# Quantum learning: optimal classification of qubit states

**Mădălin Guţă**[1] **and Wojciech Kotłowski**[2]

[1]School of Mathematical Sciences, University of Nottingham, University Park, NG7 2RD Nottingham, United Kingdom
[2] CWI, Science Park 123, 1098 XG Amsterdam
E-mail: madalin.guta@nottingham.ac.uk

**Abstract.** Pattern recognition is a central topic in Learning Theory with numerous applications such as voice and text recognition, image analysis, computer diagnosis. The statistical set-up in classification is the following: we are given an i.i.d. training set $(X_1, Y_1), \dots (X_n, Y_n)$ where $X_i$ represents a feature and $Y_i \in \{0, 1\}$ is a label attached to that feature. The underlying joint distribution of $(X, Y)$ is unknown, but we can learn about it from the training set and we aim at devising low error classifiers $f : X \to Y$ used to predict the label of new incoming features.

Here we solve a quantum analogue of this problem, namely the classification of two arbitrary *unknown* qubit states. Given a number of 'training' copies from each of the states, we would like to 'learn' about them by performing a measurement on the training set. The outcome is then used to design mesurements for the classification of future systems with unknown labels. We find the asymptotically optimal classification strategy and show that typically, it performs strictly better than a plug-in strategy based on state estimation.

The figure of merit is the *excess risk* which is the difference between the probability of error and the probability of error of the optimal measurement when the states are known, that is the Helstrom measurement. We show that the excess risk has rate $n^{-1}$ and compute the exact constant of the rate.

**Contents**



## 1. Introduction

Statistical learning theory [1, 2, 3, 4] is a broad research field stretching over statistics and computer science, whose general goal is to devise algorithms which have the ability to learn from data. One of the central learning problems is how to recognise patterns [5], with practical applications in speech and text recognition, image analysis, computer-aided diagnosis, data mining.

The paradigm of Quantum Information theory is that quantum systems carry a new type of information with potentially revolutionary applications such as faster computation and secure communication [6]. Motivated by these theoretical challenges, Quantum Engineering is developing new tools to control and accurately measure individual quantum systems [7]. In the process of engineering exotic quantum states, statistical validation has become a standard experimental procedure [8, 9] and Quantum Statistical Inference has passed from its purely theoretical status in the 70's [10, 11] to a more practically oriented theory at the interface between the classical and quantum worlds [12, 13, 14, 15].

In this paper we put forward a new type of quantum statistical problem inspired by learning theory, namely *quantum state classification*. Similar ideas have already appeared in the

physics [16, 17, 18, 19] and learning [20, 21, 22] literature but here we emphasise the close connection with learning and we aim at going beyond the special models based on group symmetry and pure states. However, we limit ourselves to a two dimensional state which could be regarded as a toy model from the viewpoint of learning theory, but hope that more interesting applications will follow.

Before explaining what quantum classification is, let us briefly mention the classical set-up we aim at generalising. In *supervised learning* the goal is to learn to predict an output $y \in \mathcal{Y}$, given the input (object) $x \in \mathcal{X}$, where input and output are assumed to be correlated and have an *unknown* joint distribution $\mathbb{P}$ over $\mathcal{X} \times \mathcal{Y}$. To do this, we are first provided with a set of $n$ previously observed inputs with known output variables (called *training examples*), i.e. independent random pairs $(X_i, Y_i), i = 1, \ldots, n$ drawn from $\mathbb{P}$. Using the training set, we construct a function $h_n \colon \mathcal{X} \to \mathcal{Y}$ to predict the output for future, yet unseen objects. When $\mathcal{Y} = \{0, 1\}$, i.e. the output is a binary variable, this is called *binary classification* and is the typical set-up in pattern recognition. The input space is usually considered to be a subset of $p$-dimensional space $\mathbb{R}^p$, so that the object $x$ can be described by $p$ measurement values often called *features*. This description is very general as it allows e.g. to handle categorical (non-numerical) values (encoded as integer numbers), images (e.g. measured brightness of each pixel corresponds to a separate feature), time series (features corresponds to the values of the signal at given times), etc.

In this paper, we consider the classification problem in which the objects to be classified are quantum states. Simply, we have a quantum system prepared in either of two *unknown* quantum states and we want to know which one it is. As in the classical case, this only makes sense if we are also provided with training examples from both states, with their respective labels, from which we can learn about the two alternatives. How could such a scenario occur? Suppose we send one bit of information through a noisy quantum channel which is not known. To decode the information (the input in this case) we need to be able to classify the output states corresponding to the two inputs. Alternatively, the binary variable may be related to a coupling of the channel which we want to detect.

Needless to say, quantum systems are *intrinsically statistical* and can be 'learned' only by repeated preparation, so that the problem is really the quantum extension of the classical classification problem. On the other hand this is related to the problem of state discrimination which in the case of two hypotheses, has an explicit solution known as the Helstrom measurement [11]. The point is that when the states are unknown, the Helstrom measurements is itself unknown and has to be learned from the training set. An intuitive solution would be a *plug-in* procedure: first estimate the two states, and then apply the Helstrom measurement corresponding to the estimates on any new to-be-classified state. This indeed gives a reasonable classification strategy, but as we will see, this is *not* the best one. The optimal strategy in the asymptotic framework is to directly estimate the Helstrom measurement without intermediate states estimation. The optimality is defined by the natural figure of merit called *excess risk*, which is the difference between the expected error probability and the error probability of the Helstrom measurement. We show that the excess risk converges to zero with the size of the training set as $n^{-1}$ and the ratio between the optimal and state estimation plug-in risk is a constant factor.

Our analysis is valid for arbitrary mixed states and is performed in a *pointwise, local minimax* (rather than Bayesian) setting which captures the behaviour of the risk around any pair of states. The key theoretical tool is the recently developed theory of local asymptotic normality (LAN) for quantum states [23, 24, 25, 26] which is an extension of the classical concept in mathematical statistics introduced by Le Cam [27]. Roughly, LAN says that the

collective state $\rho_\theta^{\otimes n}$ of $n$ i.i.d. quantum systems can be approximated by a simple Gaussian state of some classical variables and quantum oscillators. This was used to derive optimal state estimation strategies for arbitrary mixed states of arbitrary finite dimension, and also in finding quantum teleportation benchmarks for multiple qubit states [28]. In this paper, LAN is used to identify the (asymptotically) optimal measurement on the training set as *linear measurement* on two harmonic oscillators. Similarly to the case of state estimation such collective measurements perform strictly better than the local ones [29, 30]. Moreover, optimal learning collective measurement is different from the optimal measurement for state estimation, showing once again that generically, different quantum decision problems cannot be solved optimally simultaneously.

*Related work.* Sasaki and Carlini [16] defined a *quantum matching machine* which aims at pairing a given 'feature' state with the closest out of a set of 'template' states. The problem is formulated in a Bayesian framework with uniform priors over the feature and template pure states which are considered to be unknown. Bergou and Hillery [17] introduced a discrimination machine, which corresponds to our set-up in the special case when the training set is of size $n = 1$. The papers [18, 19] deal with the problem of quantum state identification as defined in this paper. The special case of Bayesian risk with uniform priors over pure states was solved in [18], with the small difference that the learning and classification steps are done in a single measurement over $n + 1$ systems. However, as in the case of state estimation [31], the proof relies on the special symmetry of the prior and does not cover mixed states. Finally, the concept of quantum classification was already proposed in a series of papers [20, 21, 22]. However, the authors mostly focused on problem formulation, reduction between different problem classes and general issues regarding learnability. Other related papers which fall outside the scope of our investigation are [32, 33].

This paper is organised as follows. Section 2 gives a short overview of the classical classification set-up and introduces its quantum analogue. Section 3 discusses the LAN theory with emphasis on the qubit case. In section 4 we reformulate the classification problem in the asymptotic (local) framework, as an estimation problem with quadratic loss for the training set. The main result is Theorem 5.1 of Section 5 which gives the mimimax excess risk for the case of known priors. The case of unknown priors is treated Section 5.2. The optimal classifier is compared to the plug-in procedure based on optimal state estimation in Section 5.1. The geometry of the problem is captured by the Bloch ball illustrated in Figure 4. We conclude the paper with discussions.

## 2. Classical and quantum learning

### *2.1. Classical Learning*

Let $(X, Y)$ be a pair of random variables with joint distribution $\mathbb{P}$ over the measure space $(\mathcal{X} \times \{0, 1\}, \Sigma)$. In the classical setting $\mathcal{X}$ is usually a subset of $\mathbb{R}^p$ and $Y$ is a binary variable. In a first stage we are given a *training* set of $n$ i.i.d. pairs $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ with distribution $\mathbb{P}$, from which we would like to 'learn' about $\mathbb{P}$. In the second stage we are presented with a new sample $X$ and we are asked to guess its unseen label $Y$. For this we construct a (random) *classifier*

$$\hat{h}_n \colon \mathcal{X} \to \{0, 1\}$$

which depends on the data $(X_1, Y_1), \ldots, (X_n, Y_n)$. Its overall accuracy is measured in terms

of the *expected error rate* according the data distribution $\mathbb{P}$,

$$\mathbb{P}_e(\hat{h}_n) = \mathbb{P}(\hat{h}_n(X) \neq Y) = \mathbb{E}[1_{\hat{h}_n(X)\neq Y}],$$

where $1_C$ is the indicator function equal to 1 if $C$ is true, and 0 otherwise. However the error rate itself does not give a good indication on the performance of the learning method. Indeed, even an 'oracle' who knows $\mathbb{P}$ exactly has typically a non-zero error: in this case the optimal $\hat{h}$ is the *Bayes classifier* which chooses the label that is more probable with respect to conditional distribution $\mathbb{P}(y|x)$

$$h^*(X) = \begin{cases} 0 & if \quad \eta(X) \leq 1/2 \\ 1 & if \quad \eta(X) > 1/2 \end{cases} \tag{1}$$

where $\eta(x) := \mathbb{P}(Y = 1|x)$. The *Bayes risk* is

$$\mathbb{P}_e(h^*) = \mathbb{E}[\mathbb{E}[1_{\hat{h}^*(X)\neq Y}|X]] = \frac{1}{2}\left(1 - \mathbb{E}[|1 - 2\eta(X)|]\right).$$

An alternative view of the Bayes classifier which fits more naturally in the quantum set-up is the following. We are given data $X$ whose probability distribution is either $\mathbb{P}_0(X) := \mathbb{P}(X|Y = 0)$ or $\mathbb{P}_1(X) := \mathbb{P}(X|Y = 1)$ and we would like to test between the two hypotheses. We are in a Bayesian set-up where the hypotheses are chosen randomly with prior distributions $\pi_i = \mathbb{P}(Y = i)$. The optimal solution of this problem is the well known likelihood ratio test: we choose the hypothesis with higher likelihood

$$h^*(X) = \begin{cases} 0 & if \quad \pi_0\mathbb{P}_0(X) > \pi_1\mathbb{P}_1(X) \\ \\ 1 & if \quad \pi_0\mathbb{P}_0(X) \leq \pi_1\mathbb{P}_1(X) \end{cases}$$

which can be easily verified to be identical to the previously defined Bayes classifier. The Bayes risk can be written as

$$\mathbb{P}_e^* = \frac{1}{2}(1 - \|\pi_0 p_0 - \pi_1 p_1\|_1), \tag{2}$$

where $p_i$ are the densities of $\mathbb{P}(X|Y = i)$ with respect to some common reference measure.

Returning to the classfication set-up where $\mathbb{P}$ is unknown, we see that a more informative performance measure for $\hat{h}_n$ is the *excess risk*:

$$R(\hat{h}_n) = \mathbb{P}_e(\hat{h}_n) - \mathbb{P}_e(h^*) \geq 0 \tag{3}$$

which measures how much worse the procedure $\hat{h}_n$ performs compared to the performance of the oracle classifier. In statistical learning theory one is primarily interested in consistent classifiers, for which the excess risk converges to 0 as $n \to \infty$, and then in finding classifiers with fast convergence rates [2, 3]. But how to compare different learning procedures? One can always design algorithms which work well for certain distributions and badly for others. Here we take the statistical approach and consider that all prior information about the data is encoded in the *statistical model* $\{\mathbb{P}_\theta : \theta \in \Theta\}$ i.e. the data comes from a distribution which depends on some unknown parameter $\theta$ belonging to a parameter space $\Theta$. The later may be a subset of $\mathbb{R}^k$ (parametric) or a large class of distributions with certain 'smoothness' properties (non-parametric). One can then define the maximum risk of $\hat{h}_n$

$$R_{max}(\hat{h}_n) := \sup_{\theta\in\Theta} R_\theta(\hat{h}_n)$$

where $R_\theta$ denotes the excess risk when the underlying distribution is $\mathbb{P}_\theta$. A procedure $\tilde{h}_n$ is called minimax if its maximum risk is smaller than that of any other procedure

$$R_{max}(\tilde{h}_n) = \inf_{\hat{h}_n} R_{max}(\hat{h}_n) = \inf_{\hat{h}_n} \sup_{\theta\in\Theta} R_\theta(\hat{h}_n). \tag{4}$$

Alternatively one can take a Bayesian approach and optimise the average risk with respect to a given prior over $\Theta$.

**Example 2.1.** *Let $(X,Y) \in \{0,1\}^2$ with unknown parameters $\eta(0), \eta(1)$ and $\mathbb{P}(X=0)$, satisfying $\eta(0) < 1/2$ and $\eta(1) > 1/2$. Then the Bayes classifier is $h^*(0) = 0$ and $h^*(1) = 1$. On the other hand, from the training sample one can estimate $\eta(i)$ and obtain the concentration result*

$$\mathbb{P}[\hat{\eta}_n(0) < 1/2 \text{ and } \hat{\eta}_n(1) > 1/2] = 1 - O(\exp(-cn)).$$

*Thus the plug-in estimator $\hat{h}_n$ obtained by replacing $\eta$ by $\hat{\eta}_n$ in* (1) *is equal to $h^*$ with high probability and the excess risk is exponentially small.*

The crucial feature leading to exponentially small risk was the fact that the regression function $\eta(X)$ is bounded away from the critical value $1/2$. This situation is rather special but shows that the behaviour of the excess risk depends on the properties of $\eta$ around the value $1/2$. Let us look at another simple example with a different behaviour.

**Example 2.2.** *Let $(X,Y) \in \mathbb{R}\times\{0,1\}$ with*

$$\mathbb{P}(X|Y=0) = N(a,1), \qquad \mathbb{P}(X|Y=1) = N(b,1)$$

*for some unknown means $a < b$, and $\mathbb{P}(Y=0) = 1/2$. From Figure 2.1 we can see that $p_0(x) \le p_1(x)$ if and only if $x \ge (a+b)/2$ so that the Bayes classifier is*

$$h^*(x) = \begin{cases} 0 & if \quad x < (a+b)/2 \\ 1 & if \quad x \ge (a+b)/2 \end{cases}$$

*The Bayes risk is equal to the orange area under the two curves. Again a natural classifier is obtained by estimating the midpoint $(a+b)/2$ and plugging into the above formula. The additional error is the area of the green triangle. Since $(\hat{a}+\hat{b})/2 - (a+b)/2 \approx 1/\sqrt{n}$ one can deduce that*

$$R(\hat{h}_n) = O(n^{-1}),$$

*and it can be shown that this rate of convergence is optimal [34].*

From this example we see that the rate is determined by the behaviour of the regression function $\eta$ around $1/2$, namely in this case

$$\mathbb{P}(|\eta(x) - 1/2| \le t) = O(t), \quad t \ge 0$$

which is called the *margin condition*. Roughly speaking, in a parametric model satisfying the margin condition, the excess risk goes to zero as $O\left(n^{-1}\right)$. In non-parametric models (which are the main focus of learning theory), arbitrarily slow rates are possible depending on the complexity of the model and the behaviour of the regression function [34].

According to Vapnik [3], one of the principles of statistical learning is: "when solving a problem of interest, do not solve a more general problem as an intermediate step." This is interpreted as saying that learning procedures which estimate first the statistical model (or

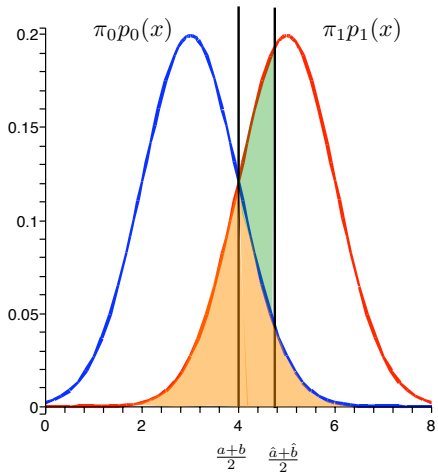


**Figure 1.** Likelihood functions for two normal distributions with means $a, b$. The Bayes risk is the area of the orange triangle. The excess risk is the area of the green triangle

regression function) and then plug this estimate into the Bayes classifier, are less efficient than methods which aim at constructing $\hat{h}(x)$ directly. Recently it has been shown [34] that this is not necessarily the case if some type of margin condition is assumed, and that plug-in estimators

$$\hat{h}_{\text{PLUG-IN}}(x) = 1_{\hat{\eta}(x)\geq 1/2}. \tag{5}$$

can perform close to, or at 'fast $n^{-1}$ rates'. In this paper we show that at least in what concerns the *constant* in front of the rate, *direct quantum learning performs better than plug in methods based on optimal state estimation*. This is a purely quantum phenomenon which stems from the incompatibility between the optimal measurements for estimation and learning.

### *2.2. Quantum Learning*

We now consider the quantum counterpart of the learning problem, the classification of quantum states. In this case, $\mathcal{X}$ is replaced by a Hilbert space of dimension $d$. To find the counterpart of $\mathbb{P}$ we write $\mathbb{P}(dx, y) = \mathbb{P}(dx|y)\mathbb{P}(y)$ and replace the conditional distributions $\mathbb{P}(dx|y = 0)$ and $\mathbb{P}(dx|y = 1)$ by density matrices $\rho$ and $\sigma$, while $\mathbb{P}(y)$ describes prior probabilities over the states, usually denoted by $\pi_y := \mathbb{P}(Y = y)$. There is no direct counterpart of the object $x$, since the quantum state is *identified* with its description in terms of a density matrix; however, one can think of $x$ as a set of values obtained by measuring the state $\rho$.

The training set consists of $n$ i.i.d. pairs $\{(\tau_1, Y_1), \ldots, (\tau_n, Y_n)\}$, where $\tau_i = \rho$ if $Y_i = 0$ and $\tau_i = \sigma$ if $Y_i = 1$. Thus we are randomly given copies of $\rho$ and $\sigma$ together with their labels, but *we do not know* what $\rho$ and $\sigma$ are. After a permutation the joint state of the training set can be concisely written as $\rho^{\otimes n_0} \otimes \sigma^{\otimes n_1}$, where $n_y$ is the number of copies for which $Y_j = y$.

The experimenter is allowed to make any physical operations on the training set (such as unitary evolution or measurements) and outputs a binary-valued measurement $\mathbb{C}^2$ with POVM elements $\widehat{M}_n := (\widehat{P}_n, \mathbf{1} - \widehat{P}_n)$. This (random) POVM plays the role of the classical classifier $\hat{h}_n$: given a new copy of the quantum state whose label is unknown, we apply the measurement

**Table 1.** Comparison of classical and quantum learning.

| element | classical learning | quantum learning |
|---|---|---|
| distribution | $\mathbb{P}$ | $(\rho, \sigma)$ with priors $(\pi_0, \pi_1)$ |
| training example | $(x, y)$ | $(\rho, 0)$ or $(\sigma, 1)$ |
| training set | $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ | $\rho^{\otimes n_0} \otimes \sigma^{\otimes n_1}$ |
| function | classifier $\hat{h}$ | measurement $\widehat{P}$ |
| optimal function | $h^*(x) = 1_{\eta(x) \geq 1/2}$ | $P^* = [\pi_0\rho - \pi_1\sigma_1]_+$ |
| minimum risk | $\frac{1}{2}\left(1 - \mathbb{E}[\lvert 1 - 2\eta(X)\rvert]\right)$ | $\frac{1}{2}\left(1 - \mathrm{Tr}[\lvert\pi_1\sigma - \pi_0\rho\rvert]\right)$ |
| risk | $\mathbb{P}(\hat{h}(X) \neq Y) - \mathbb{P}_e^*$ | $\mathbb{E}\mathrm{Tr}\left[(\pi_1\sigma - \pi_0\rho)(\widehat{P} - P^*)\right]$ |

$\widehat{M}_n$ to guess whether the state is $\rho$ or $\sigma$. The accuracy is measured in terms of the expected misclassification error:

$$\mathbb{P}_e(\widehat{M}_n) = \mathbb{E}\left[\pi_0 \mathrm{Tr}[\rho(\mathbf{1} - \widehat{P}_n)] + \pi_1 \mathrm{Tr}[\sigma \widehat{P}_n]\right]$$

where the expectation is taken over the outcomes $\widehat{P}_n$.

The Bayes classifier $M^*$ is nothing but the Helstrom measurement [11] which optimally discriminates between *known* states $\rho, \sigma$ with priors $\pi_0, \pi_1$. In this case $M^* = (P^*, \mathbf{1} - P^*)$ where $P^*$ is the projection onto the subspace of positive eigenvalues of the operator $\pi_0\rho - \pi_1\sigma$, i.e. $P^* = [\pi_0\rho - \pi_1\sigma]_+$. Note that if both eigenvalues are of the same sign, the optimal procedure is to choose the state with higher $\pi_i$ without making any measurement at all. The *Helstrom risk* can be expressed as: $\mathbb{P}_e^* = \frac{1}{2}\left(1 - \mathrm{Tr}[\lvert\pi_1\sigma - \pi_0\rho\rvert]\right)$. which is the quantum extension of (2).

As before, the performance of an arbitrary classifier $\widehat{M}_n$ is measured by the excess risk:

$$R(\widehat{M}_n) = \mathbb{P}_e(\widehat{M}_n) - \mathbb{P}_e^* = \mathbb{E}\mathrm{Tr}\left[(\pi_1\sigma - \pi_0\rho)(\widehat{P}_n - P^*)\right], \tag{6}$$

which is expected to vanish asymptotically with $n$.

In Table 1 we summarise the analogous concepts in the classical and the quantum learning set-up. Besides these obvious correspondences we would like to point out some interesting differences. Based on the coin toss example 2.1 one may expect that the classification of two qubit states should exhibit similar exponentially fast rates. In fact as we will show in this paper, the rate is $n^{-1}$ as in example 2.2 where the data is not discrete but continuous and the regression function is not bounded away from $1/2$. A possible explanation is the fact that in the quantum case the 'data' to be labelled is a quantum system and the distribution of the outcome depends on the measurement. A helpful way to think about it is illustrated in Figure 2.2. The unknown label is the input of a black box which outputs the data $X$ with conditional distribution $\mathbb{P}(X|Y)$. In the quantum case the box has an additional input, the measurement choice which appears as a parameter in the conditional distribution and is controlled by the experimenter. The game is to learn from the training set the optimal value of this parameter, for which the identification of the label $Y$ is most facile. This set-up resembles that of *active learning* [35] where the training data $X_i$ are actively chosen rather than collected randomly.

### *2.3. Local minimax formulation of optimality*

We now give the precise formulation of what we mean by asymptotic optimality of a learning strategy $\{\widehat{M}_n : n \in \mathbb{N}\}$. As in the classical case we construct a model which contains all

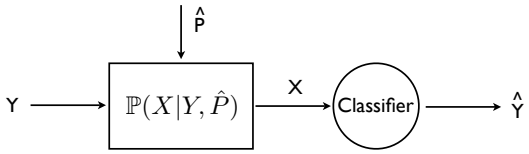


**Figure 2.** Quantum learning seen as classical learning with data distribution depending on on additional parameter controlled by the experimenter

unknown parameters of the problem: the two states $\rho, \sigma$ and the prior $\pi_0$. We denote these parameters collectively by $\theta$ which belongs to a parameter space $\Theta \subset \mathbb{R}^k$. When some prior information is available about the model, it can be included by restricting to a sub-model of the general one. As in the classical case we denote by $R_\theta(\widehat{M}_n)$, the risk of $\widehat{M}_n$ at $\theta$, and we can define the maximum risk as in (4). However, assuming for the moment that that the optimal rate of classification is $n^{-1}$, we use a more refined performance measure which is the local version of the maximum risk $R_{max}$ around a fixed parameter $\theta_0$

$$R^{(l)}_{max}(\widehat{M}_n\ ;\ \theta_0) := \sup_{\|\theta-\theta_0\|\leq n^{-1/2+\epsilon}} nR_\theta(\widehat{M}_n) \tag{7}$$

where $\epsilon > 0$ is a small number. Note that in the above definition the usual risk was multiplied by the inverse of its rate $n$ so that we can expect $R^{(l)}_{max}$ to have a non-trivial limit when $n \to \infty$. The reason for choosing the local maximum risk is that it reflects better the difficulty of the problem in different regions of the parameter space while the maximum risk captures the worst possible behavior over the whole parameter space. We can think of the local ball $\|\theta - \theta_0\| \leq n^{-1/2+\epsilon}$ as the intrinsic parameter space when the training set consists of $n$ samples. Indeed a simple estimator $\theta_0$ on a small proportion $\tilde{n} = n^{1-\epsilon}$ of the sample locates the true parameter in such a ball with high probability (see Lemma 2.1 in [24]).

**Definition 2.1.** *The local minimax risk at $\theta_0$ is defined as*

$$R^{(l)}_{minmax}(\theta_0) := \limsup_{n\to\infty} \inf_{\widetilde{M}_n} R^{(l)}_{max}(\widehat{M}_n\ ;\ \theta_0).$$

*A sequence of classifiers $\{\tilde{M}_n : n \in \mathbb{N}\}$ is called locally asymptotic minimax if*

$$\limsup_{n\to\infty} R^{(l)}_{max}(\tilde{M}_n\ ;\ \theta_0) = R^{(l)}_{minmax}(\theta_0).$$

We identify two general learning strategies. The first one consists in estimating the states $\rho, \sigma$ and prior $\pi_0$ (optimally) to get $\hat{\rho}, \hat{\sigma}, \hat{\pi}_0$ and then constructing the classifier (measurement) as:

$$\widehat{P}_{\text{PLUG-IN}} = [\hat{\pi}_0\hat{\rho} - \hat{\pi}_1\hat{\sigma}]_+. \tag{8}$$

The second strategy aims at estimating the Helstrom projection $P^*$ directly from the training set without passing through state estimation. As we will see, it turns out that in general the latter performs better than the former.

In section 3 we review the concept of local asymptotic normality which means that locally, the training set can be efficiently approximated by a simple Gaussian model consisting of displaced thermal equlibrium states and classical Gaussian random variables. In section 4 we

show how to reduce the local classification risk for qubits to an expectation of a quadratic form in the local parameters. This will simplify the problem of finding the optimal measurement of the training set, to that of finding the optimal measurement of a Gaussian state for a quadratic loss function [10].

## 3. Local asymptotic normality

In a series of papers [23, 24, 25] Guţă and Kahn and Guţă and Jencova [26] developed a new approach to state estimation based on the extension of the classical statistical concept of local asymptotic normality [27]. Using this tool one can cast the problem of (asymptotically) optimal state estimation into a much simpler one of estimating the mean of a Gaussian state with known variance.

Local asymptotic normality provides a convenient description of quantum statistical models involving i.i.d. quantum states which can also be applied to the present learning problem. In this section we will give a brief introduction to this subject in as much as it is necessary for this paper and we refer to [25] for proofs and a more in depth analysis.

### *3.1. Local asymptotic normality in classical statistics*

A typical statistical problem is the estimation of some unknown parameter $\theta$ from a sample $X_1, \ldots, X_n \in \mathcal{X}$ of independent, identically distributed random variables drawn from a distribution $\mathbb{P}_\theta$ over a measure space $(\mathcal{X}, \Sigma)$. If $\theta$ belongs to an open subset of $\mathbb{R}^k$ for some finite dimension $k$ and if the map $\theta \to \mathbb{P}_\theta$ is sufficiently smooth, then widely used estimators $\hat{\theta}_n(X_1, \ldots, X_n)$ such as the maximum likelihood are asymptotically optimal in the sense that they converge to $\theta$ at a rate $n^{-1/2}$ and the error has an asymptotically normal distribution

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} N(0, I^{-1}(\theta)), \tag{9}$$

where the right side is the lower bound set by the Cramér-Rao inequality for unbiased estimators. To give a simple example, if $X_i \in \{0, 1\}$ is the result of a coin toss with $\mathbb{P}[X_i = 1] = \theta$ and $\mathbb{P}[X_i = 0] = 1 - \theta$ then the sufficient statistic

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

satisfies (9) by the Central Limit Theorem (CLT).

Naturally, the first inquiries into quantum statistics concentrated on generalising the Cramér-Rao inequality to unbiased measurements, and on finding asymptotically optimal estimators which achieve the quantum version of the Fisher information matrix [11, 10, 36]. However it was found that due to the additional uncertainty introduced by the non-commutative nature of quantum mechanics the situation is essentially different from the classical case. A summary of these finding is

(i) the multi-dimensional version of the Cramér-Rao bound is in general not achievable;

(ii) the optimal measurement depends on the loss function, i.e. the quadratic form $(\hat{\theta} - \theta)^t G(\hat{\theta} - \theta)$ and different weight matrices $G$ lead in general to incompatible measurements.

As we will see, these issues can be overcome by adopting a more modern perspective to asymptotic statistics provided by the technique of local asymptotic normality [27, 37]. Instead of analysing particular estimation problems, the idea is to consider the structure of the statistical model underlying the data and to approximate it by a simpler model for which the statistical problems are easy to solve. In order to obtain a non-trivial limit model it makes sense to rescale the parameters according to their uncertainty, so we assume that $\theta$ is localised in a region of size $n^{-1/2}$ and we can write $\theta = \theta_0 + h/\sqrt{n}$ with $\theta_0$ known and $h \in \mathbb{R}^k$ the local parameter to be estimated. Such an assumption does not restrict the generality of the problem since one can use an adaptive two-steps procedure where a rough estimate $\theta_0$ is obtained in the first step using a small part of the sample, and the rest is used for the accurate estimation of the local parameter $h$.

Local asymptotic normality means that the sequence of (local) statistical models

$$\mathcal{P}_n := \left\{ \mathbb{P}^n_{\theta_0 + h/\sqrt{n}} : \|h\| < C \right\}, \qquad n \in \mathbb{N} \tag{10}$$

depending 'smoothly' on $h$, converges to the *Gaussian shift model*

$$\mathcal{G} := \left\{ N(h, I^{-1}(\theta_0)) : \|h\| < C \right\} \tag{11}$$

where we observe a single Gaussian variable with mean $h$ and fixed and known variance. The convergence has a precise mathematical definition in terms of the Le Cam distance between two statistical models which quantifies the extent to which each model can be 'simulated' by randomising data from the other.

**Definition 3.1.** *A positive linear map*

$$T : L^1(\mathcal{X}, \mathcal{A}, \mathbb{P}) \to L^1(\mathcal{Y}, \mathcal{B}, \mathbb{Q})$$

*is called a stochastic operator (or randomisation) if* $\|T(p)\|_1 = \|p\|_1$ *for every* $p \in L^1_+(\mathcal{X})$.

For simplicity we consider only dominated models for which all distributions have densities with respect to some fixed reference distribution. In this case a randomisation is the classical analogue of a quantum channel.

**Definition 3.2.** *Let* $\mathcal{P} := \{\mathbb{P}_\theta : \theta \in \Theta\}$ *and* $\mathcal{Q} := \{\mathbb{Q}_\theta : \theta \in \Theta\}$ *be two dominated statistical models with distributions having probability densities* $p_\theta := d\mathbb{P}_\theta / d\mathbb{P}$ *and* $q_\theta := d\mathbb{Q}_\theta / d\mathbb{Q}$. *The deficiencies* $\delta(\mathcal{P}, \mathcal{Q})$ *and* $\delta(\mathcal{Q}, \mathcal{P})$ *are defined as*

$$\delta(\mathcal{P}, \mathcal{Q}) := \inf_T \sup_{\theta \in \Theta} \|T(p_\theta) - q_\theta\|_1$$
$$\delta(\mathcal{Q}, \mathcal{P}) := \inf_S \sup_{\theta \in \Theta} \|S(q_\theta) - p_\theta\|_1$$

*where the infimum is taken over all randomisations* $T, S$. *The Le Cam distance between* $\mathcal{P}$ *and* $\mathcal{Q}$ *is*

$$\Delta(\mathcal{P}, \mathcal{Q}) := \max(\delta(\mathcal{Q}, \mathcal{P}),\, \delta(\mathcal{P}, \mathcal{Q})).$$

With this definitions the local asymptotic normality for i.i.d. parametric models can be formulated as

**Theorem 3.3.** *The sequence of local models* (10) *converges in the Le Cam distance to the Gaussian shift model* (11)

$$\lim_{n \to \infty} \Delta(\mathcal{P}_n, \mathcal{G}) = 0.$$

This statement can be extended to slowly increasing local neighbourhoods $\|h\| \leq n^\epsilon$ with precise convergence rate for the Le Cam distance.

### *3.2. Local asymptotic normality in quantum statistics*

We will now describe the quantum version of local asymptotic normality for the simplest case of a family of spin states. The general result valid for arbitrary finite dimensional systems can be found in [25].

We are given $n$ spins independent identically prepared in the state

$$\rho_{\vec{r}} = \frac{1}{2}(\mathbf{1} + \vec{r}\vec{\sigma})$$

where $\vec{r}$ is the unknown Bloch vector of the state and $\vec{\sigma} = (\sigma_x, \sigma_y, \sigma_z)$ are the Pauli matrices in $M(\mathbb{C}^2)$. Following the methodology of the previous section, we concentrate on the structure of the statistical model itself rather than optimal state estimation. The latter, and other statistical problems can be solved easily once the convergence to a Gaussian model is established.

By measuring a small proportion $n^{1-\epsilon} \ll n$ of the systems we can devise an initial rough estimator $\rho_0 := \rho_{\vec{r}_0}$ so that with high probability the state is in a ball of size $n^{-1/2+\epsilon}$ around $\rho_0$ [23]. We label the states in this ball by the local parameter $\vec{u}$

$$\rho_{\vec{u}/\sqrt{n}} = \frac{1}{2}\left(\mathbf{1} + (\vec{r}_0 + \vec{u}/\sqrt{n})\vec{\sigma}\right)$$

and define the local statistical model by

$$\mathcal{Q}_n := \{\rho_{\vec{u}}^n : \|\vec{u}\| \leq n^{\epsilon}\}, \qquad \rho_{\vec{u}}^n := \rho_{\vec{u}/\sqrt{n}}^{\otimes n}. \tag{12}$$

By choosing a coordinate system $(\vec{a}_1, \vec{a}_2, \vec{a}_3)$ with $\vec{a}_3$ along $\vec{r}_0$ and writing $\vec{u} = u_1\vec{a}_1 + u_2\vec{a}_2 + u_3\vec{a}_3$ we observe that $\rho_{\vec{u}/\sqrt{n}}$ is essentially obtained by perturbing the eigenvalues of $\rho_0$ by $u_3/2\sqrt{n}$ and rotating it with a 'small' unitary

$$U := \exp(i(-u_2\vec{a}_1 + u_1\vec{a}_2)\vec{\sigma}/2r_0\sqrt{n}), \qquad r_0 := \|\vec{r}_0\|.$$

The splitting into 'classical' and 'quantum' parameters $u_3$ and $(u_1, u_2)$ can be intuitively explained through the 'big Bloch sphere' picture commonly used to describe spin coherent [38] and spin squeezed states [39]. Let

$$L_j := \sum_{i=1}^{n} \vec{a}_j \cdot \vec{\sigma}^{(i)}, \qquad j = 1, 2, 3$$

be the collective spin components along the directions $\vec{a}_j$. By the Central Limit Theorem, the distributions of $L_i$ with respect to $\rho_0^{\otimes n}$ converge as

$$\frac{1}{\sqrt{n}}(L_3 - nr_0) \xrightarrow{\mathcal{D}} N(0, 1 - r_0^2), \qquad \frac{1}{\sqrt{n}}L_{1,2} \xrightarrow{\mathcal{D}} N(0, 1),$$

so that the joint spins state can be pictured as a vector of length $nr_0$ whose tip has a Gaussian blob of size $\sqrt{n}$ representing the uncertainty in the collective variables (see Figure 3.2). Furthermore, by a law of large numbers heuristic we estimate the commutators

$$\left[\frac{1}{\sqrt{n}}L_1, \frac{1}{\sqrt{n}}L_2\right] = 2i\frac{1}{n}L_3 \approx 2ir_0\mathbf{1}, \qquad \left[\frac{1}{\sqrt{n}}L_{1,2}, \frac{1}{\sqrt{n}}L_3\right] \approx 0.$$

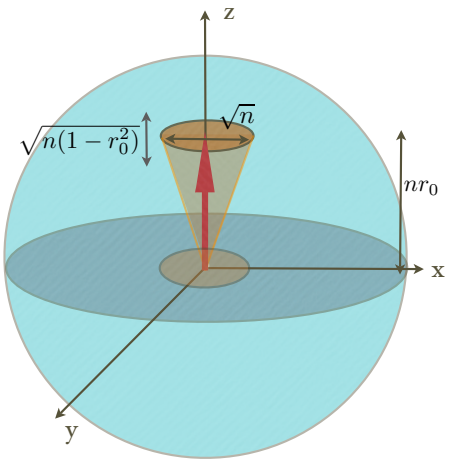


**Figure 3.** Big ball picture of the collective state of identical mixed spins. The total spin is represented as a vector of length $nr_0$ with a 3D uncertainty blob of size $\sqrt{n}$ in the $x, y$ directions and $\sqrt{n(1-r_0^2)}$ in the $z$ direction.

This suggests that $L_1/\sqrt{2r_0 n}$ and $L_2/\sqrt{2r_0 n}$ converge to the canonical coordinates $Q$ and $P$ of a quantum harmonic oscillator in a thermal equilibrium state

$$\Phi := (1-p)\sum_{k=0}^{\infty} p^k |k\rangle\langle k|, \qquad p = \frac{1-r_0}{1+r_0},$$

where $\{|k\rangle : k \geq 0\}$ represents the Fock basis. Moreover the (rescaled) component $\frac{1}{\sqrt{n}}(L_3 - nr_0)$ converges to a *classical* Gaussian variable $X \sim N := N(0, 1-r_0^2)$ which is independent of the quantum state. Note that the Gaussian limit state has both quantum and classical components and should be identified with the state $\Phi \otimes N$ on the von Neumann algebra $\mathcal{B}(\ell^2(\mathbb{N})) \otimes L^\infty(\mathbb{R})$.

What is the Gaussian state when the spins are in the 'perturbed' state $\rho_{\vec{u}}^n$ ? By applying the same argument we obtain that the variables $Q, P, X$ pick up expectations which (in the first order in $n^{-1/2}$) are proportional to the local parameters $(u_1, u_2, u_3)$ while the variances remain unchanged. More precisely the oscillator is in a displaced thermal equilibrium state $\Phi_{\vec{u}} := D(\vec{u})\Phi D(\vec{u})^*$, where $D(\vec{u})$ is the displacement operator

$$D(\vec{u}) := \exp\left(i(-u_2 Q + u_1 P)/\sqrt{2r_0}\right),$$

and the classical bit has distribution $N_{\vec{u}} := N(u_3, 1-r_0^2)$.

**Definition 3.4.** *The quantum Gaussian shift model $\mathcal{G}$ is defined by the family of quantum-classical states*

$$\mathcal{G} := \{\Phi_{\vec{u}} \otimes N_{\vec{u}} : \vec{u} \in \mathbb{R}^3\} \tag{13}$$

*on $\mathcal{B}(\ell^2(\mathbb{N})) \otimes L^\infty(\mathbb{R})$.*

Having defined the sequence of local models $\mathcal{Q}_n$ and the Gaussian shift model, we need to define the quantum counterparts of randomisations and convergence of models. The natural

analogue of a classical randomisation is a quantum channel, i.e. completely positive, trace preserving map $C : \mathcal{T}_1(\mathcal{H}) \to \mathcal{T}_1(\mathcal{K})$ where $\mathcal{T}_1(\mathcal{H})$ represents the trace class operators on $\mathcal{H}$. However, as we saw above, a sequence of quantum statistical models may converge to a quantum-classical one. The mathematical framework covering randomisations of both classical and quantum statistical models is that of von Neuman algebras and channels between their preduals. In finite dimensions this simply means that we deal with channels between block diagonal matrix algebras. We can now define the Le Cam distance between two quantum models in the same way as in definition 3.2 with classical randomisation replaced by quantum ones and the $\|\cdot\|_1$ representing the norm on the predual, which is the trace norm in the case of density matrices.

**Theorem 3.5.** *Let $\mathcal{Q}_n$ be the sequence of statistical models* (12) *for $n$ i.i.d. local spin states. and let $\mathcal{G}_n$ be the restriction of the Gaussian shift model* (13) *to the range of parameters $\|\vec{u}\| \leq n^{\epsilon}$. Then*

$$\lim_{n\to\infty} \Delta(\mathcal{Q}_n, \mathcal{G}_n) = 0,$$

*i.e. there exist sequences of channels $T_n$ and $S_n$ such that*

$$\begin{aligned} \lim_{n\to\infty} \sup_{\|\mathbf{u}\|\leq n^{\epsilon}} \|\Phi_{\vec{u}} \otimes N_{\vec{u}} - T_n\left(\rho_{\vec{u}}^n\right)\|_1 = 0, \\ \lim_{n\to\infty} \sup_{\|\mathbf{u}\|\leq n^{\epsilon}} \|\rho_{\vec{u}}^n - S_n\left(\Phi_{\vec{u}} \otimes N_{\vec{u}}\right)\|_1 = 0. \end{aligned} \tag{14}$$

To conclude this section we would like to make a few comments on the significance of the above result. The first point is that although it was intuitively illustrated using the Central Limit Theorem, the concept of local asymptotic normality provides a stronger characterisation of the 'Gaussian approximation'. Indeed the convergence in Theorem 3.5 is strong (in $L_1$) rather than weak (in distribution), it is *uniform* over a range of local parameters rather than at a single point, and has an operational meaning based on quantum channels.

Secondly, one can exploit these features to devise asymptotically optimal measurement strategies for state estimation and prove that the Holevo bound [10] is asymptotically attainable [40].

Thirdly, the result can be applied to other quantum statistical problems involving i.i.d. qubit states such as cloning, teleportation benchmarks, quantum learning, and can serve as a mathematical framework for analysing quantum state transfer protocols.

## 4. Local formulation of the classification problem

In this section we reformulate the problem of quantum state classification in the 'local' set-up. This allows us to replace, on the one hand the excess error probability by a *quadratic* form in local parameters, and on the other hand the training set consisting of i.i.d. spins by a simpler Gaussian shift model.

Throughout the section we restrict to the case where the priors $\pi_0, \pi_1$ are known. In Section 5.2 we show that the results for known priors can easily be extended to unknown ones by simply estimating them from the counts of $\rho$ and $\sigma$ states in the training sample.

### *4.1. The loss function*

Recall that the classification problem is to discriminate between two unknown states $\rho$ and $\sigma$ by learning from a training set of $n$ *labelled* systems prepared randomly in one of the states with probabilities $\pi_0$ and $\pi_1$. For this we measure the training set and produce an outcome which is itself a measurement $\widehat{M}_n := (\widehat{P}_n, \mathbf{1} - \widehat{P}_n)$ on $\mathbb{C}^2$. The accuracy of the procedure is measured by the excess risk (6):

$$R(\widehat{M}_n) = \mathbb{E}\mathrm{Tr}\left[(\pi_1\sigma - \pi_0\rho)(\widehat{P}_n - P^*)\right], \tag{15}$$

with $P^* = [\pi_0\rho - \pi_1\sigma]_+$. Since any binary measurement is a mixture of projective POVM's [41], we can assume without loss of generality that $\widehat{P}_n$ is a projection and pull back the randomness into the definition of the training set measurement.

As explained in section 3.2, the a priori unknown states $\rho$ and $\sigma$ can be localised with high probability in $n^{-1/2+\epsilon}$ neighbourhoods of $\rho_0$ and $\sigma_0$ by sacrificing a small proportion of the training set systems; this means that $\rho_0$ and $\sigma_0$ are known and can be used by the classification procedure. Let $\vec{r}_0$ and $\vec{s}_0$ be the Bloch vectors of $\rho_0$ and $\sigma_0$ and let us parametrise their neighbourhoods as follows

$$\begin{aligned}\rho = \rho_{\vec{u}/\sqrt{n}} &= \frac{1}{2}\left(\mathbf{1} + \left(\vec{r}_0 + \frac{\vec{u}}{\sqrt{n}}\right)\vec{\sigma}\right),\\ \sigma = \sigma_{\vec{v}/\sqrt{n}} &= \frac{1}{2}\left(\mathbf{1} + \left(\vec{s}_0 + \frac{\vec{v}}{\sqrt{n}}\right)\vec{\sigma}\right).\end{aligned} \tag{16}$$

Let $P_0 := [\pi_0\rho_0 - \pi_1\sigma_0]_+$ be the optimal projection corresponding to the pair $(\rho_0, \sigma_0)$ and note that it can have dimension one, or it can be zero or identity. In the second case, the optimal measurement is *trivial*, one can guess the state without measuring by checking whether the operator $\pi_0\rho_0 - \pi_1\sigma_0$ is positive or negative.

**Lemma 4.1.** *Let* $(\rho_0, \sigma_0)$ *and* $(\pi_0, \pi_1)$ *satisfy*

$$\|\pi_0\vec{r}_0 - \pi_1\vec{s}_0\| < |\pi_0 - \pi_1|.$$

*Then* $P_0$ *is either zero or identity and the local minimax excess risk satisfies*

$$\inf_{\widehat{M}_n} \sup_{\|\vec{u}\|,\|\vec{v}\| \le n^\epsilon} \mathbb{P}_e(\widehat{M}_n) - P_e^* = O(\exp(-cn))$$

*for some* $c > 0$.

*Proof.* Note that the inequality is satisfied only if $\pi_0 \neq \pi_1$ and it implies that

$$\pi_0\rho_0 - \pi_1\sigma_0 = \frac{\pi_0 - \pi_1}{2}\left(\mathbf{1} + \frac{\pi_0\vec{r}_0 - \pi_1\vec{s}_0}{\pi_0 - \pi_1}\vec{\sigma}\right)$$

it a positive or negative operator depending on the sign of $\pi_0 - \pi_1$.

Since both eigenvalues of $\pi_0\rho_0 - \pi_1\sigma_0$ are non-zero, there exists a constant $\eta > 0$ such that

$$\|\pi_0\rho_0 - \pi_1\sigma_0 - A\|_2 \le \eta$$

implies that $A$ is also a positive or negative operator. In fact, when $n$ is large enough all $\pi_0\rho_{\vec{u}/\sqrt{n}} - \pi_1\sigma_{\vec{v}/\sqrt{n}}$ with $\|\vec{u}\|, \|\vec{v}\| \le n^\epsilon$ have this property for some other constant $\tilde{\eta}$.

Consider a simple measurement on the training set where the states are measured separately in the three bases of the Pauli matrices and the outcomes averages are used to construct a estimators of the states $\rho_{\vec{u}/\sqrt{n}}$ and $\sigma_{\vec{v}/\sqrt{n}}$. Then by basic concentration inequalities we get

$$\mathbb{P}\left(\left\|\pi_0\left(\rho_{\vec{u}/\sqrt{n}}-\rho_{\hat{\vec{u}}/\sqrt{n}}\right)+\pi_1\left(\sigma_{\vec{v}/\sqrt{n}}-\sigma_{\hat{\vec{v}}/\sqrt{n}}\right)\right\|_2\geq\tilde{\eta}\right)\leq exp(-cn)$$

which means that with exponentially small probability error the plug-in estimator of $P^*:=[\pi_0\rho_{\vec{u}/\sqrt{n}}-\pi_1\sigma_{\vec{v}/\sqrt{n}}]_+$ will be equal to $P^*$ which is zero or identity.

□

From now on we will work under the assumption that

$$\|\pi_0\vec{r}_0-\pi_1\vec{s}_0\|>|\pi_0-\pi_1|, \tag{17}$$

so that $P_0:=[\pi_0\rho_0-\pi_1\sigma_0]_+$ is a one dimensional projection whose Bloch vector is

$$\vec{p}_0=\frac{\vec{d}_0}{\|\vec{d}_0\|}:=\frac{\pi_0\vec{r}_0-\pi_1\vec{s}_0}{\|\pi_0\vec{r}_0-\pi_1\vec{s}_0\|}.$$

The Helstrom projection $P^*$ for the pair of unknown states $(\rho,\sigma)$ has Bloch vector

$$\vec{p}=\frac{\vec{d}}{\|\vec{d}\|}=\frac{\pi_0\left(\vec{r}_0+\frac{\vec{u}}{\sqrt{n}}\right)-\pi_1\left(\vec{s}_0+\frac{\vec{v}}{\sqrt{n}}\right)}{\left\|\pi_0\left(\vec{r}_0+\frac{\vec{u}}{\sqrt{n}}\right)-\pi_1\left(\vec{s}_0+\frac{\vec{v}}{\sqrt{n}}\right)\right\|}=\frac{\vec{d}_0+\frac{\vec{z}}{\sqrt{n}}}{\left\|\vec{d}_0+\frac{\vec{z}}{\sqrt{n}}\right\|}, \tag{18}$$

where $\vec{z}:=\pi_0\vec{u}-\pi_1\vec{v}$ is a relative parameter and $\vec{d}:=\vec{d}_0+\frac{\vec{z}}{\sqrt{n}}$.

As discussed before, we can take the estimator $\widehat{M}_n$ to be a projective measurement $\widehat{M}_n:=(\widehat{P}_n,\mathbf{1}-\widehat{P}_n)$, so to minimise the risk (15) we aim at producing an estimator $\widehat{P}_n$ which is close to $P^*$. Since the latter is obtained by rotating $P_0$ with angle of order $n^{-1/2+\epsilon}$, we can assume without loss of generality that $\widehat{P}_n$ has a Bloch vector $\hat{\vec{p}}_n$ which is a small rotation of $\vec{p}_0$ so that

$$\hat{\vec{p}}_n=\frac{\vec{p}_0+\hat{\vec{z}}_n/\sqrt{n}}{\|\vec{p}_0+\hat{\vec{z}}_n/\sqrt{n}\|}, \tag{19}$$

with $\hat{\vec{z}}_n=O(n^\epsilon)$ a vector in the plane orthogonal to $\vec{p}_0$.
Expanding (18) and (19) in powers of $n^{-1/2}$ we get

$$\begin{aligned}\vec{p}-\hat{\vec{p}}_n=&\frac{1}{\sqrt{n}}\left[\frac{\vec{z}-\hat{\vec{z}}_n}{\|\vec{d}_0\|}-\frac{\vec{d}_0(\vec{d}_0\cdot(\vec{z}-\hat{\vec{z}}_n))}{\|\vec{d}_0\|^3}\right]\\&+\frac{1}{n}\left[-\frac{\vec{d}_0(\|\vec{z}\|^2-\|\hat{\vec{z}}_n\|^2)}{2\|\vec{d}_0\|^3}+\frac{3\vec{d}_0((\vec{d}_0\cdot\vec{z})^2-(\vec{d}_0\cdot\hat{\vec{z}}_n)^2}{2\|\vec{d}_0\|^5}\right]+o(n^{-1}).\end{aligned}$$

We now plug these expressions back into into (15) taking into account that $\hat{\vec{z}}_n$ is perpendicular to $\vec{d}_0$ and obtain

$$\begin{aligned}\mathbb{P}_e(\widehat{M}_n)-P_e^*&=\mathbb{E}\mathrm{Tr}\left((\pi_0\rho-\pi_1\sigma)(P-\widehat{P}_n)\right)\\&=\frac{1}{2}\mathbb{E}\vec{d}\cdot(\vec{p}-\hat{\vec{p}}_n)\\&=\frac{1}{4n\|\vec{d}_0\|}\mathbb{E}\|\vec{z}_\perp-\hat{\vec{z}}_n\|^2+o(n^{-1})\end{aligned}$$

where $\vec{z}_\perp = \vec{z} - \vec{d}_0(\vec{z}\cdot\vec{d}_0)/\|\vec{d}_0\|^2$ is the projection of $\vec{z}$ onto the plane orthogonal to $\vec{d}_0$.

It is clear now that the rate of convergence of the excess risk (15) is $n^{-1}$, so it is meaningful to optimise the quantity $nR^{(l)}_{max}(\widehat{M}_n)$, and the contribution coming from the $o(n^{-1})$ term can be dropped.

Since $\widehat{M}_n$ is uniquely determined by $\hat{\vec{z}}_n$ by (19), we define the *quadratic* loss function for the measurement on the training set in terms of local variables

$$L((\vec{u},\vec{v}),\hat{\vec{z}}_n) := \frac{1}{4\|\vec{d}_0\|}\|\vec{z}_\perp - \hat{\vec{z}}_n\|^2, \quad \vec{z} := (\pi_0\vec{u} - \pi_1\vec{v}) \tag{20}$$

and the associated renormalised risk is $R_{\vec{u},\vec{v}}(\hat{\vec{z}}_n) := \mathbb{E}L((\vec{u},\vec{v}),\hat{\vec{z}}_n)$. The local maximum risk (7) around $(\rho_0,\sigma_0)$ is then

$$\begin{aligned} R^{(l)}_{max}(\hat{\vec{z}}_n\ ;\ \rho_0,\sigma_0) &:= \sup_{\|\vec{u}\|,\|\vec{v}\|\le n^\epsilon} R_{\vec{u},\vec{v}}(\hat{\vec{z}}_n) \\ &= \sup_{\|\vec{u}\|,\|\vec{v}\|\le n^\epsilon} \frac{1}{4\|\vec{d}_0\|}\mathbb{E}\|\vec{z}_\perp - \hat{\vec{z}}_n\|^2. \end{aligned} \tag{21}$$

In conclusion, we need to find the optimal measurement strategy on the training set with respect to the above quadratic form of the local parameters.

### *4.2. The training set*

To solve the above problem we employ the machinery of local asymptotic normality. As before, let $\rho$ and $\sigma$ be states in local neighbourhood of $\rho_0$ and respectively $\sigma_0$ described by (16). We write their local Bloch vectors $(\vec{u},\vec{v})$ as

$$\vec{u} = u_1\vec{a_1} + u_2\vec{a_2} + u_2\vec{a_3} \qquad \text{and} \qquad \vec{v} = v_1\vec{b_1} + v_2\vec{b_2} + v_2\vec{b_3}$$

where $(\vec{a}_1,\vec{a}_2,\vec{a}_3)$ and $(\vec{b}_1,\vec{b}_2,\vec{b}_3)$ are two coordinate systems which satisfy the conditions (see Figure 4)

(i) $\vec{a}_3$ is parallel to $\vec{r}_0$,
(ii) $\vec{b}_3$ is parallel to $\vec{s}_0$ ,
(iii) $\vec{a}_1,\vec{b}_1$ are in the plane $(\vec{r}_0,\vec{s}_0)$,
(iv) $\vec{a}_2 = \vec{b}_2$ is perpendicular to the plane $(\vec{r},\vec{s})$.

With these notations the local statistical model for the training set is

$$\mathcal{T}_n := \{\rho_{\vec{u}}^{n\pi_0}\otimes\sigma_{\vec{v}}^{n\pi_1} : \|\vec{u}\|,\|\vec{v}\| \le n^\epsilon\}$$

and the corresponding Gaussian shift model is

$$\mathcal{G}^{(2)} := \{N_{\vec{u}}\otimes N_{\vec{v}}\otimes\Phi_{\vec{u}}\otimes\Phi_{\vec{v}} : \vec{u},\vec{v}\in\mathbb{R}^3\} \tag{22}$$

where

$$\begin{aligned} N_{\vec{u}} &:= N(\sqrt{\pi_0}u_3, 1-r_0^2), \\ N_{\vec{v}} &:= N(\sqrt{\pi_1}v_3, 1-s_0^2), \\ \Phi_{\vec{u}} &:= \Phi\left(\sqrt{\frac{\pi_0}{2r_0}}u_1,\ \sqrt{\frac{\pi_0}{2r_0}}u_2;\ \frac{\mathbf{1}}{2r_0}\right), \\ \Phi_{\vec{v}} &:= \Phi\left(\sqrt{\frac{\pi_1}{2s_0}}v_1,\ \sqrt{\frac{\pi_1}{2s_0}}v_2;\ \frac{\mathbf{1}}{2s_0}\right) \end{aligned} \tag{23}$$

and $\Phi(q,p,v)$ is a displaced thermal equilibrium state with means $(q,p)$ and variance $v$.

The following technical lemma shows that local asymptotic normality can be used to transfer the problem of the optimal classification from a training set consisting of qubits, to a Gaussian one. The arguments are rather standard though tedious, and since the same method has been used for finding the optimal estimation procedure for qubits [24], we refer to that paper for the proof.

**Lemma 4.2.** *Consider the problems of finding asymptotically optimal strategies for the models $\mathcal{T}_n$ and respectively $\mathcal{G}_n^{(2)}$ with respect to the loss function* (20)*. Then the local minimax risks of both problems converge to the same constant which is the the minimax risk of the unrestricted Gaussian shift model $\mathcal{G}^{(2)}$.*

In conclusion, the measurement of the training set should be aimed at optimally estimating the two parameter vector $\vec{z}_\perp$ directly, rather than using a 'plug-in' strategy where the three dimensional local parameters $(\vec{u},\vec{v})$ are first (optimally) estimated and then the measurement $\widehat{P}_n$ is constructed as in (8). We will come back to this point later on when the two methods will be compared.

## 5. Optimal classifier

In this section we formulate our main result characterising the asymptotically optimal measurement on the training set and derive the expression of the optimal excess risk. Summarising the previous section, we transformed the original problem into a parameter estimation one for the Gaussian shift model (22) with parameters $(\vec{u},\vec{v}) \in \mathbb{R}^3 \times \mathbb{R}^3$. The parameter to be estimated $\vec{z}_\perp \in \mathbb{R}^2$ is a linear transformation of $(\vec{u},\vec{v})$

$$\vec{z}_\perp = \vec{z} - \vec{d_0}(\vec{z}\cdot\vec{d_0})/\|\vec{d_0}\|^2, \qquad \vec{z} := \pi_0\vec{u} - \pi_1\vec{v}$$

i.e. we would like to minimise the risk

$$R_{max}(\hat{\vec{z}};\rho_0,\sigma_0) := \sup_{\vec{u},\vec{v}} \mathbb{E}L((\vec{u},\vec{v}),\hat{\vec{z}}) = \sup_{\vec{u},\vec{v}} \frac{1}{4\|\vec{d_0}\|}\mathbb{E}\|\hat{\vec{z}} - \vec{z}_\perp\|^2.$$

.

Since the local parameters contain both classical and quantum components it is convenient to express the loss function $L((\vec{u},\vec{v}),\hat{\vec{z}})$ in terms of these components. Let $(\vec{p_0},\vec{l_0},\vec{k_0})$ be the reference frame with $\vec{l_0}$ in the plane $(\vec{r_0},\vec{s_0})$. Denote by $\varphi_0,\varphi_1$ the angles between $(\vec{r_0},\vec{l_0})$ and respectively $(\vec{s_0},\vec{l_0})$ (see Figure 4). Then $\vec{z}_\perp = z_l\vec{l_0} + z_k\vec{k_0}$ with components

$$\begin{aligned} z_l &= (\pi_0\cos\varphi_0 u_3 - \pi_1\cos\varphi_1 v_3) + (\pi_0\sin\varphi_0 u_1 + \pi_1\sin\varphi_1 v_1) := z_l^{(c)} + z_l^{(q)}, \\ z_k &= \pi_0 u_2 - \pi_1 v_2 \end{aligned}$$

where $z_l$ was split into a contribution coming from the 'classical' parameters $(u_3,v_3)$, and another one from the 'quantum' parameters. Since the classical and quantum parts of the Gaussian model are *independent* it is easy to verify that the optimal estimator $\hat{\vec{z}}$ can be written as

$$\hat{\vec{z}} = (\hat{z}_l^{(c)} + \hat{z}_l^{(q)})\vec{l_0} + \hat{z}_k\vec{k_0}$$

where $\hat{z}_l^{(c)}$ is the optimal estimator of $z_l^{(c)}$ and $(\hat{z}_l^{(q)},\hat{z}_k)$ are optimal estimators of $(z_l^{(q)},z_k)$ obtained by (jointly) measuring the two quantum Gaussian components. The excess risk can

be written as

$$4\|\vec{d}_0\|\mathbb{E}\left[L\left((\vec{u},\vec{v}),\hat{\vec{z}}\right)\right] = \mathbb{E}\left[(z_l^{(c)}-\hat{z}_l^{(c)})^2\right] + \mathbb{E}\left[(z_l^{(q)}-\hat{z}_l^{(q)})^2+(z_k-\hat{z}_k)^2\right]$$

where the classical and quantum contributions separate and can be optimised separately. The optimal choice for the classical estimator is

$$\hat{z}_l^{(c)} = \sqrt{\pi_0}\cos\varphi_0 X_r - \sqrt{\pi_1}\cos\varphi_1 X_s$$

where $(X_r, X_s) \sim N_{\vec{u}} \otimes N_{\vec{v}}$ denote the random variables making up the classical part of the limit Gaussian model. Its contribution to the excess risk is

$$\mathbb{E}\left[\left(z_l^{(c)}-\hat{z}_l^{(c)}\right)^2\right] = \pi_0(1-r_0^2)\cos^2\varphi_0 + \pi_1(1-s_0^2)\cos^2\varphi_1. \tag{24}$$

On the other hand $(z_l^{(q)}, z_k)$ are the means of the canonical coordinates

$$\begin{aligned} Q^{(l)} &:= \sqrt{2r_0\pi_0}\sin\varphi_0 Q_1 + \sqrt{2s_0\pi_1}\sin\varphi_1 Q_2, \\ Q^{(k)} &:= \sqrt{2r_0\pi_0}P_1 - \sqrt{2s_0\pi_1}P_2 \end{aligned} \tag{25}$$

whose commutator is

$$[Q^{(l)}, Q^{(k)}] = i(2r_0\pi_0\sin\varphi_0 - 2s_0\pi_1\sin\varphi_1)\mathbf{1} := ic\mathbf{1}.$$

Now, the optimal joint measurement of canonical variables is the heterodyne type where the non-commuting coordinates are combined with the coordinates of an additional oscillator prepared in a squeezed state [10, 24]. The optimal mean square error is

$$\begin{aligned} \mathbb{E}\left[(z_l^{(q)}-\hat{z}_l^{(q)})^2+(z_k-\hat{z}_k)^2\right] &= Var(Q^{(l)}) + Var(Q^{(k)}) + |c| \\ &= \pi_0\sin^2\varphi_0 + \pi_1\sin^2\varphi_1 + 1 \\ &\quad + 2|\pi_0 r_0\sin\varphi_0 - \pi_1 s_0\sin\varphi_1| \end{aligned} \tag{26}$$

Adding the classical and quantum contributions (24) and (26) we obtain the minimax risk

$$R^{(l)}_{minmax}(\rho_0,\sigma_0) = \frac{2+2|\pi_0 r_0\sin\varphi_0 - \pi_1 s_0\sin\varphi_1| - r_0 s_0\cos\varphi_0\cos\varphi_1}{4\|\vec{d}_0\|} \tag{27}$$

which only depends on the states $(\rho_0, \sigma_0)$, for given priors $(\pi_0, \pi_1)$.

**Theorem 5.1.** *Consider the quantum classification problem with training set $\rho^{\otimes\pi_0 n}\otimes\sigma^{\otimes\pi_1 n}$ where $\rho,\sigma$ are unknown qubit states and $(\pi_0,\pi_1)$ are known.*

*Let $R^{(l)}_{minmax}(\rho_0,\sigma_0)$ be the local minimax risk as defined in Section 2.3. Under the assumption* (17)*, $R^{(l)}_{minmax}(\rho_0,\sigma_0)$ is given by* (27)*.*

*The optimal measurement consists of the following steps:*

- *(i) construct rough estimators of $\rho$ and $\sigma$ by measuring $n^{1-\epsilon}$ systems;*
- *(ii) transfer the localised spins state by $T_n$ as in Theorem 3.5 ;*
- *(iii) perform the optimal coherent measurement of $(Q^{(l)}, Q^{(k)})$ and combine with classical estimator $\hat{z}_c^l$ to produce estimator $\widehat{P}_n$.*

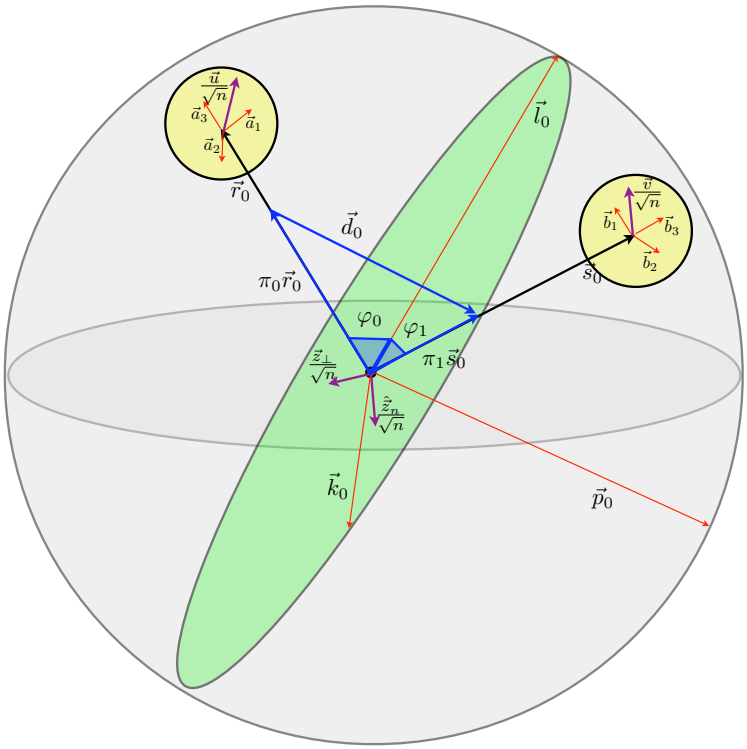


**Figure 4.** Bloch ball geometry of the learning problem.

The unknown states are localised in the two yellow balls centred at $\vec{r}_0$ and $\vec{s}_0$ and have local vectors $\vec{u}/\sqrt{n}$ and $\vec{v}/\sqrt{n}$ coloured in purple.

The three reference systems $(\vec{a}_1, \vec{a}_2, \vec{a}_3)$, $(\vec{b}_1, \vec{b}_2, \vec{b}_3)$ and $(\vec{p}_0, \vec{l}_0, \vec{k}_0)$ are coloured in red.

The green equatorial plane is orthogonal to $\vec{p}_0$ and contains the estimator $\hat{\vec{z}}_n$ and the vector to be estimated $\vec{z}_\perp$ (coloured in purple).

### *5.1. Plug-in classifier based on optimal state estimation*

Here we compute the asymptotics of the renormalised risk of the plug-in classifier based on optimal state estimation.

The problem of optimal state estimation for mixed i.i.d. qubits was solved in the asymptotic local minimax setting in [24]. The optimal measurement procedure is adaptive and the first two steps are identical to those of Theorem 5.1

(i) construct rough estimators of $\rho$ and $\sigma$ by measuring $n^{1-\epsilon}$ systems;

(ii) transfer the localised spins state by $T_n$ as in Theorem 3.5 ;

(iii) Perform separate heterodyne measurements on the modes $(Q_1, P_1)$ and $(Q_2, P_2)$ and observe the classical components to obtain the estimators $\tilde{\vec{u}}_n$ and $\tilde{\vec{v}}_n$.

Once the states (local parameters) have been estimated we can classify new states by applying

the plug-in measurement $\widetilde{M}_n := (\widetilde{P}_n, \mathbf{1} - \widetilde{P}_n)$ where $\widetilde{P}_n$ has Bloch vector

$$\tilde{\vec{p}} = \frac{\tilde{\vec{d}}}{\|\tilde{\vec{d}}\|} = \frac{\vec{d}_0 + \frac{\tilde{\vec{z}}_n}{\sqrt{n}}}{\left\|\vec{d}_0 + \frac{\tilde{\vec{z}}}{\sqrt{n}}\right\|}, \qquad \tilde{\vec{z}}_n := \tilde{\vec{z}}_\perp := (\pi_0 \tilde{\vec{u}}_n - \pi_1 \tilde{\vec{v}}_n)_\perp. \tag{28}$$

Note that $\tilde{\vec{z}}_n$ was chosen to be the orthogonal component of $\tilde{\vec{z}}$ onto the vector $\vec{p}_0$ rather than $\tilde{\vec{z}}$ itself. However a simple Taylor expansion shows that the two estimators give the same leading order contribution to the risk. As before, the minimax risk is the expectation of the quadratic loss function $L((\vec{u}, \vec{v}), \tilde{\vec{z}})$ defined in (20), but now with $\tilde{\vec{z}}$ having a different distribution compared with the optimal $\hat{\vec{z}}$. Again, we write $\tilde{z}$ as

$$\tilde{z} = \tilde{z}_l \vec{l}_0 + \tilde{z}_k \vec{k}_0 = (\tilde{z}_l^c + \tilde{z}_l^q)\vec{l}_0 + \tilde{z}_k \vec{k}_0$$

and the risk is

$$R_{max}(\tilde{\vec{z}}\,;\, \rho_0, \sigma_0) = \frac{1}{4\|\vec{d}_0\|} \mathbb{E}\left[(z_l^{(c)} - \tilde{z}_l^{(c)})^2 + (z_l^{(q)} - \tilde{z}_l^{(q)})^2 + (z_k - \tilde{z}_k)^2\right].$$

While the contribution from the first term is given by (24), the 'quantum components' have different variances due to the fact that we used a different heterodyne measurement. By using (25) and the fact that heterodyne adds a factor $1/2$ to the variance of canonical coordinates we obtain

$$\begin{aligned} \mathbb{E}\left[(z_l^{(q)} - \tilde{z}_l^{(q)})^2\right] &= \pi_0 \sin^2 \varphi_0 (r_0 + 1) + \pi_1 \sin^2 \varphi_1 (s_0 + 1) \\ \mathbb{E}\left[(z_k - \tilde{z}_k)^2\right] &= \pi_0 (r_0 + 1) + \pi_1 (s_0 + 1) \end{aligned} \tag{29}$$

Adding the three contributions we get

$$\begin{aligned} 4\|d_0\| R_{max}(\tilde{\vec{z}}\,;\, \rho_0, \sigma_0) = 2 &+ \pi_0 (r_0 \sin^2 \varphi_0 + r_0 - r_0^2 \cos^2 \varphi_0) \\ &+ \pi_1 (s_0 \sin^2 \varphi_1 + s_0 - s_0^2 \cos^2 \varphi_1). \end{aligned} \tag{30}$$

**Theorem 5.2.** *Consider the quantum classification problem with training set $\rho^{\otimes \pi_0 n} \otimes \sigma^{\otimes \pi_1 n}$ where $\rho, \sigma$ are unknown qubit states and $(\pi_0, \pi_1)$ are known.*

*Under the assumption* (17)*, the asymptotic renormalised maximum risk $R_{max}(\tilde{\vec{z}}\,;\, \rho_0, \sigma_0)$ of the plug-in classifier* (28) *is given by* (30)*.*

Comparing the minimax risk (27) with the risk (30) of the plug-in classifier we get

$$R_{max}(\tilde{\vec{z}}\,;\, \rho_0, \sigma_0) - R^{(l)}_{minmax}(\rho_0, \sigma_0) = \pi_0 r_0 (1 \pm \sin \varphi_0)^2 + \pi_1 s_0 (1 \mp \sin \varphi_1)^2,$$

with the signs are chosen according to the sign of $\pi_0 r_0 \sin \varphi_0 - \pi_1 s_0 \sin \varphi_1$. This quantity is equal to zero if and only if $\sin \varphi_0 = \mp 1$ and $\sin \varphi_1 = \pm 1$ which means that the vectors $\vec{r}_0$ and $\vec{s}_0$ are parallel and point in the same direction. For fixed priors, the difference is maximum when the $\vec{r}_0$ and $\vec{s}_0$ point in opposite directions and have length one.

This can be easily understood from the Gaussian model. When the vectors are parallel then learning requires an optimal joint measurement of *non-commuting* variables $(Q_1 - Q_2, P_1 - P_2)$ whose risk is the same as that of heterodyning the oscillators first and constructing linear combinations. In the anti-parallel case we need to measure *commuting* variables $(Q_1 + Q_2, P_1 - P_2)$ which can be done directly, without any loss.

### *5.2. The case of unknown priors*

The analysis so far deals with known priors $\pi_0, \pi_1$, which is the standard set-up usually considered in quantum statistics. In general, the priors may be unknown but can be estimated from the training set with a standard $n^{-1/2}$ error. Since the Helstrom measurement depends also on $(\pi_0, \pi_1)$, this uncertainty will bring an additional contribution to the excess risk. To find it, one needs to go back to the derivation of the quadratic loss function and add another unknown local parameter $\delta$ for the prior: $\pi_0 = q_0 + \delta/\sqrt{n}$.

Then (18) becomes

$$\vec{p} = \frac{\vec{d}}{\|\vec{d}\|} = \frac{\vec{d}_0 + \frac{\vec{z}}{\sqrt{n}} + \frac{\delta(\vec{r}_0+\vec{s}_0)}{\sqrt{n}}}{\left\|\vec{d}_0 + \frac{\vec{z}}{\sqrt{n}} + \frac{\delta(\vec{r}_0+\vec{s}_0)}{\sqrt{n}}.\right\|}, \tag{31}$$

By going through the same steps, we get to the quadratic loss function

$$L((\vec{u},\vec{v},\delta),\hat{\vec{z}}_n) := \frac{1}{4\|\vec{d}_0\|}\|\vec{z}_\perp + \delta(\vec{r}_0+\vec{s}_0)_\perp - \hat{\vec{z}}_n\|^2, \tag{32}$$

where $(\vec{r}_0+\vec{s}_0)_\perp$ is the component orthogonal to $\vec{p}_0$.

As before, the training set can be cast into a Gaussian model, with an additional independent component $Z \sim N(\delta, \pi_0\pi_1)$. This means that when taking the expectation of $L$ we get an additional factor

$$\frac{\|(\vec{r}_0+\vec{s}_0)_\perp\|^2}{4\|\vec{d}_0\|} Var(Z) = \frac{\pi_0\pi_1\|(\vec{r}_0+\vec{s}_0)_\perp\|^2}{4\|\vec{d}_0\|}.$$

## 6. Conclusions

We solved the problem of classifying two qubit states in the asymptotic local minimax statistical framework. Asymptotically the problem reduces to that of optimally estimating a sub-parameter of a quantum Gaussian model consisting of two independent oscillators in displaced thermal states with unknown means. The estimator is then used to construct an approximation of the (unknown) Helstrom measurement which is used to classify unlabelled states. The optimal procedure has excess risk of order $n^{-1}$ and we computed the exact constant factor $R^{(l)}_{minmax}(\rho_0,\sigma_0)$ as function of the two unknown states. Except in the special case of states with parallel Bloch vectors, the optimal procedure performs strictly better than the plug-in classifier obtained by estimating the states and applying the corresponding Helstrom measurement. The difference is only a constant factor, but it would probably become significant in more interesting infinite dimensional models.

Finally let us briefly discuss the Bayesian analogue of our result. In the Bayesian framework one would choose a 'regular' prior $\mu(d\rho \times d\sigma)$ over the two types of states and try to find the (asymptotically) optimal Bayes risk for this prior

$$R^\mu_{opt} := \limsup_{n\to\infty} \inf_{\widehat{M}_n} n \int \mu(d\rho \times d\sigma) R_{(\rho,\sigma)}(\widehat{M}_n).$$

When the states are pure and the prior is uniform, this has been done (even non-asymptotically) in [18], but the proof relies on the symmetry of the prior and cannot be applied to general ones, and mixed states. Based on a similar analysis done for state estimation[42],

we expect that our result can be used to prove that

$$R^{\mu}_{opt} = \int R^{(l)}_{minmax}(\rho_0, \sigma_0)\mu(d\rho_0 \times d\sigma_0).$$

The intuitive explanation is that when $n \to \infty$ the features of the prior $\mu$ are washed out and the posterior distribution concentrates in a local neighbourhood of the true parameter, where the behaviour of the classifiers is governed by the local minimax risk. Proving this relation is however beyond the scope of this paper.

## Acknowledgments

We thank Richard Gill for useful discussions. M.G. was supported by the EPSRC Fellowship EP/E052290/1.

## References

[1] Mitchell T 1997 *Machine Learning* 1st ed (McGraw-Hill Education)
[2] Devroye L, Györfi L and Lugosi G 1996 *A Probabilistic Theory of Pattern Recognition* 1st ed (Springer)
[3] Vapnik V 1998 *Statistical Learning Theory* (Wiley)
[4] Friedman J H, Hastie T and Tibshirani R 2003 *Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer)
[5] Bishop C M 2006 *Pattern recognition and machine learning* (Springer)
[6] Nielsen M and Chuang I 2000 *Quantum Computation and Quantum Information* (Cambridge: Cambridge University Press)
[7] Wiseman H M and J M G 2009 *Quantum measurements and control* (Cambridge University Press)
[8] Smithey D T, Beck, M, Raymer, M G and Faridani, A 1993 *Phys. Rev. Lett.* **70** 1244–1247
[9] Häffner H, Hänsel W, Roos C F, Benhelm J, Chek-al kar D, Chwalla M, Körber T, Rapol U D, Riebe M, Schmidt P O, Becher C, Gühne O, Dür W and Blatt R 2005 *Nature* **438** 643–646
[10] Holevo A S 1982 *Probabilistic and Statistical Aspects of Quantum Theory* (North-Holland)
[11] Helstrom C W 1976 *Quantum Detection and Estimation Theory* (Academic Press, New York)
[12] Leonhardt U 1997 *Measuring the Quantum State of Light* (Cambridge University Press)
[13] Hayashi M (ed) 2005 *Asymptotic theory of quantum statistical inference: selected papers* (World Scientific)
[14] Paris M G A and Řeháček J (eds) 2004 *Quantum State Estimation*
[15] Barndorff-Nielsen O E, Gill, R and Jupp, P E 2003 *J. R. Statist. Soc. B* **65** 775–816
[16] Sasaki M and Carlini A 2002 *Physical Review A* **66** 022303
[17] Bergou J and Hillery M 2005 *Phys. Rev. Lett.* **94** 160501
[18] Hayashi A, Horibe M and Hashimoto T 2005 *Phys. Rev. A* **72** 052306
[19] A Hayashi M Horibe T H 2006 *Phys. Rev. A* **93** 012328
[20] Aïmeur E, Brassard G and Gambs S 2006 *Proc. of the 19th Canadian Conference on Artificial Intelligence (Canadian AI'06)* (Québec City, Canada: Springer) pp 433–444
[21] Aïmeur E, Brassard G and Gambs S 2007 *Proc. of the 24th International Conference of Machine Learning (ICML'07)* (Corvallis, USA) pp 1–8
[22] Gambs S 2008 Quantum classification arXiv:0809.0444
[23] Guţă M and Kahn J 2006 *Phys. Rev. A* **73** 052108
[24] Guţă M, Janssens B and Kahn J 2008 *Commun. Math. Phys.* **277** 127–160
[25] Kahn J and Guţă M 2009 *Commun. Math. Phys.* **289** 597–652
[26] Guţă M and Jençová A 2007 *Commun. Math. Phys.* **276** 341–379
[27] Le Cam L 1986 *Asymptotic Methods in Statistical Decision Theory* (Springer Verlag, New York)
[28] Guţă M, Adesso G and Bowles P Quantum teleportation benchmarks for independent and identically-distributed spin states and displaced thermal states in preparation
[29] Bagan E, Baig M, Muñoz Tapia R and Rodriguez A 2004 *Phys. Rev. A* **69** 010304
[30] Bagan E, Ballester M A, Gill R D, Muñoz Tapia R and Romero-Isart O 2006 *Phys. Rev. Lett.* **97** 130501
[31] Bagan E, Ballester, M A, Gill, R D, Monras, A and Munõz-Tapia, R 2006 *Phys. Rev. A* **73** 032301
[32] Gammelmark S and Molmer K 2009 *New Journal of Physics* **11** 033017
[33] Bisio A, Chiribella G, D'Ariano G M, Facchini S and Perinotti P 2010 *Phys. Rev. A* **81** 032324
[34] Audibert J Y and Tsybakov A B 2007 *Annals of Statistics* **35** 608–633

[35] Cohn D A, Ghahramani, Z and Jordan M I 1996 *Journal of Artificial Intelligence Research* **4** 129–145
[36] Belavkin V P 1976 *Theor. Math. Phys.* **26** 213–222
[37] van der Vaart A 1998 *Asymptotic Statistics* (Cambridge University Press)
[38] Radcliffe J M 1971 *J. Phys. A* **4** 313–323
[39] Kitagawa M and Ueda M 1993 *Phys. Rev. A* **47** 5138–5143
[40] Guţă M and Kahn J Optimal state estimation: attainability of the Holevo bound in preparation
[41] D'Ariano G M, Presti P L and Perinotti P 2005 *Journal of Physics A* **38** 5979
[42] Gill R D 2008 *Quantum Stochastics and Information: Statistics, Filtering and Control* ed Belavkin V P and Guta M (World Scientific, Singapore) pp 239–261