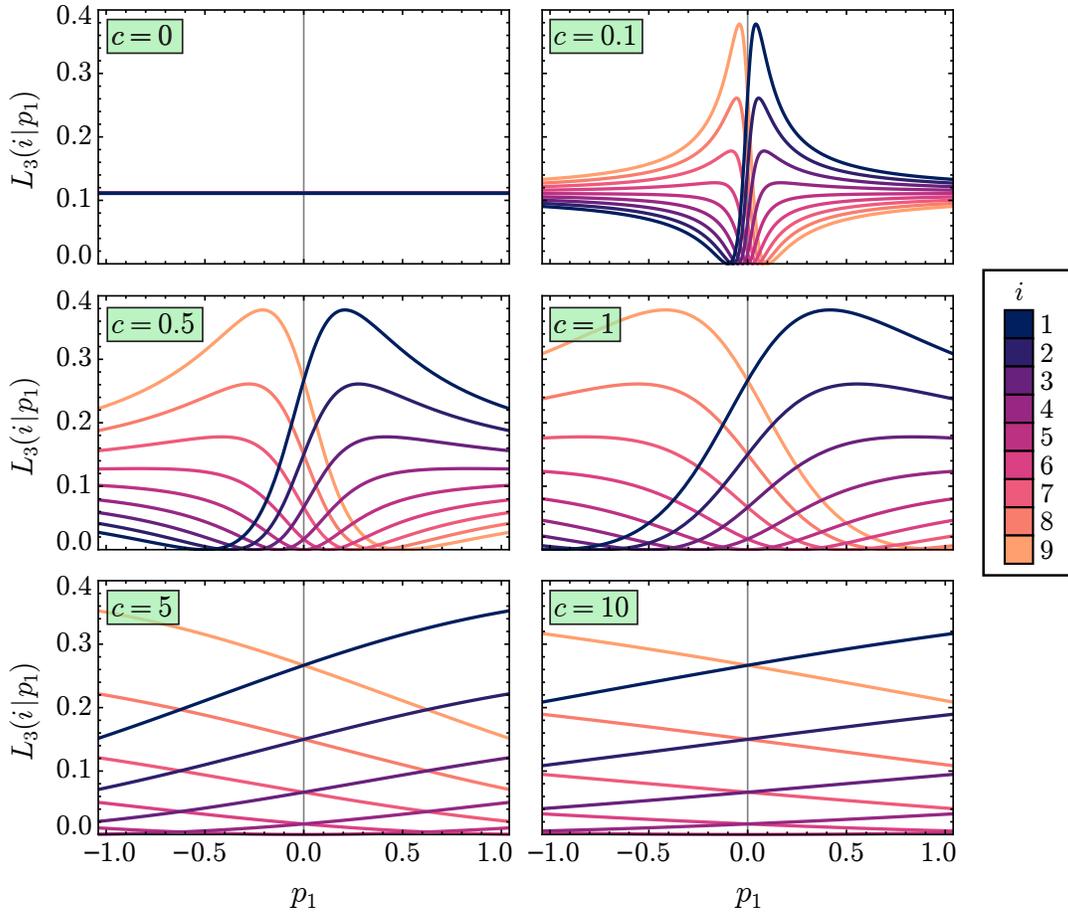


Figure 18: Likelihood functions $L_3(i|p_1)$ associated with each pixel i in a measurement with theoretical intensity distribution $I_3(p_1)$, plotted for several nonnegative values of c . Symmetric results are obtained for the corresponding negative values of c , with each plot flipped about the vertical $p_1 = 0$ axis.

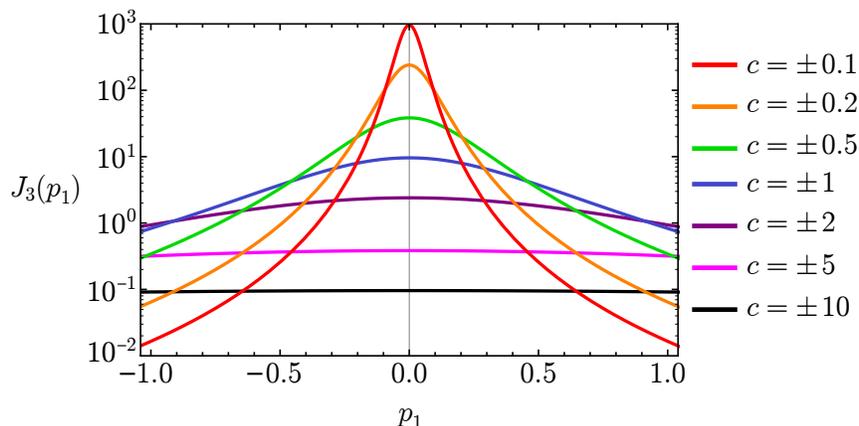


Figure 19: Expected unit Fisher information $J_3(p_1)$ for a measurement of $I_3(x; p_1)$, plotted on a logarithmic scale for several values of c .

formation, recall that the magnitude of the local slope of L_3 is an indicator of the information content of a measurement of p_1 . In agreement with the observations made above, for $|c| \ll 1$, the likelihood generally has a very large slope when $|p_1|$ is small (enabling a precise estimate of p_1), but it becomes nearly flat for larger parameter values. Meanwhile, for $|c| \gg 1$, the likelihood is relatively flat over the entire range of interest, making parameter estimation difficult. Qualitatively, it is evident that the best balance between these two extremes is achieved when c is on the order of unity, so that $L_3(i|p_1)$ exhibits a similar amount of variation over the full range of interest of p_1 .

For a measurement containing a large number of photons, the uncertainty of the MLE can be calculated from the expected unit Fisher information; a somewhat lengthy but straightforward calculation shows that

$$J_3(p_1) = \sum_{i=1}^9 \frac{16c^2[5c + 3(i-5)p_1^2]^2}{3(12p_1^2 + 5c^2)^3} = \frac{240c^2}{(12p_1^2 + 5c^2)^2}. \quad (37)$$

This function is plotted in Fig. 19 for several values of c . Notice that the Fisher information is the same for positive and negative c ; the $c = 0$ case does not appear on the plot since $J_3(p_1)$ goes to zero. Suppose that we are designing an experiment where the output intensity takes the form of $I_3(x; p_1)$, and we wish to determine the optimal value of c that, on average, will produce the best parameter estimate for any true value of p_1 within the range of interest, i.e., the smallest expected error $\sigma(p_1) = J_3(p_1)^{-1/2}$. One approach to do so is by minimizing the average value of the

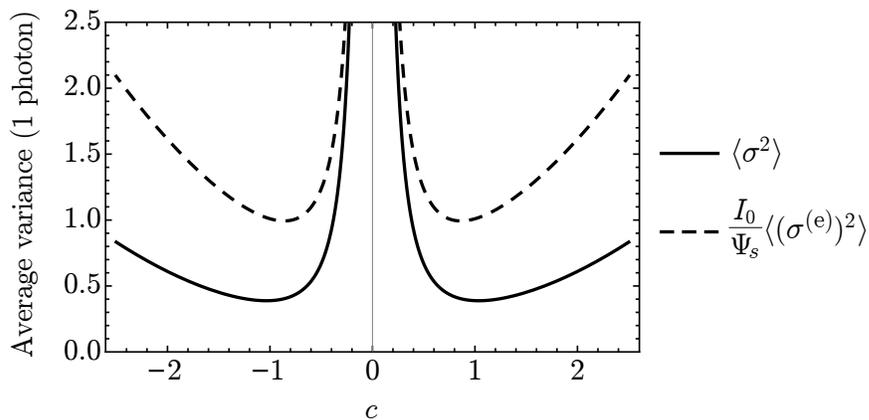


Figure 20: Expected variances (averaged over p_1) for parameter estimates based on measurements of $I_3(x; p_1)$ containing one detected photon (solid line) and one emitted photon (dashed line), plotted as a function of c . For the latter case, the error is scaled by the ratio between I_0 and the source power Ψ_s , which can be treated as a unitless quantity (see footnote 10 on page 42).

variance $\sigma(p_1)^2$ over the interval $p_1 \in [-1, 1]$, which is given by

$$\begin{aligned}
 \langle \sigma^2 \rangle &= \frac{1}{2} \int_{-1}^1 \sigma(p_1)^2 dp_1 \\
 &= \frac{1}{240c^2} \int_{-1}^1 (12p_1^2 + 5c^2)^2 dp_1 \\
 &= \frac{5}{24} c^2 + \frac{6}{25} \frac{1}{c^2} + \frac{1}{3}.
 \end{aligned} \tag{38}$$

This function is plotted as a solid line in Fig. 20. (The dashed line will be explained shortly). Note that for a multi-photon measurement, the variance scales as $1/\mathcal{N}$. The average error $\langle \sigma^2 \rangle$ is minimized when $c = \pm(144/125)^{1/4} \approx \pm 1.036$, in close agreement with the above prediction that the optimal value of c is on the order of unity.

As alluded to in the previous section, all of the statistics and performance metrics discussed thus far have pertained exclusively to photons detected by the sensor. However, the information contained in each detected photon is not the only thing to take into consideration when designing an experiment. In a typical experiment, the light source emits a constant optical power Ψ_s , of which some fraction reaches the detector. The power incident on the detector, which is given by

$$\Psi_d(p_1) = \int_{-1}^1 I_3(x; p_1) dx = I_0 \frac{2(3p^2 + c^2)}{3(|c| + 1)^2} \tag{39}$$

in this example¹⁰, is typically smaller than Ψ_s by some ratio that is influenced by power losses (e.g., due to reflection, absorption, or scattering) and/or the choice of measurement scheme (e.g., an off-null configuration). During the exposure time of the sensor, the number of detected photons is (on average) equal to $\mathcal{N} = (\Psi_d/\Psi_s)\mathcal{N}_s$, where \mathcal{N}_s is the number of photons emitted by the source. If the speed of the measurement is a priority, then it is important to make efficient use of the source, i.e., to maximize the information acquired per emitted photon. To that end, let us define the expected unit Fisher information per emitted photon as

$$J^{(e)}(p_1) = \frac{\Psi_d}{\Psi_s} J(p_1), \quad (40)$$

so that the total information acquired in a given time interval is $\mathcal{N}J(p_1) = \mathcal{N}_s J^{(e)}(p_1)$. (Obviously, this is not to suggest that each photon carries information about p_1 at the moment that it is emitted from the source; rather, $J^{(e)}(p_1)$ is the average information acquired at the detector plane per photon emitted by the source.)

For the present example, using Eqs. (37) and (39), the Fisher information per emitted photon is found to be

$$J_3^{(e)}(p_1) = \frac{I_0}{\Psi_s} \frac{160c^2(3p_1^2 + c^2)}{(12p_1^2 + 5c^2)^2(|c| + 1)^2}. \quad (41)$$

This result is plotted in Fig. 21 for several values of c . In comparison to Fig. 19, notice that the peak in $J_3^{(e)}(p_1)$ when $|c| \ll 1$ is much less pronounced than that of $J_3(p_1)$. This is because as $|c|$ decreases, the amount of information per detected photon increases, but the number of detected photons decreases by nearly the same ratio. From Eq. (41), the minimum expected variance $\sigma^{(e)}(p_1)^2 = J_3^{(e)}(p_1)^{-1}$ can be calculated for a measurement of one emitted photon, averaged over the range of

¹⁰The right-hand side of Eq. (39) implicitly has units of I_0 times the unitless coordinate x (acquired from the integration), i.e., units of power.

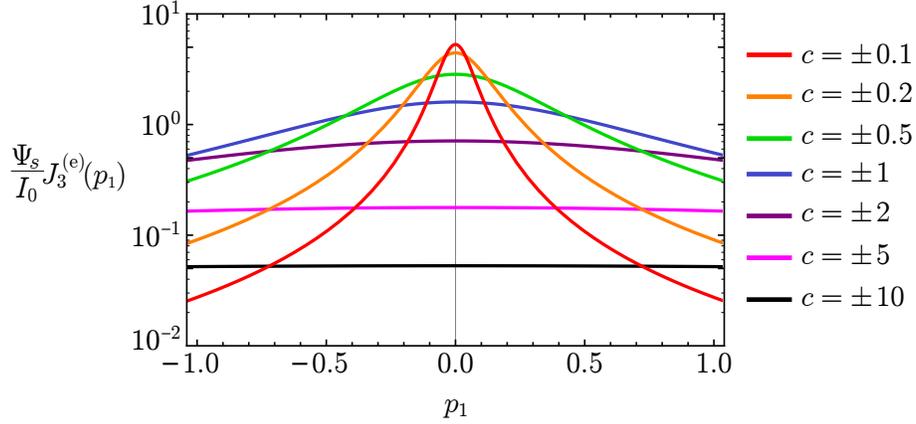


Figure 21: Expected unit Fisher information $J_3^{(e)}(p_1)$ per emitted photon for a measurement of $I_3(x; p_1)$, scaled by the ratio of source power to I_0 and plotted on a logarithmic scale for several values of c .

interest of p_1 :

$$\begin{aligned}
\langle (\sigma^{(e)})^2 \rangle &= \frac{1}{2} \int_{-1}^1 \sigma^{(e)}(p_1)^2 dp_1 \\
&= \frac{\Psi_s}{I_0} \frac{(|c| + 1)^2}{320c^2} \int_{-1}^1 \frac{(12p_1^2 + 5c^2)^2}{3p_1^2 + c^2} dp_1 \\
&= \frac{\Psi_s}{I_0} \frac{(|c| + 1)^2}{480c^2} \left[\sqrt{3} c^3 \arctan\left(\frac{\sqrt{3}}{c}\right) + 72c^2 + 48 \right]. \tag{42}
\end{aligned}$$

This function is plotted as a dashed line in Fig. 20, shown in comparison to the average variance per detected photon derived earlier. A numerical calculation shows that the expected error per emitted photon is minimized when $c = \pm 0.863$, which is slightly smaller than the optimal value $c = \pm 1.036$ for detected photons. This is due to the fact that for parameter values near $|p_1| = 1$, the power on the detector is up to 10% larger for $|c| = 0.863$ than for $|c| = 1.036$, compensating for the slight reduction in information per detected photon in the former case.

Recall that in this example the intensity is normalized to have a peak value of I_0 regardless of the value of c . This is not particularly realistic, since in an actual off-null measurement, a change in the (spatially varying) off-null condition is likely to be accompanied by a global scaling factor in the measured intensity. In some cases, this could result in a much more dramatic difference between the Fisher information per emitted and detected photon than in this example. On a separate note, in situations

where $\sigma(p_1)^2$ and $\sigma^{(e)}(p_1)^2$ cannot be calculated analytically, the integral over p_1 can be evaluated numerically. If the numerical integration is too computationally expensive, a simpler merit function could be constructed by summing the variance over some appropriately chosen set of parameter values.

6.4 Far-from-null (high intensity) measurement

For the final one-parameter example, consider the intensity distribution

$$I_4(x; p_1) = \Pi(x) \frac{1}{(|d| + 2)^2} (p_1 - x - d)^2, \quad (43)$$

where the constant d introduces a spatially uniform offset from the off-null condition considered in the previous example. When $d = 0$, the intensity is identical to $I_3(x; p_1)$ with $c = 1$, which was plotted previously in Fig. 17(c). For comparison, Fig. 22 contains plots of $I_4(x; p_1)$ and the corresponding PMF for several positive values of d . (Symmetric results are obtained for negative d .) The likelihood functions $L_4(i|p_1)$ for each case are plotted in Fig. 23. Observe that when $d = 1$, the intensity profile and likelihood function are translated in parameter space so that they are symmetric about $p_1 = 1$. As d increases, the distribution continues to shift farther away from the off-null condition of $I_3(x; p_1)$, so that the intensity becomes large and uniform over the range of interest of p_1 and the likelihood function becomes very flat. As seen in Fig. 24, the expected Fisher information per detected photon¹¹ decreases rapidly as d increases. Following the same procedure as in the previous example, it can be shown that the average estimation error over the parameter range is minimized when $d = 0$. (This holds true when optimizing for detected or emitted photons, though as noted before, the latter result is in part due to the choice of normalization of the intensity.)

The takeaway from this example is that it illustrates the statistical advantage of off-null measurements over a “far-from-null” experimental configuration in which the parameter of interest causes a small fractional change in the output intensity. Although the parameter estimation technique outlined in Section 3 is only useful for imaging experiments where the off-null condition (and thus the output intensity) varies with position, by looking at Fig. 22 one can also appreciate the principle of traditional off-null ellipsometry, in which only the total power is measured. In that

¹¹Henceforth, all mentions of the Fisher information refer to the expected information per detected photon unless specified otherwise.

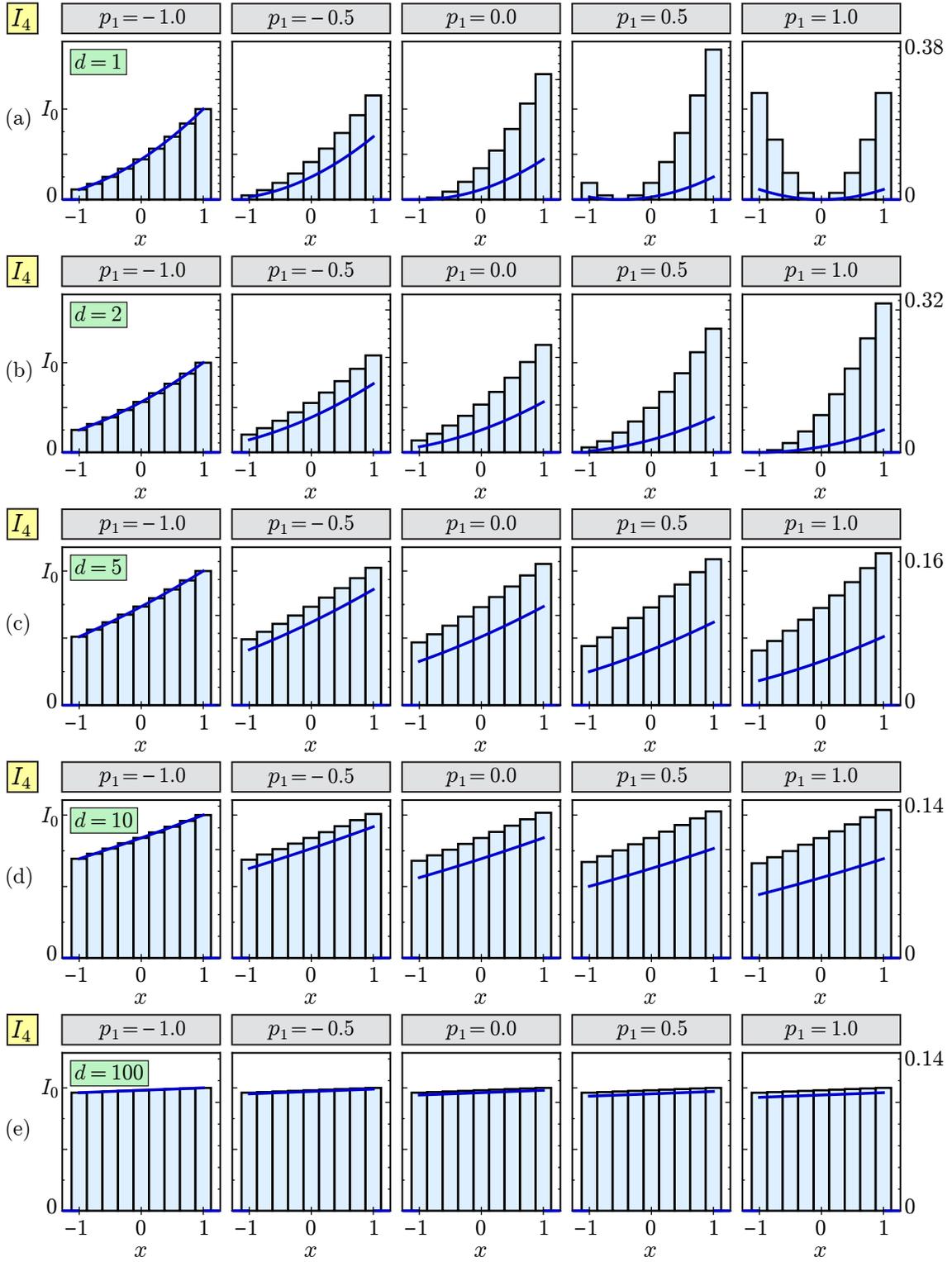


Figure 22: Plots of $I_4(x; p_1)$ (left axes) and $P_4(i|p_1)$ (right axes) for several values of p_1 . Each row of plots corresponds to a different value of d , as indicated in the leftmost plot.

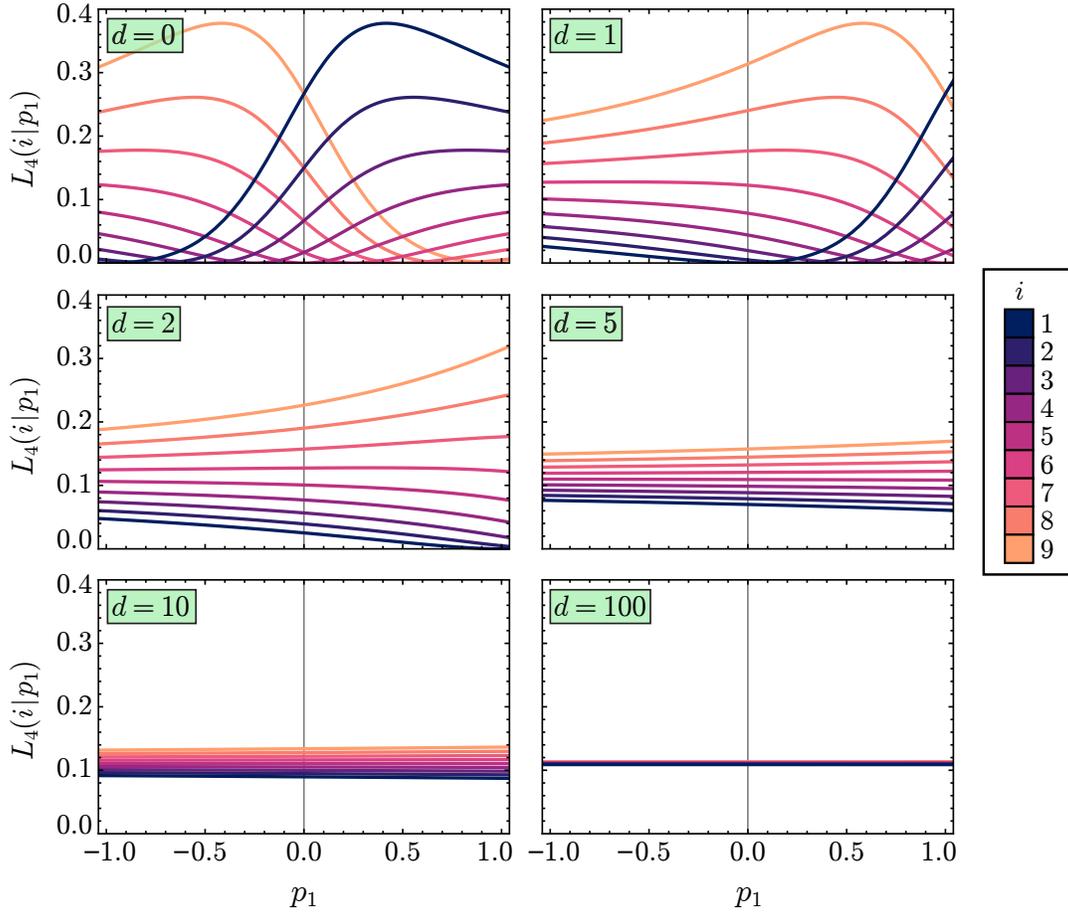


Figure 23: Likelihood functions $L_4(i|p_1)$ associated with each pixel i in a measurement with theoretical intensity distribution $I_4(p_1)$, plotted for several nonnegative values of d . Notice that the effect of d is simply a horizontal translation; when $d \gg 1$, the range of interest $p_1 \in [-1, 1]$ only contains a small portion of the left tail of the distribution. Symmetric results are obtained for the corresponding negative values of d , for which the curves are translated in the opposite direction (with respect to the $d = 0$ case).

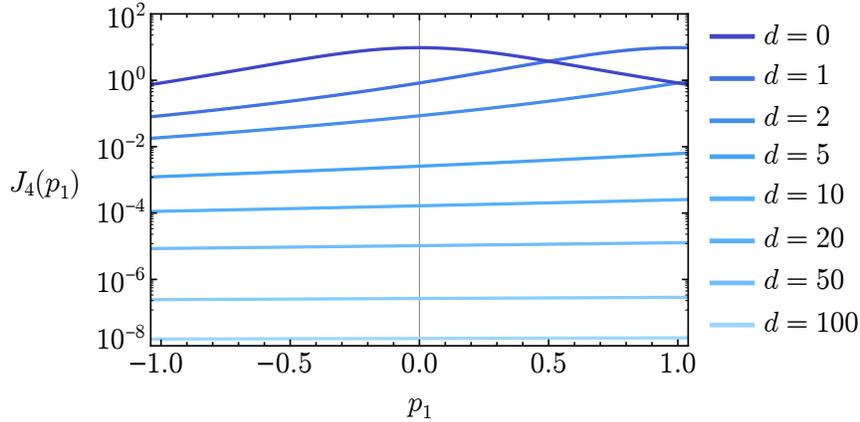


Figure 24: Expected unit Fisher information $J_4(p_1)$ for a measurement of $I_4(x; p_1)$, plotted on a logarithmic scale for several values of d . The $d = 0$ case is identical to $J_3(p_1)$ with $c = 1$ (see Fig. 19). For negative values of d , each curve is flipped about the vertical $p_1 = 0$ axis.

case, the off-null configuration greatly increases the contrast of the variation in power with respect to p_1 , enabling a more accurate measurement while placing less stringent requirements on the fidelity of the sensor.

7 Two-parameter optical MLE examples

To illustrate the use of MLE in the multiple-parameter case, this section contains several intensity distributions that depend on two parameters $\mathbf{p} = (p_1, p_2)$. The procedures for calculating the PMF, FIM, and expected error are fundamentally the same as in the one-parameter case, although the algebra is more complicated. Rather than dwelling on the mathematical details, numerical results are presented in the following discussion. This is representative of most real-world applications, in which MLE techniques are typically implemented numerically.

The intensity distributions discussed in Sections 7.1 through 7.6 are summarized in Table 5. Similarly to the one-parameter examples, each intensity distribution is normalized so that it attains a maximum value of I_0 over the region of interest $-1 \leq p_1, p_2 \leq 1$. The distributions considered in Sections 7.1 and 7.2 each have a p_1 term with linear spatial variation and a p_2 term with sinusoidal spatial variation, serving as simple examples for the two-parameter case. Sections 7.3 and 7.4 contain two thought-provoking (albeit unrealistic) examples that illustrate the mathematical

Section	Intensity distribution
7.1	$I_5(x; \mathbf{p}) = 0.563\Pi(x)[2 + p_1x + p_2 \sin(\pi x)]$
7.2	$I_6(x; \mathbf{p}) = 0.250\Pi(x)[2 + p_1x + p_2 \cos(\pi x)]$
7.3	$I_7(x; \mathbf{p}) = \begin{cases} 0.5\Pi(x)(1 + p_1x), & x < 0 \\ 0.5\Pi(x)(1 + p_2x), & x \geq 0 \end{cases}$
7.4	$I_8(x; \mathbf{p}) = \begin{cases} 0.5\Pi(x) [1 + 2p_1(x + 0.625)], & x < -0.125 \\ 0.5\Pi(x), & -0.125 \leq x < 0.125 \\ 0.5\Pi(x) [1 + 2p_2(x - 0.625)], & x \geq 0.125 \end{cases}$
7.5	$I_9(x; \mathbf{p}) = 0.125\Pi(x) [(p_1 - x)^2 + (p_2 - \cos(\pi x))^2]$
7.6	$I_{10}(x; \mathbf{p}) = 0.320\Pi(x) [(p_1 - 0.25x)^2 + (p_2 - 0.25 \cos(\pi x))^2]$

Table 5: Intensity distributions for each example considered in Section 7.

mechanisms that can lead to statistical correlations between the parameter estimates for p_1 and p_2 . Finally, a pair of two-parameter off-null measurements are discussed in Sections 7.5 and 7.6.

7.1 Linear and sinusoidal variations (case 1)

For the first two-parameter example, consider the intensity distribution

$$I_5(x; \mathbf{p}) = 0.563\Pi(x)[2 + p_1x + p_2 \sin(\pi x)], \quad (44)$$

which is valid over the region of interest $-1 \leq p_1, p_2 \leq 1$. Similarly to the first example in Section 6, $I_5(x; \mathbf{p})$ depends linearly on the product of p_1 and x . The dependence on p_2 is also linear, but this additional term varies sinusoidally across the sensor. Therefore, variations in p_1 and p_2 result in distinct changes in the shape of the intensity $I_5(x; \mathbf{p})$ and the PMF $P_5(i|\mathbf{p})$, as shown in Fig. 25. For instance, when $p_2 = 0$ (the third row of plots), the intensity is strictly a linear function of x with slope p_1 . When $p_1 = 0$ (the third column of plots), it is a sine function with a DC offset. For all other cases, the intensity is a linear combination of the two.

For the two-parameter case, the likelihood $L_5(\mathbf{p}|i) = P_5(i|\mathbf{p})$ can be plotted in two dimensions as a function of p_1 and p_2 . The likelihood functions associated with each

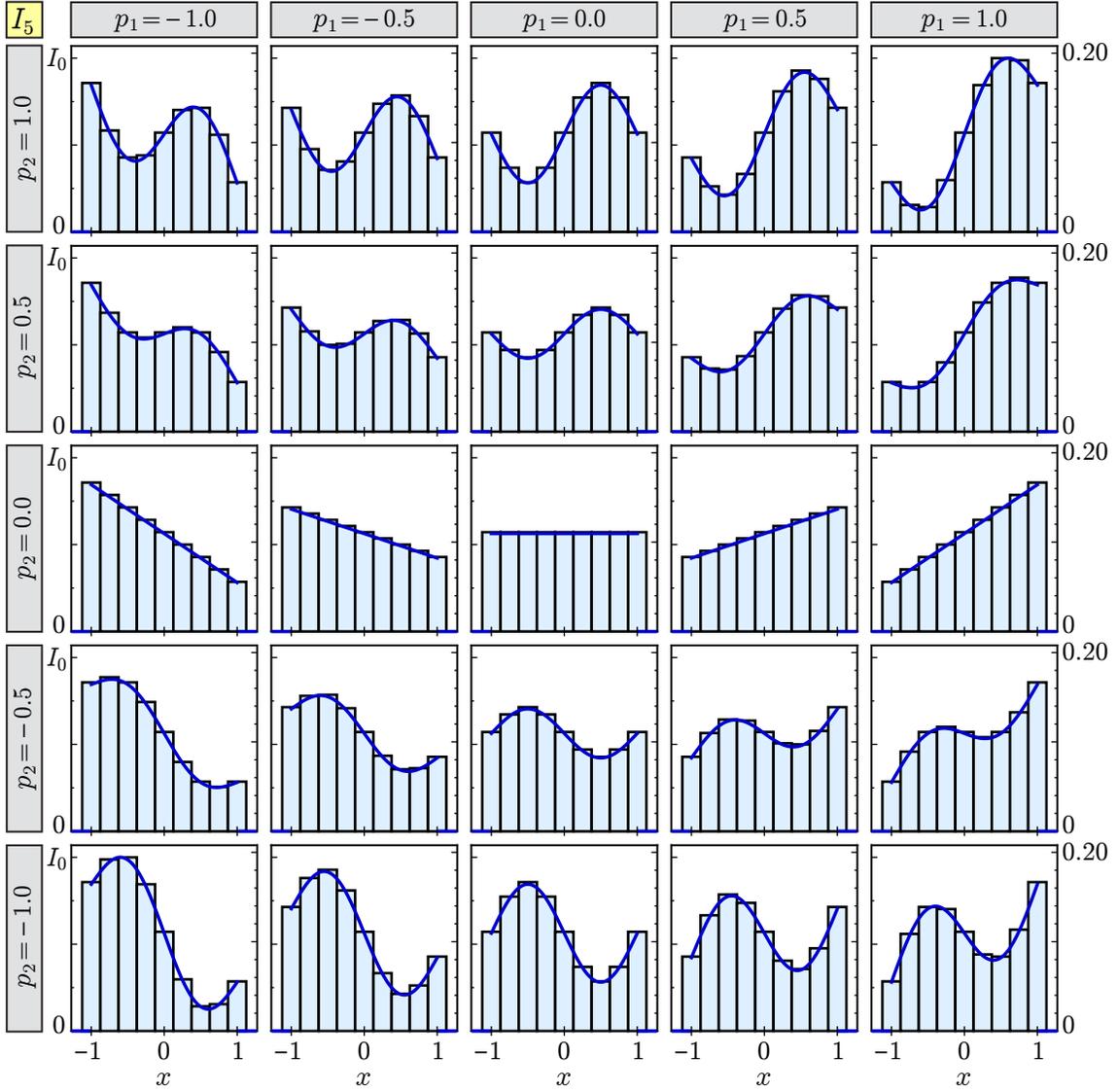


Figure 25: Plots of $I_5(x; \mathbf{p})$ (left axes) and $P_5(i; \mathbf{p})$ (right axes) for several values of p_1 and p_2 .

pixel are shown in Fig. 26, with contour lines drawn as a visual aid to identify paths of constant likelihood. These plots have several interesting features. First, notice that $L_5(\mathbf{p}|i = 5)$ is constant, meaning that pixel 5 provides no useful information about p_1 and p_2 . (Incidentally, this was also the case for the one-parameter intensity distributions I_1 and I_2 . Since the signal from pixel 5 has no effect on the MLE, it can be ignored.) Secondly, the likelihood functions for pixels 1 and 9 are independent of p_2 (as evident from the vertical contour lines) since $\sin(\pi x) = 0$ for $x = \pm 1$. In contrast, the likelihood functions associated with pixels 4 and 6 depend more strongly

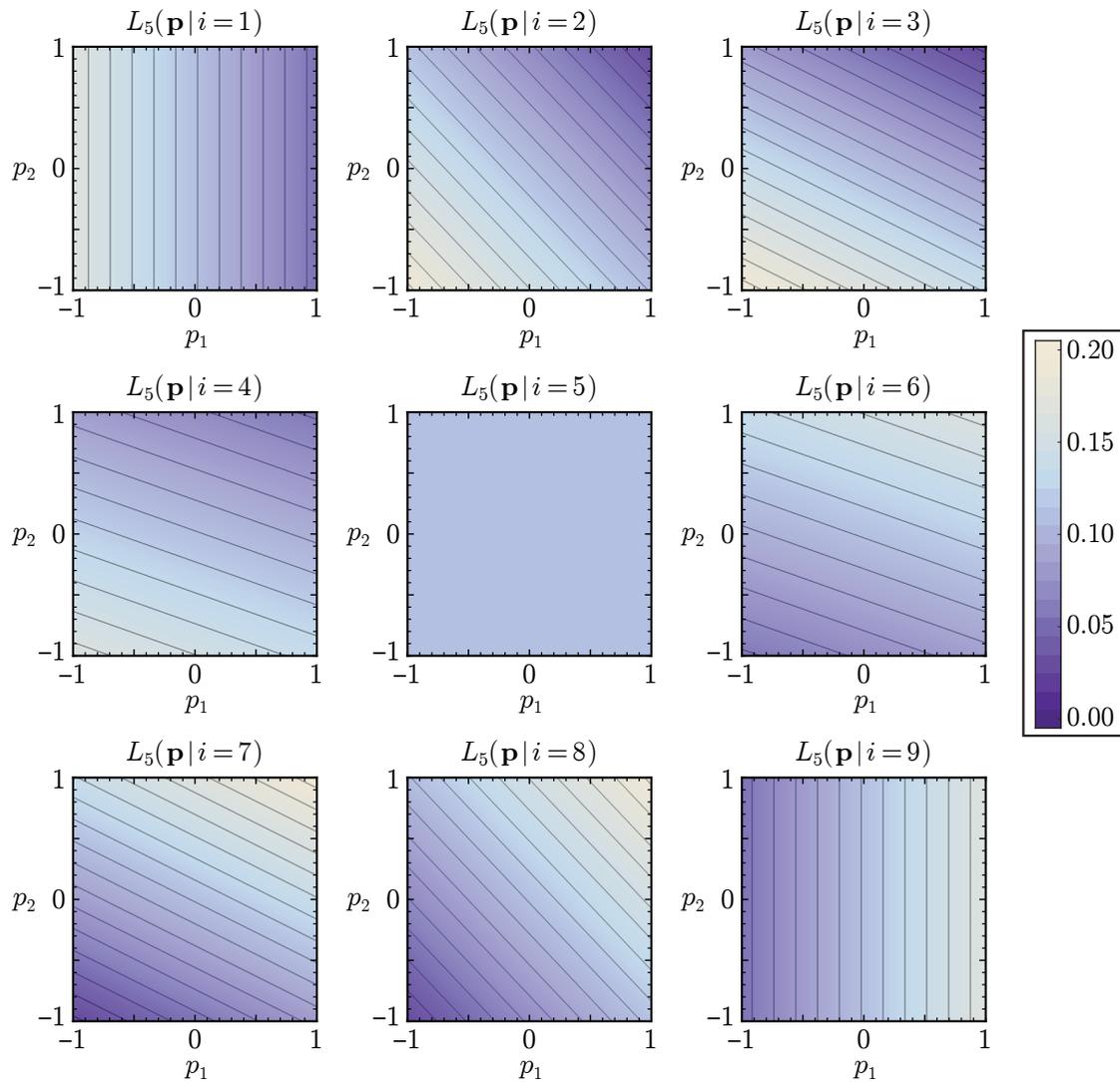


Figure 26: Likelihood functions $L_5(\mathbf{p}|i)$ associated with each pixel i for a measurement of $I_5(x; \mathbf{p})$. Contour lines are shown in increments of 0.01.

on p_2 than p_1 as a consequence of the fact that $\sin(\pi x)$ has a larger slope near the center of the sensor than the linear term x . Lastly, note that the paths of constant likelihood generally have negative (or vertical) slopes in parameter space. Roughly speaking, this means that if p_1 increases and p_2 decreases by a similar amount (or if p_2 increases and p_1 decreases), the likelihood function will only change slightly, making it difficult to distinguish linear combinations of parameters along this direction. On the other hand, a simultaneous increase (or simultaneous decrease) in p_1 and p_2 will tend to cause a more significant change in the likelihood function, making it easier to distinguish this type of variation in \mathbf{p} .

The patterns described above can be quantified by calculating the estimation error based on the 2×2 expected Fisher information matrix, whose elements may be computed using either form of Eq. (9). For a measurement of $\mathcal{N} = 1000$ photons with true parameter values $\mathbf{p} = (0, 0)$, the FIM and its inverse are found to be

$$\mathcal{N}\mathbb{J}_5 = \begin{bmatrix} 104.2 & 67.1 \\ 67.1 & 111.1 \end{bmatrix}, \quad (\mathcal{N}\mathbb{J}_5)^{-1} = \begin{bmatrix} 0.0157 & -0.0095 \\ -0.0095 & 0.0147 \end{bmatrix}. \quad (45)$$

As discussed in Section 2, $(\mathcal{N}\mathbb{J}_5)^{-1}$ places a lower limit on the covariance matrix for a 1000-photon measurement of p_1 and p_2 . Since its off-diagonal elements are fairly large in relation to its diagonal elements, a strong coupling between parameters (i.e., large covariance) is expected. Indeed, the principal axes of the error ellipse are given by the eigenvectors $[0.69; 0.72]$ and $[0.72; -0.69]$, and the axis lengths (the square roots of the corresponding eigenvalues) are 0.076 and 0.157, respectively. Thus, the major axis of the ellipse is oriented at approximately -45° in parameter space, and the standard deviation error is about twice as large along the -45° direction as the $+45^\circ$ direction.¹² In this example, it turns out that similar results are obtained for all values of \mathbf{p} within the region of interest. The error ellipses for a selection of true parameter values are plotted in Fig. 27.

Given a measured intensity $\tilde{\mathbf{I}}$, the magnitude and orientation of the uncertainty of the MLE are also manifested in the shape of the likelihood function $L_5(\mathbf{p}|\tilde{\mathbf{I}})$ and its logarithm $\ell_5(\mathbf{p}|\tilde{\mathbf{I}})$. Fig. 28 contains two examples of the log-likelihood functions obtained for simulated 1000-photon measurements with true parameter values $\mathbf{p} = (0, 0)$ and $\mathbf{p} = (0.63, -0.25)$. Again, these plots contain several interesting features. First,

¹²It is only meaningful to refer to angles in parameter space when p_1 and p_2 have the same units and are normalized to their respective ranges of interest, as they are in this discussion.

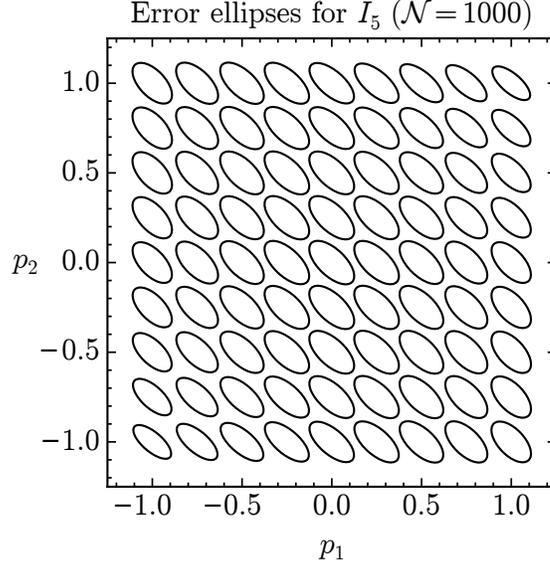


Figure 27: Ellipses representing the expected standard deviation error of a 1000-photon measurement of $I_5(x; \mathbf{p})$ with true parameter values p_1 and p_2 , sampled over a 9×9 grid in parameter space.

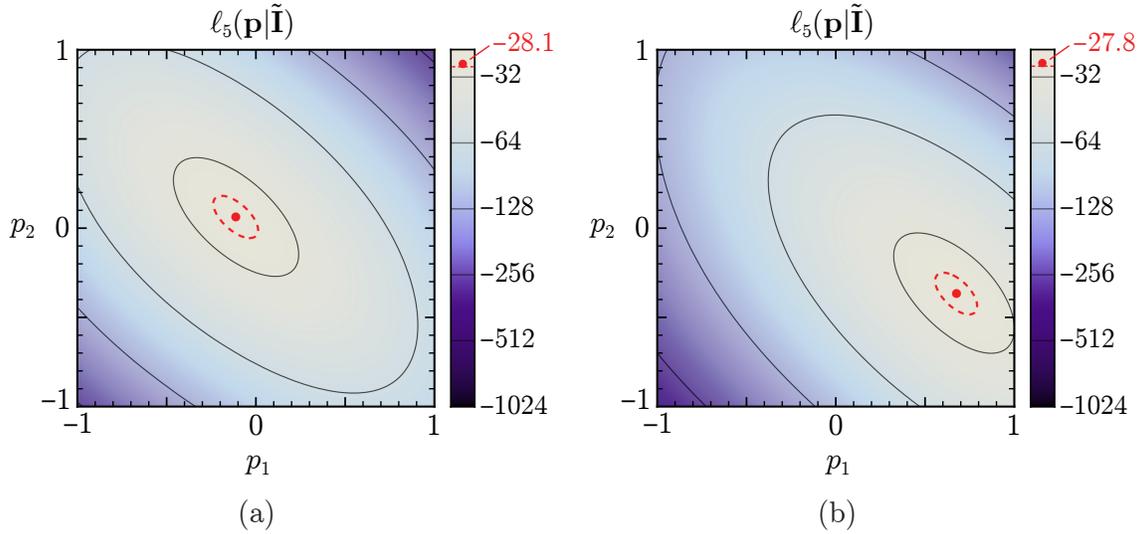


Figure 28: Log-likelihood functions $\ell_5(\mathbf{p}|\tilde{\mathbf{I}})$ for simulated 1000-photon measurements of $I_5(x; \mathbf{p})$ with true parameter values (a) $\mathbf{p} = (0, 0)$ and (b) $\mathbf{p} = (0.63, -0.25)$. The plots are shaded on a logarithmic scale with solid contour lines drawn at powers of 2, as indicated in the legend. The peak of each distribution is marked with a red dot. The locations of these maxima (i.e., the MLEs for each measurement) are $\mathbf{p} = (-0.115, 0.064)$ and $\mathbf{p} = (0.673, -0.366)$, respectively. The dashed contour line indicates where the likelihood $L_5(\mathbf{p}|\tilde{\mathbf{I}})$ drops to $1/\sqrt{e}$ times its peak value, representing the standard deviation confidence interval for the MLE.

notice that the contours of equal likelihood are approximately elliptical. This behavior is characteristic of a bivariate Gaussian distribution $f(\mathbf{p}) = f_0 \exp(-\frac{1}{2}\mathbf{p}^T \boldsymbol{\Sigma}^{-1} \mathbf{p})$ with covariance matrix $\boldsymbol{\Sigma}$, for which the locus of points satisfying $\mathbf{p}^T \boldsymbol{\Sigma}^{-1} \mathbf{p} = \kappa^2$ (for some constant κ) traces out an ellipse [42]. Thus, the shape of $\ell_5(\mathbf{p}|\tilde{\mathbf{I}})$ supports the claim made earlier (see Eq. (14)) that the posterior probability distribution $P(\mathbf{p}|\tilde{\mathbf{I}})$, which is a scaled version of the likelihood if no prior distribution is assumed, closely approximates a Gaussian distribution when a large number of photons are measured. Comparing Figs. 27 and 28, one can also see that the likelihood function is elongated along the direction with the largest expected estimation error. In Section 6 it was noted that the estimation error is largest when the likelihood function is nearly flat; for the multiple-parameter case, it can be further specified that the error is largest along the *direction* where the likelihood function is flattest, i.e., the direction perpendicular to the local gradient of ℓ with respect to \mathbf{p} .

Each plot in Fig. 28 contains a red dot representing the MLE for the measurement, i.e., the location of the peak of $\ell_5(\mathbf{p}|\tilde{\mathbf{I}})$. The estimated parameter values (which are listed in the figure caption) differ considerably from the true values, with errors as large as ~ 0.11 for each parameter. The standard deviation confidence interval for the MLE, which is outlined by a red dashed line, consists of the region where the likelihood function $L_5(\mathbf{p}|\tilde{\mathbf{I}})$ is greater than or equal to $1/\sqrt{e}$ times its peak value.¹³ This is equivalent to an additive decrease in the log-likelihood by $\ln(e^{-1/2}) = -0.5$. Notice that this region is elliptical, and its size and shape are virtually identical to the nearest ellipse in Fig. 27. In fact, by evaluating the expected FIM at the MLE with $\mathcal{N} = 1000$, an extremely close agreement is found between the predicted covariance matrix $(\mathcal{N}\mathbb{J}_5)^{-1}$ and the standard deviation confidence interval of $\ell_5(\mathbf{p}|\tilde{\mathbf{I}})$. (When plotted together, the ellipses are virtually indistinguishable even when zoomed in.) In general, the correlation between the two grows stronger as the number of photons increases. In this example, 1000 photons are sufficient to obtain a very close agreement; in an experiment with smaller expected error, fewer photons would be required.

To conclude this example, similarly to Sections 6.1 and 6.2, a Monte Carlo simulation was performed for 50,000 trials of a 1000-photon simulated measurement of $I_5(x; \mathbf{p})$ for which the true parameter values are given by $\mathbf{p} = (0, 0)$. A histogram

¹³For the Gaussian distribution $f(\mathbf{p})$ mentioned above, the $\kappa = 1$ ellipse encloses one standard deviation. Along this contour, the function value drops to $f_0 \exp(-\frac{1}{2}) = f_0/\sqrt{e}$.

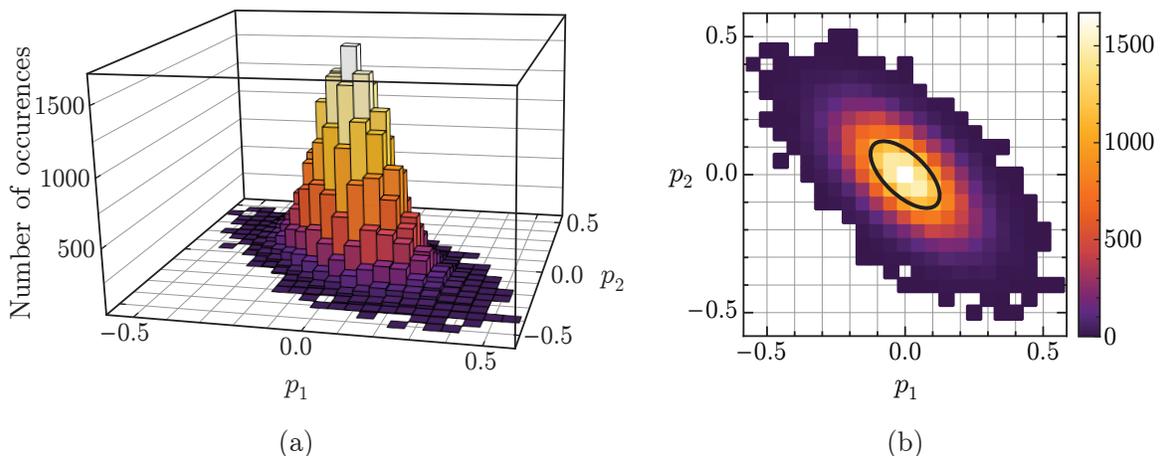


Figure 29: (a) Histogram of the maximum likelihood estimates obtained from 50,000 trials of a simulated 1000-photon measurement of $I_5(x; \mathbf{p})$ with true parameter value $\mathbf{p} = (0, 0)$. (b) Overhead view of the distribution shown in plot (a), with the color of each pixel indicating the number of trials for which the MLE was within a given interval. The black ellipse at the center of the plot represents the expected standard deviation error based on the Fisher information matrix.

of the maximum likelihood estimates obtained in all trials is shown in Fig. 29(a); an overhead view of the distribution is also shown in Fig. 29(b). The data closely resembles a Gaussian distribution with the same orientation as the expected error ellipse, which is shown in black in the overhead view. The statistical covariance matrix of the data matches the matrix $(\mathcal{N}\mathbb{J}_5)^{-1}$ given in Eq. (45) to within three significant digits.

7.2 Linear and sinusoidal variations (case 2)

For the second two-parameter example, consider the intensity distribution

$$I_6(x; \mathbf{p}) = 0.250\Pi(x)[2 + p_1x + p_2 \cos(\pi x)], \quad (46)$$

which is similar to $I_5(x; \mathbf{p})$, but with the sine term replaced by a cosine. The intensity and PMF are plotted for several parameter values in Fig. 30, and the likelihood functions for each pixel are shown in Fig. 31. In this example, it can be seen that the paths of constant likelihood have different orientations for each pixel. This implies, for instance, that a simultaneous increase in p_1 and p_2 will cause a significant change in $L_6(\mathbf{p}|i = 1)$, but very little change in $L_6(\mathbf{p}|i = 9)$; meanwhile, a simultaneous increase in p_1 and decrease in p_2 will do just the opposite. The reason for this can

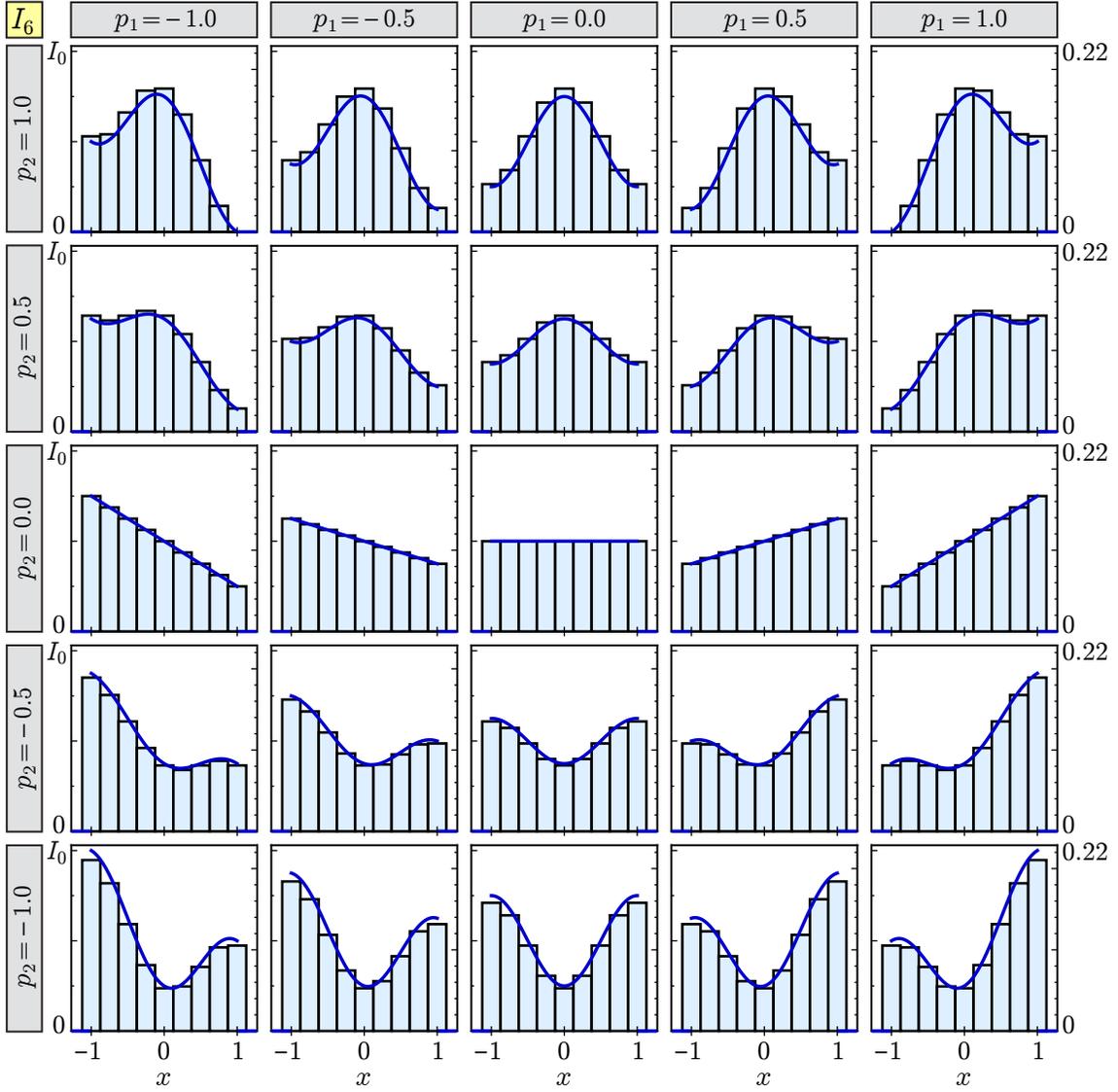


Figure 30: Plots of $I_6(x; \mathbf{p})$ (left axes) and $P_6(i; \mathbf{p})$ (right axes) for several values of p_1 and p_2 .

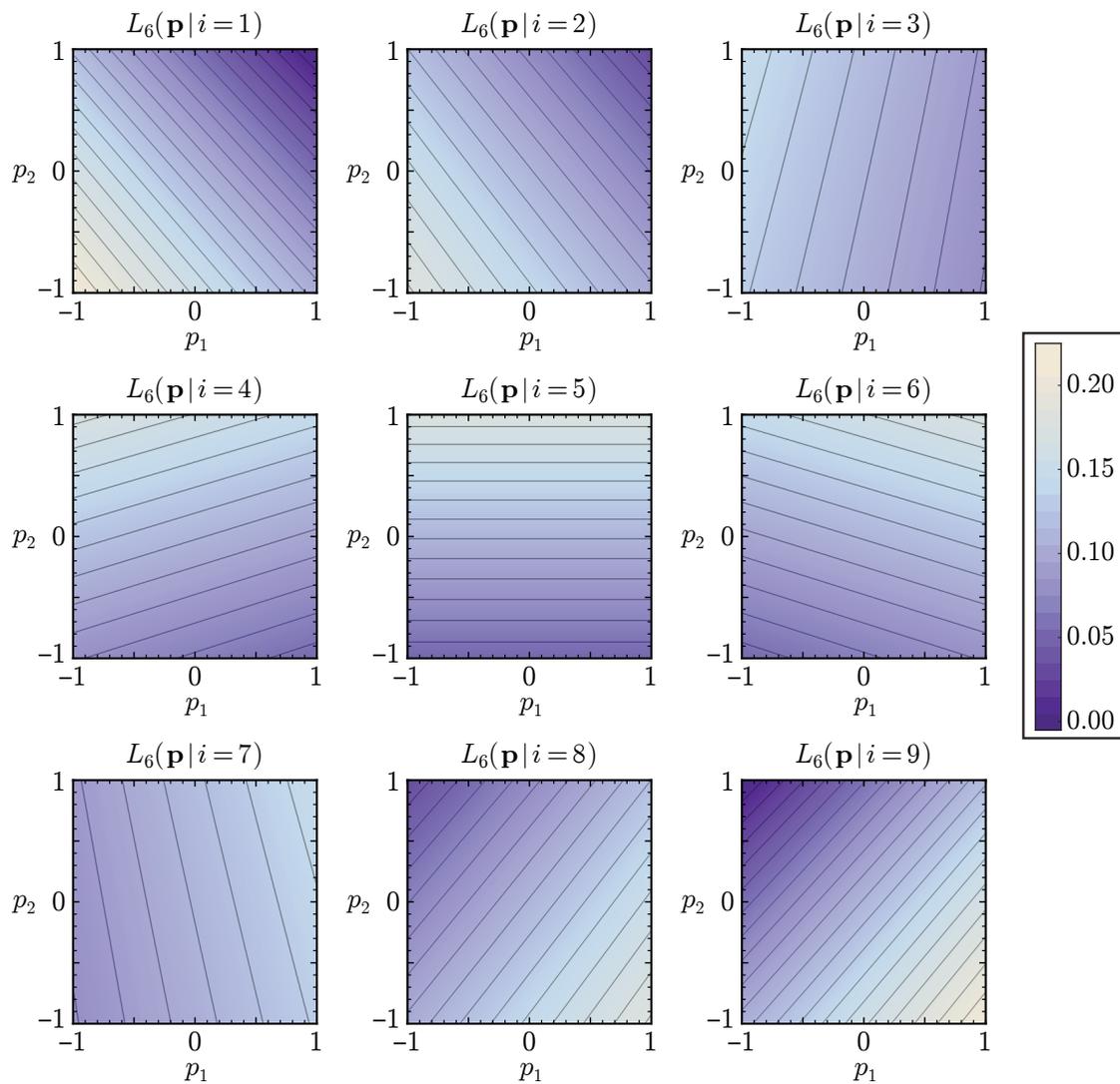


Figure 31: Likelihood functions $L_6(\mathbf{p} | i)$ associated with each pixel i for a measurement of $I_6(x; \mathbf{p})$. Contour lines are shown in increments of 0.01.

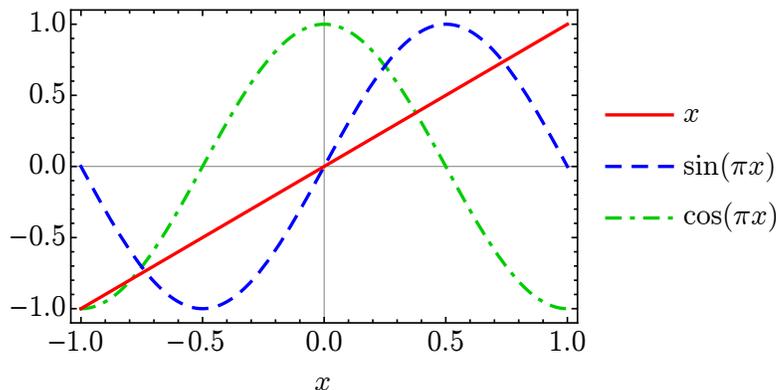


Figure 32: Spatial variations of each term appearing in intensity distributions $I_5(x; \mathbf{p})$ and $I_6(x; \mathbf{p})$.

be understood by examining the plots of x , $\sin(\pi x)$, and $\cos(\pi x)$ shown in Fig. 32. Whereas x and $\sin(\pi x)$ always have the same sign, this is not the case for x and $\cos(\pi x)$. Therefore, for the intensity distribution $I_5(x; \mathbf{p})$, an increase in p_1 can be compensated (to a certain extent) by a decrease in p_2 . The distribution $I_6(x; \mathbf{p})$ is less prone to this situation since any linear combination of p_1 and p_2 produces distinct fluctuations at different pixels. However, correlations can still arise in cases where very few photons are incident on one or more pixels (for example, when $p_1 = p_2 = 1$), since the contributions of each pixel to the log-likelihood function $\ell_6(\mathbf{p}|\tilde{\mathbf{I}})$ associated with a measured intensity $\tilde{\mathbf{I}}$ may be imbalanced.

Based on the above observations, one can reasonably expect there to be a smaller correlation between the estimated parameters from a measurement of $I_6(x; \mathbf{p})$ than in the previous example. As a matter of fact, for $\mathbf{p} = (0, 0)$, the FIM and its inverse are diagonal, indicating that there is zero covariance:

$$\mathcal{N}\mathbb{J}_6 = \begin{bmatrix} 104.2 & 0 \\ 0 & 135.8 \end{bmatrix}, \quad (\mathcal{N}\mathbb{J}_6)^{-1} = \begin{bmatrix} 0.0096 & 0 \\ 0 & 0.0074 \end{bmatrix}, \quad (47)$$

where $\mathcal{N} = 1000$. The eigenvectors of $(\mathcal{N}\mathbb{J}_6)^{-1}$ are $[1; 0]$ and $[0; 1]$, and the square roots of the corresponding eigenvalues are 0.098 and 0.086, respectively. Thus, the error ellipse is nearly circular, with its principal axes oriented along the p_1 and p_2 axes. The error ellipses for a selection of parameter values are shown in Fig. 33. As seen in the plot, the expected error is relatively uniform over the entire parameter range, with the smallest error occurring when p_2 is close to 1. The covariance between p_1 and p_2 is also generally small, with one notable exception: as $|p_1| \rightarrow 1$ and $p_2 \rightarrow 1$,

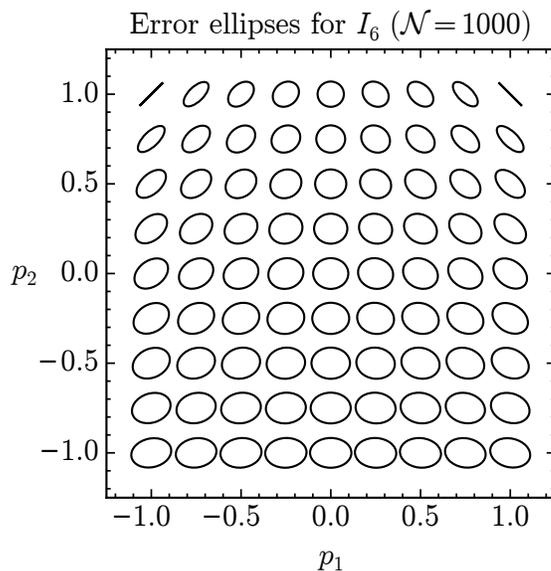


Figure 33: Ellipses representing the expected standard deviation error of a 1000-photon measurement of $I_6(x; \mathbf{p})$ with true parameter values p_1 and p_2 , sampled over a 9×9 grid in parameter space.

the two parameters become highly correlated. At the far upper corners of the region of interest, the error ellipse resembles a straight line, indicating complete correlation between p_1 and p_2 . (Even so, the magnitude of the uncertainty of each parameter is still smaller than the expected errors for other parameter values.) From the uppermost plots in Fig. 30, it can be seen that this correlation arises when the intensity drops to zero at either edge of the sensor (near pixel 1 or pixel 9). This happens because the intensity distribution and the likelihood functions $L(\mathbf{p}|i)$ are distributed such that the remaining pixels cannot easily distinguish between all possible combinations of p_1 and p_2 , as alluded to in the previous paragraph.¹⁴

The log-likelihood functions $\ell_6(\mathbf{p}|\tilde{\mathbf{I}})$ for simulated 1000-photon measurements of $I_6(x; \mathbf{p})$ with true parameter values $\mathbf{p} = (0, 0)$ and $\mathbf{p} = (0.63, -0.25)$ are shown in Fig. 34. As in the previous example, the contours of equal likelihood are highly elliptical near the peak, indicating that the likelihood is approximately a Gaussian

¹⁴The astute reader might wonder why the expected error is asymmetric with respect to p_2 despite the fact that the last term of $I_6(x; \mathbf{p})$ exhibits symmetry with respect to both p_2 and x . The answer is that the asymmetry is a sampling artifact of the 9-pixel array, since pixels 1 and 9 sample the periodic function $\cos(\pi x)$ at points that are offset by 2π radians. This causes the total measured intensity to vary with p_2 despite the fact that $\int_{-1}^1 \cos(\pi x) dx = 0$. As is often the case, the error is smallest in this example when the total intensity is minimized, which occurs when $p_2 = 1$.

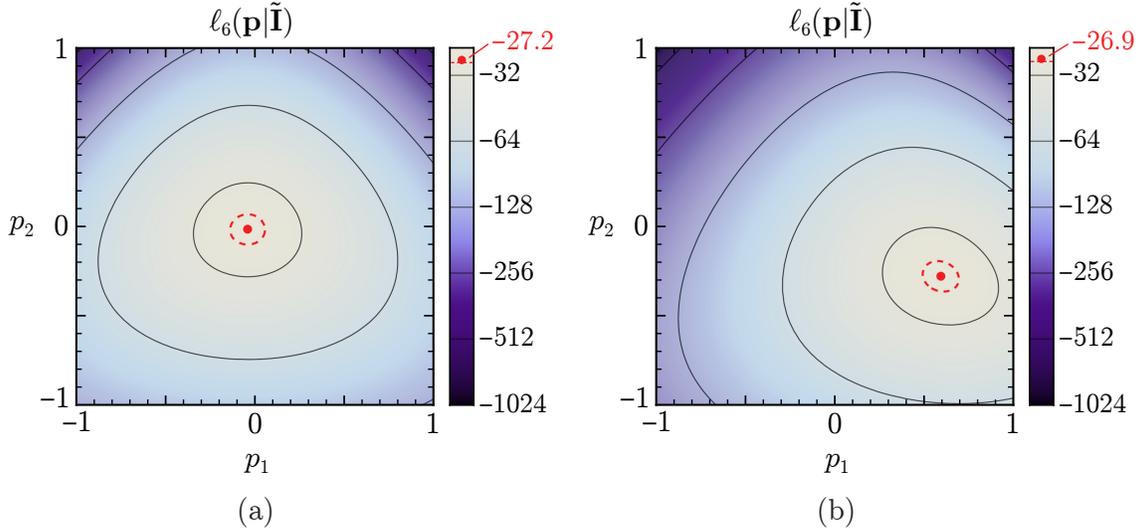


Figure 34: Log-likelihood functions $\ell_6(\mathbf{p}|\tilde{\mathbf{I}})$ for simulated 1000-photon measurements of $I_6(x;\mathbf{p})$ with true parameter values (a) $\mathbf{p} = (0,0)$ and (b) $\mathbf{p} = (0.63, -0.25)$. The plots are shaded on a logarithmic scale with solid contour lines drawn at powers of 2, as indicated in the legend. The peak of each distribution is marked with a red dot. The locations of these maxima (i.e., the MLEs for each measurement) are $\mathbf{p} = (-0.043, -0.014)$ and $\mathbf{p} = (0.591, -0.278)$, respectively. The dashed contour line indicates where the likelihood $L_6(\mathbf{p}|\tilde{\mathbf{I}})$ drops to $1/\sqrt{e}$ times its peak value, representing the standard deviation confidence interval for the MLE.

distribution. The Gaussian approximation weakens away from the peak, with the contours of $\ell_6(\mathbf{p}|\tilde{\mathbf{I}})$ becoming slightly distorted. Compared to $\ell_5(\mathbf{p}|\tilde{\mathbf{I}})$, the distribution is much more symmetric due to the small covariance between p_1 and p_2 (for these particular true parameter values). The standard deviation confidence interval, indicated by the dashed red line, is also highly symmetric and slightly narrower than it was in the previous example, matching the expected error based on the FIM. The uncertainty is also reflected in the distribution of the MLEs obtained from 50,000 trials of a 1000-photon measurement of $I_6(\mathbf{p}|\tilde{\mathbf{I}})$, as shown in Fig. 35. The diagonal elements of the covariance matrix of the simulated data agree with the matrix $(\mathcal{N}\mathbb{J}_6)^{-1}$ given in Eq. (47) to within two significant digits; the off-diagonal elements of the matrix are very close to zero (approximately 500 times smaller than the diagonal elements).

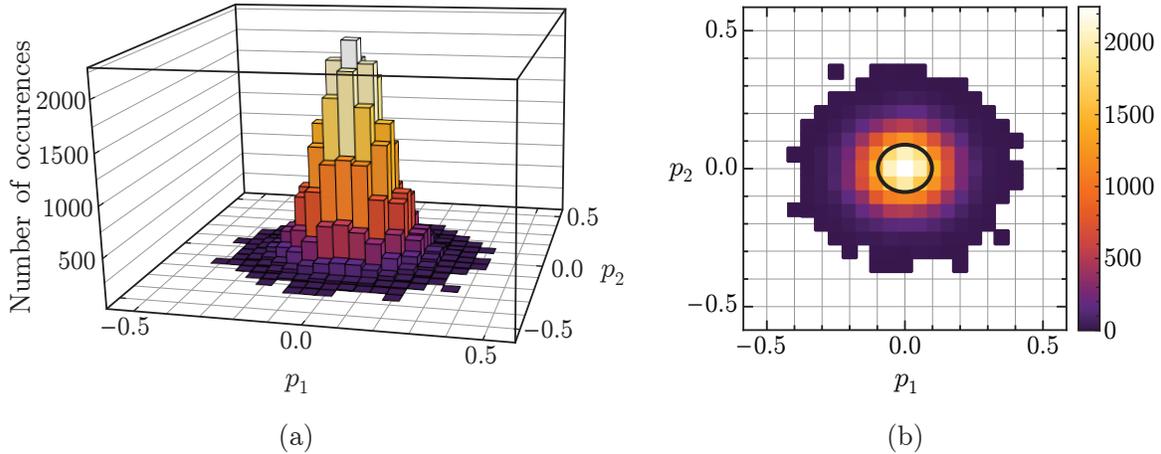


Figure 35: (a) Histogram of the maximum likelihood estimates obtained from 50,000 trials of a simulated 1000-photon measurement of $I_6(x; \mathbf{p})$ with true parameter value $\mathbf{p} = (0, 0)$. (b) Overhead view of the distribution shown in plot (a), with the color of each pixel indicating the number of trials for which the MLE was within a given interval. The black ellipse at the center of the plot represents the expected standard deviation error based on the Fisher information matrix.

7.3 Piecewise linear dependence (nonzero covariance)

The next two examples involve intensity distributions for which fluctuations due to p_1 and p_2 occur in completely separate portions of the sensor. Although this is not a particularly common real-world scenario, some interesting insight can be gained from the analysis. First, consider the piecewise intensity distribution

$$I_7(x; \mathbf{p}) = \begin{cases} 0.5\Pi(x)(1 + p_1x), & x < 0, \\ 0.5\Pi(x)(1 + p_2x), & x \geq 0, \end{cases} \quad (48)$$

which is plotted in Fig. 36. This distribution is similar to the one-parameter linear intensity profile $I_1(x; p_1)$, except that the slopes on the left and right halves of the sensor are proportional to p_1 and p_2 , respectively. Since the intensities on each half of the sensor only depend on a single parameter, one would expect the parameters to be completely uncoupled, enabling an estimate with zero covariance. However, this turns out not to be the case when applying the MLE approach outlined in Section 3. (Note: the MLE formalism only requires the PMF to be twice differentiable with respect to \mathbf{p} , so the discontinuity in the derivative of $I_7(\mathbf{x}; \mathbf{p})$ with respect to \mathbf{x} is not problematic.) As established previously, this treatment relies on the information

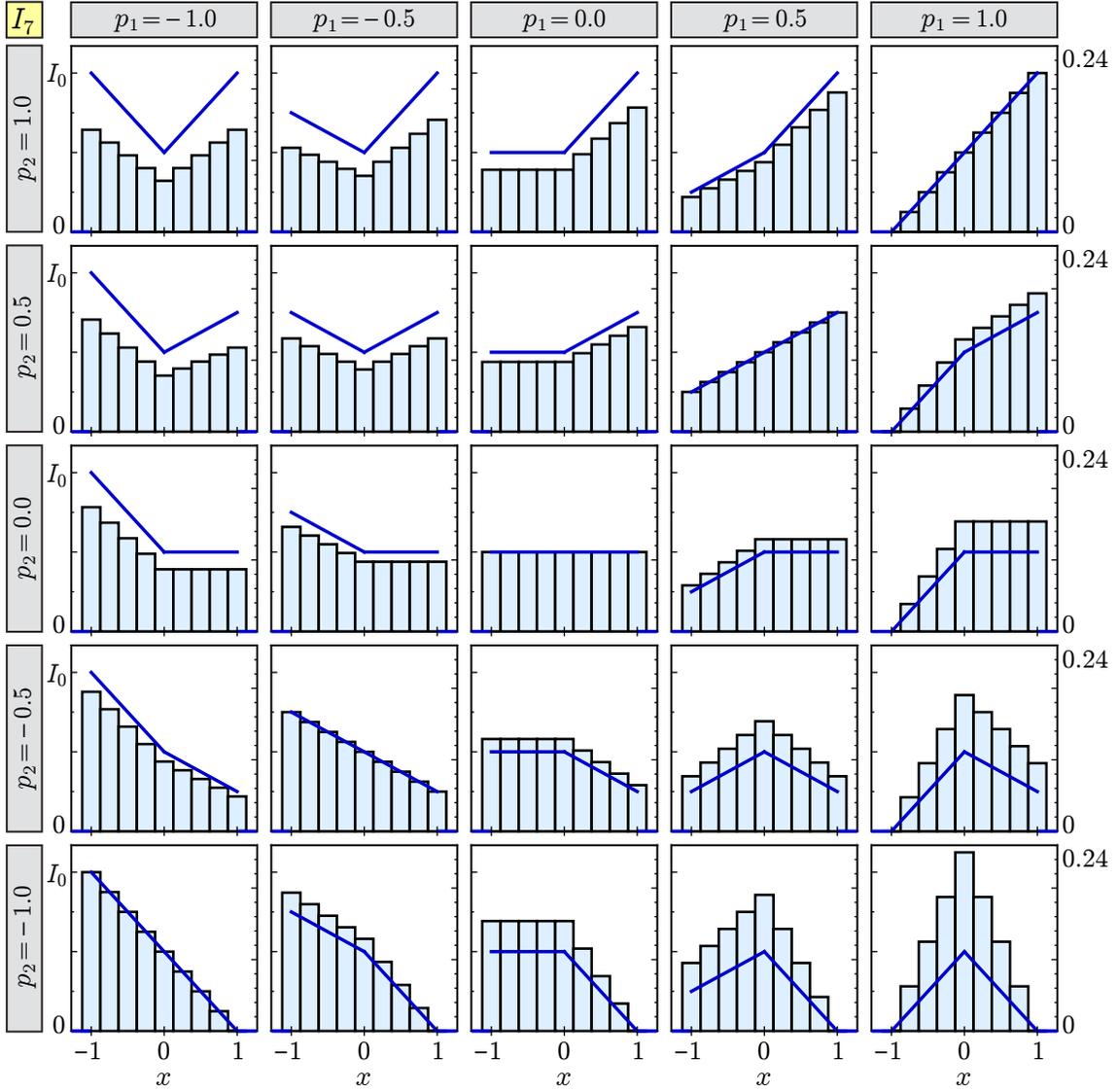


Figure 36: Plots of $I_7(x; \mathbf{p})$ (left axes) and $P_7(i; \mathbf{p})$ (right axes) for several values of p_1 and p_2 .

contained in the shape of the intensity distribution, that is, the relative intensity or the PMF. Clearly, the value of p_1 impacts the probability $P_7(i|\mathbf{p})$ of detecting a photon at each pixel on the left half of the sensor ($i = 1, \dots, 5$); what is perhaps less obvious, however, is that it also affects the probabilities for pixels 6 through 9. Indeed, within any given row of Fig. 36 (for which p_2 has a fixed value), the intensity on the right half of the sensor is always the same, yet the PMF changes depending on the value of p_1 . This is possible because the total intensity $\sum_i I_7(x_i|\mathbf{p})$, which appears in the denominator of $P_7(i|\mathbf{p})$, varies with p_1 and p_2 so that each parameter affects the relative number of photons incident on every pixel i . Therefore, the estimates for p_1 and p_2 based on the PMF will generally be correlated to some degree. (In this particular example, the best workaround is to treat the signals from each half of the detector as completely separate measurements — more on this later.)

As usual, these effects can also be visualized by plotting the likelihood functions $L_7(i|\mathbf{p})$ for each pixel, which are shown in Fig. 37. Notice that the likelihood function for pixel 1 is most heavily influenced by p_1 , while that of pixel 9 is mostly influenced by p_2 . Nevertheless, every pixel contains information about both p_1 and p_2 , since the partial derivatives of $\ell_7(i|\mathbf{p})$ with respect to each parameter are nonzero. Interestingly, this even implies that photons measured at pixel 5 (the center of the sensor, where $I(x_5|\mathbf{p}) = 0.5$ for any \mathbf{p}) provide information about p_1 and p_2 when considered in relation to the number of photons measured at the other eight pixels.

The error ellipses for several values of p_1 and p_2 are shown in Fig. 38. Unlike the prior two examples, the expected estimation error for a measurement of $I_7(x; \mathbf{p})$ is strongly dependent on \mathbf{p} , with the largest error (and substantial covariance between p_1 and p_2) occurring in the upper left quadrant where $p_1 < 0$ and $p_2 > 0$. The distributions of the log-likelihood functions obtained for two 1000-photon measurements with different true parameter values, shown in Fig. 39, are consistent with this trend. The magnitude of the expected error is inversely proportional to the total intensity $\sum_i I_7(x_i|\mathbf{p})$, which is minimized when $p_1 = 1$ and $p_2 = -1$. Not coincidentally, the errors in p_1 and p_2 approach zero as $p_1 \rightarrow 1$ and $p_2 \rightarrow -1$, respectively. (As in Section 6.1, this expectation of zero error is only meaningful in the limit of large \mathcal{N} .) The dramatic variations in error with respect to \mathbf{p} can also be understood by revisiting Fig. 37, in which the contours of equal likelihood for each pixel tend to be most closely spaced in the lower right quadrant (where $p_1 > 0$ and $p_2 < 0$), indicating high information content. Pixel 5 in particular provides extremely useful information in

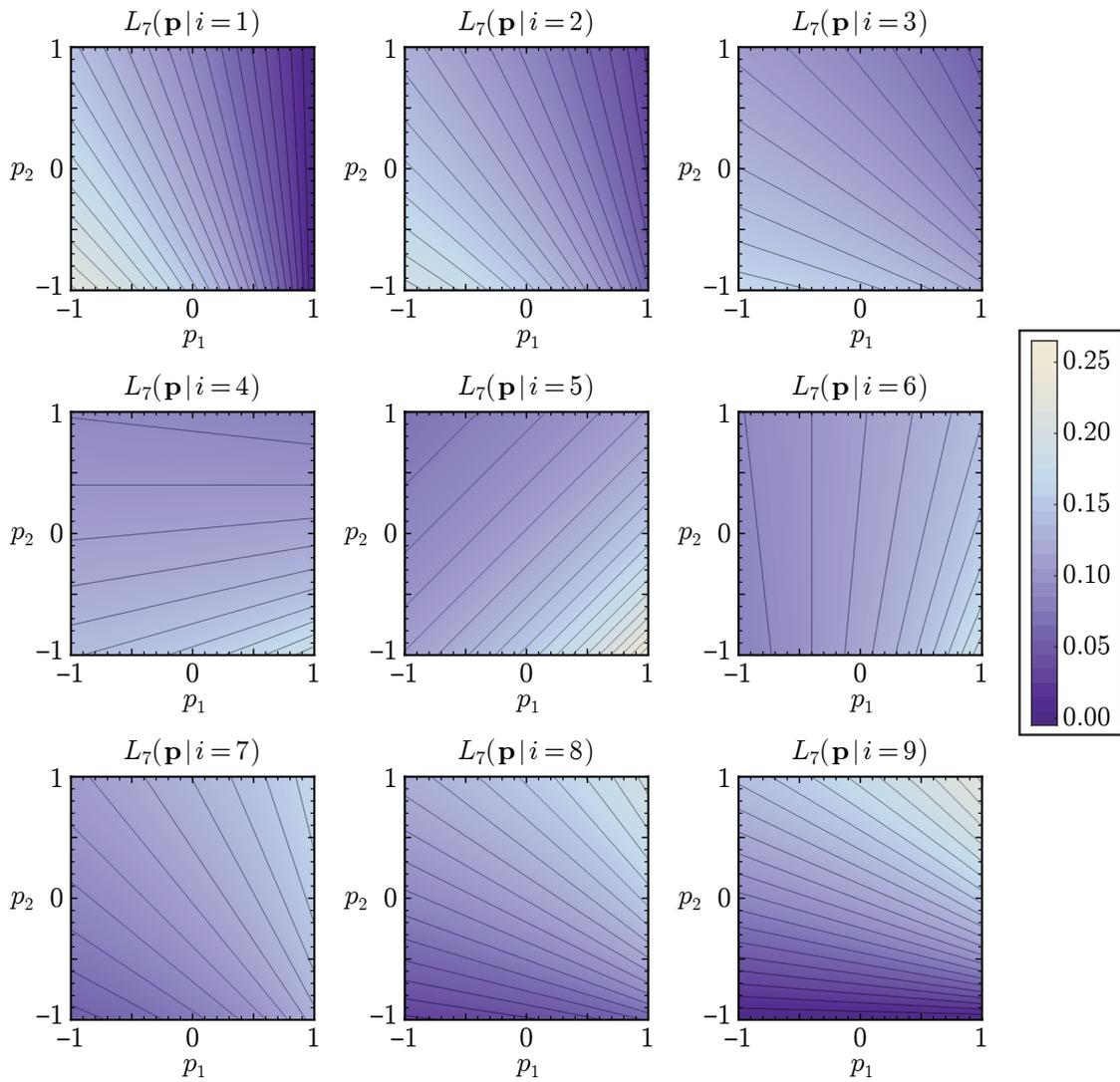


Figure 37: Likelihood functions $L_7(\mathbf{p} | i)$ associated with each pixel i for a measurement of $I_7(x; \mathbf{p})$. Contour lines are shown in increments of 0.01.

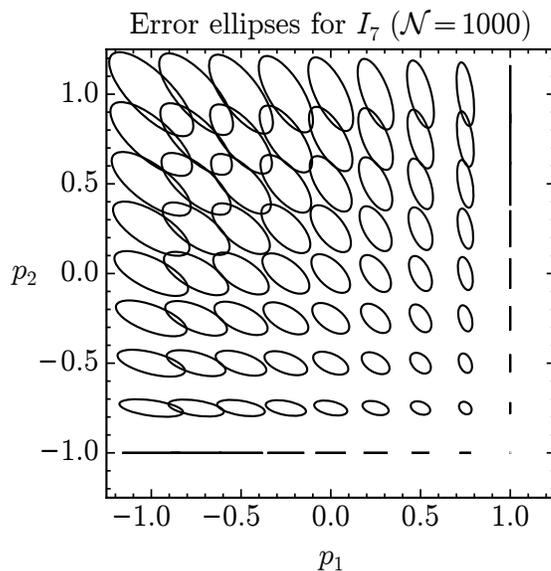


Figure 38: Ellipses representing the expected standard deviation error of a 1000-photon measurement of $I_7(x; \mathbf{p})$ with true parameter values p_1 and p_2 , sampled over a 9×9 grid in parameter space.

this quadrant, not only due to the large slope of $L_7(\mathbf{p}|i=5)$, but also because the direction of maximum variation (i.e., the gradient with respect to \mathbf{p}) opposes that of pixels 1 and 9. In contrast, pixel 5 is nearly useless in the upper left quadrant of the parameter space since the likelihood changes very slowly with respect to \mathbf{p} .

As mentioned before, in practice, the best way to deal with an intensity distribution such as $I_7(x; \mathbf{p})$ would be to treat it as two separate measurements: one involving pixels 1 through 5 (for which the intensity only depends on p_1), and another involving pixels 5 through 9 (for which the intensity only depends on p_2). The MLE approach could then be applied separately to each set of data, producing independent estimates for each parameter. In general, whenever it is possible to set up an experiment such that independent measurements can be made in this manner, it is probably best to do so, at least from a statistical standpoint. However, in cases where one does not have this luxury, the above example illustrates how subtle interactions between parameters (of either a physical or mathematical nature) can affect the accuracy of the measurement. Therefore, extra care should be taken to design the experiment such that the error obtained using the chosen statistical method is minimized.

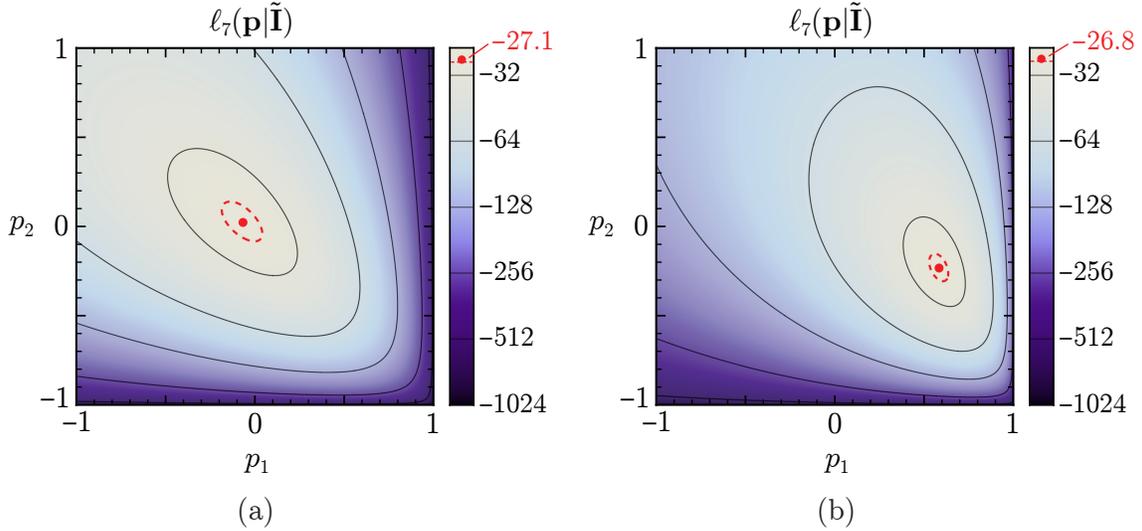


Figure 39: Log-likelihood functions $\ell_7(\mathbf{p}|\tilde{\mathbf{I}})$ for simulated 1000-photon measurements of $I_7(x; \mathbf{p})$ with true parameter values (a) $\mathbf{p} = (0, 0)$ and (b) $\mathbf{p} = (0.63, -0.25)$. The plots are shaded on a logarithmic scale with solid contour lines drawn at powers of 2, as indicated in the legend. The peak of each distribution is marked with a red dot. The locations of these maxima (i.e., the MLEs for each measurement) are $\mathbf{p} = (-0.067, 0.024)$ and $\mathbf{p} = (0.582, -0.232)$, respectively. The dashed contour line indicates where the likelihood $L_7(\mathbf{p}|\tilde{\mathbf{I}})$ drops to $1/\sqrt{e}$ times its peak value, representing the standard deviation confidence interval for the MLE.

7.4 Piecewise linear dependence (zero covariance)

Next, in comparison to the previous example, consider the intensity distribution

$$I_8(x; \mathbf{p}) = \begin{cases} 0.5\Pi(x) [1 + 2p_1(x + 0.625)], & x < -0.125, \\ 0.5\Pi(x), & -0.125 \leq x < 0.125, \\ 0.5\Pi(x) [1 + 2p_2(x - 0.625)], & x \geq 0.125, \end{cases} \quad (49)$$

which is plotted in Fig. 40. As with $I_7(x; \mathbf{p})$, this intensity varies linearly with p_1 or p_2 in either half of the sensor. The key difference in this example is that $I_8(x; \mathbf{p})$ is contrived in such a way that the total intensity $\sum_i I_8(x_i|\mathbf{p})$ is independent of \mathbf{p} . As a result, the PMF (relative intensity) $P_8(i|\mathbf{p})$ only depends on p_1 on the left half of the sensor and p_2 on the right half of the sensor. Naturally, the same is true of the likelihood function $L_8(\mathbf{p}|i)$, as seen in Fig. 41. Since the gradient of $L_8(\mathbf{p}|i)$ always points along p_1 or p_2 (when it is nonzero), the FIM and its inverse are always diagonal, indicating that there is zero covariance between the parameters. For any value of \mathbf{p} ,

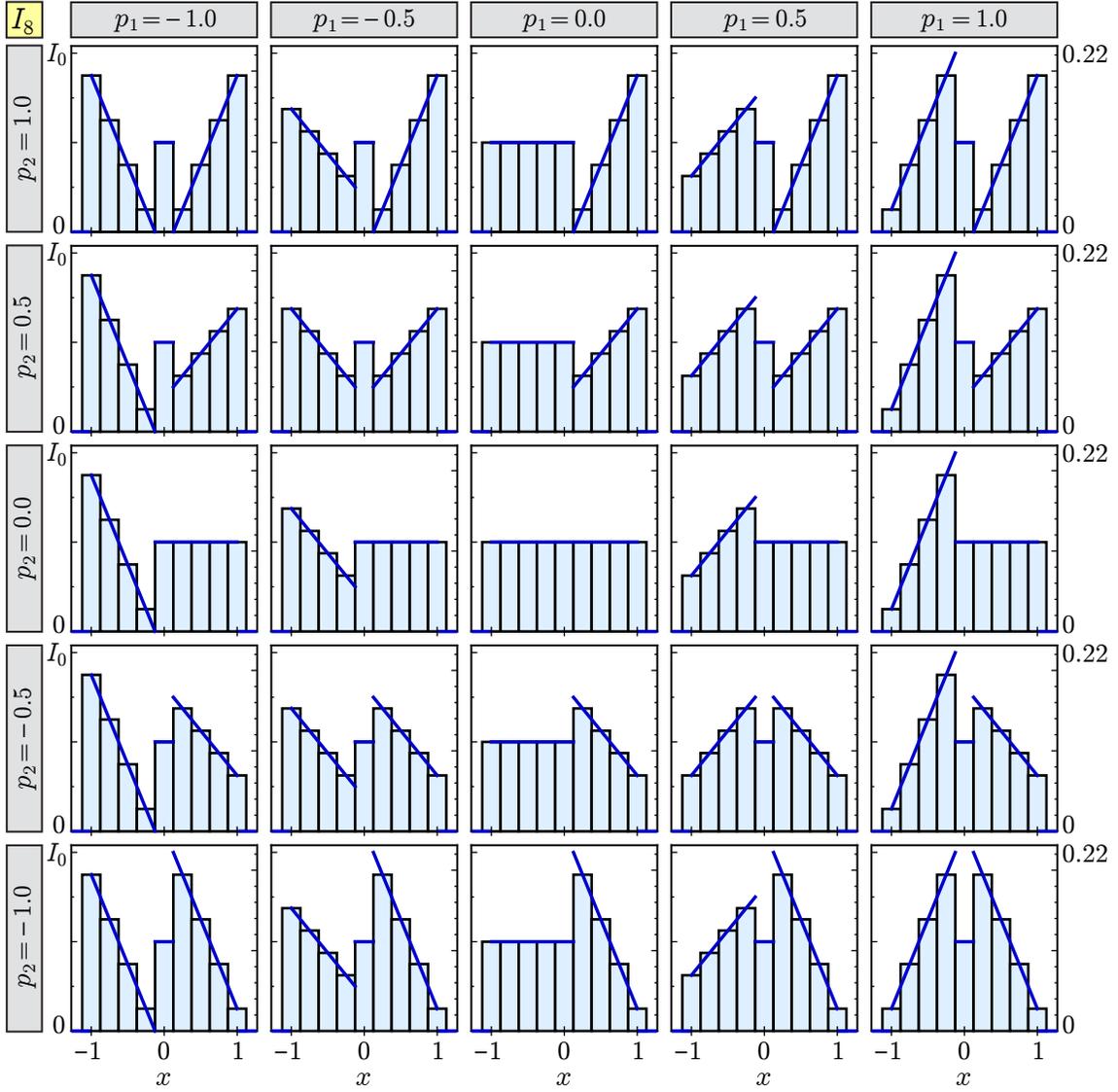


Figure 40: Plots of $I_8(x; \mathbf{p})$ (left axes) and $P_8(i; \mathbf{p})$ (right axes) for several values of p_1 and p_2 .

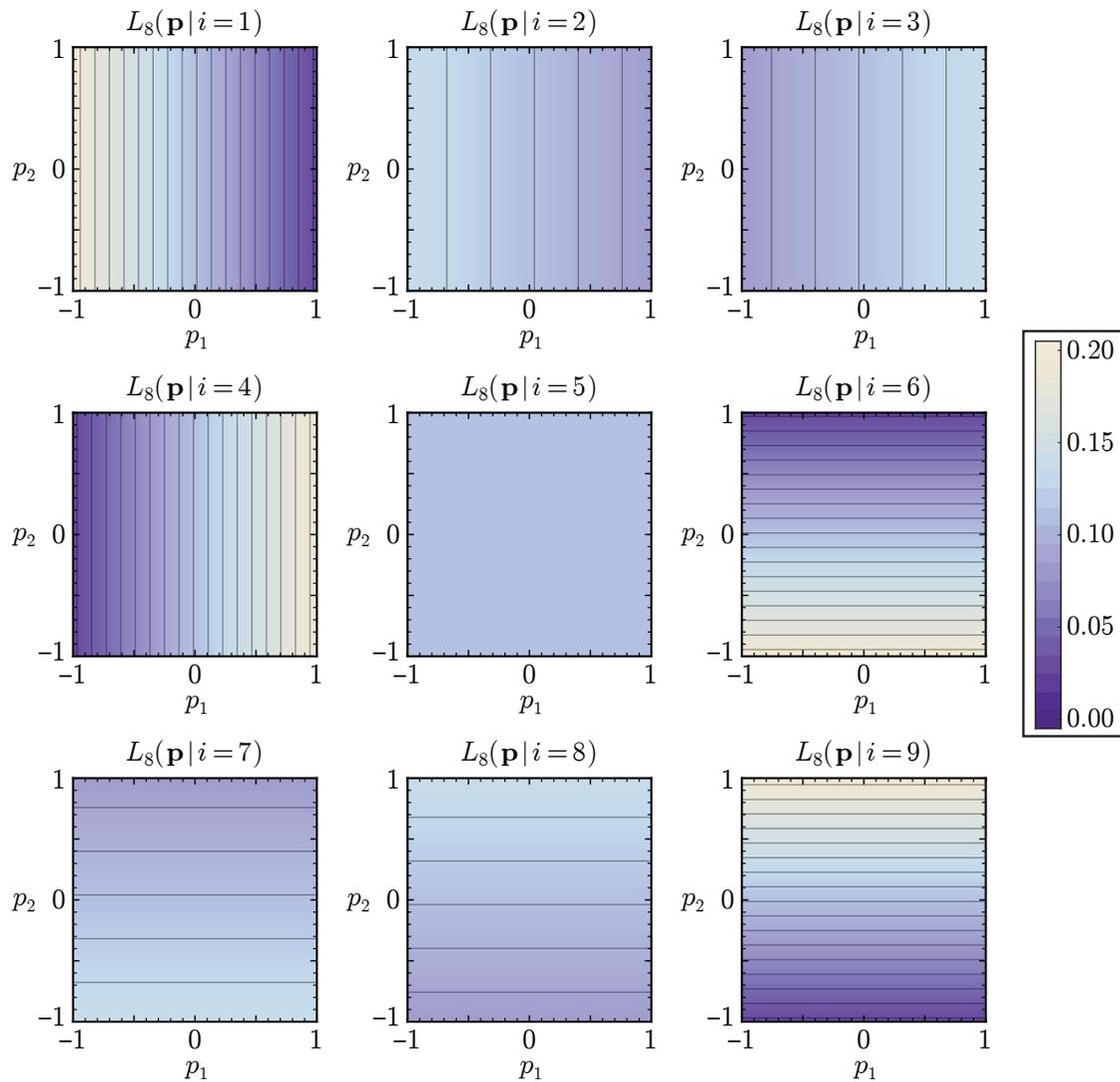


Figure 41: Likelihood functions $L_8(\mathbf{p} | i)$ associated with each pixel i for a measurement of $I_8(x; \mathbf{p})$. Contour lines are shown in increments of 0.01.

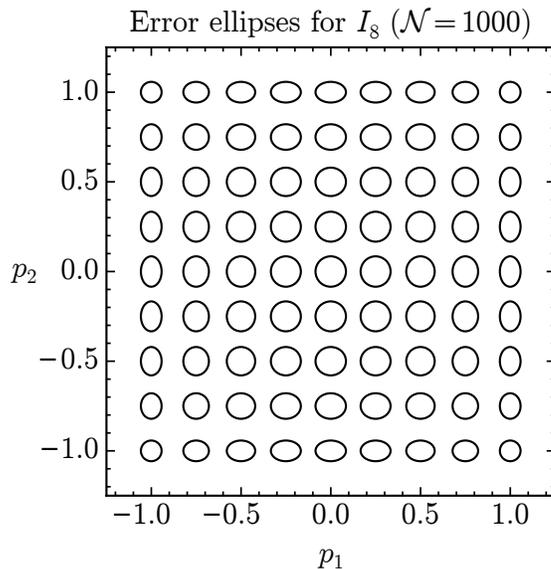


Figure 42: Ellipses representing the expected standard deviation error of a 1000-photon measurement of $I_8(x; \mathbf{p})$ with true parameter values p_1 and p_2 , sampled over a 9×9 grid in parameter space.

the principal axes of the error ellipse are oriented along the p_1 and p_2 axes, as seen in Fig. 42. When $\mathbf{p} = (0, 0)$, the error ellipse is circular, meaning that the expected error is identical for each parameter. For other values of \mathbf{p} , the relative errors of the two parameters vary in a symmetric fashion over the region of interest. Fig. 43 contains plots of the log-likelihood functions $\ell_8(\mathbf{p}|\tilde{\mathbf{I}})$ for simulated 1000-photon measurements of $I_8(x; \mathbf{p})$ with true parameter values $\mathbf{p} = (0, 0)$ and $\mathbf{p} = (0.63, -0.25)$. In light of the above observations, it should come as no surprise that the distribution is highly symmetric about the MLE in each case.

To recap, the contrast between $I_7(x; \mathbf{p})$ and $I_8(x; \mathbf{p})$ illustrates a limitation of the MLE approach described in Section 3, as well as one of its key strengths. The shortcoming is that the sole reliance of the parameter estimate on the relative intensity can introduce correlations between parameters that are not present in the absolute (unnormalized) intensity; furthermore, any additional information contained within the overall scale of the intensity is ignored. On the other hand, the advantage of the method is that with good experimental design, the relative intensity can be tailored for optimal sensitivity and minimal coupling between parameters, so that there is no need to analyze the unnormalized intensity. Conveniently, the MLE formalism includes a straightforward error metric (the FIM) that can be used to predict and

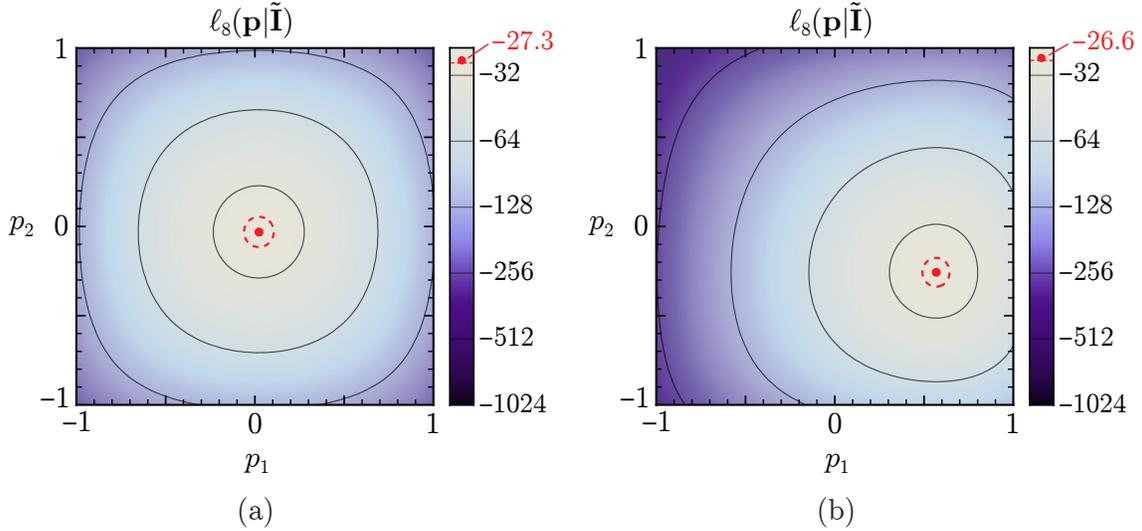


Figure 43: Log-likelihood functions $\ell_8(\mathbf{p}|\tilde{\mathbf{I}})$ for simulated 1000-photon measurements of $I_8(x; \mathbf{p})$ with true parameter values (a) $\mathbf{p} = (0, 0)$ and (b) $\mathbf{p} = (0.63, -0.25)$. The plots are shaded on a logarithmic scale with solid contour lines drawn at powers of 2, as indicated in the legend. The peak of each distribution is marked with a red dot. The locations of these maxima (i.e., the MLEs for each measurement) are $\mathbf{p} = (0.021, -0.029)$ and $\mathbf{p} = (0.565, -0.256)$, respectively. The dashed contour line indicates where the likelihood $L_8(\mathbf{p}|\tilde{\mathbf{I}})$ drops to $1/\sqrt{e}$ times its peak value, representing the standard deviation confidence interval for the MLE.

optimize the sensitivity of the measurement. As stated earlier, the lack of reliance on total intensity has the added benefit of reducing or eliminating errors arising from fluctuations of the source power.

7.5 Two-parameter off-null measurement

The final two examples correspond to a pair of off-null measurements involving two parameters, following Eq. (20). For the sake of simplicity, we consider a case in which \bar{p}_1 is linear in x , \bar{p}_2 varies as a cosine, and both T_1 and T_2 are constant in magnitude and are mutually related by $T_1 = \pm iT_2$. (Recall that it was mentioned after Eq. (21) that it is convenient to tailor the illumination and filtering so that $\text{Im}(T_1^*T_2)$ is nonzero.) With this, Eq. (20) becomes

$$I_9(x; \mathbf{p}) = 0.125\Pi(x)[(p_1 - x)^2 + (p_2 - \cos(\pi x))^2]. \quad (50)$$

Notice that the p_1 term of $I_9(x; \mathbf{p})$ is identical to that of the one-parameter example $I_3(x; p_1)$ considered in Section 6.3, with $c = 1$. The p_2 term introduces an additional departure from the null condition, which varies sinusoidally over the sensor. These spatial variations were chosen to allow comparison between $I_9(x; \mathbf{p})$ and the earlier two-parameter example $I_6(x; \mathbf{p})$, for which the terms with x and $\cos(\pi x)$ dependences were linear in p_1 and p_2 , respectively. The intensity and PMF for $I_9(x; \mathbf{p})$ are shown in Fig. 44. Compared to $I_6(x; \mathbf{p})$, observe that the off-null configuration employed in the present example produces more dramatic variations in the shape of the intensity profile with respect to p_1 and p_2 , particularly for parameter values close to zero.

The likelihood functions $L_9(\mathbf{p}|i)$ for each pixel, which are plotted in Fig. 45, have a far more complex structure than the ones seen in the previous examples. The contributions of each pixel have similar shapes, consisting of a peaked distribution that rotates clockwise and changes scale as i runs from 1 to 9. The balance between different pixels and the densely spaced contours of constant likelihood suggest that the FIM is likely to be large and diagonal, which would result in a small and diagonal covariance matrix. As indicated by the ellipse map shown in Fig. 46, the expected error is indeed quite small, particularly for parameter values near $\mathbf{p} = (0, 0)$, for which the total measured intensity tends to be the lowest. This symmetric ellipse pattern, with the error growing as the departure from null increases, is typical for an off-null measurement. There is a considerable covariance between p_1 and p_2 near the edge of the parameter range, but in nearly all cases, the error is still smaller (often significantly so) than it would be for a measurement of $I_6(x; \mathbf{p})$ (see Fig. 33 for comparison).

The log-likelihood functions $\ell_9(\mathbf{p}|\tilde{\mathbf{I}})$ obtained for two simulated measurements of $I_9(x; \mathbf{p})$ with true parameter values $\mathbf{p} = (0, 0)$ and $\mathbf{p} = (0.63, -0.25)$ can be found in Fig. 47. For the $\mathbf{p} = (0, 0)$ case, the likelihood is a sharply peaked distribution, with the location of the peak (the MLE) nearly coinciding with the true value of \mathbf{p} . (The numerical results are provided in the figure caption.) The distribution is considerably wider and less symmetric for the $\mathbf{p} = (0.63, -0.25)$ case, but the standard deviation uncertainty is still quite small. These results demonstrate the usefulness of an off-null measurement, which enables the simultaneous estimate of multiple parameters with high precision.

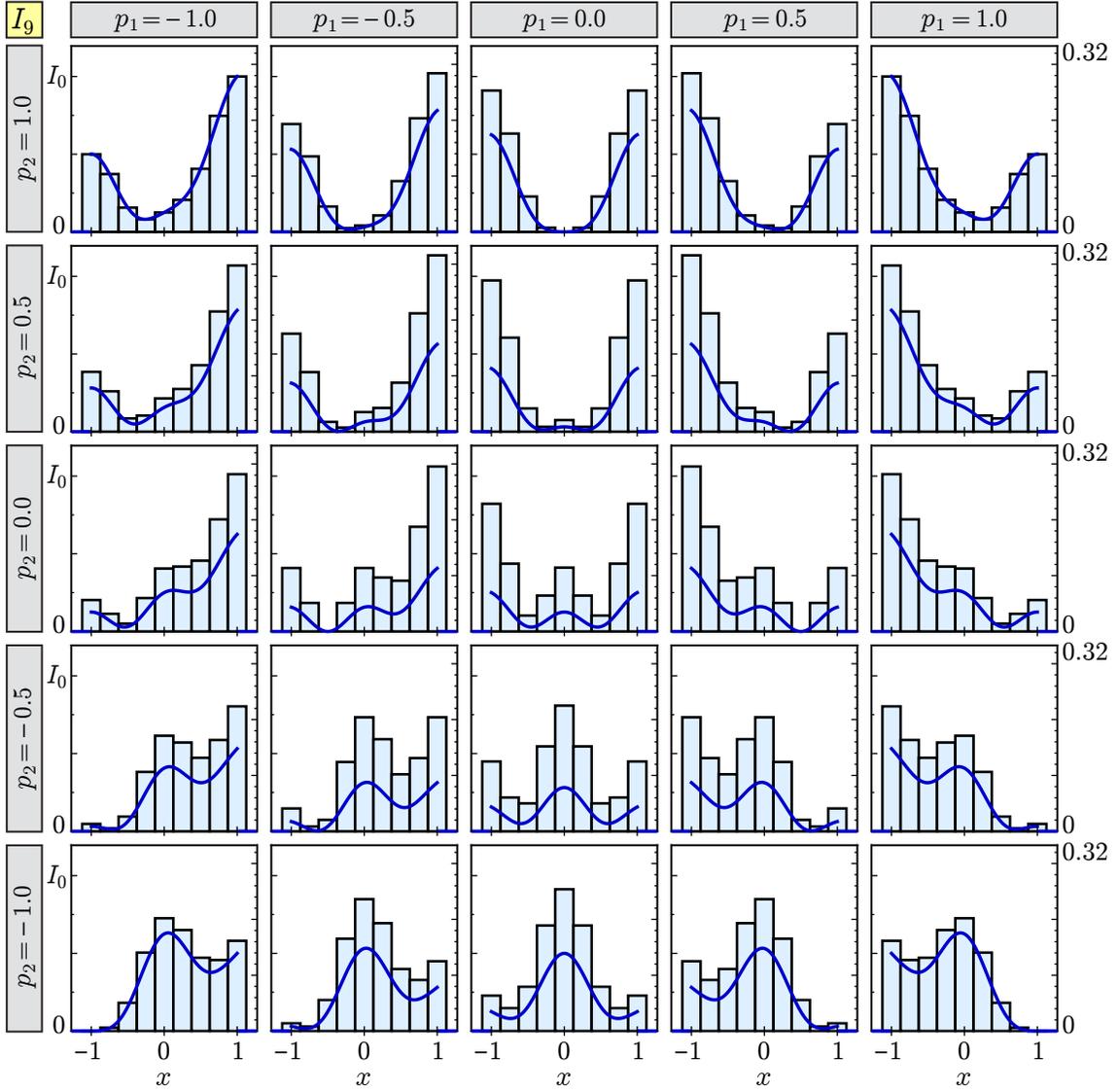


Figure 44: Plots of $I_9(x; \mathbf{p})$ (left axes) and $P_9(i|\mathbf{p})$ (right axes) for several values of p_1 and p_2 .

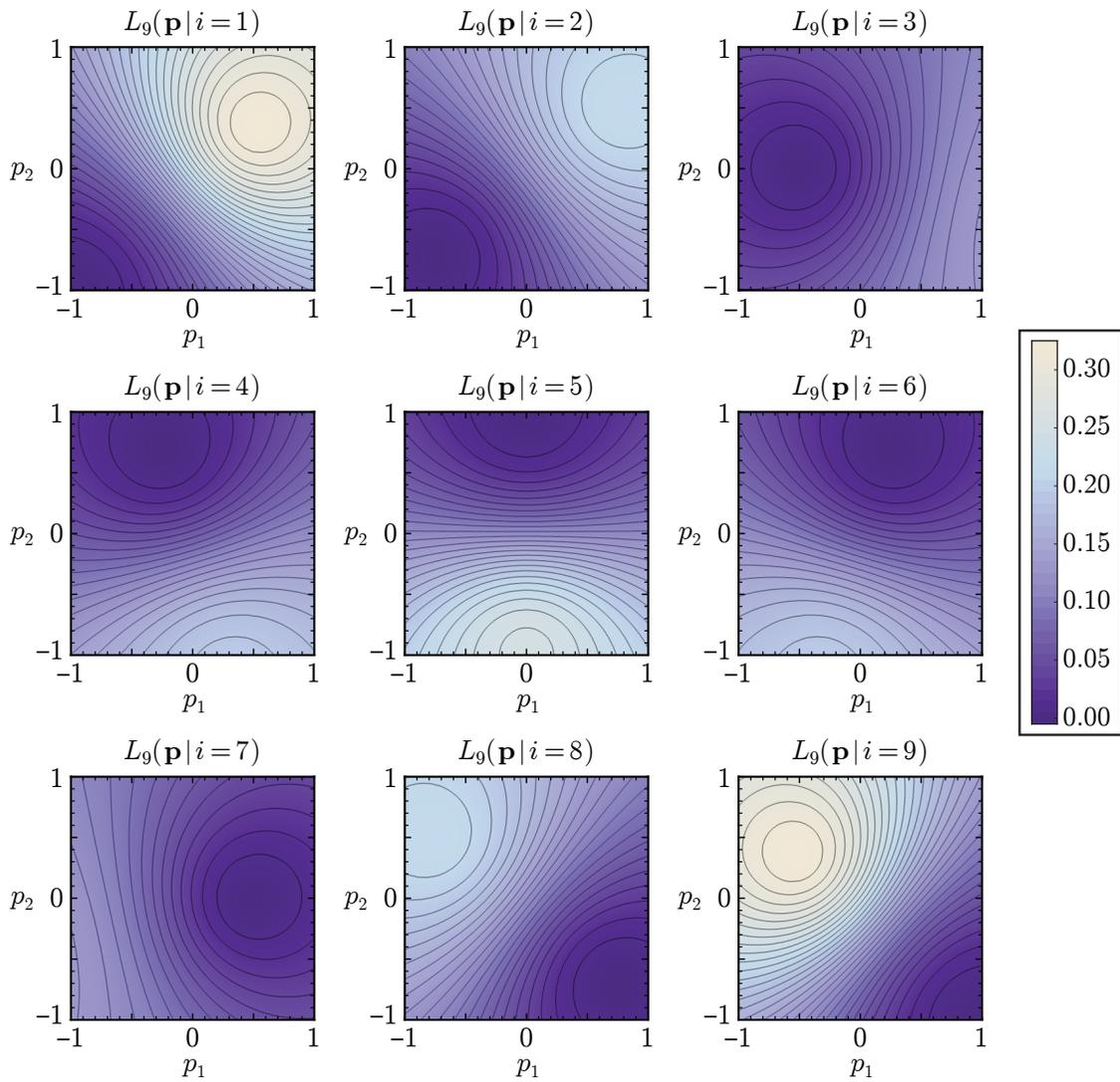


Figure 45: Likelihood functions $L_9(\mathbf{p}|i)$ associated with each pixel i for a measurement of $I_9(x; \mathbf{p})$. Contour lines are shown in increments of 0.01.

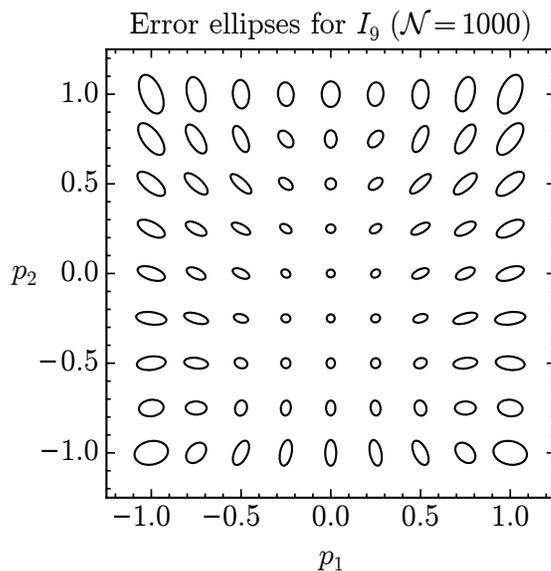


Figure 46: Ellipses representing the expected standard deviation error of a 1000-photon measurement of $I_9(x; \mathbf{p})$ with true parameter values p_1 and p_2 , sampled over a 9×9 grid in parameter space.

7.6 Two-parameter off-null measurement with smaller departure from null

For the final example, consider the intensity distribution

$$I_{10}(x; \mathbf{p}) = 0.320\Pi(x)\left[(p_1 - 0.25x)^2 + (p_2 - 0.25\cos(\pi x))^2\right]. \quad (51)$$

Notice that the x dependence of $I_{10}(x; \mathbf{p})$ is identical to the previous case except that the departure from null associated with each parameter is four times smaller. As seen in the plots of the intensity profile (Fig. 48) and the likelihood functions for each pixel (Fig. 49), the measurement is very sensitive to variations in p_1 and p_2 when both parameters are close to zero. However, similarly to the $|c| \ll 1$ case in Section 6.3, this comes at the expense of greatly reduced sensitivity (i.e., slower variations in likelihood) near the edges of the region of interest.

The expected error ellipses based on the FIM are plotted for several parameter values in Fig. 50. The error for a measurement of $I_{10}(x; \mathbf{p})$ exhibits the same pattern as that of $I_9(x; \mathbf{p})$ (see Fig. 46), but with a larger disparity between the magnitudes of the errors near the center and edges of the parameter range. More precisely, for a

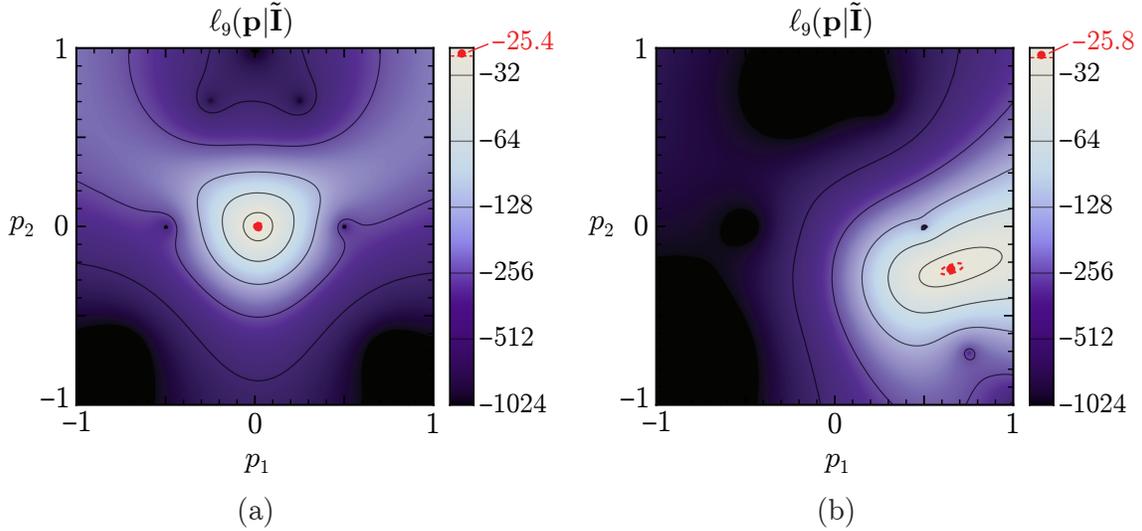


Figure 47: Log-likelihood functions $\ell_9(\mathbf{p}|\tilde{\mathbf{I}})$ for simulated 1000-photon measurements of $I_9(x; \mathbf{p})$ with true parameter values (a) $\mathbf{p} = (0, 0)$ and (b) $\mathbf{p} = (0.63, -0.25)$. The plots are shaded on a logarithmic scale with solid contour lines drawn at powers of 2, as indicated in the legend. (Values smaller than -1024 are shown in black.) The peak of each distribution is marked with a red dot. The locations of these maxima (i.e., the MLEs for each measurement) are $\mathbf{p} = (0.016, 0.001)$ and $\mathbf{p} = (0.648, -0.237)$, respectively. The dashed contour line indicates where the likelihood $L_9(\mathbf{p}|\tilde{\mathbf{I}})$ drops to $1/\sqrt{e}$ times its peak value, representing the standard deviation confidence interval for the MLE. (The dashed contour in plot (a) is too small to be seen.)

true parameter value of $\mathbf{p} = (0, 0)$, the expected error is exactly four times smaller for a measurement of I_{10} as it is for a measurement of I_9 ; conversely, the errors near the far corners of the parameter range (where $|p_1| \approx |p_2| \approx 1$) are about two to three times larger for I_{10} than for I_9 .

Finally, the log-likelihood functions $\ell_{10}(\mathbf{p}|x)$ for simulated measurements of I_{10} with true parameter values $\mathbf{p} = (0, 0)$ and $\mathbf{p} = (0.63, -0.25)$ are shown in Fig. 51. As expected, the likelihood for the $\mathbf{p} = (0, 0)$ case is extremely narrowly distributed about its peak, producing an estimate with error on the order of 0.001. In contrast, the distribution for $\mathbf{p} = (0.63, -0.25)$ is substantially wider; for parameter values with magnitudes closer to 1, the width of the distribution would continue to grow.

The practical implication of this example is that an off-null measurement can be tailored for high sensitivity over an arbitrarily small range of parameter values. Therefore, it is possible to design an iterative experiment for which the parameter estimate is refined through a series of successive measurements. As an example of

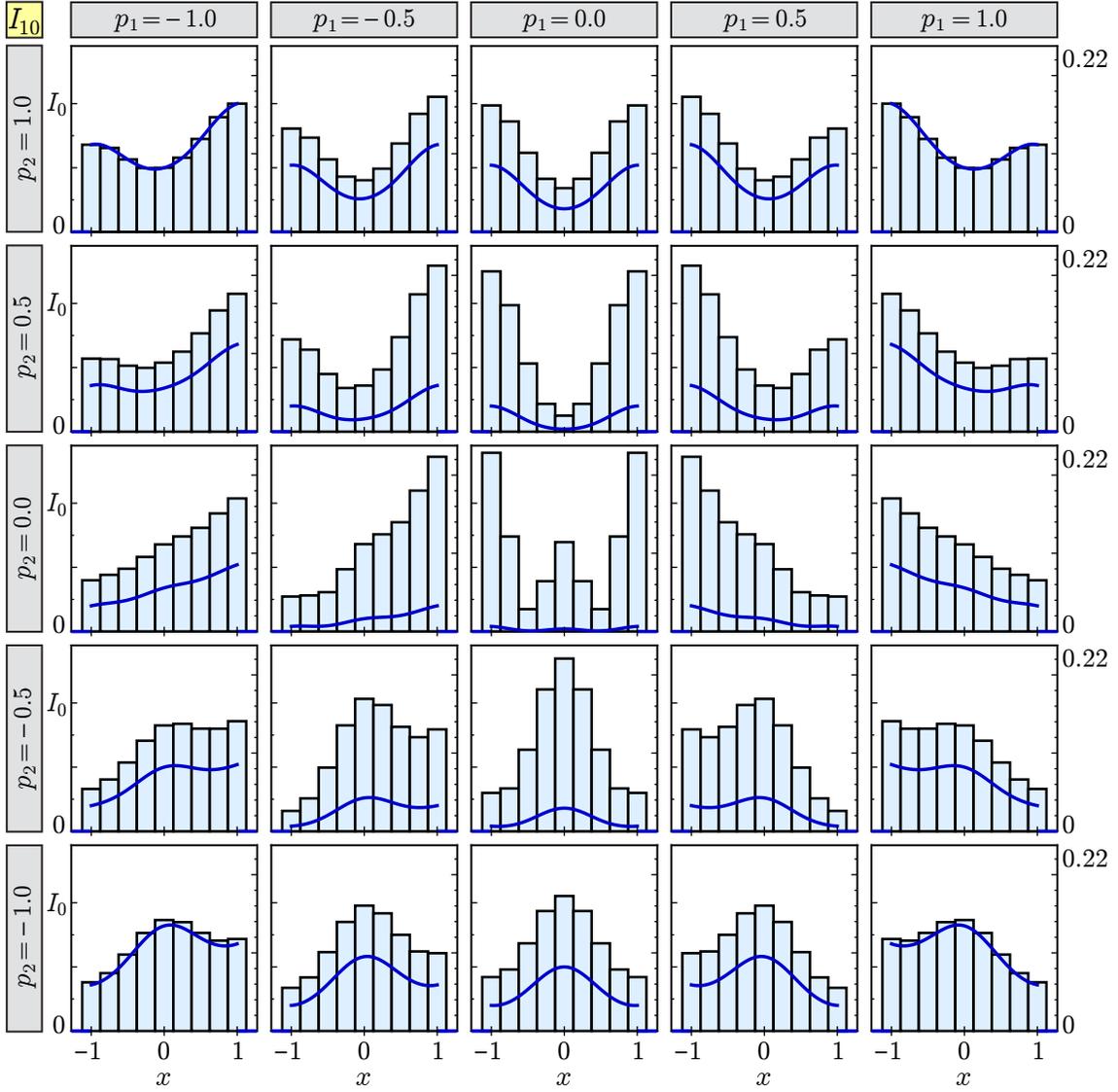


Figure 48: Plots of $I_{10}(x; \mathbf{p})$ (left axes) and $P_{10}(i|\mathbf{p})$ (right axes) for several values of p_1 and p_2 .

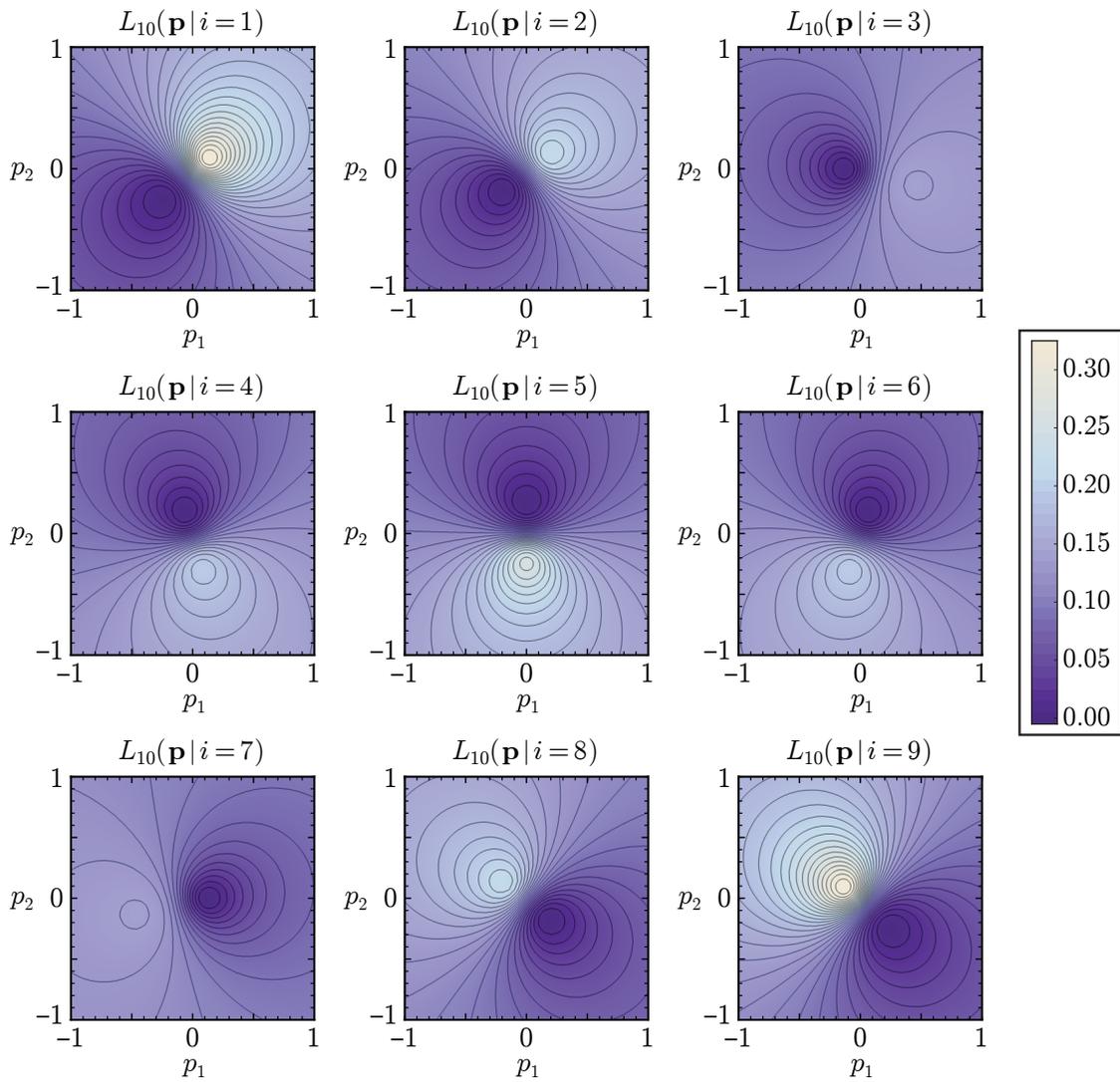


Figure 49: Likelihood functions $L_{10}(\mathbf{p} | i)$ associated with each pixel i for a measurement of $I_{10}(x; \mathbf{p})$. Contour lines are shown in increments of 0.01.

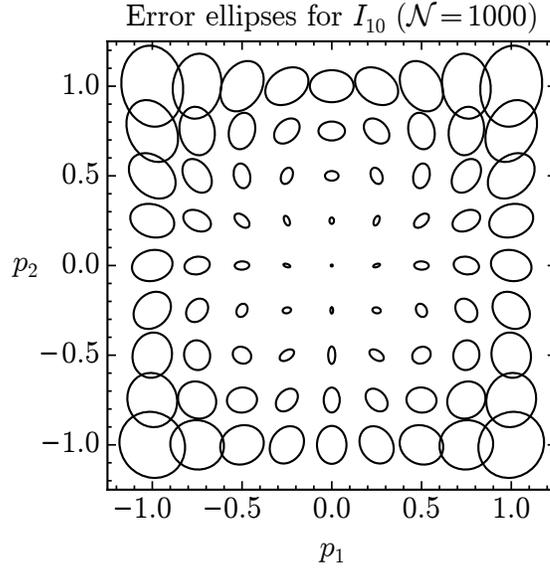


Figure 50: Ellipses representing the expected standard deviation error of a 1000-photon measurement of $I_{10}(x; \mathbf{p})$ with true parameter values p_1 and p_2 , sampled over a 9×9 grid in parameter space.

this iterative procedure, suppose that we wish to refine the measurement of $I_9(x; \mathbf{p})$ with true parameter values $\mathbf{p} = (0.63, -0.25)$ obtained in Section 7.5. The plot of the log-likelihood function $\ell_9(\mathbf{p}|\tilde{\mathbf{I}})$ for this measurement is shown again in Fig. 52(a); the MLE based on this initial measurement is $\mathbf{p} = (0.648, -0.237)$. To refine the parameter estimate, the experimental configuration could be altered such that the output intensity follows the distribution

$$I_9^{(2)}(x; \mathbf{p}) = \Pi(x) [(p_1 - 0.648 - 0.5x)^2 + (p_2 + 0.237 - 0.5 \cos(\pi x))^2], \quad (52)$$

where the constant normalization factor in front of $\Pi(x)$ has been omitted for simplicity.¹⁵ This distribution is designed so that the departure from null is half as large and centered at the previous MLE. The resulting log-likelihood function $\ell_9^{(2)}(\mathbf{p}|\tilde{\mathbf{I}})$ for a simulated measurement of 1000 photons, shown in Fig. 52(b), is much more narrowly distributed than $\ell_9(\mathbf{p}|\tilde{\mathbf{I}})$. The MLE based on the refined measurement is found to be $\mathbf{p} = (0.644, -0.255)$. This process can be applied repeatedly to obtain an estimate with arbitrary precision (barring experimental limitations, as discussed in the

¹⁵In a real experiment, the leading factor (which determines the peak intensity) would typically vary under different experimental configurations. Since the MLE approach ignores any information contained in this scaling factor, it is not important for this discussion.

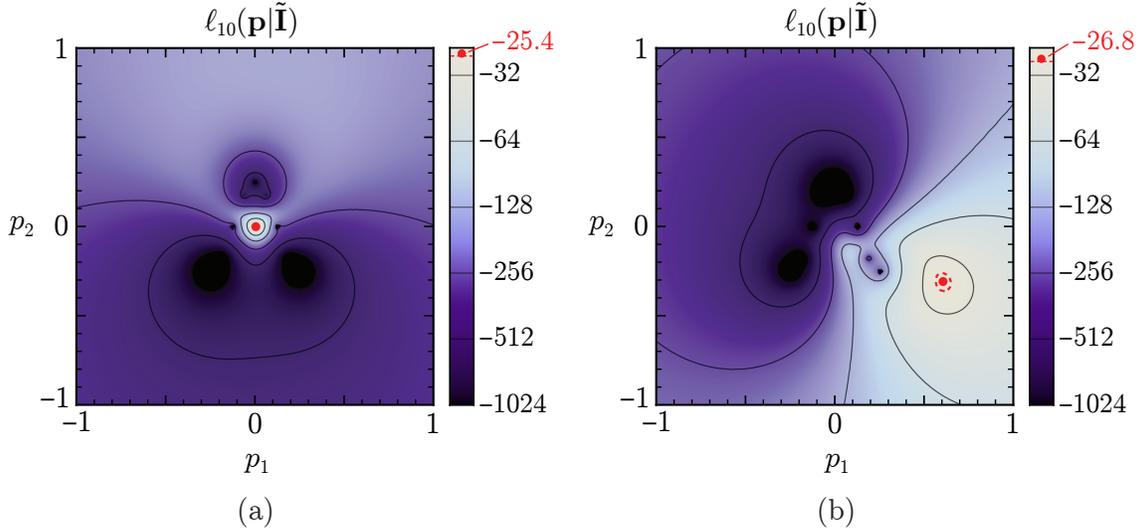


Figure 51: Log-likelihood functions $\ell_{10}(\mathbf{p}|\tilde{\mathbf{I}})$ for simulated 1000-photon measurements of $I_{10}(x; \mathbf{p})$ with true parameter values (a) $\mathbf{p} = (0, 0)$ and (b) $\mathbf{p} = (0.63, -0.25)$. The plots are shaded on a logarithmic scale with solid contour lines drawn at powers of 2, as indicated in the legend. (Values smaller than -1024 are shown in black.) The peak of each distribution is marked with a red dot. The locations of these maxima (i.e., the MLEs for each measurement) are $\mathbf{p} = (0.004, -2.6 \times 10^{-4})$ and $\mathbf{p} = (0.602, -0.308)$, respectively. The dashed contour line indicates where the likelihood $L_{10}(\mathbf{p}|\tilde{\mathbf{I}})$ drops to $1/\sqrt{e}$ times its peak value, representing the standard deviation confidence interval for the MLE. (The dashed contour in plot (a) is too small to be seen.)

Concluding Remarks). The intensity distributions and resulting MLEs for the first four iterations of the process, including the two mentioned above, are listed in Table 6, and the log-likelihood functions for simulated measurements of $I_9^{(3)}(x; \mathbf{p})$ and $I_9^{(4)}(x; \mathbf{p})$ are plotted in Fig. 52(c,d). As seen in the table, the MLE gets closer to the true value with each iteration, leading to a final estimate of $\mathbf{p} = (0.631, -0.249)$. As this happens, the likelihood function becomes increasingly compact with an exceptionally sharp peak, which is the reason for the improvement in accuracy. However, note that the calculation of the MLE must be performed carefully in this case since the likelihood function may contain local maxima or regions with very small slopes, which can cause problems with the numerical search procedure. These issues can generally be mitigated by using the previous MLE as the starting point for the search.

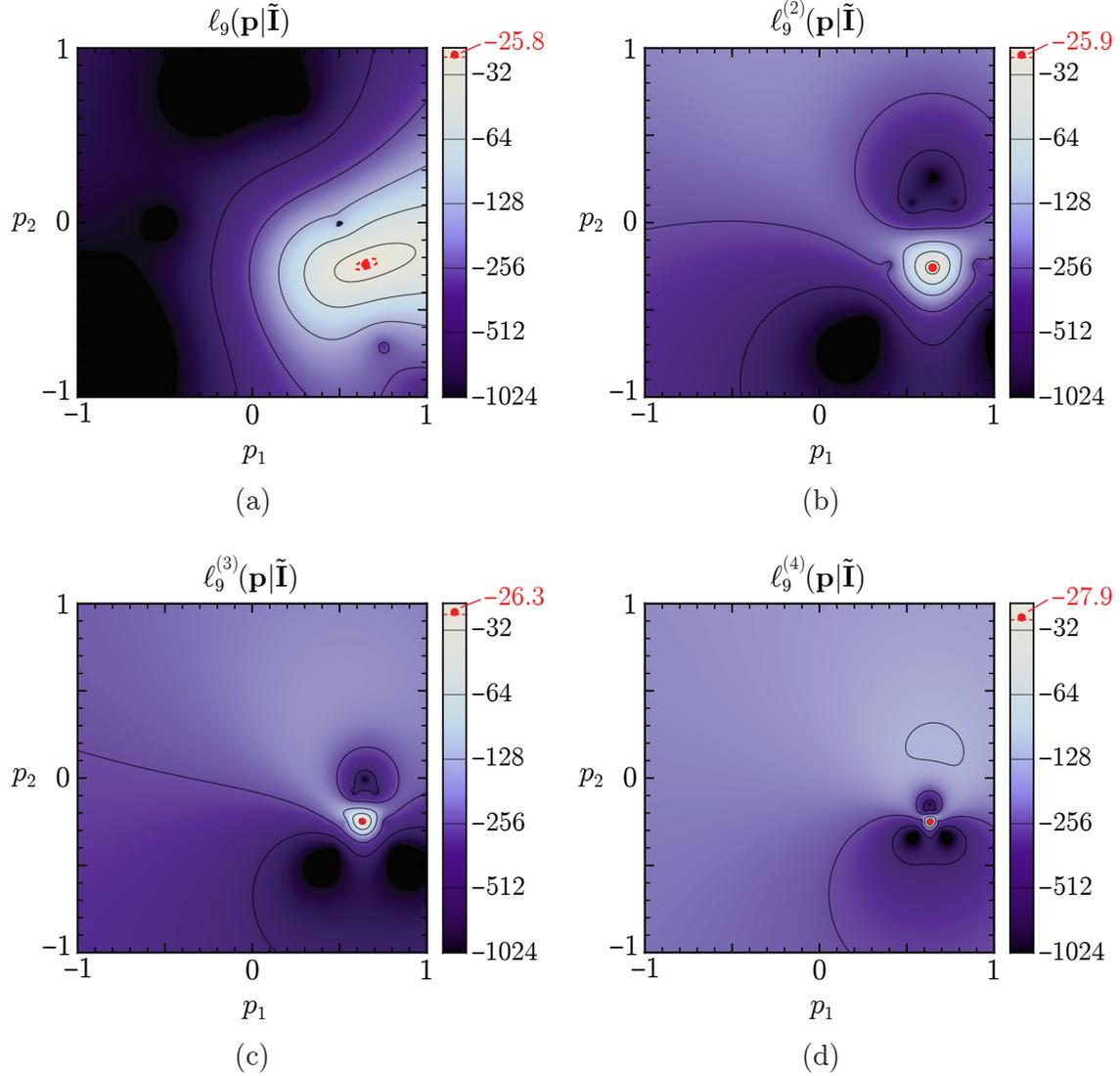


Figure 52: Log-likelihood functions for simulated 1000-photon measurements of intensity distributions (a) $I_9(x; \mathbf{p})$, (b) $I_9^{(2)}(x; \mathbf{p})$, (c) $I_9^{(3)}(x; \mathbf{p})$, and (d) $I_9^{(4)}(x; \mathbf{p})$ obtained throughout a four-step iterative measurement with true parameter values $\mathbf{p} = (0.63, -0.25)$. The peaks of each distribution are indicated with a red dot, and their locations are listed in the rightmost column of Table 6. The dashed red contour in plot (a) represents the standard deviation confidence interval; the confidence intervals in plots (b-d) are too small to be seen.

Intensity distribution	MLE for \mathbf{p}
$I_9(x; \mathbf{p}) \propto [(p_1 - x)^2 + (p_2 - \cos(\pi x))^2]$	(0.648, -0.237)
$I_9^{(2)}(x; \mathbf{p}) \propto [(p_1 - 0.648 - 0.50x)^2 + (p_2 + 0.237 - 0.50 \cos(\pi x))^2]$	(0.644, -0.255)
$I_9^{(3)}(x; \mathbf{p}) \propto [(p_1 - 0.644 - 0.25x)^2 + (p_2 + 0.255 - 0.25 \cos(\pi x))^2]$	(0.628, -0.246)
$I_9^{(4)}(x; \mathbf{p}) \propto [(p_1 - 0.628 - 0.10x)^2 + (p_2 + 0.246 - 0.10 \cos(\pi x))^2]$	(0.631, -0.249)

Table 6: Intensity distributions used for a simulated four-step iterative measurement with true parameter values $\mathbf{p} = (0.63, -0.25)$, along with the MLEs obtained from the simulated intensities at each step. The off-null departures for iterations 2 through 4 are each centered at the MLE from the previous iteration. The magnitude of the departure from null decreases with each iteration in order to refine the accuracy of the estimate.

8 Concluding remarks

This tutorial has summarized the fundamental concepts of maximum likelihood estimation and their application to the measurement of an optical intensity distribution. In this treatment, one or more parameters are estimated from the shape of the intensity profile, without regard for the total measured power. However, the power incident on the detector is still relevant because it determines the uncertainty of the parameter estimate, which scales as the inverse of the square root of the number of detected photons. Depending on the needs of a given application, the methods discussed here may be used to optimize the performance of an experiment for minimal estimation error per photon detected by the sensor or per photon emitted by the source. Some sample code for calculating and evaluating the uncertainty of the maximum likelihood estimate in such an experiment can be found in the appendix.

Given that our emphasis was in the information encoded in the shape of the measured intensity distribution, we devoted a section and four examples to the approach referred to as off-null or weak measurements, whose goal is precisely to boost the dependence of the measured intensity shape on the parameter(s) of interest. Particular attention was paid to the case of two measured parameters, where it was discussed that the two functions $T_n = \langle f | \hat{T}_n | i \rangle$ should preferably be out of phase. In general, whether the measured parameters are two or more, it is advisable to make not only the real functions \bar{p}_n as different as possible, but also the complex functions T_n func-

tionally different and not only different in phase. The reader can verify, for example, that by including different functional forms for T_1 and T_2 in the example in Section 7.5 while keeping them at a relative phase of $\pm\pi/2$, the standard deviation ellipses can be made considerably smaller.

As mentioned in Section 7.6, from a statistical standpoint, an iterative MLE approach can be employed to obtain a parameter estimate with arbitrary precision. That is, for any fixed, reasonably large number of detected photons \mathcal{N} , the measurement experiment can be designed to make the Cramér-Rao bound arbitrarily small, meaning that there is no fundamental limit to the sensitivity of the measurement. In practice, the accuracy is determined by experimental factors, including but not limited to:

- the bit depth and signal-to-noise ratio of the sensor;
- the power of the source (which affects the number of photons detected in a given time interval);
- the level of precision and temporal stability of the experimental configuration (e.g., SLM control in the application mentioned above);
- the validity of the theoretical model and any approximations made;
- other sources of random or systematic error (e.g., thermal fluctuations or ghost images).

(Note that the second point above can be addressed by optimizing the FIM for emitted photons, as in Section 6.3.) In any case, the statistical methods discussed in this tutorial are still useful for determining the best nominal design for an experiment, as well as for obtaining parameter estimates from measured data based on a theoretical or empirical model.

Appendix

A Mathematica code

This appendix describes a simple implementation of the MLE approach described in Section 3 using the Wolfram Mathematica [59] programming language. The code includes functions to calculate the PMF, likelihood function, simulated intensity, Fisher information, and MLE for an optical measurement, as well as functions to plot the expected error ellipse(s) for a two-parameter measurement. For simplicity, the code was written for the one-and-two parameter cases explored in Sections 6 and 7; as necessary, it could readily be extended for higher-dimensional problems. The code also assumes a one-dimensional spatial coordinate.

Section A.1 below contains a list of the functions defined in this package and the syntax for their use. The function definitions are provided in Section A.2. Finally, a few example calculations are shown in Section A.3.

A.1 Syntax and usage

The functions defined in this package are detailed in Table 7.

Table 7: Summary of symbols and functions created to perform MLE calculations in Mathematica. When applicable, the relevant equations from the main text are listed in the second column.

Symbol	Eq.	Syntax and description
<code>xv</code>	(23)	One-dimensional array of spatial coordinates x_i of each pixel. Can be modified to simulate different pixel arrays.
<code>I_{sim}</code>	(25), (44)	<code>I_{sim}[j][p₁][x]</code> evaluates the one-parameter intensity distribution $I_j(x; p_1)$ at coordinate x and parameter value p_1 . <code>I_{sim}[j][{p₁, p₂}[x]</code> evaluates the two-parameter intensity distribution $I_j(x; \mathbf{p})$ at coordinate x and parameter values $\mathbf{p} = (p_1, p_2)$. In each function below, the argument j identifies which distribution I_j should be used; the corresponding function <code>I_{sim}</code> should be defined beforehand. See Section A.2 for a few examples.
<code>I_{photonsim}</code>	N/A	<code>I_{photonsim}[j][p, N]</code> randomly generates a simulated measurement of $I_j(x; \mathbf{p})$ containing N photons, such as those shown in Tables 3 and 4. The output is an array with the same length as <code>xv</code> containing the number of photons detected at each pixel.
<code>P</code>	(5)	<code>P[j][i, p]</code> calculates the PMF $P_j(i \mathbf{p})$, evaluated at pixel i for true parameter value \mathbf{p} , or equivalently the likelihood function $L_j(\mathbf{p} i)$ at \mathbf{p} associated with pixel i . The argument \mathbf{p} should be specified in the form p_1 or $\{p_1, p_2\}$ for the one- and two-parameter cases, respectively.
<code>PInt</code>	(7)	<code>PInt[j][I-tilde, p]</code> calculates the probability $P_j(\tilde{\mathbf{I}} \mathbf{p})$ of measuring an intensity distribution $\tilde{\mathbf{I}}$ given true parameter value \mathbf{p} , or equivalently the likelihood function $L_j(\mathbf{p} \tilde{\mathbf{I}})$. The argument $\tilde{\mathbf{I}}$ is an array with the same length as <code>xv</code> . The argument \mathbf{p} should be specified in the form p_1 or $\{p_1, p_2\}$ for the one- and two-parameter cases, respectively.
<code>LLIntSum</code>	(8)	<code>LLIntSum[j][p, I-tilde]</code> calculates the sum appearing in the log-likelihood function $\ell_j(\mathbf{p} \tilde{\mathbf{I}})$. The constant term $\ln P_0$ in Eq. (8) is ignored to improve computational efficiency when calculating the MLE. The arguments \mathbf{p} and $\tilde{\mathbf{I}}$ are the same as for <code>PInt</code> above.

Continued on next page

Symbol	Eq.	Syntax and description
Fisher1D	(9), (10)	<code>Fisher1D[j] [p₁]</code> calculates the (scalar) expected unit Fisher information $J_j(p_1)$, evaluated at parameter value p_1 , for the one-parameter intensity distribution $I_j(x; p_1)$. <code>Fisher1D[j] [p₁, $\tilde{\mathbf{I}}$]</code> calculates the (scalar) observed Fisher information $J_j^{(\text{obs})}(p_1; \tilde{\mathbf{I}})$ for a measured intensity $\tilde{\mathbf{I}}$, which should be specified as an array with the same length as <code>xv</code> .
Fisher2D	(9), (10)	<code>Fisher2D[j] [{p₁, p₂}]</code> calculates the 2×2 expected unit Fisher information matrix $\mathbb{J}_j(\mathbf{p})$, evaluated for parameter values $\mathbf{p} = (p_1, p_2)$, for the two-parameter intensity distribution $I_j(x; \mathbf{p})$. <code>Fisher2D[j] [{p₁, p₂}, $\tilde{\mathbf{I}}$]</code> calculates the 2×2 observed Fisher information matrix $\mathbb{J}_j^{(\text{obs})}(\mathbf{p}; \tilde{\mathbf{I}})$ for a measured intensity $\tilde{\mathbf{I}}$, which should be specified as an array with the same length as <code>xv</code> .
MLE1D	N/A	<code>MLE1D[j] [$\tilde{\mathbf{I}}$, <i>cons</i>]</code> finds the maximum likelihood estimate for p_1 based on a measurement $\tilde{\mathbf{I}}$ of the one-parameter intensity distribution $I_j(x; p_1)$. The optional argument <i>cons</i> may be used to specify a constraint on p_1 , for example, $-1 \leq p_1 \leq 1$.
MLE2D	N/A	<code>MLE2D[j] [$\tilde{\mathbf{I}}$, <i>cons</i>]</code> finds the maximum likelihood estimate for $\mathbf{p} = (p_1, p_2)$ based on a measurement $\tilde{\mathbf{I}}$ of the one-parameter intensity distribution $I_j(x; p_1)$. The optional argument <i>cons</i> may be used to specify constraints on p_1 and/or p_2 , for example, $\{-1 \leq p_1 \leq 1, -1 \leq p_2 \leq 1\}$.
ErrorEllipse	N/A	<code>ErrorEllipse[\mathbb{J}, {p₁, p₂}]</code> produces a graphics primitive for the error ellipse associated with the 2×2 Fisher information matrix \mathbb{J} . The ellipse is centered at $\mathbf{p} = (p_1, p_2)$.
EllipseGrid	N/A	<code>EllipseGrid[j] [\mathcal{N}, M]</code> plots an $M \times M$ grid of error ellipses based on the expected Fisher information matrix for a measurement of \mathcal{N} photons based on the two-parameter intensity distribution $I_j(x; \mathbf{p})$. Each ellipse corresponds to a different pair of true parameter values $\mathbf{p} = (p_1, p_2)$. The grid is sampled over the region $-1 \leq p_1, p_2 \leq 1$ at M equally spaced points in each dimension.

A.2 Code

The first step of the calculation is to define an array of spatial coordinates xv and the intensity distribution I_{sim} . For purpose of demonstration, consider the examples discussed in Sections 6 and 7. The coordinates of the 9-pixel array given in Eq. (23) may be defined as follows:

```
In[1]:= xv = Array[#&, 9, {-1., 1}]
```

```
Out[1]= {-1., -0.75, -0.5, -0.25, 0., 0.25, 0.5, 0.75, 1.}
```

(For users familiar with MATLAB, this command is the Mathematica equivalent of `linspace(-1, 1, 9)`.) The next step is to define the one- and two-parameter intensity distributions I_1 through I_{10} , which are summarized in Tables 2 and 5 in the main text. Their definitions (with normalization constant $I_0 = 1$) are below.

```
In[2]:= (*For one-parameter intensity distributions, use the following function
to extract p1 in the event that multiple parameters are supplied*)
getp1[p_]:=First[Flatten[{p}]];

(*Hard aperture with radius 1*)
rect[x_]:= UnitBox[x/2];

(*Intensity distributions*)
Isim[1][p_][x_]:= rect[x]*(0.5+0.5 getp1[p]*x);
Isim[2][p_][x_]:= rect[x]*(0.9+0.1 getp1[p]*x);
Isim[{3, c_}][p_][x_]:= rect[x]/(Abs[c]+1)^2*(getp1[p]-c*x)^2;
Isim[{4, d_}][p_][x_]:= rect[x]/(Abs[d]+2)^2*(getp1[p]-x-d)^2;
Isim[5][p_][x_]:= 0.563 rect[x]*(2+p[[1]]*x+p[[2]] Sin[Pi*x]);
Isim[6][p_][x_]:= 0.25 rect[x]*(2+p[[1]]*x+p[[2]] Cos[Pi*x]);
Isim[7][p_][x_]:= 0.5 rect[x]*(1+p[[1]]*x*Boole[x<0]+p[[2]]*x*Boole[x>0]);
Isim[8][p_][x_]:= 0.5 rect[x]*(1+2p[[1]]*(x+.625)*Boole[x<-1/8] +
2p[[2]]*(x-.625)*Boole[x>1/8]);
Isim[9][p_][x_]:= 0.125 rect[x]*((p[[1]]-x)^2+(p[[2]]-Cos[Pi*x])^2);
Isim[10][p_][x_]:=
0.320 rect[x]*((p[[1]]-0.25x)^2+(p[[2]]-0.25Cos[Pi*x])^2);
```

The remaining functions listed in Table 7 are defined as follows:

```

In[3]:= Iphotonsim[case_][p_, nphotons_] :=
  BinCounts[
    RandomChoice[(Isim[case][p]/@xv)->Range[Length[xv]], nphotons],
    {1, Length[xv]+1}];

P[case_][i_, p_] := Isim[case][p][xv[[i]]]/Total[Isim[case][p]/@xv];

PInt[case_][Imeas_, p_] :=
  (Factorial[Total[Imeas]]/Apply[Times, (Factorial/@Imeas)])*
  Product[If[Imeas[[i]]==0, 1, P[case][i, p]^Imeas[[i]]], {i, Length[xv]}];

LLIntSum[case_][p_, Imeas_] :=
  Total[ Imeas * Log[P[case][#, p] &/@ Range[Length[xv]]] ];

(*Use the following function to avoid errors from 0*Log[0]*)
prod[xv_, yv_] := MapThread[If[#1==0, 0, #1*#2]&, {xv, yv}];

Fisher1D[case_][p_, Imeas_:0] :=
  Module[{pvar, Pv, PorI},
    Pv=Table[P[case][i, pvar], {i, Length[xv]}];
    PorI=If[Imeas===0, Pv, Imeas];
    Total[prod[PorI, D[Log[Pv], pvar]^2]/.pvar->p];

Fisher2D[case_][p_, Imeas_:0] :=
  Module[{Pv, PorI, p1var, p2var, J11, J12, J22},
    Pv=Table[P[case][i, {p1var, p2var}], {i, Length[xv]}];
    PorI=If[Imeas===0, Pv, Imeas];
    J11=Total[prod[PorI, D[Log[Pv], p1var]^2]];
    J12=Total[prod[PorI, D[Log[Pv], p1var]*D[Log[Pv], p2var]]];
    J22=Total[prod[PorI, D[Log[Pv], p2var]^2]];
    {{J11, J12}, {J12, J22}}/.{p1var->p[[1]], p2var->p[[2]]};

MLE1D[case_][Imeas_, cons_:{}] :=
  p1/.Last@Maximize[{LLIntSum[case][p1, Imeas], cons}, p1];

MLE2D[case_][Imeas_, cons_:{}] :=
  {p1, p2}/.Last@Maximize[{LLIntSum[case][{p1, p2}, Imeas], cons}, {p1, p2}];

ErrorEllipse[J_, p_] :=
  Rotate[Circle[p, Sqrt[#[[1]]]], ArcTan@@#[[2, 1]]] &@
  Eigensystem[Inverse[J]];

EllipseGrid[case_][nphotons_, numpts_] :=
  Module[{pvals, eig, ellipses},
    pvals=Array[Identity, numpts, {-1, 1}];
    ellipses=Table[
      ErrorEllipse[nphotons*Fisher2D[case][{pp1, pp2}], {pp1, pp2}],
      {pp1, pvals}, {pp2, pvals}];
    Graphics[{Black, Thick, ellipses}, Frame->True] ];

```

A.3 Examples

The following examples demonstrate some calculations for the one- and two-parameter intensity distributions $I_1(x; p_1)$ and $I_9(x; \mathbf{p})$ from Sections 6.1 and 7.5, respectively. (Note: numerical results will vary depending on the seed for Mathematica's random number generator.)

```
In[4]:= (*Simulate an intensity measurement with 1000 photons, given true
parameter values p1=0.63 and p2=-0.25*)
I meas1=Iphotonsim[1][.63,1000]
I meas9=Iphotonsim[9][{.63,-.25},1000]
```

```
Out[4]= {45,56,71,98,99,109,161,177,184}
```

```
Out[5]= {255,160,124,131,155,88,7,14,66}
```

```
In[6]:= (*Find the maximum likelihood estimates based on each simulated
intensity from above. For the one-parameter example, constrain
p1 to the range over which I1 is valid*)
MLE1=MLE1D[1][I meas1,{-1<=p1<=1}]
MLE9=MLE2D[9][I meas9]
```

```
Out[6]= 0.638811
```

```
Out[7]= {0.618961, -0.257141}
```

```
In[8]:= (*Calculate the expected and observed Fisher information for the
one-parameter measurement of I1. Then invert and take the square root
to calculate the standard deviation error for each case*)
J1=1000*Fisher1D[1][.63];
J1obs=Fisher1D[1][.63,I meas1];
Sqrt[1/J1]
Sqrt[1/J1obs]
```

```
Out[8]= 0.0406453
```

```
Out[9]= 0.0397809
```

```

In[10]:= (*Calculate the expected and observed Fisher information matrices for
the two-parameter measurement of I9. Then invert the matrices to
obtain the covariance matrices for each case*)
J9=1000*Fisher2D[9][{.63,-.25}];
J9obs=Fisher2D[9][{.63,-.25},Imeas9];
Inverse[J9]
Inverse[J9obs]

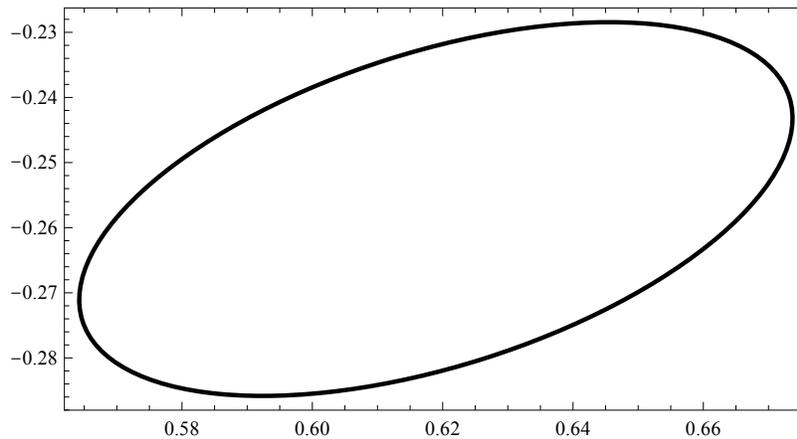
Out[10]= {{0.00299494, 0.000763432}, {0.000763432, 0.000823367}}

Out[11]= {{0.00283021, 0.000765918}, {0.000765918, 0.000839408}}

In[12]:= (*Plot the error ellipse for the measurement of I9, centered at the MLE*)
Graphics[{{Thick,ErrorEllipse[J9,MLE9]}},Frame->True]

```

Out[12]=



```

In[13]:= (*The following command reproduces Fig.~46 (sans formatting), which
contains a grid of error ellipses representing the expected error for
a measurement of I9.*)
EllipseGrid[9][1000,9]

```

Acknowledgments

The authors would like to thank Philippe Réfrégier, Thomas G. Brown, Valentine Wasik, Stephen Head, Jonathan Ellis, Steven Gillmer, Alan Rosenbluth, Tim Farrell, Andrew Jordan, and John Howell for helpful discussions and suggestions. This work was supported by funding from the National Science Foundation (NSF) (PHY-1507278). MAA received funding from the Excellence Initiative of Aix-Marseille University - A*MIDEX, a French “Investissements d’Avenir” programme.

References

- [1] M. Born and E. Wolf, *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light* (Elsevier, 2013).
- [2] L. Mandel and E. Wolf, *Optical Coherence and Quantum Optics* (Cambridge university press, 1995).
- [3] E. Wolf, *Introduction to the Theory of Coherence and Polarization of Light* (Cambridge University Press, 2007), p. 157.
- [4] A. Vella, S. T. Head, T. G. Brown, and M. A. Alonso, “Simultaneous measurement of multiple parameters of a subwavelength structure based on the weak value formalism,” *Phys. Rev. Lett.* **122**, 123603 (2019).
- [5] A. Vella and M. A. Alonso, “Optimal birefringence distributions for star test polarimetry,” arXiv preprint arXiv:1908.02142 (2019).
- [6] R. A. Fisher, “On the mathematical foundations of theoretical statistics,” *Phil. Trans. R. Soc. Lond. A* **222**, 309–368 (1922).
- [7] R. A. Fisher, “Theory of statistical estimation,” in “Mathematical Proceedings of the Cambridge Philosophical Society” (Cambridge University Press, 1925), Vol. 22, No. 5, pp. 700–725.
- [8] R. A. Fisher, “The logic of inductive inference,” *Journal of the Royal Statistical Society* **98**, 39–82 (1935).
- [9] H. Cramér, “Mathematical methods of statistics,” Princeton U. Press, Princeton p. 500 (1946).
- [10] J. L. Melsa, “Decision and estimation theory,” Tech. rep. (1978).
- [11] P. Réfrégier, *Noise Theory and Application to Physics: From Fluctuations to Information* (Springer, New York, 2003).
- [12] H. L. Van Trees Jr, A. B. Baggeroer, L. Collins, R. Kurth, and T. Cruise, “Detection and estimation theory,” Tech. rep., Research Laboratory of Electronics (RLE) at the Massachusetts Institute of Technology (1968).

- [13] L. L. Scharf, *Statistical Signal Processing*, Vol. 98 (Addison-Wesley Reading, MA, 1991).
- [14] S. M. Kay, *Fundamentals of Statistical Signal Processing* (Prentice Hall PTR, 1993).
- [15] R. Frieden, *Probability, Statistical Optics, and Data Testing*, Vol. 10 (Springer-Verlag, 2001), 3rd ed.
- [16] G. King, *Unifying Political Methodology: The Likelihood Theory of Statistical Inference* (University of Michigan Press, 1998).
- [17] S. Gailmard, *Statistical Modeling and Inference for Social Science* (Cambridge University Press, 2014).
- [18] I. J. Myung, “Tutorial on maximum likelihood estimation,” *Journal of Mathematical Psychology* **47**, 90–100 (2003).
- [19] A. Ly, M. Marsman, J. Verhagen, R. P. Grasman, and E.-J. Wagenmakers, “A tutorial on Fisher information,” *Journal of Mathematical Psychology* **80**, 40–55 (2017).
- [20] F. Davidson and J. Amoss, “Sequential photon-counting statistics and maximum-likelihood estimation techniques for gaussian optical fields,” *JOSA* **63**, 30–37 (1973).
- [21] H. H. Barrett and K. J. Myers, *Foundations of Image Science* (John Wiley & Sons, 2013).
- [22] I. Moon and B. Javidi, “Three dimensional imaging and recognition using truncated photon counting model and parametric maximum likelihood estimator,” *Opt. Express* **17**, 15709–15715 (2009).
- [23] D. Aloni, A. Stern, and B. Javidi, “Three-dimensional photon counting integral imaging reconstruction using penalized maximum likelihood expectation maximization,” *Opt. Express* **19**, 19681–19687 (2011).
- [24] T. J. Holmes and Y.-H. Liu, “Acceleration of maximum-likelihood image restoration for fluorescence microscopy and other noncoherent imagery,” *JOSA A* **8**, 893–907 (1991).

- [25] M. Bertero, P. Boccacci, G. Desiderà, and G. Vicidomini, “Image deblurring with poisson data: from cells to galaxies,” *Inverse Problems* **25**, 123006 (2009).
- [26] D. Kasaragod, S. Makita, S. Fukuda, S. Beheregaray, T. Oshika, and Y. Yasuno, “Bayesian maximum likelihood estimator of phase retardation for quantitative polarization-sensitive optical coherence tomography,” *Opt. Express* **22**, 16472–16492 (2014).
- [27] M.-A. Henn, H. Gross, F. Scholze, M. Wurm, C. Elster, and M. Bär, “A maximum likelihood approach to the inverse problem of scatterometry,” *Opt. Express* **20**, 12771–12786 (2012).
- [28] A. Aiello, G. Puentes, D. Voigt, and J. Woerdman, “Maximum-likelihood estimation of mueller matrices,” *Opt. Lett.* **31**, 817–819 (2006).
- [29] H. Hu, R. Ossikovski, and F. Goudail, “Performance of maximum likelihood estimation of mueller matrices taking into account physical realizability and gaussian or poisson noise statistics,” *Opt. Express* **21**, 5117–5129 (2013).
- [30] R. D. Ramkhalawon, T. G. Brown, and M. A. Alonso, “Imaging the polarization of a light field,” *Opt. Express* **21**, 4106–4115 (2013).
- [31] E. W. Rogala and H. H. Barrett, “Phase-shifting interferometry and maximum-likelihood estimation theory,” *Appl. Opt.* **36**, 8871–8876 (1997).
- [32] E. W. Rogala and H. H. Barrett, “Phase-shifting interferometry and maximum-likelihood estimation theory. II. A generalized solution,” *Appl. Opt.* **37**, 7253–7258 (1998).
- [33] P. Su, J. H. Burge, and R. E. Parks, “Application of maximum likelihood reconstruction of subaperture data for measurement of large flat mirrors,” *Appl. Opt.* **49**, 21–31 (2010).
- [34] H. H. Barrett, C. Dainty, and D. Lara, “Maximum-likelihood methods in wave-front sensing: stochastic models and likelihood functions,” *JOSA A* **24**, 391–414 (2007).
- [35] P. Réfrégier, C. Scotté, H. B. de Aguiar, H. Rigneault, and F. Galland, “Precision of proportion estimation with binary compressed raman spectrum,” *JOSA A* **35**, 125–134 (2018).

- [36] P. Réfrégier and F. Galland, “Bhattacharyya bound for raman spectrum classification with a couple of binary filters,” *Optics letters* **44**, 2228–2231 (2019).
- [37] S. S. Rao, “Advanced statistical inference,” <http://www.stat.tamu.edu/~suhasini/teaching613/STAT613.pdf> (2017). Accessed: 2018-03-31.
- [38] B. Efron and D. V. Hinkley, “Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information,” *Biometrika* **65**, 457–483 (1978).
- [39] X. Cao, “Relative performance of expected and observed Fisher information in covariance estimation for maximum likelihood estimates,” Ph.D. thesis, John Hopkins University, Baltimore, MD (2013).
- [40] W. K. Newey and D. L. McFadden, “Chapter 36: Large sample estimation and hypothesis testing”, in R. F. Engle and D. L. McFadden, editors, *Handbook of Econometrics*, Vol. 4 (1994).
- [41] E. Naftali and N. C. Makris, “Necessary conditions for a maximum likelihood estimate to become asymptotically unbiased and attain the Cramér–Rao Lower Bound. Part I. General approach with an application to time-delay and Doppler shift estimation,” *The Journal of the Acoustical Society of America* **110**, 1917–1930 (2001).
- [42] M. Friendly, G. Monette, J. Fox *et al.*, “Elliptical insights: understanding statistical methods through elliptical geometry,” *Statistical Science* **28**, 1–39 (2013).
- [43] H. Arwin, S. Welin-Klintström, and R. Jansson, “Off-null ellipsometry revisited: basic considerations for measuring surface concentrations at solid/liquid interfaces,” *Journal of Colloid and Interface Science* **156**, 377–382 (1993).
- [44] F. Zernike, “Phase contrast, a new method for the microscopic observation of transparent objects,” *Physica* **9**, 686–698 (1942).
- [45] F. Zernike, “Phase contrast, a new method for the microscopic observation of transparent objects part II,” *Physica* **9**, 974–986 (1942).
- [46] D. Murphy, *Fundamentals of Light Microscopy and Digital Imaging* (Wiley-Liss, New York, 2001).

- [47] Y. Aharonov, D. Z. Albert, and L. Vaidman, “How the result of a measurement of a component of the spin of a spin-1/2 particle can turn out to be 100,” *Phys. Rev. Lett.* **60**, 1351 (1988).
- [48] B. Tamir and E. Cohen, “Introduction to weak measurements and weak values,” *Quanta* **2**, 7–17 (2013).
- [49] B. E. Svensson, “Pedagogical review of quantum measurement theory with an emphasis on weak measurements,” *Quanta* **2**, 18–49 (2013).
- [50] O. Hosten and P. Kwiat, “Observation of the spin hall effect of light via weak measurements,” *Science* **319**, 787–790 (2008).
- [51] P. B. Dixon, D. J. Starling, A. N. Jordan, and J. C. Howell, “Ultrasensitive beam deflection measurement via interferometric weak value amplification,” *Phys. Rev. Lett.* **102**, 173601 (2009).
- [52] Y. Qin, Y. Li, H. He, and Q. Gong, “Measurement of spin Hall effect of reflected light,” *Optics letters* **34**, 2551–2553 (2009).
- [53] M. R. Dennis and J. B. Götte, “The analogy between optical beam shifts and quantum weak measurements,” *New J. Phys.* **14**, 073013 (2012).
- [54] J. B. Götte and M. R. Dennis, “Generalized shifts and weak values for polarization components of reflected light beams,” *New J. Phys.* **14**, 073016 (2012).
- [55] G. Jayaswal, G. Mistura, and M. Merano, “Observing angular deviations in light-beam reflection via weak measurements,” *Opt. Lett.* **39**, 6257–6260 (2014).
- [56] K. Pedersen and O. Keller, “Photoelastic properties of metals measured by off-null ellipsometry,” *Appl. Opt.* **25**, 226–234 (1986).
- [57] A. Van der Vaart and J. A. Wellner, “Existence and consistency of maximum likelihood in upgraded mixture models,” *Journal of Multivariate Analysis* **43**, 133–146 (1992).
- [58] E. Demidenko and H. Massam, “On the existence of the maximum likelihood estimate in variance components models,” *Sankhyā: The Indian Journal of Statistics, Series A* pp. 431–443 (1999).

[59] Wolfram Research, Inc., “Mathematica, Version 11.1,” Champaign, IL, 2018.