Asymptotic information bounds in quantum statistics

Richard D. Gill*
Mathematical Institute, University Utrecht
The Netherlands

February 18, 2009

Abstract

We derive an asymptotic lower bound on the Bayes risk when N identical quantum systems whose state depends on a vector of unknown parameters are jointly measured in an arbitrary way and the parameters of interest estimated on the basis of the resulting data. The bound is an integrated version of a quantum Cramér-Rao bound due to Holevo (1982), and it thereby links the fixed N exact Bayesian optimality usually pursued in the physics literature with the pointwise asymptotic optimality favoured in classical mathematical statistics. By heuristic arguments the bound can be expected to be sharp. This does turn out to be the case in various important examples, where it can be used to prove asymptotic optimality of interesting and useful measurement-and-estimation schemes. On the way we obtain a new family of "dual Holevo bounds" of independent interest.

1 Introduction

The aim of this paper is to derive asymptotic information bounds for "quantum i.i.d. models" in quantum statistics. That is to say, one has N copies of a quantum system each in the same state depending on an unknown vector of parameters θ, and one wishes to estimate θ, or more generally a vector function of the parameters $\psi(\theta)$, by making some measurement on the N

*URL: `www.math.uu.nl/people/gill`. Also affiliated with EURANDOM, Eindhoven, the Netherlands, `www.eurandom.nl`.

systems together. This yields data whose distribution depends on θ and on the choice of the measurement. Given the measurement, we therefore have a classical parametric statistical model, though not necessarily an i.i.d. model, since we are allowed to bring the N systems together before measuring the resulting joint system as one quantum object. In that case the resulting data need not consist of (a function of) N i.i.d. observations, and a key quantum feature is that we can generally extract more information about θ using such "collective" or "joint" measurements than when we measure the systems separately. What is the best we can do as $N \to \infty$, when we are allowed to optimize both over the measurement and over the ensuing data-processing?

A heuristic, statistically motivated, approach to deriving methods with good properties for large N is to choose the measurement to optimize the Fisher information in the data, leaving it to the statistician to process the data efficiently, using for instance maximum likelihood or related methods, including Bayesian. This heuristic principle has already been shown to work in a number of special cases in quantum statistics. Since the measurement maximizing the Fisher information typically depends on the unknown parameter value this often has to be implemented in a two-step approach, first using a small fraction of the N systems to get a first approximation to the true parameter, and then optimizing on the remaining systems using this rough guess.

The approach favoured by many physicists is to choose a prior distribution and loss function on grounds of symmetry and physical interpretation, and then to *exactly* optimize the Bayes risk over all measurements and estimators, for any given N. This approach succeeds in producing attractive methods on those rare occasions when a felicitous combination of all the mathematical ingredients leads to a simple and analytically tractable solution. Now it has been observed in a number of problems that the two approaches result in asymptotically equivalent estimators, though the measurement schemes can be strikingly different. Heuristically, this can be understood to follow from the fact that, in the physicists' approach, for large N the prior distribution should become increasingly irrelevant and the Bayes optimal estimator close to the maximum likelihood estimator. Moreover, we expect those estimators to be asymptotically normal with variances corresponding to inverse Fisher information.

Here we link the two approaches by deriving a sharp asymptotic lower bound on the Bayes risk of the physicists' approach, in terms of the optimal Fisher information of the statisticians' approach. This enables us to conclude the asymptotic optimality of some heuristically motivated measurement-and-estimation schemes by showing that they attain the asymptotic bound. Sometimes one can find in this way asymptotically optimal solutions which

are much easier to implement than the exactly optimal solution of the physicists' approach. On the other hand, it also shows (if only heuristically) that the physicists' approach, when successful, leads to procedures which are *asymptotically* optimal for other prior distributions than those used in the computation, also for loss functions only locally equivalent to their loss function of choice, and also asymptotically optimal in a pointwise rather than a Bayesian sense.

We derive our main result by combining an existing quantum Cramér-Rao bound (Holevo, 1982) with the van Trees inequality, a Bayesian Cramér-Rao bound from classical statistics (van Trees, 1968; Gill and Levit, 1995). The former can be interpreted as a bound on the Fisher information in an arbitrary measurement on a quantum system, the latter is a bound on the Bayes risk (for a quadratic loss function) in terms of the Fisher information in the data. This means that our result and its proof can be understood without any familiarity with quantum statistics. Of course, to appreciate the applications of the result, some further appreciation of "what is a quantum statistical model" is needed. The paper contains a brief summary of this; for more information the reader is referred to the papers of Barndorff-Nielsen et al. (2003), and Gill (2001). For an overview of the "state of the art" in quantum asymptotic statistics see Hayashi (2005) which reprints papers of many authors together with introductions by the editor.

Let us develop enough notation to state the main result of the paper and compare it with the comparable result from classical statistics. Starting on familar ground with the latter, suppose we want to estimate a function $\psi(\theta)$ of a parameter θ, both represented by real column vectors of possibly different dimension, based on N i.i.d. observations from a distribution with Fisher information matrix $I(\theta)$. Let π be a prior density on the parameter space and let $\widetilde{G}(\theta)$ be a symmetric positive-definite matrix defining a quadratic loss function $l(\widehat{\psi}^{(N)}, \theta) = (\widehat{\psi}^{(N)} - \psi(\theta))^\top \widetilde{G}(\theta)(\widehat{\psi}^{(N)} - \psi(\theta))$. (Later we will use $G(\theta)$, without the tilde, in the special case when ψ is θ itself). Define the mean square error matrix $V^{(N)}(\theta) = \mathbb{E}_\theta(\widehat{\psi}^{(N)} - \psi(\theta))(\widehat{\psi}^{(N)} - \psi(\theta))^\top$ so that the risk can be written $R^{(N)}(\theta) = \mathrm{trace}\, \widetilde{G}(\theta) V^{(N)}(\theta)$. The Bayes risk is $R^{(N)}(\pi) = \mathbb{E}_\pi \mathrm{trace}\, \widetilde{G} V^{(N)}$. Here, $\mathbb{E}_\theta$ denotes expectation over the data for given θ, $\mathbb{E}_\pi$ denotes averaging over θ with respect to the prior π. The estimator $\widehat{\psi}^{(N)}$ is completely arbitrary. We assume the prior density to be smooth, compactly supported and zero on the smooth boundary of its support. Furthermore a certain quantity roughly interpreted as "information in the prior" must be finite. Then it is very easy to show (Gill and Levit, 1995), using the van Trees inequality, that under minimal smoothness conditions on the statistical

model,

$$\liminf_{N\to\infty} NR^{(N)}(\pi) \;\geq\; \mathbb{E}_\pi \mathrm{trace}\, GI^{-1} \tag{1}$$

where $G = \psi' \widetilde{G} \psi'^{\top}$ and ψ' is the matrix of partial derivatives of elements of ψ with respect to those of θ.

Now in quantum statistics the data depends on the choice of measurement and the measurement should be tuned to the loss function. Given a measurement $M^{(N)}$ on N copies of the quantum system, denote by $\overline{I}_M^{(N)}$ the average Fisher information (i.e., Fisher information divided by N) in the data. The Holevo (1982) quantum Cramér-Rao bound, as extended by Hayashi and Matsumoto (2004) to the quantum i.i.d. model, can be expressed as saying that, for all θ, G, N and $M^{(N)}$,

$$\mathrm{trace}\, G(\theta)(\overline{I}_M^{(N)}(\theta))^{-1} \;\geq\; \mathcal{C}_G(\theta) \tag{2}$$

for a certain quantity $\mathcal{C}_G(\theta)$, which depends on the specification of the quantum statistical model (state of one copy, derivatives of the state with respect to parameters, and loss function G) *at the point* θ only, i.e., on local or pointwise model features (see (7) below). According to as yet unpublished work of M. Hayashi the bound is asymptotically sharp. The idea behind his work is that locally, the quantum i.i.d. model is well approximated by a quantum Gaussian location model, a quantum statistical problem for which the Holevo bound is sharp (Holevo, 1982).

We aim to prove that under minimal smoothness conditions on the quantum statistical model, and conditions on the prior similar to those needed in the classical case, but under essentially no conditions on the estimator-and-measurement sequence,

$$\liminf_{N\to\infty} NR^{(N)}(\pi) \;\geq\; \mathbb{E}_\pi \mathcal{C}_G \tag{3}$$

where, as before, $G = \psi' \widetilde{G} \psi'^{\top}$. The main result (3) is exactly the bound one would hope for, from heuristic statistical principles, and one may also expect it to be sharp, for the reasons mentioned above. In specific models of interest, the right hand side is often easy to calculate. Various specific measurement-and-estimator sequences, motivated by a variety of approaches, can also be shown in interesting examples to achieve the bound. The restrictions on the prior can often be relaxed by approximating the prior of interest, as we will show in our examples.

It was also shown in Gill and Levit (1995), how—in the classical statistical context—one can replace a fixed prior π by a sequence of priors indexed by N, concentrating more and more on a fixed parameter value θ_0, at rate

$1/\sqrt{N}$. Following their approach would, in the quantum context, lead to the pointwise asymptotic lower bounds

$$\liminf_{N\to\infty} NR^{(N)}(\theta) \;\geq\; \mathcal{C}_G(\theta) \tag{4}$$

for each θ, for *regular* estimators, and to local asymptotic minimax bounds

$$\lim_{M\to\infty} \liminf_{N\to\infty} \sup_{\|\theta-\theta_0\|\leq N^{-1/2}M} NR^{(N)}(\theta) \;\geq\; \mathcal{C}_G(\theta_0) \tag{5}$$

for *all* estimators, but we do not further develop that theory here. In classical statistics the theory of Local Asymptotic Normality is the way to unify, generalise, and understand this kind of result. We do not yet have a theory of "Q-LAN" though there are indications that it may be possible to build such a theory. The results we obtain here using more elementary tools do give further support to the distant aim of building a Q-LAN theory.

The basic tools used in this paper have now all been mentioned, but as we shall see, the proof is not a routine application of the van Trees inequality. The missing ingredient will be provided by the following new *dual* bound to (2): for all θ, K, N and $M^{(N)}$,

$$\operatorname{trace} K(\theta)\overline{I}_M^{(N)}(\theta) \;\leq\; \mathcal{C}^K(\theta) \tag{6}$$

where $\mathcal{C}^K(\theta)$ actually equals $\mathcal{C}_G(\theta)$ for a certain G defined in terms of K (as explained in Theorem 2 below). This is an *upper* bound on Fisher information, in contrast to (2) which is a lower bound on inverse Fisher information. The new inequality (6) follows from the convexity of the sets of information matrices and of inverse information matrices for arbitrary measurements on a quantum system, and these convexity properties have a simple statistical explanation. Such dual bounds have cropped up incidentally in quantum statistics, for instance in Gill and Massar (2000), but this is the first time a connection is established.

The argument for (6), and given that, for (3), is based on some general structural features of quantum statistics, and hence it is not necessary to be familiar with the technical details of the set-up. In the next section we will summarize the i.i.d. model in quantum statistics, focussing on the key facts which will be used in the proof of the dual Holevo bound (6) and of our main result, the asymptotic lower bound (3). These proofs are given in a subsequent section, where no further "quantum" arguments will be used. In a final section we will give three applications, leading to new results on some much studied quantum statistical estimation problems.

2 Quantum statistics: the i.i.d. parametric case.

The basic objects in quantum statistics are *states* and *measurements*, defined in terms of certain operators on a complex Hilbert space. To avoid technical complications we restrict attention to the finite-dimensional case, already rich in structure and applications, when operators are represented by ordinary (complex) matrices.

States and measurement The state of a d-dimensional system is represented by a $d \times d$ matrix ρ, called the *density matrix* of the state, having the following properties: $\rho^* = \rho$ (self-adjoint or Hermitian), $\rho \geq \mathbf{0}$ (non-negative), $\mathrm{trace}(\rho) = 1$ (normalized). "Non-negative" actually implies "self-adjoint" but it does no harm to emphasize both properties. $\mathbf{0}$ denotes the zero matrix; $\mathbf{1}$ will denote the identity matrix.

Example: when $d = 2$, every density matrix can be written in the form $\rho = \frac{1}{2}(\mathbf{1} + \theta_1\sigma_1 + \theta_2\sigma_2 + \theta_3\sigma_3)$ where

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

are the three Pauli matrices and where $\theta_1^2 + \theta_2^2 + \theta_3^2 \leq 1$. □

"Quantum statistics" concerns the situation when the state of the system $\rho(\theta)$ depends on a (column) vector θ of p unknown (real) parameters.

Example: a completely unknown two-dimensional quantum state depends on a vector of three real parameters, $\theta = (\theta_1, \theta_2, \theta_3)^\top$, known to lie in the unit ball. Various interesting submodels can be described geometrically: e.g., the equatorial plane; the surface of the ball; a straight line through the origin. More generally, a completely unknown d-dimensional state depends on $p = d^2 - 1$ real parameters. □

Example: in the previous example the two-parameter case obtained by demanding that $\theta_1^2 + \theta_2^2 + \theta_3^2 = 1$ is called the case of a two-dimensional pure state. In general, a state is called pure if $\rho^2 = \rho$ or equivalently ρ has rank one. A completely unknown pure d-dimensional state depends on $p = 2(d-1)$ real parameters. □

A measurement on a quantum system is characterized by the outcome space, which is just a measurable space $(\mathcal{X}, \mathcal{B})$, and a *positive operator valued measure* (POVM) M on this space. This means that for each $B \in \mathcal{B}$ there corresponds a $d \times d$ non-negative self-adjoint matrix $M(B)$, together

having the usual properties of an ordinary (real) measure (sigma-additive), with moreover $M(\mathcal{X}) = \mathbf{1}$. The probability distribution of the outcome of doing measurement M on state $\rho(\theta)$ is given by the Born law, or trace rule: $\Pr(\text{outcome} \in B) = \text{trace}(\rho(\theta)M(B))$. It can be seen that this is indeed a bona-fide probability distribution on the sample space $(\mathcal{X}, \mathcal{B})$. Moreover it has a density with respect to the finite real measure $\text{trace}(M(B))$.

Example: the most simple measurement is defined by choosing an orthonormal basis of $\mathbb{C}^d$, say $\psi_1, \ldots, \psi_d$, taking the outcome space to be the discrete space $\mathcal{X} = \{1, \ldots, d\}$, and defining $M(\{x\}) = \psi_x \psi_x^*$ for $x \in \mathcal{X}$; or in physicists' notation, $M(\{x\}) = |\psi_x\rangle\langle\psi_x|$. One computes that $\Pr(\text{outcome} = x) = \psi_x^* \rho(\theta) \psi_x = \langle\psi_x|\rho|\psi_x\rangle$. If the state is pure then $\rho = \phi\phi^* = |\phi\rangle\langle\phi|$ for some $\phi = \phi(\theta) \in \mathbb{C}^d$ of length 1 and depending on the parameter θ. One finds that $\Pr(\text{outcome} = x) = |\psi_x^* \phi|^2 = |\langle\psi_x|\phi\rangle|^2$. □

So far we have discussed state and measurement for a single quantum system. This encompasses also the case of N copies of the system, via a tensor product construction, which we will now summarize. The joint state of N identical copies of a single system having state $\rho(\theta)$ is $\rho(\theta)^{\otimes N}$, a density matrix on a space of dimension d^N. A joint or collective measurement on these systems is specified by a POVM on this large tensor product Hilbert space. An important point is that joint measurements give many more possibilities than measuring the separate systems independently, or even measuring the separate systems adaptively.

Fact to remember 1. *State plus measurement determines probability distribution of data.*

Quantum Cramér-Rao bound. Our main input is going to be the Holevo (1982) quantum Cramér-Rao bound, with its extension to the i.i.d. case due to Hayashi and Matsumoto (2004).

Precisely because of quantum phenomena, different measurements, incompatible with one another, are appropriate when we are interested in different components of our parameter, or more generally, in different loss functions. The bound concerns estimation of θ itself rather than a function thereof, and depends on a quadratic loss function defined by a symmetric real non-negative matrix $G(\theta)$ which may depend on the actual parameter value θ. For a given estimator $\widehat{\theta}^{(N)}$ computed from the outcome of some measurement $M^{(N)}$ on N copies of our system, define its mean square error matrix $V^{(N)}(\theta) = \mathbb{E}_\theta(\widehat{\theta}^{(N)} - \theta)(\widehat{\theta}^{(N)} - \theta)^\top$. The risk function when using the quadratic loss determined by G is $R^{(N)}(\theta) = \mathbb{E}_\theta(\widehat{\theta}^{(N)} - \theta)^\top G(\theta)(\widehat{\theta}^{(N)} - \theta) = \text{trace}(G(\theta)V^{(N)}(\theta))$.

One may expect the risk of good measurements-and-estimators to decrease like N^{-1} as $N \to \infty$. The quantum Cramér-Rao bound confirms that this is the best rate to hope for: it states that for unbiased estimators of a p-dimensional parameter θ, based on arbitrary joint measurements on N copies,

$$NR^{(N)}(\theta) \;\geq\; \mathcal{C}_G(\theta) \;=\; \inf_{\vec{X},V:V\geq Z(\vec{X})} \mathrm{trace}(G(\theta)V) \tag{7}$$

where $\vec{X} = (X_1, \ldots, X_p)$, the X_i are $d \times d$ self-adjoint matrices satisfying $\partial/\partial\theta_i \,\mathrm{trace}(\rho(\theta)X_j) = \delta_{ij}$; Z is the $p \times p$ self-adjoint matrix with elements $\mathrm{trace}(\rho(\theta)X_iX_j)$; and V is a real symmetric matrix. It is possible to solve the optimization over V for given $\vec{X}$ leading to the formula

$$\mathcal{C}_G(\theta) \;=\; \inf_{\vec{X}} \mathrm{trace}\big(\Re(G^{1/2}Z(\vec{X})G^{1/2}) + \mathrm{abs}\Im(G^{1/2}Z(\vec{X})G^{1/2})\big) \tag{8}$$

where $G = G(\theta)$. The absolute value of a matrix is found by diagonalising it and taking absolute values of the eigenvalues. We'll assume that the bound is finite, i.e., there exists $\vec{X}$ satisfying the constraints. A sufficient condition for this is that the Helstrom quantum information matrix H introduced in (17) below is nonsingular.

For specific interesting models, it often turns out not difficult to compute the bound $\mathcal{C}_G(\theta)$. Note, it is a bound which depends only on the density matrix of one system ($N = 1$) and its derivative with the respect to the parameter, and on the loss function, both at the given point θ. It can be found by solving a finite-dimensional optimization problem.

We will not be concerned with the specific form of the bound. What we are going to need, are just two key properties.

Firstly: the bound is local, and applies to the larger class of *locally unbiased estimators*. This means to say that *at the given point* θ, $\mathbb{E}_\theta\widehat{\theta}^{(N)} = \theta$, and at this point also $\partial/\partial\theta_i \,\mathbb{E}_\theta\widehat{\theta}_j^{(N)} = \delta_{ij}$. Now, it is well known that the "estimator" $\theta_0 + I(\theta_0)^{-1}S(\theta_0)$, where $I(\theta)$ is Fisher information and $S(\theta)$ is score function, is locally unbiased at $\theta = \theta_0$ *and achieves the Cramér-Rao bound there*. Thus the Cramér-Rao bound for *locally* unbiased estimators is sharp. Consequently, we can rewrite the bound (7) in the form (2) announced above, where $\overline{I}_M^{(N)}(\theta)$ is the *average* (divided by N) Fisher information in the outcome of an arbitrary measurement $M = M^{(N)}$ on N copies and the right hand side is defined in (7) or (8).

Fact to remember 2. *We have a family of computable lower bounds on the inverse average Fisher information matrix for an arbitrary measurement on N copies, given by (2) and (7) or (8),*

Secondly, for given θ, define the following two sets of positive-definite symmetric real matrices, in one-to-one correspondence with one another through the mapping "matrix inverse". The matrices G occurring in the definition are also taken to be positive-definite symmetric real.

$$\mathcal{V} = \{V : \mathrm{trace}(GV) \geq \mathcal{C}_G \ \forall \ G\}, \tag{9}$$

$$\mathcal{I} = \{I : \mathrm{trace}(GI^{-1}) \geq \mathcal{C}_G \ \forall \ G\}. \tag{10}$$

In the appendix to this paper, we give an algebraic proof that that the set $\mathcal{I}$ is convex (for $\mathcal{V}$, convexity is obvious), and that the inequalities defining $\mathcal{V}$ define supporting hyperplanes to that convex set, i.e., all the inequalities are achievable in $\mathcal{V}$, or equivalently $\mathcal{C}_G = \inf_{V \in \mathcal{V}} \mathrm{trace}(GV)$.

In fact, these properties have a statistical explanation, connected to the fact that the quantum statistical problem of collective measurements on N identical quantum systems approaches a quantum Gaussian problem as $N \to \infty$, see Guţă and Kahn (2006). It can be shown (Hayashi, 2003; Hayashi, personal communication; Guţă, 2005, unpublished manuscript). that $\mathcal{V}$ *consists of all covariance matrices of locally unbiased estimators achievable (by suitable choice of measurement) on a certain p-parameter quantum Gaussian statistical model. The inequalities defining* $\mathcal{V}$ *are the Holevo bounds for that model, and each of those bounds is attainable.* Thus, for each G, there exists a $V \in \mathcal{V}$ achieving equality in $\mathrm{trace}(GV) \geq \mathcal{C}_G$. It follows from this that $\mathcal{I}$ consists of all non-singular information matrices together with any non-singular matrix smaller than some information matrix, achievable by choice of measurement on the same quantum Gaussian model. Consider the set of information matrices attainable by some measurement together with all smaller matrices; and consider the set of variance matrices of locally unbiased estimators based on arbitrary measurements. Note that adding zero mean noise to a locally unbiased estimator preserves its local unbiasedness, so adding larger matrices to this set does not change it. The set of information matrices is convex: choosing measurement 1 with probability p and measurement 2 with probability q (and remembering your choice) gives a measurement whose Fisher information is the convex combination of the informations of measurements 1 and 2. Augmenting the set with all matrices smaller than something in the set, preserves convexity. (The set of variances of locally unbiased estimators is convex, by a similar randomization argument). Putting this together, we obtain

Fact to remember 3. *For given θ, both $\mathcal{V}$ and $\mathcal{I}$ defined in (9) and (10) are convex, and all the inequalities defining these sets are achieved by points in the sets.*

See the appendix for a direct algebraic proof.

3 An asymptotic Bayesian information bound

We will now introduce the van Trees inequality, a Bayesian Cramér-Rao bound, and combine it with the Holevo bound (2) via derivation of a dual bound following from the convexity of the sets (7) and (8). We return to the problem of estimating the (real, column) vector function $\psi(\theta)$ of the (real, column) vector parameter θ of a state $\rho(\theta)$ based on collective measurements of N identical copies. The dimensions of ψ and of θ need not be the same. The sample size N is largely suppressed from the notation. Let V be the mean square error matrix of an arbitrary estimator $\widehat{\psi}$, thus $V(\theta) = \mathbb{E}_\theta(\widehat{\psi}-\psi(\theta))(\widehat{\psi}-\psi(\theta))^\top$. Often, but not necessarily, we'll have $\widehat{\psi} = \psi(\widehat{\theta})$ for some estimator of θ. Suppose we have a quadratic loss function $(\widehat{\psi} - \psi(\theta))^\top \widetilde{G}(\theta)(\widehat{\psi} - \psi(\theta))$ where $\widetilde{G}$ is a positive-definite matrix function of θ, then the Bayes risk with respect to a given prior π can be written $R(\pi) = \mathbb{E}_\pi \mathrm{trace}\, \widetilde{G}V$. We are going to prove the following theorem:

Theorem 1. *Suppose $\rho(\theta) : \theta \in \Theta \subseteq \mathbb{R}^p$ is a smooth quantum statistical model and suppose π is a smooth prior density on a compact subset $\Theta_0 \subseteq \Theta$, such that Θ_0 has a piecewise smooth boundary, on which π is zero. Suppose moreover the quantity $\mathcal{J}(\pi)$ defined in (15) below, is finite. Then*

$$\liminf_{N\to\infty} NR^{(N)}(\pi) \;\geq\; \mathbb{E}_\pi \mathcal{C}_{G_0} \tag{11}$$

where $G_0 = \psi' \widetilde{G} \psi'^\top$ (and assumed to be positive-definite), ψ' is the matrix of partial derivatives of elements of ψ with respect to those of θ, and $\mathcal{C}_{G_0}$ is defined by (7) or (8).

"Once continuously differentiable" is enough smoothness. Smoothness of the quantum statistical model implies smoothness of the classical statistical model following from applying an arbitrary measurement to N copies of the quantum state. Slightly weaker but more elaborate smoothness conditions on the statistical model and prior are spelled out in Gill and Levit (1995). The restriction that G_0 be non-singular can probably be avoided by a more detailed analysis.

Let $\overline{I}_M$ denote the average Fisher information matrix for θ based on a given collective measurement on the N copies. Then the van Trees inequality states that for all matrix functions C of θ, of size $\dim(\psi) \times \dim(\theta)$,

$$N\mathbb{E}_\pi \mathrm{trace}\, \widetilde{G}V \;\geq\; \frac{(\mathbb{E}_\pi \mathrm{trace}\, C\psi'^\top)^2}{\mathbb{E}_\pi \mathrm{trace}\, \widetilde{G}^{-1} C \overline{I}_M C^\top + \frac{1}{N}\mathbb{E}_\pi \frac{(C\pi)'^\top \widetilde{G}^{-1} (C\pi)'}{\pi^2}} \tag{12}$$

where the primes in ψ' and in $(C\pi)'$ both denote differentiation, but in the first case converting the vector ψ into the matrix of partial derivatives of elements of ψ with respect to elements of θ, of size $\dim(\psi) \times \dim(\theta)$, in the second case converting the matrix $C\pi$ into the column vector, of the same length as ψ, with row elements $\sum_j (\partial/\partial\theta_j)(C\pi)_{ij}$. To get an optimal bound we need to choose $C(\theta)$ cleverly.

First though, note that the Fisher information appears in the denominator of the van Trees bound. This is a nuisance since we have a Holevo's *lower* bound (2) to the *inverse* Fisher information. We would like to have an *upper* bound on the information itself, say of the form (6), together with a recipe for computing $\mathcal{C}^K$.

All this can be obtained from the convexity of the sets $\mathcal{I}$ and $\mathcal{V}$ defined in (10) and (9) and the non-redundancy of the inequalities appearing in their definitions. Suppose V_0 is a boundary point of $\mathcal{V}$. Define $I_0 = V_0^{-1}$. Thus I_0 (though not necessarily an attainable average information matrix $\overline{I}_M^{(N)}$) satisfies the Holevo bound for each positive-definite G, and attains equality in one of them, say with $G = G_0$. In the language of convex sets, and "in the V-picture", $\operatorname{trace} G_0 V = \mathcal{C}_{G_0}$ is a supporting hyperplane to $\mathcal{V}$ at $V = V_0$.

Under the mapping "matrix-inverse" the hyperplane $\operatorname{trace} G_0 V = \mathcal{C}_{G_0}$ in the V-picture maps to the smooth surface $\operatorname{trace} G_0 I^{-1} = \mathcal{C}_{G_0}$ touching the set $\mathcal{I}$ at I_0 in the I-picture. Since $\mathcal{I}$ is convex, the tangent plane to the smooth surface at $I = I_0$ must be a supporting hyperplane to $\mathcal{I}$ at this point. The matrix derivative of the operation of matrix inversion can be written $\mathrm{d}A^{-1}/\mathrm{d}x = -A^{-1}(\mathrm{d}A/\mathrm{d}x)A^{-1}$. This tells us that the equation of the tangent plane is $\operatorname{trace} G_0 I_0^{-1} I I_0^{-1} = \operatorname{trace} G_0 I_0^{-1} = \mathcal{C}_{G_0}$. Since this is simultaneously a supporting hyperplane to $\mathcal{I}$ we deduce that for all $I \in \mathcal{I}$, $\operatorname{trace} G_0 I_0^{-1} I I_0^{-1} \le \mathcal{C}_{G_0}$. Defining $K_0 = I_0^{-1} G_0 I_0^{-1}$ and $\mathcal{C}^{K_0} = \mathcal{C}_{G_0}$ we rewrite this inequality as $\operatorname{trace} K_0 I \le \mathcal{C}^{K_0}$.

A similar story can be told when we start in the I-picture with a supporting hyperplane (at $I = I_0$) to $\mathcal{I}$ of the form $\operatorname{trace} K_0 I = \mathcal{C}^{K_0}$ for some symmetric positive-definite K_0. It maps to the smooth surface $\operatorname{trace} K_0 V^{-1} = \mathcal{C}^{K_0}$, with tangent plane $\operatorname{trace} K_0 V_0^{-1} I V_0^{-1} = \mathcal{C}^{K_0}$ at $V = V_0 = I_0^{-1}$. By strict convexity of the function "matrix inverse", the tangent plane touches the smooth surface only at the point V_0. Moreover, the smooth surface lies above the tangent plane, but below $\mathcal{V}$. This makes V_0 the unique minimizer of $\operatorname{trace} K_0 V_0^{-1} I V_0^{-1}$ in $\mathcal{V}$.

It would be useful to extend these computations to allow singular I, G and K. Anyway, we summarize what we have so far in a theorem.

Theorem 2. *Dual to the Holevo family of lower bounds on average inverse information,* $\operatorname{trace} G\overline{I}_M^{-1} \ge \mathcal{C}_G$ *for each positive-definite* G*, we have a family*

of upper bounds on information,

$$\operatorname{trace} K\overline{I}_M \le \mathcal{C}^K \quad \textit{for each} \quad K. \tag{13}$$

If $I_0 \in \mathcal{I}$ *satisfies* $\operatorname{trace} G_0 I_0^{-1} = \mathcal{C}_{G_0}$ *then with* $K_0 = I_0^{-1} G_0 I_0^{-1}$, $\mathcal{C}^{K_0} = \mathcal{C}_{G_0}$. *Conversely if* $I_0 \in \mathcal{I}$ *satisfies* $\operatorname{trace} K_0 I_0 = \mathcal{C}^{K_0}$ *then with* $G_0 = I_0 K_0 I_0$, $\mathcal{C}_{G_0} = \mathcal{C}^{K_0}$. *Moreover, none of the bounds is redundant, in the sense that for all positive-definite* G *and* K, $\mathcal{C}_G = \inf_{V \in \mathcal{V}} \operatorname{trace}(GV)$ *and* $\mathcal{C}^K = \sup_{I \in \mathcal{I}} \operatorname{trace}(KI)$. *The minimizer in the first equation is unique.*

Now we are ready to apply the van Trees inequality. First we make a guess for what the left hand side of (12) should look like, at its best. Suppose we use an estimator $\widehat{\psi} = \psi(\widehat{\theta})$ where $\widehat{\theta}$ makes optimal use of the information in the measurement M. Denote now by I_M the asymptotic normalized Fisher information of a sequence of measurements. Then we expect that the asymptotic normalized covariance matrix V of $\widehat{\psi}$ is equal to $\psi' I_M^{-1} \psi'^{\top}$ and therefore the asymptotic normalized Bayes risk should be $\mathbb{E}_\pi \operatorname{trace} \widetilde{G} \psi' I_M^{-1} \psi'^{\top} = \mathbb{E}_\pi \operatorname{trace} \psi'^{\top} \widetilde{G} \psi' I_M^{-1}$. This is bounded below by the integrated Holevo bound $\mathbb{E}_\pi \mathcal{C}_{G_0}$ with $G_0 = \psi'^{\top} \widetilde{G} \psi'$. Let $I_0 \in \mathcal{I}$ satisfy $\operatorname{trace} G_0 I_0^{-1} = \mathcal{C}_{G_0}$; its existence and uniqueness are given by Theorem 2. (Heuristically we expect that I_0 is asymptotically attainable). By the same Theorem, with $K_0 = I_0^{-1} G_0 I_0^{-1}$, $\mathcal{C}^{K_0} = \mathcal{C}_{G_0} = \operatorname{trace} G_0 I_0^{-1} = \operatorname{trace} \psi'^{\top} \widetilde{G} \psi' I_0^{-1}$.

Though these calculations are informal, they lead us to try the matrix function $C = \widetilde{G} \psi' I_0^{-1}$. Define $V_0 = I_0^{-1}$. With this choice, in the numerator of the van Trees inequality, we find the square of $\operatorname{trace} C \psi'^{\top} = \operatorname{trace} \widetilde{G} \psi' I_0^{-1} \psi'^{\top} = \operatorname{trace} G_0 V_0 = \mathcal{C}_{G_0}$. In the main term of the denominator, we find $\operatorname{trace} \widetilde{G}^{-1} \widetilde{G} \psi' I_0^{-1} \overline{I}_M I_0^{-1} \psi'^{\top} \widetilde{G} = \operatorname{trace} I_0^{-1} G_0 I_0^{-1} \overline{I}_M = \operatorname{trace} K_0 \overline{I}_M \le \mathcal{C}^{K_0} = \mathcal{C}_{G_0}$ by the dual Holevo bound (13). This makes the numerator of the van Trees bound equal to the square of this part of the denominator, and using the inequality $a^2/(a+b) \ge a - b$ we find

$$N \mathbb{E}_\pi \operatorname{trace} GV \;\ge\; \mathbb{E}_\pi \mathcal{C}_{G_0} - \frac{1}{N} \mathcal{J}(\pi) \tag{14}$$

where

$$\mathcal{J}(\pi) \;=\; \mathbb{E}_\pi \frac{(C\pi)'^{\top} \widetilde{G}^{-1} (C\pi)'}{\pi^2} \tag{15}$$

with $C = \widetilde{G} \psi' V_0$ and V_0 uniquely achieving in $\mathcal{V}$ the bound $\operatorname{trace} G_0 V \ge \mathcal{C}_{G_0}$, where $G_0 = \psi'^{\top} \widetilde{G} \psi'$. Finally, provided $\mathcal{J}(\pi)$ is finite (which depends on the prior distribution and on properties of the model), we obtain the asymptotic lower bound

$$\liminf_{N \to \infty} N \mathbb{E}_\pi \operatorname{trace} \widetilde{G} V \;\ge\; \mathbb{E}_\pi \mathcal{C}_{G_0}. \tag{16}$$

4 Examples

In the three examples discussed here, the loss function is derived from a very popular (among the physicists) figure-of-merit in state estimation called *fidelity*. Suppose we wish to estimate a state $\rho = \rho(\theta)$ by $\widehat{\rho} = \rho(\widehat{\theta})$. Fidelity measures the closeness of the two states, being maximally equal to 1 when the estimate and truth coincide. It is defined as $\mathrm{Fid}(\widehat{\rho}, \rho) = \left(\mathrm{trace}(\sqrt{\rho^{\frac{1}{2}}\widehat{\rho}\rho^{\frac{1}{2}}})\right)^2$ (some authors would call this *squared* fidelity). When both states are pure, thus $\rho = |\phi\rangle\langle\phi|$ and $\widehat{\rho} = |\widehat{\phi}\rangle\langle\widehat{\phi}|$ where ϕ and $\widehat{\phi}$ are unit vectors in $\mathbb{C}^d$, then $\mathrm{Fid}(\widehat{\phi}, \phi) = |\langle\widehat{\phi}|\phi\rangle|^2$. There is an important characterization of fidelity due to Fuchs (1995) which both explains its meaning and leads to many important properties. Suppose M is a measurement on the quantum system. Denote by $M(\rho)$ the probability distribution of the outcome of the measurement M when applied to a state ρ. For two probability distributions P, $\widehat{P}$ on the same sample space, let p and $\widehat{p}$ be their densities with respect to a dominating measure μ and define the fidelity between these probability measures as $\mathrm{Fid}(\widehat{P}, P) = \left(\int \widehat{p}^{\frac{1}{2}} p^{\frac{1}{2}} \mathrm{d}\mu\right)^2$. In usual statistical language, this is the *squared Hellinger affinity* between the two probability measures. It turns out that $\mathrm{Fid}(\widehat{\rho}, \rho) = \inf_M \mathrm{Fid}(M(\widehat{\rho}), M(\rho))$, thus two states have small fidelity when there is a measurement which distinguishes them well, in the sense that the Hellinger affinity between the outcome distributions is small, or in other words, the L_2 distance between the root densities of the data under the two models is large.

Now suppose states are smoothly parametrized by a vector parameter θ. Consider the fidelity between two states with close-by parameter values θ and $\widehat{\theta}$, and suppose they are measured with the same measurement M. From the relation $\int p^{\frac{1}{2}}\widehat{p}^{\frac{1}{2}}\mathrm{d}\mu = 1 - \frac{1}{2}\|\widehat{p}^{\frac{1}{2}} - p^{\frac{1}{2}}\|^2$ and by a Taylor expansion to second order one finds $1 - \mathrm{Fid}(\widehat{P}, P) \approx \frac{1}{4}(\widehat{\theta} - \theta)^\top I_M(\theta)(\widehat{\theta} - \theta)$ where $I_M(\theta)$ is the Fisher information in the outcome of the measurement M on the state $\rho(\theta)$. We will define the *Helstrom* quantum information matrix $H(\theta)$ by the analogous relation

$$1 - \mathrm{Fid}(\widehat{\rho}, \rho) \approx \frac{1}{4}(\widehat{\theta} - \theta)^\top H(\theta)(\widehat{\theta} - \theta). \tag{17}$$

It turns out that $H(\theta)$ is the smallest "information matrix" such that $I_M(\theta) \leq H(\theta)$ for all measurements M.

Taking as loss function $l(\widehat{\theta}, \theta) = 1 - \mathrm{Fid}(\rho(\widehat{\theta}), \rho(\theta))$ we would expect (by a quadratic approximation to the loss) that $\mathbb{E}_\pi \mathcal{C}_{\frac{1}{4}H}$ is a sharp asymptotic lower bound on N times the Bayes risk. We will prove this result for a number of special cases, in which by a fortuitous circumstance, the fidelity-loss function

is *exactly* quadratic in a (sometimes rather strange) function of the parameter. The first two examples concern a two-dimensional quantum system and are treated in depth in Bagan et al. (2006a); below we just outline some important features of the application. In the second of those two examples our asymptotic lower bound is an essential part of a proof of asymptotic optimality of a certain measurement-and-estimation scheme.

The third example concerns an unknown pure state of arbitrary dimension. Here we are present a short and geometric proof of a surprising but little known result of Hayashi (1998) which shows that an extraordinarily simple measurement scheme leads to an asymptotically optimal estimator (providing the data is processed efficiently). The analysis also links the previously unconnected Holevo and Gill-Massar bounds (Holevo, 1982; Gill and Massar, 2000).

4.1 Completely unknown spin half (d=2, p=3)

Recall that a completely unknown 2-dimensional quantum state can be written $\rho(\theta) = \frac{1}{2}(\mathbf{1} + \theta_1\sigma_1 + \theta_2\sigma_2 + \theta_3\sigma_3)$, where θ lies in the unit ball in $\mathbb{R}^3$. It turns out that $\mathrm{Fid}(\widehat{\rho}, \rho) = \frac{1}{2}(1 + \widehat{\theta} \cdot \theta + (1 - \|\widehat{\theta}\|^2)^{\frac{1}{2}}(1 - \|\theta\|^2)^{\frac{1}{2}})$. Define $\psi(\theta)$ to be the four-dimensional vector obtained by adjoining $(1 - \|\theta\|^2)^{\frac{1}{2}}$ to θ_1, θ_2, θ_3. Note that this vector has constant length 1. It follows that $1 - \mathrm{Fid}(\widehat{\rho}, \rho) = \frac{1}{4}\|\widehat{\psi} - \psi\|^2$. This is a quadratic loss-function for estimation of $\psi(\theta)$ with $\widetilde{G} = \mathbf{1}$, the 4×4 identity matrix. By Taylor expansion of both sides, we find that $\frac{1}{4}H = \psi'^{\top}\widetilde{G}\psi' = G$ and conclude from Theorem 1 that N times 1− mean fidelity is indeed asymptotically lower bounded by $\mathbb{E}_\pi \mathcal{C}_{\frac{1}{4}H}$.

In Bagan, Ballester, Gill, Monras and Muñoz-Tapia (2006a) the exactly optimal measurement-and-estimation scheme is derived and analysed in the case of a rotationally invariant prior distribution over the unit ball. The optimal *measurement* turns out not to depend on the (arbitrary) radial part of the prior distribution, and separates into two parts, one used for estimating the direction $\theta/\|\theta\|$, the other part for estimating the length $\|\theta\|$. The Bayes optimal estimator of the length of θ naturally depends on the prior. Because of these simplifications it is feasible to compute the asymptotic value of N times the (optimal) Bayes mean fidelity, and this value is $(3 + 2\mathbb{E}_\pi\|\theta\|)/4$.

The Helstrom quantum information matrix H and the Holevo lower bound $\mathcal{C}_{\frac{1}{4}H}$ are also computed. It turns out that $\mathcal{C}_{\frac{1}{4}H}(\theta) = (3+2\|\theta\|)/4$. Our asymptotic lower bound is not only correct but also, as expected, sharp.

The van Trees approach does put some non-trivial conditions on the prior density π. The most restrictive conditions are that the density is zero at the boundary of its support and that the quantity (15) be finite. Within the unit

ball everything is smooth, but there are some singularities at the boundary of the ball. So our main theorem does not apply directly to many priors of interest. However there is an easy approximation argument to extend its scope, as follows.

Suppose we start with a prior π supported by the whole unit ball which does not satisfy the conditions. For any $\epsilon > 0$ construct $\widetilde{\pi} = \widetilde{\pi}_\epsilon$ which is smaller than $(1+\epsilon)\pi$ everywhere, and 0 for $\|\theta\| \geq 1 - \delta$ for some $\delta > 0$. If the original prior π is smooth enough we can arrange that $\widetilde{\pi}$ satisfies the conditions of the van Trees inequality, and makes (15) finite. N times the Bayes risk for $\widetilde{\pi}$ cannot exceed $1+\epsilon$ times that for π, and the same must also be true for their limits. Finally, $\mathbb{E}_{\widetilde{\pi}_\epsilon} \mathcal{C}_{\frac{1}{4}H} \to \mathrm{E}_\pi \mathcal{C}_{\frac{1}{4}H}$ as $\epsilon \to 0$.

Some last remarks on this example: first of all, it is known that *only* collective measurements can asymptotically achieve this bound. Separate measurements on separate systems lead to strictly worse estimators. In fact, by the same methods one can obtain the sharp asymptotic lower bound 9/4 (independent of the prior), see Bagan, Ballester, Gill, Muñoz-Tapia and Romero-Isart (2006b), when one allows the measurement on the nth system to depend on the data obtained from the earlier ones. Instead of the Holevo bound itself, we use here a bound of Gill and Massar (2000), which is actually has the form of a dual Holevo bound. (We give some more remarks on this at the end of the discussion of the third example). Secondly, our result gives strong heuristic support to the claim that the measurement-and-estimation scheme developed in Bagan, Ballester, Gill, Monras and Muñoz-Tapia (2006a) for a specific prior and specific loss function is also pointwise optimal in a minimax sense, or among regular estimators, for loss functions which are locally equivalent to fidelity-loss; and also asymptotically optimal in the Bayes sense for other priors and locally equivalent loss functions. In general, if the physicists' approach is successful in the sense of generating a measurement-and-estimation scheme which can be analytically studied and experimentally implemented, then this scheme will have (for large N) good properties independent of the prior and only dependent on local properties of the loss.

4.2 Spin half: equatorial plane (d=2, p=2)

Bagan, Ballester, Gill, Monras and Muñoz-Tapia (2006a) also considered the case where it is known that $\theta_3 = 0$, thus we now have a two-dimensional parameter. The prior is again taken to be rotationally symmetric. The exactly Bayes optimal measurement turns out (at least, for some N and for some priors) to depend on the radial part of the prior. Analysis of the exactly optimal measurement-and-estimation procedure is not feasible since we do not know if this phenomenon persists for all N. However there is a

natural measurement, which is exactly optimal for some N and some priors, which one might conjecture to be asymptotically optimal for all priors. This sub-optimal measurement, combined with the Bayes optimal estimator given the measurement, can be analysed and it turns out that N times $1-$ mean fidelity converges to $1/2$ as $N \to \infty$, independently of the prior. Again, the Helstrom quantum information matrix H and the Holevo lower bound $\mathcal{C}_{\frac{1}{4}H}$ are computed. It turns out that $\mathcal{C}_{\frac{1}{4}H}(\theta) = 1/2$. This time we can use our asymptotic lower bound to prove that the natural sub-optimal measurement-and-estimator is in fact asymptotically optimal for this problem.

For a p-parameter model the best one could every hope for is that for large N there are measurements with $\overline{I}_M$ approaching the Helstrom upper bound H. Using this bound in the van Trees inequality gives the asymptotic lower bound on N times $1-$ mean fidelity of $p/4$. The example here is a special case where this is attainable. Such a model is called *quasi-classical*.

If one restricts attention to separate measurements on separate systems the sharp asymptotic lower bound is 1, twice as large, see Bagan, Ballester, Gill, Muñoz-Tapia and Romero-Isart (2006b).

4.3 Completely unknown d dimensional pure state

In this example we make use of the dual Holevo bound and symmetry arguments to show that in this example, the original Holevo bound for a natural choice of G (corresponding to fidelity-loss) is attained by an extremely large class of measurements, including one of the most basic measurements around, known as "standard tomography".

For a pure state $\rho = |\phi\rangle\langle\phi|$, fidelity can be written $|\langle\widehat{\phi}|\phi\rangle|^2$ where $|\phi\rangle \in \mathbb{C}^d$ is a vector of unit length. The state-vector can be multiplied by e^{ia} for an arbitrary real phase a without changing the density matrix. The constraint of unit length and the arbitrariness of the phase means that one can parametrize the density matrix ρ corresponding to $|\phi\rangle$ by $2(d-1)$ real parameters which we take to be our underlying vector parameter θ (we have d real parts and d imaginary parts of the elements of $|\phi\rangle$, but one constraint and one parameter which can be fixed arbitrarily).

For a pure state, $\rho^2 = \rho$ so $\mathrm{trace}(\rho^2) = 1$. Another way to write the fidelity in this case is as $\mathrm{trace}(\widehat{\rho}\rho) = \sum_{ij}(\Re(\widehat{\rho}_{ij})\Re(\rho_{ij}) + \Im(\widehat{\rho}_{ij})\Im(\rho_{ij}))$. So if we take $\psi(\theta)$ to be the vector of length $2d^2$ and of length 1 containing the real and the imaginary parts of elements of ρ we see that $1-\mathrm{Fid}(\widehat{\rho}, \rho) = \frac{1}{2}\|\widehat{\psi} - \psi\|^2$. It follows that $1-$ fidelity is a quadratic loss function in $\psi(\theta)$ with again $\widetilde{G} = \mathbf{1}$.

Define again the Helstrom quantum information matrix $H(\theta)$ for θ by $1 - \mathrm{Fid}(\widehat{\rho}, \rho) \approx \frac{1}{4}(\widehat{\theta} - \theta)^\top I_M(\theta)(\widehat{\theta} - \theta)$. Just as in the previous two examples

we expect the asymptotic lower bound $\mathbb{E}_\pi \mathcal{C}_{\frac{1}{4}H}$ to hold for N times Bayes mean fidelity-loss, where $G = \frac{1}{4}H = \psi'^{\top}\widetilde{G}\psi'$.

Some striking facts are known about estimation of a pure state. First of all, from Matsumoto (2002), we know that the Holevo bound is attainable, for all G, already at $N = 1$. Secondly, from Gill and Massar (2000) we have the following inequality

$$\text{trace} H^{-1}\overline{I}_M \leq d - 1 \tag{18}$$

with *equality* (in the case that the state is completely unknown) for all *exhaustive* measurements $M^{(N)}$ on N copies of the state. Exhaustivity means, for a measurement with discrete outcome space, that $M^{(N)}(\{x\})$ is a rank one matrix for each outcome x. The meaning of exhaustivity in general is by the same property for the density $m(x)$ of the matrix-valued measure $M^{(N)}$ with respect to a real dominating measure, e.g., $\text{trace}(M^{(N)}(\cdot))$. This tells us that (18) is one of the "dual Holevo inequalities". We can associate it with an original Holevo inequality once we know an information matrix of a measurement attaining the bound. We will show that there is an information matrix of the form $\overline{I}_M = cH$ attaining the bound. Since the number of parameters (and dimension of H) is $2(d-1)$ it follows by imposing equality in (18) that $c = \frac{1}{2}$. The corresponding Holevo inequality must be $\text{trace}\frac{1}{2}HH^{-1}\frac{1}{2}H\overline{I}_M^{-1} \geq d - 1$ which tells us that $\mathcal{C}_{\frac{1}{4}H} = d - 1$.

The proof uses an invariance property of the model. For any unitary matrix U (i.e., $UU^* = U^*U = \mathbf{1}$) we can convert the pure state ρ into a new pure state $U\rho U^*$. The unitary matrices form a group under multiplication. Consequently the group can be thought to act on the parameter θ used to describe the pure state. Clearly the fidelity between two states (or the fidelity between their two parameters) is invariant when the same unitary acts on both states. This group action possesses the "homogenous two point property": for any two pairs of states such that the fidelities between the members of each pair are the same, there is a unitary transforming the first pair into the second pair.

We illustrate this in the case $d = 2$ where (first example, section 2), the pure states can be represented by the surface of the unit ball in $\mathbb{R}^3$. It turns out that the action of the unitaries on the density matrices translates into the action of the group of orthogonal rotations on the unit sphere. Two points at equal distance on the sphere can be transformed by some rotation into any other two points at the same distance from one another; a constant distance between points on the sphere corresponds to a constant fidelity between the underlying states.

In general, the pure states of dimension d can be identified with the

Riemannian manifold CP^{d-1} whose natural Riemannian metric corresponds locally to fidelity (locally, $1-$ fidelity is squared Riemannian distance) and whose isometries correspond to the unitaries. This space posseses the homogenous two point property, as we argued above. It is easy to show that the *only* Riemannian metrics invariant under isometries on such a space are proportional to one another. Hence the quadratic forms generating those metrics with respect to a particular parametrization must also be proportional to one another.

Consider a measurement whose outcome is actually an estimate of the state, and suppose that this measurement is *covariant* under the unitaries. This means that transforming the state by a unitary, doing the measurement on the transformed state, and transforming the estimate back by the inverse of the same unitary, is the same (has the same POVM) as the original measurement. The information matrix for such a measurement is generated from the squared Hellinger affinity between the distributions of the measurement outcomes under two nearby states, just as the Helstrom information matrix is generated from the fidelity between the states. If the measurement is covariant then the Riemannian metric defined by the information matrix of the measurement outcome must be invariant under unitary transformations of the states. Hence: *the information matrix of any covariant measurement is proportional to the Helstrom information matrix.*

Exhaustive covariant measurements certainly do exist. A particularly simple one is that, for each of the N copies of the quantum system, we independently and uniformly choose a basis of $\mathbb{C}^d$ and perform the simple measurement (given in an example in Section 2) corresponding to that basis.

The first conclusion of all this is: any exhaustive covariant measurement has information matrix $\overline{I}_M^{(N)}$ equal to one half the Helstrom information matrix. All such measurements attain the Holevo bound $\mathrm{trace} \frac{1}{4} H (\overline{I}_M^{(N)})^{-1} \geq d-1$. In particular, this holds for the i.i.d. measurement based on repeatedly choosing a uniformly distributed random basis of $\mathbb{C}^d$.

The second conclusion is that an asymptotic lower bound on N times $1-$ mean fidelity is $d-1$. Now the exactly Bayes optimal measurement-and-estimation strategy is known to achieve this bound. The measurement involved is a mathematically elegant collective measurement on the N copies together, but hard to realise in the laboratory. Our results show that one can expect to asymptotically attain the bound by decent information processing (maximum likelihood? optimal Bayes with uniform prior and fidelity loss?) following an arbitrary *exhaustive covariant measurement*, of which the most simple to implement is the standard tomography measurement consisting of an independent random choice of measurement basis for each separate

system.

In Gill and Massar (2000) the same bound as (18) was shown to hold for separable (and in particular, for adaptive sequential) measurements also in the mixed state case. Moreover in the case $d = 2$, any information matrix satisfying the bound is attainable already at $N = 1$. This is used in Bagan et al. (2006b) to obtain sharp asymptotic bounds to mean fidelity for separable measurements on mixed qubits.

5 Acknowledgements

Thanks to Peter Jupp for advice on Riemannian manifolds, Manuel Ballester, Madalin Guţă and Jonas Kahn for their help in understanding the Holevo bound, and to Emili Bagan, Ramon Muñez-Tapia and Alex Monras whose work led to the questions studied here.

I acknowledge financial support from the European Community project RESQ, contract no IST-2001-37559.

References

Bagan, E., M. A. Ballester, R. D. Gill, A. Monras, and R. Muñoz-Tapia (2006a). Optimal full estimation of qubit mixed states. *Phys. Rev. A* **73** 032301; `arXiv:quant-ph/0510158` .

Bagan, E., M. A. Ballester, R. D. Gill, R. Muñoz-Tapia, and O. Romero-Isart (2006b). Separable-measurement estimation of density matrices and its fidelity gap with collective protocols. *Phys. Rev. Lett.* (in print); `arXiv:quant-ph/0512177`

Barndorff-Nielsen, O. E., R. D. Gill, and P. E. Jupp (2003). On quantum statistical inference. *J. Roy. Statist. Soc. (B)* **65**, 775–816. With discussion and reply by the authors. `arXiv:quant-ph/0307191`.

Fuchs, C. A. (1995). Distinguishability and Accessible Information in Quantum theory. PhD thesis, University of New Mexico; `arXiv:quant-ph/9601020`.

Gill, R. D. (2001). Asymptotics in quantum statistics. In A. W. van der Vaart, M. de Gunst, and C. A. J. Klaassen (Eds.), *State of the Art in Probability and Statistics (Leiden, 1999)*, Volume **36** of *IMS Lecture notes Monogr. Ser.*, pp. 255–285. Beachwood, OH: Inst. Math. Statist. `arXiv:math.ST/0405571`.

Gill, R. D. and B. Y. Levit (1995). Applications of the Van Trees inequality: a Bayesian Cramér-Rao bound. *Bernoulli* **1**, 59–79.

Gill, R. D. and S. Massar (2000). State estimation for large ensembles. *Phys. Rev. A* **61**, 042312–042335. `arXiv:quant-ph/9902063` .

Guţă, M. and J. Kahn (2006). Local asymptotic normality for qubit states. *Phys. Rev. A* **73**, 052108. `arXiv:quant-ph/0512075`.

Hayashi, M. (1998). Asymptotic estimation theory for a finite dimensional pure state model. *J. Phys. A* **31**, 4633–4655. `arXiv:quant-ph/9704041` .

Hayashi, M. (2003). Quantum estimation and quantum central limit theorem (in Japanese). *Sugaku* **55**(4), 368–391. New, English translation: `arXiv:quant-ph/0608198`.

Hayashi, M. (2005) (editor). *Asymptotic Theory of Quantum Statistical Inference: Selected Papers*. Singapore: World Scientific.

Hayashi, M. and K. Matsumoto (2004). Asymptotic performance of optimal state estimation in quantum two level system. `arXiv:quant-ph/0411073`.

Holevo, A. S. (1982). *Probabilistic and Statistical Aspects of Quantum Theory*. Amsterdam: North-Holland. First appeared, in Russian, 1980.

Matsumoto, K. (2002). A new approach to the Cramér-Rao-type bound of the pure state model. *J. Phys. A* **35**, 3111–3123. `arXiv:quant-ph/9711008`.

van Trees, H. (1968). *Detection, Estimation and Modulation Theory, Part 1*. New York: Wiley.

Appendix: proof of convexity

The first step is to show that

$$\mathcal{V} \;=\; \mathrm{clos}\{V : V \geq Z(\vec{X}) \text{ for some } \vec{X}\} \tag{19}$$

where, as before, $\vec{X} = (X_1, \ldots, X_p)$, the X_i are $d \times d$ self-adjoint matrices satisfying $\partial/\partial\theta_i \,\mathrm{trace}(\rho(\theta)X_j) = \delta_{ij}$; Z is the $p \times p$ self-adjoint matrix with elements $\mathrm{trace}(\rho(\theta)X_iX_j)$; and V is a real symmetric matrix.

An easy computation shows that $Z(p\vec{X}+(1-p)\vec{Y}) \leq pZ(\vec{X})+(1-p)Z(\vec{Y})$ (check that the second derivative w.r.t. p of $\langle\psi|Z(p\vec{X}+(1-p)\vec{Y})|\psi\rangle$ is non-negative, for any complex vector ψ.) This makes $\{V : V \geq Z(\vec{X}) \text{ for some } \vec{X}\}$, where V is self-adjoint, a convex set. Restricting to the real matrices in this set preserves convexity, as does taking the closure of the set. By convexity, the definition (7) tells us that the equations $\text{trace}(GV) = \mathcal{C}_G$ define supporting hyperplanes to the set defined on the right hand side of (19). Since a closed convex set is the intersection of the closed halfspaces defined by its supporting hyperplanes, it follows that $\mathcal{V}$ as defined by (9) can also be specified as (19), and that all the Holevo bounds $\text{trace}(GV) \geq \mathcal{C}_G$ are attained in $\mathcal{V}$.

The convexity of $\mathcal{I}$, the set of inverses of elements of $\mathcal{V}$, is a lot more subtle. In the following argument I will suppose that the state $\rho(\theta)$ is strictly positive. The proof is easily adapted to the case of a model for a pure state. (More generally we need the notion of D-invariant model and the $\mathcal{L}^2$ spaces defined by a quantum state, see Holevo, 1982 or Hayashi and Matsumoto, 2004).

We can consider our model with p parameters and a strictly positive density matrix as a submodel of the model of a completely unknown mixed state, which has $d^2 - 1$ parameters. Denote the parameter vector of the full model by ϕ. The submodel is parametrized by θ, a subvector of ϕ. I'll use the terminology interest parameter, nuisance parameter for the two subvectors of ϕ corresponding to submodel parameters and auxiliary parameters. Subscripts 1, 2 will be also used when we partition matrices or vectors according to these two parts. By the strict positivity of ρ we are working at a point in the interior of the full model (this is one of the reasons why the argument needs to be adapted for a pure-state model). Since $\text{trace}\,\rho = 1$, the partial derivatives of ρ with respect to the components of θ in submodel and ϕ in fullmodel are traceless (i.e., have trace zero). It is easy to see from this that we may restrict the elements X of $\vec{X}$, entering into the Holevo bounds for the submodel, and elements Y of $\vec{Y}$, entering into the Holevo bounds for the full model, to be such that $\text{trace}\,\rho Y = 0$. Such Y form a $d^2 - 1$ dimensional real Hilbert space $\mathcal{L}_0^2(\rho)$ under the innerproduct $\langle X, Y\rangle_\rho = \Re\,\text{trace}\,\rho XY$.

Let ρ_i' denote the partial derivative of ρ with respect to θ_i at the fixed parameter value under consideration. For the submodel, define the symmetric logarithmic derivatives $\lambda_i \in \mathcal{L}_0^2(\rho)$ by $\langle\lambda_i, X\rangle_\rho = \text{trace}\,\rho_i' X$ for all $X \in \mathcal{L}_0^2(\rho)$. The constraints $\text{trace}\,\rho_i' X_j = \delta_{ij}$ translate into constraints $\langle\lambda_i, X_j\rangle_\rho = \delta_{ij}$ for all $i, j \leq p$. In the full model, I'll use the notation $\vec{\mu}$ for the vector of symmetric logarithmic derivatives, and $\vec{Y}$ for a candidate vector of Y_i, each of length $d^2 - 1$. Of course, $\vec{\lambda}$ is a subvector of $\vec{\mu}$. In the full model, the

constraints on $\vec{Y}$ translate into $\langle \mu_i, Y_j \rangle_\rho = \delta_{ij}$ for all $i, j \leq d^2 - 1$. The μ_i form a basis of $\mathcal{L}_0^2(\rho)$ of linearly independent vectors.

Now in the full model, the constraints on the Y_i make them uniquely defined. Thus for the full model, the set $\mathcal{V}_{\text{full}}$ is the set of all $(d^2-1) \times (d^2-1)$ real matrices W exceeding the fixed self-adjoint matrix $Z_{\text{full}} = Z(\vec{Y})$. Unfortunately, Z_{full} is singular. But we may describe $\mathcal{I}_{\text{full}}$ as the closure of the set of all real matrices less than or equal to $(Z_{\text{full}} + \delta \mathbf{1})^{-1}$ for some $\delta > 0$. The convexity of both sets is trivial. This suggests that we try to deal with the case of a p parameter model by considering it a submodel of the full d^2-1 parameter model.

The relation between inverse information matrices for full models and submodels is complicated, but that between the information matrices themselves is simple: the information matrix for a submodel is a submatrix of the information matrix of a full model. Thus we might conjecture that for every $I \in \mathcal{I}$, there exists a $W \geq Z_{\text{full}}$ such that $I \leq (W^{-1})_{11}$, the subscript "11" indicating the submodel submatrix. However, it could be that we have positive information for the submodel parameters, but zero information for the auxiliary parameters. This would make the corresponding inverse information matrix W^{-1} for the full model undefined. This problem can be solved by approximating singular information matrices by nonsingular ones. We will prove the following theorem:

Theorem 3. *$V^{-1} \in \mathcal{I}$ if and only if there exist real matrices $W^{(n)} > Z_{\text{full}}$, with $((W^{(n)})^{-1})_{11} = (V^{(n)})^{-1} \to V^{-1}$ as $n \to \infty$.*

In words, $\mathcal{I}$ is the closure of the set of 11 submatrices of real symmetric non-nonsingular matrices less than or equal to $(Z_{\text{full}} + \delta \mathbf{1})^{-1}$ for some $\delta > 0$. Consequently $\mathcal{I}$ is convex.

Proof. The proof will work by frequent reparametrizations of the nuisance part of the full model. By this we mean that ϕ is transformed smoothly and one-to-one into, say, ψ, in such a way that the interest component of ϕ is unaltered. Under such a transformation, the vector of symmetric logarithmic derivatives $\vec{\mu}$ transforms by premultiplication by an invertible matrix C whose 11 block is the identity and whose 12 block is zero, so the 'interest" part of $\vec{\mu}$ is unchanged. (Subject to C being nonsingular, for which it is just necessary that the 22 block is nonsingular, the 21 block of C can be arbitrary). At the same time the vector of operators $\vec{Y}$ transforms by premultiplication by the transposed inverse of C. Consequently, Z_{full} is transformed into $(C^\top)^{-1} Z_{\text{full}} C^{-1}$, $W \geq Z_{\text{full}}$ is transformed the same way, while W^{-1} is transformed into $C W^{-1} C^\top$. We therefore see that the 11 block (i.e., the submatrix corresponding to the submodel) of W^{-1} *remains invariant under reparametrization of the auxiliary or nuisance parameters.*

In the statement of the theorem the choice of parametrization of the auxiliary parameters is arbitrary, and so can be chosen in any convenient way. We take advantage of this possibility immediately, in the proof of the the forwards implication of the theorem.

Suppose $V \geq Z(\vec{X})$ for some $\vec{X}$ satisfying the usual constraints. Augment $\vec{\lambda}$ to a vector $\vec{\mu}$ of $d^2 - 1$ linearly independent elements $\mu_i \in \mathcal{L}_0^2(\rho)$ such that $\langle \mu_i, X_j \rangle_\rho = \delta_{ij}$ for all $i \leq d^2 - 1$, $j \leq p$. (The extra elements can be an arbitrary basis of the orthocomplement of the X_j, it is easy to check that together with the old elements they are linearly independent, hence because of their number, a basis). Next augment $\vec{X}$ to $\vec{Y}$, so that the the orthogonality relation, with X_j replaced by Y_j, also holds for $p < j \leq d^2 - 1$.

For square matrices A, B write $\mathrm{diag}(A, B)$ for the block diagonal matrix with A and B as diagonal blocks corresponding to interest and nuisance parts of the full model. Let $D_\epsilon = \mathrm{diag}(\mathbf{1}, \epsilon\mathbf{1})$, this is the diagonal matrix with 1's on the interest parameter part of the diagonal, ϵ's on the nuisance part.

We have $D_\epsilon Z_{\text{full}} D_\epsilon \to \mathrm{diag}(Z(\vec{X}), \mathbf{0}) \leq \mathrm{diag}(V, \mathbf{0})$ as $\epsilon \to 0$. Therefore, for each $\epsilon > 0$ we can find $\delta = \delta(\epsilon) > 0$ such that $D_\epsilon Z_{\text{full}} D_\epsilon < \mathrm{diag}(V, \mathbf{0}) + \delta\mathbf{1}$ and moreover such that $\delta \to 0$ as $\epsilon \to 0$. Thus for each ϵ, $Z_{\text{full}} < D_\epsilon^{-1}(\mathrm{diag}(V, \mathbf{0}) + \delta\mathbf{1}) D_\epsilon^{-1} = W_\epsilon$ where $((W_\epsilon)^{-1})_{11} \to V^{-1}$ as $\epsilon \to 0$.

Choosing a sequence $\epsilon_n \to 0$ as $n \to \infty$ we have found $W^{(n)} > Z_{\text{full}}$ for all n with $((W^{(n)})^{-1})_{11} \to V^{-1}$ as $n \to \infty$. Going back to the original parametrization does not alter $((W^{(n)})^{-1})_{11}$ so the forwards implication of the theorem is proved.

Now for the backwards implication. Suppose I am given $W > Z_{\text{full}}$, $(W^{-1})_{11} = V^{-1}$. Reparametrize the nuisance part of the full model so that $(W^{-1})_{12} = \mathbf{0}$. This does not alter $(W^{-1})_{11}$ but does alter both interest and nuisance parts of $\vec{Y}$. Denote the interest part of the transformed $\vec{Y}$ by $\vec{X}$. The inequality $W > Z_{\text{full}}$ remains true after the transformation, hence $W_{11} > Z(\vec{X})$. Since W is block diagonal, we obtain from this $(W^{-1})_{11} \leq (Z(\vec{X}) + \delta\mathbf{1})^{-1}$ for some $\delta > 0$. Taking the closure completes the proof. □