State estimation for large ensembles

R. D. Gill[a] and S. Massar[b]
[a] *Mathematical Institute, University Utrecht, Box 80010, 3508 TA Utrecht, Netherlands and Eurandom, Box 513, 5600 MB Eindhoven, Netherlands;*
[b] *Service de Physique Théorique, Université Libre de Bruxelles, CP 225, Bvd du Triomphe, B1050 Brussels, Belgium*
(18 February 1999)

We consider the problem of estimating the state of a large but finite number N of identical quantum systems. In the limit of large N the problem simplifies. In particular the only relevant measure of the quality of the estimation is the mean quadratic error matrix. Here we present a bound on the mean quadratic error which is a new quantum version of the Cramér-Rao inequality. This new bound expresses in a succinct way how in the quantum case one can trade information about one parameter for information about another parameter. The bound holds for arbitrary measurements on pure states, but only for separable measurements on mixed states—a striking example of non-locality without entanglement for mixed but not for pure states. Cramér-Rao bounds are generally derived under the assumption that the estimator is unbiased. We also prove that under additional regularity conditions our bound also holds for biased estimators. Finally we prove that when the unknown states belong to a 2 dimensional Hilbert space our quantum Cramér-Rao bound can always be attained and we provide an explicit measurement strategy that attains our bound. This therefore provides a complete solution to the problem of estimating as efficiently as possible the unknown state of a large ensemble of qubits in the same pure state. For qubits in the same mixed state, this also provides an optimal estimation strategy if one only considers separable measurements.

I. INTRODUCTION

One of the essential problems of quantum measurement theory is the estimation of an unknown quantum state of which one possesses a finite number N of copies. An often used approach to this problem is to specify a cost function that measures how much the estimation differs from the true state. One then tries to devise a measurement and estimation strategy which minimizes the mean cost. However optimal strategies have only been found in some simple highly symmetric cases (the covariant measurements of [1]).

When the number of copies N becomes large the problem simplifies considerably and one can hope to find all the optimal strategies in this limit. The solution of this problem would not only be of interest theoretically but also experimentally. Indeed the problem of estimating the state of a quantum system of which one has a large number of copies (quantum tomography) is of growing experimental importance. In some situations the major experimental limitation may be limited statistics (finite but large N) and then these optimal strategies could be applied directly. On the other hand the noise of the measuring apparatus often cannot be neglected, and then the optimal strategies only provide an upper bound on the quality of the estimation.

The reason why one can hope to solve the state estimation problem in the large N limit is that it ceases to be a "global" problem and becomes "local". Indeed for small N the estimated state will often be very different from the true state. Hence the optimal measurement strategy must take into account the behavior of the cost function for large estimation errors. On the other hand in the limit of an infinite number of copies any two states can be distinguished with certainty. So the relevant question to ask about the estimation strategy is at what *rate* it distinguishes neighboring states. That is we are only concerned with the behavior of the estimator and of the cost function very close to the true value.

To formulate the problem with precision, let us suppose that the unknown state $\rho(\theta^i)$ depends on some unknown parameters $\theta^1, \ldots, \theta^p$. After carrying out a measurement on the N copies of ρ, one will guess what are the values of the parameters θ^i. Call $\hat{\theta}^i_N$ the guessed values. For a good estimation strategy we expect the mean quadratic error (m.q.e.) to decrease as $1/N$:

$$\mathrm{E}_\theta \left((\hat{\theta}^i_N - \theta^i)(\hat{\theta}^j_N - \theta^j) \right) = V^N_{ij}(\theta) \simeq \frac{W^{ij}(\theta)}{N} \tag{1}$$

where the rescaled covariance matrix $W^{ij}(\theta)$ does not depend on N. E_θ denotes the mean taken over repetitions of the measurement with the value of θ fixed.

Consider now a smooth cost function $f(\hat{\theta},\theta)$ that measures how much the estimated value $\hat{\theta}$ differs from the true value θ of the parameter. f has a minimum at $\hat{\theta}=\theta$, hence it can be expanded as

$$f(\hat{\theta},\theta) = f_0(\theta) + C_{ij}(\theta)(\hat{\theta}^i - \theta^i)(\hat{\theta}^j - \theta^j) + O((\hat{\theta}-\theta)^3) \tag{2}$$

where C_{ij} is a positive matrix. Thus for a good estimation strategy the mean value of the cost function will decrease as

$$\mathrm{E}_\theta(f(\hat{\theta},\theta)) = f_0(\theta) + \frac{\sum_{ij} C_{ij}(\theta)W^{ij}(\theta)}{N} + o(1/N) \tag{3}$$

since we expect the expectation value of higher order terms in $\hat{\theta}-\theta$ to decrease faster then $1/N$. Note how in the limit of large N the problem becomes local, since only the quadratic cost $C_{ij}(\theta)$ and the rescaled covariance $W^{ij}(\theta)$ at θ intervene. The essential question about state estimation for large ensembles is therefore *what conditions must the rescaled covariance matrices $W^{ij}(\theta)$ satisfy?*

In the case when there is only one parameter θ the problem of finding the minimum covariance has essentially been solved. Indeed a bound on the variance of (unbiased) estimators—the quantum Cramer-Rao bound—was given in [2]. (It is interesting to note that the minimum attainable rescaled variance W^{min} induces naturally a metric on the space of states [3] [4]). A strategy for attaining the bound was proposed in [5]. In the multiparameter case different bounds for W^{ij} have been established, but in general they are not tight [2] [6] [1].

In this paper we present a new bound for W in the multiparameter case which is inspired by a discussion in [5]. This bound expresses in a natural way how one can trade information about one parameter for information about another. The interest of this new bound depends on the precise problem one is considering:

- When $\rho(\theta) = |\psi(\theta)\rangle\langle\psi(\theta)|$ is a pure state belonging to a 2 dimensional Hilbert space, then our bound is the necessary and sufficient condition W must satisfy in order to be attainable by a measurement. Furthermore in this case the bound can be attained by carrying out separate measurements on each particle. This completely solves the problem of estimating the state of a large ensemble of spin 1/2 particles (qubits) in the same pure state.
- When $\rho(\theta)$ is a pure state belonging to a Hilbert space of dimension d larger then 2, our bound is a necessary condition W must satisfy, but is not sufficient.
- When the unknown state is mixed and belongs to a 2 dimensional Hilbert space, and if one restricts oneself to measurements that act separately on each particle, then our bound is necessary and sufficient.
- When the unknown state is mixed and belongs to a Hilbert space of dimension $d > 2$, and if one restricts oneself to measurements that act separately on each particle, then our bound on W is necessary but not sufficient.
- If the unknown state is mixed and one allows collective measurements, then our bound is neither necessary nor sufficient.

This last point is quite surprising and shows that there is a fundamental difference between measuring pure states and mixed states. Indeed it is known that carrying out measurements on several identical copies of the same pure state generally requires collective measurements on the different copies [7] [8]. This is known as "non-locality without entanglement" [9]. The first point shows that in the limit of large number of copies pure states of spin 1/2 do not exhibit non-locality without entanglement. On the other hand the last point shows that in the limit of large number of copies mixed states of spin 1/2 continue to exhibit non locality without entanglement.

To describe our bound on W, we first consider for simplicity the case of a pure state of spin 1/2 particles. Suppose the unknown state is a spin 1/2 known to be in a pure state, and the state is known to be almost pointing in the $+z$ direction:

$$|\psi(\theta^1,\theta^2)\rangle \simeq |\uparrow_z\rangle + \frac{1}{2}(\theta^1 + i\theta^2)\,|\downarrow_z\rangle \tag{4}$$

where we have written an expression valid to first order in θ^1, θ^2. Suppose we carry out a measurement of the operator σ_x. We obtain the outcome $+x$ with probability $P(+x) = (1+\theta^1)/2$ and the outcome $-x$ with probability $P(-x) = (1-\theta^1)/2$. Thus the outcome of this measurement tells us about the value of θ^1. Similarly we can carry out a measurement of σ_y. We obtain the outcome $+y$ with probability $P(+y) = (1+\theta^2)/2$ and the outcome $-y$ with probability $P(-y) = (1-\theta^2)/2$. The outcome of this measurement tells us about θ^2. But the measurements σ_x and σ_y are incompatible, i.e., the operators do not commute, so they cannot be measured simultaneously. Thus if one obtains knowledge about θ^1, it is at the expense of θ^2. Indeed suppose one has N copies of the state ψ and one measures σ_x on N_1 copies and σ_y on $N_2 = N - N_1$ copies. Our estimator for θ^1 is the fraction of $+x$ outcomes minus the fraction of $-x$ outcomes. The resulting uncertainty (at the point $\theta^1 = \theta^2 = 0$) about θ^1 is then $\mathrm{E}_\theta((\hat\theta^1 - \theta^1)^2) = \frac{1}{N_1}$. Similarly we can estimate θ^2 and the corresponding uncertainty is $\mathrm{E}_\theta((\hat\theta^2 - \theta^2)^2) = \frac{1}{N_2}$. We can combine these two expressions in the following relation:

$$\frac{1}{\mathrm{E}_\theta((\hat\theta^1 - \theta^1)^2)} + \frac{1}{\mathrm{E}_\theta((\hat\theta^2 - \theta^2)^2)} = N \tag{5}$$

which expresses in a compact form how we can trade knowledge about θ^1 for knowledge about θ^2. We shall show that in the limit of a large number N of copies of pure states of spin 1/2 particles it is impossible to do better than (5).

To generalize (5), we rewrite it in a more abstract form as follows. We use polar coordinates to parameterize the unknown state of the spin 1/2 particle: $|\psi\rangle = \cos\frac{\eta}{2}|\uparrow\rangle + \sin\frac{\eta}{2}e^{i\varphi}|\downarrow\rangle$. We introduce the tensor

$$F_{\eta\eta} = 1 \quad , \quad F_{\varphi\varphi} = \sin^2\eta \quad , \quad F_{\eta\varphi} = 0 \tag{6}$$

which is simply the Euclidean metric on the sphere. Then the bound (5) can be reexpressed as

$$\mathrm{tr}\, V^{N-1} F^{-1} \leq N \tag{7}$$

where V^N is the covariance matrix defined in (1) and -1 denotes the inverse matrix.

For mixed states belonging to a 2 dimensional Hilbert space, and upon restricting oneself to separable measurements, (7) can be generalized as follows. Let us suppose that the state $\rho(\theta)$ depends on three unknown parameters. Then we can parameterize it by $\rho(\theta) = \frac{1}{2}(I + \theta^i\sigma_i)$ where I is the identity matrix, σ_i are the Pauli matrices and the 3 parameters θ^i obey $\theta^2 = \sum_i \theta^{i2} \leq 1$. We now introduce the tensor

$$F_{ij}(\theta) = \delta_{ij} + \frac{\theta_i\theta_j}{1-\theta^2} \tag{8}$$

which generalizes the tensor (6) to the case of mixed states. Then, upon restricting oneself to separable measurements, the rescaled covariance matrix W must satisfy

$$\mathrm{tr} F(\theta)^{-1} W(\theta)^{-1} \leq 1 \ . \tag{9}$$

As an application of these results, the minimum of the cost function (3) in the case of spin 1/2 particles (and for mixed states upon restricting oneself to separable measurement) is

$$\min \mathrm{E}_\theta(f(\hat\theta, \theta)) = f_0(\theta) + \frac{\left(\mathrm{tr}\sqrt{F^{-1/2} C F^{-1/2}}\right)^2}{N} + o(1/N) \tag{10}$$

which is obtained simply by minimizing (3) subject to the constraints (7) or (9).

As an application of (10), let us recall the covariant measurements on pure states of spin 1/2 particles analyzed in [1] [8]. In this problem one is given N spin 1/2 particles polarized along the direction Ω. The directions Ω are uniformly distributed on the sphere. One must devise a measurement and estimation strategy that minimize the mean value of the cost function $\cos^2\omega/2$ (where ω is the angle between the estimated direction $\hat\Omega$ and the true direction Ω). Expanding the cost function to second order in ω, and using the quantum van Trees inequality (30), one finds

$$\mathrm{E}_\Omega \cos^2\omega/2 \geq 1 - \frac{1}{N} + O(\frac{1}{N^2}) \tag{11}$$

which in the limit for large N coincides with the results (exact for all N) of [1] [8].

Equations (7) and (9) have a simple generalization in the case of particles belonging to higher dimensional Hilbert spaces. But in these cases these bounds are no longer sufficient.

In order to understand the conceptual basis of the above results, we must first recall some results from classical statistical inference.

II. CLASSICAL CRAMÉR-RAO BOUND

Consider a random variable X with probability density $p(x,\theta^i)$. (The connection—discussed below—with the quantum problem is that we can view $p(x,\theta)$ as the probability that a quantum measurement on the system yielded outcome x given that the state was $\rho(\theta)$). We take a random sample of size N from the distribution and use it to estimate the value of the parameters θ^i. Call $\hat{\theta}^i_N$ the estimated value. The following results about the variance of the estimator are known:

1. Suppose that the estimator is unbiased, that is $\mathrm{E}_\theta(\hat{\theta}^i_N - \theta^i) = 0$ (where E_θ is the expectation value at fixed θ, i.e., the integral $\int dx p(x|\theta)$). Then for any N, the following inequalities, known as the Cramér-Rao inequalities, hold [10] [2]

$$\mathrm{E}_\theta((\hat{\theta}^i_N - \theta^i)(\hat{\theta}^j_N - \theta^j)) = V^N_{ij} \geq \frac{I^{-1}_{ij}(\theta)}{N} \tag{12}$$

   and

$$V^{N}{}^{-1}_{ij} \leq N I_{ij}(\theta) \ . \tag{13}$$

   Here -1 denotes the inverse matrix and the inequality means that the difference of the two sides is a nonnegative matrix. The Fisher information matrix I is given by

$$\begin{aligned} I_{ij} &= \mathrm{E}_\theta(\partial_{\theta^i} \ln p(X|\theta) \partial_{\theta^j} \ln p(X|\theta)) \\ &= \int dx \frac{\partial_{\theta^i} p(x|\theta) \partial_{\theta^j} p(x|\theta)}{p(x|\theta)} \ . \end{aligned} \tag{14}$$

2. The hypothesis of unbiased estimators is very restrictive since most estimators will be biased. Happily it is possible to relax this condition. Here are just two of the many results available:

   (a) First of all if there is a known prior distribution $\lambda(\theta)$ for the parameters θ, then there is a Bayesian version of the Cramér-Rao inequality, the van Trees inequality [11] [12]. In the multivariate case, upon giving oneself a cost function $C_{ij}(\theta)$, one can derive the inequality

$$\int d\theta \lambda(\theta) \mathrm{tr}\ C(\theta) V^N(\theta) \geq \frac{\int d\theta \lambda(\theta) \mathrm{tr}\ C(\theta) I^{-1}(\theta)}{N} - \frac{\alpha}{N^2} \tag{15}$$

   where α is a positive number that depends on $C(\theta)$, $I(\theta)$, $\lambda(\theta)$ but is independent of N.

   (b) The second approach is independent of any prior distribution for θ, but only holds in the limit N tending to infinity and lays a mild restriction on the estimators considered. Specifically, if the probability distribution of $\sqrt{N}(\hat{\theta}^i_N - \theta^i)$ converges uniformly in θ towards a limiting distribution, Z^i, depending continuously on θ, then the limiting mean quadratic error matrix obeys $E(Z^i Z^j) \geq I^{-1}_{ij}$.

3. Furthermore in the limit of arbitrarily large samples one can attain the Cramér-Rao bound. This is proven by explicitly constructing an estimator that attains the bound in the extended senses 2a) (apart from the $1/N^2$ term) or 2b) just indicated: the maximum likelihood estimator (m.l.e.).

Modern statistical theory contains many other results having the same flavor as point 2 above, namely that the Cramér-Rao bound holds in an approximate sense for large N, without the restriction to biased estimators. Result 2a) applies to a larger class of estimators than 2b), but only gives a result on the average behavior over different values of θ. On the other hand result 2b) tells us that the maximum likelihood estimator is for large N an optimal estimator for each value of θ separately; at least, if one restricts attention to estimators satisfying some quite reasonable regularity conditions. The reason why in 2b) additional regularity is demanded is because of the phenomenon of super-efficiency (see [13] for a recent discussion) whereby an estimator can have mean quadratic error of smaller order than $1/N$ at isolated points, or even at a collection of points of measure zero[1]. Modern statistical theory (see again [13] or [14]) has concentrated on the more difficult problem of obtaining non-Bayesian results (i.e., pointwise rather than average) making much use of the technical tool of 'local asymptotic normality'. A major challenge in the quantum case is to obtain a result of type 2b) when this technique is definitely not available.

III. QUANTUM CRAMÉR-RAO BOUND

In this paper we show that similar results to 1,2a, 2b, 3 can be obtained when one must estimate the state of an unknown quantum system $\rho(\theta^i)$ of which one possesses N copies. This problem is most simply addressed, following [4], by decomposing it into a first (quantum) step in which one carries out a measurement on $\rho^N = \rho \otimes \ldots \otimes \rho$ and a second (classical) step in which one uses the result of the measurement to estimate the value of the parameters θ.

The most general way to describe the measurement is by a positive operator measurement (POVM) (taken for simplicity to be discrete) whose elements E_ξ satisfy $E_\xi \geq 0$, $\sum_\xi E_\xi = I$. Quantum mechanics tells us the probability to obtain outcome ξ given state $\rho(\theta)$:

$$p(\xi|\theta) = \mathrm{tr}\, \rho^N(\theta) E_\xi \ . \tag{16}$$

From the outcome ξ of the measurement one can guess what are the values of the parameters θ^i. Call $\hat{\theta}^i_N$ the estimated values of the parameters. We would like to obtain bounds on the variance of the estimators $\hat{\theta}^i_N$. To proceed we make for the time being—as in the classical case—the simplifying assumption that the estimators are unbiased: $\mathrm{E}_\theta \left(\hat{\theta}^i_N - \theta^i\right) = 0$. Then we can apply the classical Cramér-Rao inequality to the probability distribution $p(\xi|\theta)$ to obtain:

$$\mathrm{E}_\theta((\hat{\theta}^i_N - \theta^i)(\hat{\theta}^j_N - \theta^j)) = V^N_{ij} \geq I^N(E_\xi, \theta)^{-1}_{ij} \tag{17}$$

and

$${V^N}^{-1}_{ij} \leq I^N_{ij}(E_\xi, \theta) \tag{18}$$

where the Fisher information is

$$\begin{aligned} I^N_{ij}(E_\xi, \theta) &= \sum_\xi \frac{\partial_i p(\xi|\theta) \partial_j p(\xi|\theta)}{p(\xi|\theta)} \\ &= \sum_\xi \frac{\mathrm{tr}\,(\rho^N_{,i} E_\xi) \mathrm{tr}\,(\rho^N_{,j} E_\xi)}{\mathrm{tr}\,(E_\xi \rho^N)} \end{aligned} \tag{19}$$

with $\rho^N_{,i} = \partial_{\theta^i} \rho^N$.

These expressions suggest the following questions:

1. is there a simple bound for the variance V^N, or equivalently for the Fisher information $I^N(E_\xi, \theta)$?
2. is the bound also valid for sufficiently well behaved but possibly biased estimators—at least in the limit of large N?

[1]It is actually quite easy to see how one can get super-efficiency at a single point. The idea (due to Hodges in 1951) is to start with some estimator having the usual $1/N$ behavior but to improve it at this special point. Simply use the estimator to carry out a statistical test of the hypothesis that $\theta = \theta_0$. If the test accepts, replace the estimated value by the value θ_0, otherwise leave it unchanged. If one chooses the critical value of the test carefully one can ensure that for $N \to \infty$, if $\theta = \theta_0$, the test accepts with probability converging very fast to one and the estimator is essentially the true value of the parameter, with mean square error much smaller than $O(1/N)$. However if $\theta \neq \theta_0$ the test rejects with probability converging very fast to one and the modified estimator has the same $O(1/N)$ behavior as the original estimator. However for θ closer and closer to θ_0 as N increases, the new estimator has rather bad behavior. Hence its limiting distribution or limiting mean square error cannot be approached uniformly in θ. By imposing uniformity of convergence and continuity of the limit one rules out such estimation procedures in 2b). Alternatively, upon averaging over θ in 2a) the isolated points where such pathological behavior can occur do not contribute.

3. can this bound be attained—at least in the limit of a large number of copies N?

Most of the work on this subject has been devoted to answering the question 1). We now recall what is known about these questions.

Suppose first that there is only one parameter θ. The symmetric logarithmic derivative (s.l.d.) λ_θ of ρ is defined implicitly by

$$\rho_{,\theta} = \frac{\lambda_\theta \rho + \rho \lambda_\theta}{2}. \tag{20}$$

In a basis where ρ is diagonal, $\rho = \sum_k p_k |k\rangle \langle k|$, this can be inverted to yield

$$(\lambda_\theta)_{kl} = (\rho_{,\theta})_{kl} \frac{2}{p_k + p_l}. \tag{21}$$

Then we have the bound

$$I^N_{\theta\theta}(E_\xi, \theta) \leq N \mathrm{tr}\, \rho \lambda_\theta \lambda_\theta. \tag{22}$$

Furthermore it was suggested in [5] how to adapt the classical m.l.e. so as to attain, in the limit of large N, the bound (22).

In the multiparameter case the bound based on the s.l.d. can be generalized in a natural way. Define the s.l.d. along direction θ^i by

$$\rho_{,i} = \frac{\lambda_i \rho + \rho \lambda_i}{2}, \tag{23}$$

and the information matrix based on the s.l.d.

$$F_{ij} = \mathrm{tr}\, \rho \frac{\lambda_i \lambda_j + \lambda_j \lambda_i}{2}. \tag{24}$$

(This is the same matrix that was introduced for spin 1/2 particles for a particular choice of parameters in (6) and (8)). Then one can prove the bound [2],

$$I^N_{ij}(E_\xi, \theta) \leq N F_{ij}(\theta). \tag{25}$$

(This can be deduced directly from (22) as proven in [4]. Indeed since (22) holds for each path in parameter space, it implies the matrix equation (25)).

However this bound is in general not achievable. Another bound has been proposed based on an asymmetric logarithmic derivative (a.l.d.) [6] which in some cases is better than (25). Finally Holevo [1] has proposed a bound that is stronger then both the s.l.d. and the a.l.d. bound, but this bound is not explicit: it requires a further minimization. As far as we know no general achievable bound is known in the multiparameter case.

The difficulty in obtaining a simple bound in the multiparameter case is that there are many inequivalent ways in which one can minimize the variance V^N_{ij}. That is, in order to build a good estimator one must make a choice of what one wants to estimate, and according to this choice the measurement strategy followed will be different. Hence a bound in the form of a matrix inequality like (25) can never be tight.

IV. RESULTS

In this paper we obtain answers to the three questions raised above in the multiparameter case. Our results are summarized in this section.

We first discuss point 1), that is bounds on the quantum Fisher information. We shall show the following:

Theorem I: When $\rho(\theta) = |\psi(\theta)\rangle\langle\psi(\theta)|$ is a pure state, then the Fisher information $I^N(E_\xi, \theta)$ defined in (19) must satisfy the following relation

$$\mathrm{tr}\, F^{-1}(\theta) I^N(E_\xi, \theta) \leq (d-1)N \tag{26}$$

where F^{-1} is the inverse of the s.l.d. information matrix defined in (24) and d is the dimension of the Hilbert space to which $\rho(\theta)$ belongs. Note that this inequality (26) is invariant under change of parameterization $\theta \rightarrow \theta'(\theta)$.

This result immediately gives an inequality for the mean quadratic error matrix of *unbiased* estimators $\hat{\theta}_N$ by invoking the classical Cramér-Rao inequality in order to replace $I^N(E_\xi, \theta)$ by the inverse of the m.q.e. $V^N(\theta)$:

$$\operatorname{tr} F^{-1}(\theta)(NV^N(\theta))^{-1} \leq (d-1). \tag{27}$$

Theorem II: When $\rho(\theta)$ is a mixed state, and if the measurement E_ξ consists of separate measurements on each particle, then the Fisher information also satisfies (26). Hence for separable measurements on a mixed state, the m.q.e. matrix of an estimator satisfies (27).

Theorem III (non additivity of quantum Fisher information): In the case of mixed states, it is in general possible to devise a collective measurement for which the Fisher information does not satisfy the inequality (26).

The second part of the paper consists in proving that the constraint (27) also holds for biased estimators under suitable additional conditions. There are two forms of this generalized form of (27) corresponding to the two forms 2a) and 2b) of the generalized classical Cramér-Rao inequality.

Consider N copies of a state $\rho(\theta)$. If ρ is pure we can make either collective or separable measurements. If ρ is mixed we restrict ourselves to separable measurements (since Theorem III shows that in this case collective measurements can beat (26)). Based on the outcome of the measurement we estimate the value of the parameters θ^i. Call $\hat{\theta}^i$ the estimated values. Denote by $V^N_{ij} = \mathrm{E}_\theta\left((\hat{\theta}^i_N - \theta^i)(\hat{\theta}^j_N - \theta^j)\right)$ the m.q.e. of the estimator.

We shall prove the following generalization of result of type 2b) concerning the behavior of the mean quadratic error matrix as N tends to infinity:

Theorem IV: Suppose that the m.q.e. V^N has the limit $NV^N \rightarrow W$ as $N \rightarrow \infty$. To eliminate the possibility of superefficiency, we suppose that the convergence is uniform in θ and that W is continuous at θ_0. Furthermore we suppose that F is bounded in a neighbourhood of θ_0. Then we shall prove in section VI that W must satisfy

$$\operatorname{tr} F^{-1}(\theta_0)W^{-1}(\theta_0) \leq (d-1). \tag{28}$$

This result gives a bound on the mean value of a quadratic cost function C as N tends to infinity. Indeed using a Lagrange multiplier to impose the condition (28), the minimum cost is readily found to be

$$\lim_{N\rightarrow\infty} N \operatorname{tr} C(\theta_0)V^N(\theta_0) \geq \left(\operatorname{tr} \sqrt{F^{-\frac{1}{2}}(\theta_0)C(\theta_0)F^{-\frac{1}{2}}(\theta_0)}\right)^2 . \tag{29}$$

In terms of a cost function, it is also possible to prove a Bayesian version of the Cramér-Rao inequality which is the analog of the classical result 2a):

Theorem V: Suppose that one is given a cost function $C(\theta)$ and a prior distribution $\lambda(\theta)$ for the parameters θ. If C, λ and F are sufficiently smooth functions of θ (continuity of the first derivatives is sufficient), while λ is zero outside a compact region with smooth boundary, then

$$\int d\theta \lambda(\theta) \operatorname{tr} C(\theta)V^N(\theta) \geq \frac{1}{N}\int d\theta\lambda(\theta)\operatorname{tr}\left(\sqrt{F^{-\frac{1}{2}}(\theta)C(\theta)F^{-\frac{1}{2}}(\theta)}\right)^2 - \frac{\alpha}{N^2} \tag{30}$$

where α is a constant independent of N but which depends on C, λ and F.

The third part of this article is devoted to showing that in the case of spin 1/2 systems ($d = 2$) then (26) and the asymptotic version (28) are both necessary and sufficient. For mixed states we also require that the measurement be separable. We first show that at any point θ_0 we can attain equality in (26).

Theorem VI: Suppose one has N spin 1/2 particles in an unknown (eventually impure) state $\rho(\theta)$. Fix any point θ_0. Give yourself a matrix $G(\theta_0)$ satisfying $\operatorname{tr} F^{-1}(\theta_0)G(\theta_0) \leq 1$. We call G the target information matrix (more properly, it is the target for limiting rescaled information). Then there exists a measurement (depending on θ_0) acting on each spin separately $E_\xi(\theta_0)$ such that $I^N(E_\xi,\theta_0) = NG(\theta_0)$. This measurement is described in detail in section VII A.

Under mild regularity conditions we can also attain equality at all points θ simultaneously.

Theorem VII:
Suppose one has N spin 1/2 particles in a completely unknown pure state $|\psi(\theta)\rangle$. By completely unknown we mean that there are 2 unknown parameters.

Or suppose that one has N spin 1/2 particles in a completely unknown mixed state $\rho(\theta)$. By completely unknown we mean in this case that there are 3 unknown parameters. In this case we also require that the state never be pure, i.e. $\mathrm{tr}\,\rho(\theta) < 1$ for all θ.

Give yourself a smooth positive matrix $G(\theta)$ satisfying $\mathrm{tr}\,F^{-1}(\theta)G(\theta) \leq 1$ for all θ. Define the target mean quadratic error matrix $W(\theta) = G(\theta)^{-1}$. Suppose that $W(\theta)$ is non singular (i.e. $G(\theta)$ never has a zero eigenvalue).

Then there exists a measurement acting on each spin separately E_ξ, and a corresponding estimator $\hat{\theta}$, such that

$$\mathrm{E}_\theta((\hat{\theta}_i - \theta_i)(\hat{\theta}_j - \theta_j)) = \frac{W_{ij}(\theta)}{N} + o(1/N) \tag{31}$$

for all values of θ simultaneously. For this estimation strategy $\sqrt{N}(\hat{\theta} - \theta)$ converges in distribution towards $N(0, W)$, the normal distribution with mean zero and covariance W. The measurement E_ξ and estimation strategy is described in detail in section VII B.

V. NEW QUANTUM CRAMÉR-RAO INEQUALITY

In this section we prove Theorems I, II, III. That is we prove (26) for general measurements in the case of pure states and for separate measurements on each particle in the case of mixed states.

A. Preliminary results

The first step in proving (26) is to show that one can restrict oneself to POVM's whose elements are proportional to one dimensional projectors. Indeed any POVM can always be refined to yield a POVM whose elements are proportional to one dimensional projectors. We call such a measurement *exhaustive*. This yields a refined set of probability distributions $p(E_\xi, \theta)$. It is well known that under such refining of the probability distributions, the Fisher information can only increase [15].[2]

The second step in proving (26) consists in increasing the number of parameters. Suppose that $\rho(\theta^i)$ depends on p parameters θ^i, $i = 1, \ldots, p$. If $\rho = |\psi(\theta)\rangle\langle\psi(\theta)|$ is a pure state, then $p \leq 2d - 2$ (since $|\psi(\theta)\rangle$ is normalized and defined up to a phase). If ρ is a mixed state, then Hermiticity and the condition $\mathrm{tr}\,\rho = 1$ impose that $p \leq d^2 - 1$. Suppose that $p < M$ is less then the maximum number of possible parameters ($M = 2d - 2$ or $M = d^2 - 1$ according to whether the state is pure or mixed). Then one can always increase the number of parameters up to the maximum. Indeed let us suppose that to the p parameters, one adds independent parameters $\theta^{i'}$, $i' = p + 1, ..., M$. We now introduce a s.l.d. information matrix for the completed set of parameters $\tilde{F}_{ij}$, $i, j = 1, ..., M$. We shall show below that

$$\sum_{i,j=1}^{p} F^{-1}_{ij}(\theta) I^N_{ij}(E_\xi, \theta) \leq \sum_{i,j=1}^{M} (\tilde{F})^{-1}_{ij}(\theta) I^N_{ij}(E_\xi, \theta). \tag{32}$$

Therefore it will be sufficient to prove (26) in the case when there are M parameters.

---

[2] This can be seen by expressing the distribution of the refined measurement in terms of the distribution of the unrefined together with the conditional distribution of the refined outcome given the unrefined. Then the Fisher information for the refined outcome turns out to be equal to the Fisher information for the unrefined plus the mean of the Fisher information for the conditional distribution of refined given unrefined.

To prove (32), fix a particular point θ_0. At this point we have the derivative $\rho_{,i}$ and s.l.d. λ_i of ρ for $i = 1,..,n$. Introduce a set of Hermitian traceless matrices $\lambda_{i'}$, $i' = p+1,..,M$ such that

$$\mathrm{tr}\,\rho(\theta_0)\frac{\lambda_i\lambda_{i'} + \lambda_{i'}\lambda_i}{2} = 0 \quad , \quad i = 1,..,n \quad , \quad i' = n+1,..,M. \tag{33}$$

This is always possible because we can view (33) as a scalar product between λ_i and $\lambda_{i'}$ and a Gram-Schmidt orthogonalization procedure will then yield the matrices $\lambda_{i'}$. Now define the matrices $\rho_{,i'}$ by $\rho_{,i'} = \frac{\rho(\theta_0)\lambda_{i'}+\lambda_{i'}\rho(\theta_0)}{2}$. The additional parameters $\theta^{i'}$ are defined by the fact that at θ_0, $\partial_{\theta^{i'}}\rho = \rho_{,i'}$. The point of this construction is that because of (33), the s.l.d. information matrix $\tilde{F}$ is block diagonal with the first block equal to F. Let $\tilde{I}(E_\xi)$ be the Fisher information matrix for the enlarged set of parameters (but the same measurement). Then $\mathrm{tr}\,\tilde{F}^{-1}I(E_\xi) = \mathrm{tr}\,F_{11}^{-1}I_{11}(E_\xi) + \mathrm{tr}\,F_{22}^{-1}I_{22}(E_\xi)$ where the indices 11 and 22 denote the blocks of these matrices corresponding to the original and the new parameters. But both terms in this sum are non-negative since all matrices involved are nonnegative, and therefore we obtain (32) at θ_0 in this particular coordinate system. Since (32) is invariant under coordinate reparameterization, it is valid everyplace, in all coordinate systems.

B. One pure state

To proceed we shall consider a POVM whose elements are proportional to one dimensional projectors and calculate explicitly the left hand side (l.h.s.) of (26) in the case where the number of parameters is the maximum $p = 2d-2$ in a basis where F is diagonal.

We first consider the case where there is only one copy of the system ($N = 1$) and we fix a point θ_0. A this point we chose a basis such that

$$\rho(\theta_0) = |1\rangle\langle 1| \ . \tag{34}$$

Consider the $2d-2$ Hermitian operators

$$\begin{aligned} \rho_{,k+} &= |1\rangle\langle k| + |k\rangle\langle 1| \quad , \quad 1 < k \leq d \ , \\ \rho_{,k-} &= i\,|1\rangle\langle k| - i\,|k\rangle\langle 1| \quad , \quad 1 < k \leq d \ . \end{aligned} \tag{35}$$

We choose a parameterisation such that in the vicinity of θ_0, it has the form $\rho = \rho(\theta_0) + \sum_{k,\pm}(\theta^{k\pm} - \theta_0^{k\pm})\rho_{,k\pm}$. One then calculates the s.l.d. of ρ and the information matrix based on the s.l.d. One verifies that in this basis F^{sld} is diagonal:

$$F_{k\pm,k'\pm'} = 4\delta_{kk'}\delta_{\pm\pm'} \ . \tag{36}$$

Consider any POVM whose elements are proportional to one dimensional projectors

$$\begin{aligned} E_\xi &= |\psi_\xi\rangle\langle\psi_\xi| \ , \\ |\psi_\xi\rangle &= \sum_k a_{\xi k}\,|k\rangle \ . \end{aligned} \tag{37}$$

The completeness condition $\sum_\xi E_\xi = I$ takes the form

$$\sum_\xi a^*_{\xi k'}a_{\xi k} = \delta_{kk'}. \tag{38}$$

Putting all together the l.h.s. of (26) can now be written as

$$\begin{aligned} \mathrm{tr}\,F^{-1}I(E_\xi) &= \sum_\xi \frac{1}{\langle\psi_\xi|\rho|\psi_\xi\rangle}\sum_{k=2}^{d}\sum_{\pm}\frac{1}{4}\langle\psi_\xi|\rho_{,k\pm}|\psi_\xi\rangle^2 \\ &= \sum_\xi \frac{1}{|a_{\xi 1}|^2}\sum_{k=2}^{d}\frac{1}{4}\{(a^*_{\xi 1}a_{\xi k} + a^*_{\xi k}a_{\xi 1})^2 + (ia^*_{\xi 1}a_{\xi k} - ia^*_{\xi k}a_{\xi 1})^2\} \end{aligned}$$

$$
\begin{aligned}
&= \sum_\xi \frac{1}{|a_{\xi 1}|^2} \sum_{k=2}^{d} |a_{\xi 1}|^2 |a_{\xi k}|^2 \\
&= \sum_{k=2}^{d} \sum_\xi |a_{\xi k}|^2 \\
&= d-1
\end{aligned}
\tag{39}
$$

where in passing to the last equality we have used the completeness relation. This proves that equality holds in (26) for arbitrary exhaustive measurements in the case of one pure state.

C. N pure states

The generalization to N pure states proceeds as follows. Fix a point θ_0. At this point

$$
\rho^N = |1\rangle\langle 1| \otimes ... \otimes |1\rangle\langle 1| \ . \tag{40}
$$

Using the same parameterization as before, the derivatives of ρ^N are

$$
\rho^N_{,k\pm} = \rho_{,k\pm} \otimes \rho ... \otimes \rho + ... + \rho \otimes ... \otimes \rho_{,k\pm} \ . \tag{41}
$$

The elements of the POVM can be written as

$$
E_\xi = |\psi_\xi\rangle \langle \psi_\xi| \quad , \quad |\psi_\xi\rangle = \sum_{k_1=1}^{d} ... \sum_{k_N=1}^{d} a_{\xi k_1 ... k_N} \, |k_1 ... k_N\rangle \tag{42}
$$

with the completeness relation

$$
\sum_\xi a^*_{\xi k_1 ... k_N} a_{\xi k'_1 ... k'_N} = \delta_{k_1 k'_1} ... \delta_{k_N k'_N} \ . \tag{43}
$$

To proceed we need the following formulae:

$$
\mathrm{tr}\, \rho(\theta_0) E_\xi = |a_{\xi 1 ... 1}|^2 \tag{44}
$$

and

$$
\mathrm{tr}\, \rho(\theta_0)_{,k+} E_\xi = \sum_{p=1}^{N} (a^*_{\xi 1 ... 1} a_{\xi 1 .. k_p = k .. 1} + a^*_{\xi 1 .. k_p = k .. 1} a_{\xi 1 ... 1}) \tag{45}
$$

and similarly for $\mathrm{tr}\, \rho(\theta_0)_{,k-} E_\xi$. Thus we obtain

$$
(\mathrm{tr}\, \rho(\theta_0)_{,k+} E_\xi)^2 + (\mathrm{tr}\, \rho(\theta_0)_{,k-} E_\xi)^2 = \sum_{p=1}^{N} 4 |a_{\xi 1 ... 1}|^2 |a_{\xi 1 .. k_p = k .. 1}|^2 \ . \tag{46}
$$

Putting everything together yields

$$
\begin{aligned}
\mathrm{tr}\, F^{-1} I(E_\xi) &= \sum_\xi \frac{1}{\mathrm{tr}\, \rho(\theta_0) E_\xi} \frac{1}{4} \sum_{k=2}^{d} \sum_{\pm} (\mathrm{tr}\, \rho(\theta_0)_{,k+} E_\xi)^2 + (\mathrm{tr}\, \rho(\theta_0)_{,k-} E_\xi)^2 \\
&= \sum_{k=2}^{d} \sum_{p=1}^{N} \sum_\xi |a_{\xi 1 .. k_p = k .. 1}|^2 \\
&= N(d-1)
\end{aligned}
\tag{47}
$$

which proves (26) for an arbitrary number of pure states.

D. One mixed state

The case of one mixed state is similar but more complicated then the case of one pure state. We first diagonalize ρ at a point θ_0: $\rho(\theta_0) = \sum_{k=1}^{d} p_k \left|k\right\rangle \left\langle k\right|$. We now introduce the following complete set of Hermitian traceless matrices:

$$
\begin{aligned}
\rho_{,kl+} &= \left|k\right\rangle \left\langle l\right| + \left|l\right\rangle \left\langle k\right| \quad , \quad k < l \ , \\
\rho_{,kl-} &= i\left|k\right\rangle \left\langle l\right| - i\left|l\right\rangle \left\langle k\right| \quad , \quad k < l \ , \\
\rho_{,m} &= \sum_{k=1}^{d} c_{mk} \left|k\right\rangle \left\langle k\right| \quad , \quad m = 1, ..., d-1
\end{aligned} \tag{48}
$$

where the coefficients c_{mk} obey

$$
\begin{aligned}
\sum_k c_{mk} &= 0 \ , \\
\sum_k \frac{1}{p_k} c_{m'k} c_{mk} &= \delta_{m'm} .
\end{aligned} \tag{49}
$$

Let us denote the matrices $\rho_{,kl\pm}$ and $\rho_{,m}$ collectively as $\rho_{,i}$. (They constitute a set of generators of su(d)).

We choose a parameterization such that in the vicinity of θ_0, it has the form $\rho = \rho(\theta_0) + \sum_i (\theta^i - \theta^i_0)\rho_{,i}$. One then calculates the s.l.d. of ρ and the information matrix based on the s.l.d. One verifies that in this basis $F^{sld}_{i,i'}$ is diagonal:

$$
\begin{aligned}
F_{kl\pm,k'l'\pm'} &= \frac{4}{p_k + p_l} \delta_{kk'} \delta_{ll'} \delta_{\pm\pm'} \ , \\
F_{kl\pm,m} &= 0 \ , \\
F_{m,m'} &= \delta_{m'm} \ .
\end{aligned} \tag{50}
$$

Consider any POVM whose elements are proportional to one dimensional projectors

$$
\begin{aligned}
E_\xi &= \left|\psi_\xi\right\rangle \left\langle \psi_\xi\right| \ , \\
\left|\psi_\xi\right\rangle &= \sum_k a_{\xi k} \left|k\right\rangle \ .
\end{aligned} \tag{51}
$$

The l.h.s. of (26) can now be written as

$$
\mathrm{tr}\, F^{-1} I(E_\xi) = \sum_\xi \frac{1}{\langle\psi_\xi|\rho|\psi_\xi\rangle} \left(\sum_{k<l} \sum_{\pm} \frac{p_k + p_l}{4} \langle\psi_\xi|\rho_{,kl\pm}|\psi_\xi\rangle^2 + \sum_m \langle\psi_\xi|\rho_{,m}|\psi_\xi\rangle^2 \right) . \tag{52}
$$

Using the following expressions

$$
\begin{aligned}
\langle\psi_\xi|\rho_{,m}|\psi_\xi\rangle &= \sum_k |a_{\xi k}|^2 c_{mk} \ , \\
\langle\psi_\xi|\rho_{,kl+}|\psi_\xi\rangle^2 + \langle\psi_\xi|\rho_{,kl-}|\psi_\xi\rangle^2 &= 4|a_{\xi k}|^2 |a_{\xi l}|^2
\end{aligned} \tag{53}
$$

one obtains

$$
\begin{aligned}
\mathrm{tr}\, F^{-1} I(E_\xi) &= \sum_\xi \frac{1}{\langle\psi_\xi|\rho|\psi_\xi\rangle} \left(\sum_{k<l} (p_k + p_l) |a_{\xi k}|^2 |a_{\xi l}|^2 + \sum_m (\sum_k |a_{\xi k}|^2 c_{mk})^2 \right) \\
&= \sum_\xi \frac{1}{\langle\psi_\xi|\rho|\psi_\xi\rangle} \left(\sum_{k\neq l} p_k |a_{\xi k}|^2 |a_{\xi l}|^2 + \sum_k \sum_l |a_{\xi k}|^2 |a_{\xi l}|^2 \sum_m c_{mk} c_{ml} \right) .
\end{aligned} \tag{54}
$$

We now use the following relation

$$
\sum_m c_{mk} c_{ml} = \delta_{kl} p_k - p_k p_l \tag{55}
$$

which is derived from (49) as follows: define $v_{mk} = c_{mk}/\sqrt{p_k}$ ($m = 1,..,d-1$) and $v_{dk} = \sqrt{p_k}$. Then (49) can be rewritten as $\sum_k v_{mk} v_{m'k} = \delta_{mm'}$. The vectors v_{mk} therefore are a complete orthonormal basis of R^d, hence they obey $\sum_m v_{mk} v_{mk'} = \delta_{kk'}$. Reexpressing in terms of c_{mk} yields (55). Inserting it in (54) we obtain

$$\begin{aligned} \mathrm{tr}\, F^{-1} I(E_\xi) &= \sum_\xi \frac{1}{\langle \psi_\xi | \rho | \psi_\xi \rangle} \left(\sum_k \sum_l p_k (1-p_l) |a_{\xi k}|^2 |a_{\xi l}|^2 \right) \\ &= \sum_k (1-p_k) \sum_\xi |a_{\xi k}|^2 = \sum_\xi \mathrm{tr}\,(I-\rho) E_\xi \\ &= d-1 \end{aligned} \tag{56}$$

as announced.

Note that this has demonstrated that equality holds in (26) whenever $N = 1$, $p = d^2 - 1$, and the POVM is exhaustive. It follows from the classical properties of the Fisher information that equality also holds for arbitrary N whenever the POVM can be considered as a sequence of N separate exhaustive measurements on each copy of the system. It also holds if the n'th measurement is chosen at random depending on the outcomes of the previous measurements.

E. Separable measurements on N mixed states

We shall now prove that if we possess N identical mixed states of spin 1/2 particles, and carry out separable measurements, then

$$\mathrm{tr}\, F^{-1} I(E_\xi) \leq N(d-1) \tag{57}$$

We recall that a separable measurement is one that can be carried out on each particle separately, although the measurement on the different particles can be refined depending on the outcomes of partial measurements on the other particles, see [7] for a discussion. It is therefore more general than the case considered at the end of the previous subsection where the measurement on the nth particle could only depend on the measurements carried out on the n-1 previous particles.

A necessary condition for a POVM to be a separable measurement is that the POVM elements E_ξ can be decomposed into a sum of terms proportional to projectors onto unentangled states

$$\begin{aligned} E_\xi &= \sum_i |\psi_{\xi i}\rangle\langle\psi_{\xi i}| \ , \\ |\psi_{\xi i}\rangle &= |\psi^1_{\xi i}\rangle \otimes ... \otimes |\psi^N_{\xi i}\rangle \ . \end{aligned} \tag{58}$$

That this is not a sufficient condition was shown in [9].

Thus by refining the separable measurement (which increases the Fisher information) one can restrict oneself to measurements whose POVM elements are proportional to projectors onto product states

$$E_\xi = |\psi_\xi\rangle\langle\psi_\xi| = |\psi^1_\xi\rangle\langle\psi^1_\xi| \otimes ... \otimes |\psi^N_\xi\rangle\langle\psi^N_\xi| \ . \tag{59}$$

We now evaluate the l.h.s. of (57) for measurements of the form (59). First recall that the N unknown states have the form

$$\rho^N = \rho \otimes ... \otimes \rho = \sum_{k_1=1}^{d} ... \sum_{k_N=1}^{d} p_{k_1} ... p_{k_N} \, |k_1 ... k_N\rangle \, \langle k_1 ... k_N| \tag{60}$$

and the derivative of ρ^N have the form

$$\rho^N_{,i} = \rho_{,i} \otimes \rho ... \otimes \rho + ... + \rho \otimes ... \otimes \rho_{,i} = \sum_{p=1}^{N} \rho \otimes ... \rho_{,i} ... \otimes \rho \tag{61}$$

where in the second rewriting it is understood that $\rho_{,i}$ is at the p'th position in the product.

Using the product form of measurement (59), one finds that

$$\langle\psi_\xi|\rho^N|\psi_\xi\rangle = \langle\psi^1_\xi|\rho|\psi^1_\xi\rangle...\langle\psi^N_\xi|\rho|\psi^N_\xi\rangle$$
$$\langle\psi_\xi|\rho^N_{,i}|\psi_\xi\rangle = \sum_{p=1}^{N}\langle\psi^1_\xi|\rho|\psi^1_\xi\rangle...\langle\psi^p_\xi|\rho_{,i}|\psi^p_\xi\rangle...\langle\psi^N_\xi|\rho|\psi^N_\xi\rangle \tag{62}$$

Inserting these expressions into the Fisher information matrix one finds

$$\begin{aligned}I_{ij}(E_\xi) &= \sum_\xi \frac{\langle\psi_\xi|\rho^N_{,i}|\psi_\xi\rangle\langle\psi_\xi|\rho^N_{,j}|\psi_\xi\rangle}{\langle\psi_\xi|\rho^N|\psi_\xi\rangle}\\ &= \sum_\xi\sum_{p\neq p'}\langle\psi^1_\xi|\rho|\psi^1_\xi\rangle...\langle\psi^p_\xi|\rho_{,i}|\psi^p_\xi\rangle...\langle\psi^{p'}_\xi|\rho_{,j}|\psi^{p'}_\xi\rangle...\langle\psi^N_\xi|\rho|\psi^N_\xi\rangle\\ &\quad+\sum_\xi\sum_p\langle\psi^1_\xi|\rho|\psi^1_\xi\rangle...\frac{\langle\psi^p_\xi|\rho_{,i}|\psi^p_\xi\rangle\langle\psi^p_\xi|\rho_{,j}|\psi^p_\xi\rangle}{\langle\psi^p_\xi|\rho|\psi^p_\xi\rangle}...\langle\psi^N_\xi|\rho|\psi^N_\xi\rangle\\ &= \sum_\xi\sum_p\langle\psi^1_\xi|\rho|\psi^1_\xi\rangle...\frac{\langle\psi^p_\xi|\rho_{,i}|\psi^p_\xi\rangle\langle\psi^p_\xi|\rho_{,j}|\psi^p_\xi\rangle}{\langle\psi^p_\xi|\rho|\psi^p_\xi\rangle}...\langle\psi^N_\xi|\rho|\psi^N_\xi\rangle\end{aligned} \tag{63}$$

Where we have used the fact that the first term in the second equality vanishes. Indeed it is equal to

$$\sum_\xi\sum_{p\neq p'}\langle\psi_\xi|\rho...\rho_{,i}...\rho_{,j}...\rho|\psi_\xi\rangle \tag{64}$$

The sum over ξ can be carried out in (64) to yield the identity matrix and the resulting trace vanishes because $\operatorname{tr}\rho...\rho_{,i}...\rho_{,j}...\rho = 0$.

We now insert (63) into $\operatorname{tr}F^{-1}I(E_\xi)$. All the operations from (52) to (56) can be carried out exactly as in the previous subsection, and one arrives at the expression

$$\begin{aligned}\operatorname{tr}F^{-1}I(E_\xi) &= \sum_p\sum_\xi\langle\psi_\xi|\rho\otimes...\otimes(I-\rho)\otimes...\otimes\rho|\psi_\xi\rangle\\ &= N(d-1)\end{aligned} \tag{65}$$

which is the sought for relation.

F. Inequality for more then one mixed state

We now provide a counterexample showing that if one carries out a collective measurement on $N>1$ mixed states one can violate (26). We take $N=2$, and suppose the unknown states belong to a 2 dimensional Hilbert space. $\rho(\theta)=\frac{I}{2}+\sum_i\theta_i\sigma_i$. We take as reference point $\theta_i=0$ corresponding to $\rho=\frac{I}{2}$. At this point $F_{ij}(\theta_i=0)=\delta_{ij}$.

We consider as measurement on the two copies the following POVM

$$\begin{aligned}\{\quad &\tfrac{1}{2}|\uparrow_x\uparrow_x\rangle\langle\uparrow_x\uparrow_x| \quad,\quad \tfrac{1}{2}|\downarrow_x\downarrow_x\rangle\langle\downarrow_x\downarrow_x| \quad,\quad \tfrac{1}{2}|\uparrow_y\uparrow_y\rangle\langle\uparrow_y\uparrow_y| \quad,\quad \tfrac{1}{2}|\downarrow_y\downarrow_y\rangle\langle\downarrow_y\downarrow_y| \quad,\\ &\tfrac{1}{2}|\uparrow_z\uparrow_z\rangle\langle\uparrow_z\uparrow_z| \quad,\quad \tfrac{1}{2}|\downarrow_z\downarrow_z\rangle\langle\downarrow_z\downarrow_z| \quad,\quad \tfrac{1}{2}|\uparrow_z\downarrow_z-\downarrow_z\uparrow_z\rangle\langle\uparrow_z\downarrow_z-\downarrow_z\uparrow_z| \quad\}\ .\end{aligned} \tag{66}$$

This POVM cannot be realized by separate measurements on each particle because of the last term that projects onto an entangled state.

For this POVM one calculates that $I_{ij}(E_\xi,\theta_i=0)=\delta_{ij}$. Hence the left hand side of (26) evaluates to $\sum_{ij}F^{-1}_{ij}(\theta_i=0)I_{ij}(E_\xi,\theta_i=0)=3>N(d-1)=2$.

This proves that the quantum Fisher information is non additive.

G. Comparison with other Quantum Cramér-Rao bounds

An important question raised by the bound (26) raises how it compares to other quantum Cramér-Rao bounds obtained in the literature.

Our most important result is that (26) is both a necessary and sufficient condition that $I(E_\xi,\theta)$ must satisfy when the dimensionality of the system d equals 2 and the state is pure. This will be proven and discussed in detail in section VII.

When $d>2$ (26) is not a sufficient condition that $I(E_\xi,\theta)$ must satisfy. To see this let us compare (26) with the bound derived by Helstrom based on the s.l.d. This bound is the matrix inequality $I^N(E_\xi,\theta)\leq NF(\theta)$, see (25).

The comparison is most easily carried out by defining the matrix $H=\frac{1}{N}F^{-\frac{1}{2}}I^N F^{-\frac{1}{2}}=\sum_{i=1}^p \gamma_i h_i\otimes h_i$ where γ_i are the eigenvalues of H and h_i its eigenvectors. Helstrom's bound can be reexpressed as $\gamma_i\leq 1$ whereas the bound (26) states that $\sum_i\gamma_i\leq d-1$. From these expressions it results that the bound (26) is always better then Helstrom's bound for $d=2$. For $d>2$ and $p\leq d-1$ Helstrom's bound is better then (26) as is seen by summing the inequalities $\gamma_i\leq 1$ to obtain $\sum_i\gamma_i\leq p$. For $p>d-1$, Helstrom's bound and the bound (26) are inequivalent.

Yuen and Lax have proposed another matrix bound for F based on an asymmetric logarithmic derivative (a.l.d.). The bound based on the a.l.d. is known to be worse then the bound based on the s.l.d. in the case of one parameter, but it can be better, for some loss functions, in the case of two or more parameters. We have however not been able to make a detailed comparison between the bound based on the a.l.d. and (26).

Although when $d>2$, the bound (26) is not a sufficient condition it can be complemented by additional constraints based on partial traces of $F^{-1}I^N(E_\xi,\theta)$ which we now exhibit.

Consider a subset $i=1,..,p'$ ($p'<p$) of the parameters. Let $\rho_{,i'}$ be the corresponding derivatives of $\rho(\theta^i)$. Let us define the effective dimension d' of the space in which these parameters act at the point θ_0 as follows. Let Π be a projector that commutes with $\rho(\theta_0)$ ($[\Pi,\rho(\theta_0)]=0$) and such that $\rho_{,i'}$, $i'=1,...,p'$ acts only within the eigenspace of Π (that is $\Pi\rho_{,i'}\Pi=\rho_{,i'}$). Then d' is the smallest dimension of the eigenspace of such a projector Π ($d'=\mathrm{tr}\,\Pi$). To be more explicit, let us reexpress the definition of d' in coordinates. First we diagonalize $\rho(\theta_0)=\sum_k p_k|k\rangle\langle k|$. If some p_k are equal this can be done in many ways. The projector Π projects onto some of the eigenvectors of ρ: $\Pi=\sum_{k=1}^{d'}|k\rangle\langle k|$. Next we write the operators $\rho_{,i'}$ in this basis: $\rho_{,i'}=\sum_{k,l=1}^{d'}(\rho_{,i'})_{kl}|k\rangle\langle l|$ where the fact that the indices k,l go from one to d' expresses the fact that $\rho_{,i'}$ acts only within the eigenspace of Π. Finally we choose the smallest such d'.

We will show that

$$\sum_{i',j'=1}^{p'}F^{-1}_{i'j'}I^N_{i'j'}(E_\xi,\theta_0)\leq N(d'-1)\ . \tag{67}$$

Before proving this result let us illustrate it by an example. Consider an unknown pure state in d dimensions. In the neighborhood of a particular point we can parameterize the state by

$$\psi=|1\rangle+(\theta^2+i\eta^2)\,|2\rangle+...+(\theta^d+i\eta^d)\,|d\rangle \tag{68}$$

where the unknown parameters are θ^i and η^i, $i=2,..,d$. There are thus $2d-2$ parameters. At the point $\theta=\eta=0$, F is diagonal in this parameterization: $F_{\theta^i\theta^j}=\delta_{ij}$, $F_{\eta^i\eta^j}=\delta_{ij}$, $F_{\theta^i\eta^j}=0$. Hence (26) takes the form

$$\sum_i I^N_{\theta^i\theta^i}(E_\xi,\theta=\eta=0)+I^N_{\eta^i\eta^i}(E_\xi,\theta=\eta=0)\leq N(d-1). \tag{69}$$

But using (67) we also find the constraints

$$I^N_{\theta^i\theta^i}(E_\xi,\theta=\eta=0)+I^N_{\eta^i\eta^i}(E_\xi,\theta=\eta=0)\leq N\quad ,\quad i=2,..,d \tag{70}$$

which are stronger then (69) since they must hold separately, but by summing them one obtains (69).

The proof of equation (67) proceeds as in section V. First we can restrict ourselves to POVM's whose elements are proportional to one dimensional projectors. Second we can restrict ourselves to the subspace Π in evaluating (67). This follows from the inequality

$$\begin{aligned}
I(E_\xi)_{i'j'} &= \sum_\xi \frac{\mathrm{tr}\,(\rho_{,i'}E_\xi)\mathrm{tr}\,(\rho_{,j'}E_\xi)}{\mathrm{tr}\,(\rho E_\xi)}\\
&= \sum_\xi \frac{\mathrm{tr}\,(\rho_{,i'}\Pi E_\xi\Pi)\mathrm{tr}\,(\rho_{,j'}\Pi E_\xi\Pi)}{\mathrm{tr}\,(\rho\Pi E_\xi\Pi)+\mathrm{tr}\,(\rho(1-\Pi)E_\xi(1-\Pi))}\\
&\leq \sum_\xi \frac{\mathrm{tr}\,(\rho_{,i'}\Pi E_\xi\Pi)\mathrm{tr}\,(\rho_{,j'}\Pi E_\xi\Pi)}{\mathrm{tr}\,(\rho\Pi E_\xi\Pi)}.
\end{aligned} \tag{71}$$

Note that equality in (71) holds when the measurement consists of one dimensional projectors and when the POVM decomposes into the sum of two POVM's acting on the subspaces spanned by Π and $1-\Pi$ separately (i.e., the POVM elements $E_\xi = |\psi_\xi\rangle\langle\psi_\xi|$ must commute with Π and $1-\Pi$). Third we can increase the number of parameters from p' to d'^2-1. We then introduce exactly as in (48) a parameterization in which the $\rho_{,i}$ are particularly simple, but in in place of (55) we use

$$\sum_{1\le m'\le d'} c_{m'k'}c_{m'l'} = \delta_{k'l'}p_{k'} - \frac{p_{k'}p_{l'}}{\mathrm{tr}\,(\Pi\rho)}. \tag{72}$$

After these preliminary steps the l.h.s. of (67) is calculated exactly as in subsections V B, V C, V D.

VI. DROPPING THE CONDITION OF UNBIASED ESTIMATORS

A. Quantum van Trees inequality

In the previous section we proved a bound on the variance of unbiased estimators $\hat\theta_N$ of N copies of the quantum system $\rho(\theta)$ (with the additional condition that if ρ is mixed the measurement should be separable). In this section we shall prove Theorems IV and V that state that under additional conditions it is possible to drop the hypothesis that the estimator is unbiased.

The starting point for the results in this section is a Bayesian form of the Cramér-Rao inequality, the van Trees inequality [11], and in particular the multivariate form of the van Trees inequality proven in [12]. Adapted to the problem of estimating a quantum state, this inequality takes the following form. Let $\hat\theta_N$ be an arbitrary estimator of the parameter θ based on a measurement E_ξ of the system $\rho^N(\theta)$. Suppose it has mean quadratic error matrix $V^N(\theta)$, and Fisher information matrix $I^N(E_\xi,\theta)$. Let $\lambda(\theta)$ be a smooth density supported on a compact region (with smooth boundary) of the parameter space, and suppose λ vanishes on the boundary. By E_λ we denote expectation over a random parameter value Θ with the probability density $\lambda(\theta)$. Let $C(\theta)$ and $D(\theta)$ be two $p\times p$ matrix valued functions of θ, the former being symmetric and positive definite. Then the multivariate van Trees inequality reads

$$\mathrm{E}_\lambda \mathrm{tr}\, C(\Theta)V^N(\Theta) \ge \frac{(\mathrm{E}_\lambda \mathrm{tr}\, D(\Theta))^2}{\mathrm{E}_\lambda \mathrm{tr}\, C(\Theta)^{-1}D(\Theta)I^N(E_\xi,\Theta)D(\Theta)^\top + \tilde{\mathcal{I}}(\lambda)} \tag{73}$$

where

$$\tilde{\mathcal{I}}(\lambda) = \int d\theta \frac{1}{\lambda(\theta)} \sum_{ijkl} C_{ij}(\theta)^{-1}\partial_{\theta_k}\{D_{ik}(\theta)\lambda(\theta)\}\partial_{\theta_l}\{D_{jl}(\theta)\lambda(\theta)\}. \tag{74}$$

As a first application of this inequality we shall bound the minimum value averaged over θ of a quadratic cost function. Let $C(\theta)$ be the quadratic cost function. Consider the matrix $W_{opt}(\theta)$ that minimizes for each value of θ the cost $\mathrm{tr}\, C(\theta)W(\theta)$ under the condition that $\mathrm{tr}\, F(\theta)^{-1}W(\theta)^{-1} \le d-1$. One easily finds that

$$W_{opt} = \frac{\mathrm{tr}\sqrt{F^{-1/2}CF^{-1/2}}}{d-1}F^{-1/2}\sqrt{F^{1/2}C^{-1}F^{1/2}}F^{-1/2} \tag{75}$$

$$= \frac{\mathrm{tr}\sqrt{C^{1/2}F^{-1}C^{1/2}}}{d-1}C^{-1/2}\sqrt{C^{1/2}F^{-1}C^{1/2}}C^{-1/2} \tag{76}$$

and that

$$\mathrm{tr}\, CW_{opt} = \frac{\left(\mathrm{tr}\sqrt{F^{-1/2}CF^{-1/2}}\right)^2}{d-1} = \frac{\left(\mathrm{tr}\sqrt{C^{1/2}F^{-1}C^{1/2}}\right)^2}{d-1}\,. \tag{77}$$

We choose in (73) $D(\theta) = C(\theta)W_{opt}(\theta)$. Thus $\mathrm{tr}\, D(\theta) = \mathrm{tr}\, C(\theta)W_{opt}(\theta)$ is given by (77). Note that

$$D(\theta)^\top C(\theta)^{-1}D(\theta) = W_{opt}(\theta)C(\theta)W_{opt}(\theta) = \frac{\mathrm{tr}\, C(\theta)W_{opt}(\theta)}{d-1}F(\theta)^{-1}. \tag{78}$$

Thus

$$\operatorname{tr} D(\theta)^\top C(\theta)^{-1} D(\theta) I^N(E_\xi,\theta) = \frac{\operatorname{tr} C(\theta) W_{opt}(\theta)}{d-1} \operatorname{tr} F(\theta)^{-1} I^N(E_\xi,\theta)$$
$$\leq N \operatorname{tr} C(\theta) W_{opt}(\theta)\ . \tag{79}$$

Inserting these expressions into (73) one obtains

$$\begin{aligned} \mathrm{E}_\lambda \operatorname{tr} C(\Theta) V^N(\Theta) &\geq \frac{(\mathrm{E}_\lambda \operatorname{tr} C(\Theta) W_{opt}(\Theta))^2}{N \mathrm{E}_\lambda \operatorname{tr} C(\Theta) W_{opt}(\Theta) + \tilde{\mathcal{I}}(\lambda)} \\ &\geq \frac{\mathrm{E}_\lambda \operatorname{tr} C(\Theta) W_{opt}(\Theta)^2}{N} - \frac{\alpha}{N^2} \end{aligned} \tag{80}$$

where

$$\alpha = \frac{\tilde{\mathcal{I}}(\lambda)}{\mathrm{E}_\lambda \operatorname{tr} C(\Theta) W_{opt}(\Theta)} \tag{81}$$

is independent of N. This proves that upon averaging over θ it is impossible (for large N) to improve over the minimum cost (29).

B. Asymptotic version of the Cramér-Rao inequality

We now prove an asymptotic version (27) of our main inequality which does not make the assumption of unbiased estimators. We must however slightly restrict the class of competing estimators since otherwise by the phenomenon of super-efficiency we can beat a given estimator at any specific value of the parameter, though we pay for this by bad behavior closer and closer to the chosen value as N becomes larger.

The restriction on the class of estimators is that N times their mean quadratic error matrix must converge uniformly in a neighborhood of the true value θ_0 of θ to a limit $W(\theta)$, continuous at θ_0. We assume that both $W(\theta_0)$ and $F(\theta_0)$ are nonsingular. Furthermore we shall require some mild smoothness conditions on $F(\theta)$ in a neighborhood of θ_0: it must be continuous at θ_0 with bounded partial derivatives with respect to the parameter in a neighborhood of θ_0.

Suppose that as $N \to \infty$:

$$N V^N(\theta) \to W(\theta)$$

uniformly in θ in a neighborhood of θ_0, with W continuous at θ_0; write $W_0 = W(\theta_0)$. Now in (73) let us make the following choices for the matrix functions C and D:

$$C(\theta) = W_0^{-1} F^{-1}(\theta) W_0^{-1},$$

$$D(\theta) = W_0^{-1} F^{-1}(\theta).$$

Then (73) (multiplied throughout by N) and (74) become

$$\begin{aligned} \mathrm{E}_\lambda \operatorname{tr} W_0^{-1} F^{-1}(\Theta) W_0^{-1} N V^N(\Theta) &\geq \frac{(\mathrm{E}_\lambda \operatorname{tr} W_0^{-1} F^{-1}(\Theta))^2}{\frac{1}{N} \mathrm{E}_\lambda \operatorname{tr} F^{-1} I^N(E_\xi, \Theta) + \frac{1}{N} \tilde{\mathcal{I}}(\lambda)} \\ &\geq \frac{(\mathrm{E}_\lambda \operatorname{tr} W_0^{-1} F^{-1}(\Theta))^2}{(d-1) + \frac{1}{N} \tilde{\mathcal{I}}(\lambda)} \end{aligned} \tag{82}$$

and

$$\tilde{\mathcal{I}}(\lambda) = \int d\theta \frac{1}{\lambda(\theta)} \sum_{ijkl} F_{ij}(\theta) \partial_{\theta_k} \{F_{ik}^{-1}(\theta)\lambda(\theta)\} \partial_{\theta_l} \{F_{jl}^{-1}(\theta)\lambda(\theta)\}, \tag{83}$$

where we have used our central inequality (26) to pass to (82). Now suppose that the quantity (83) is finite (we will give conditions for that in a moment). By the assumed uniform convergence of NV^N to W, letting $N \to \infty$ (82) becomes

$$\mathrm{E}_\lambda \mathrm{tr}\, W_0^{-1} F^{-1}(\Theta) W_0^{-1} W(\Theta) \geq \frac{(\mathrm{E}_\lambda \mathrm{tr}\, W_0^{-1} F^{-1}(\Theta))^2}{(d-1)}. \tag{84}$$

Now suppose the density λ in this equation (the probability density of Θ) is replaced by an element λ^m in a sequence of densities, concentrating on smaller and smaller neighborhoods of θ_0 as $m \to \infty$. Assume that $F(\theta)$ is continuous at θ_0. Recall our earlier assumption that $W(\theta)$ is also continuous at θ_0, with $W_0 = W(\theta_0)$. Then taking the limit as $m \to \infty$ of (84) yields

$$\mathrm{tr}\, W^{-1}(\theta_0) F^{-1}(\theta_0) \geq (\mathrm{tr}\, W^{-1}(\theta_0) F^{-1}(\theta_0))^2/(d-1)$$

or the required limiting form of (26),

$$\mathrm{tr}\, W^{-1}(\theta_0) F^{-1}(\theta_0) \leq (d-1).$$

It remains to discuss whether it was reasonable to assume that $\tilde{\mathcal{I}}(\lambda^m)$ is finite (for each m separately). Note that this quantity only depends on the prior density λ and on $F(\theta)$, where λ is one of a sequence of densities supported by smaller and smaller neighborhoods of θ_0. We already assumed that $F(\theta)$ was continuous at θ_0. It is certainly possible to specify prior densities λ^m concentrating on the ball of radius $1/m$, say, satisfying the smoothness assumptions in [12] and with, for each m, finite Fisher information matrix

$$\int d\theta \frac{1}{\lambda^m(\theta)} \partial_{\theta_k}\{\lambda^m(\theta)\} \partial_{\theta_l}\{\lambda^m(\theta)\}.$$

Consideration of (83) then shows that it suffices further just to assume that $\partial_{\theta_k}\{F^{-1}_{ik}(\theta)\}$ is, for each i, k, bounded in a neighborhood of θ_0.

In conclusion we have shown that under mild smoothness conditions on $F(\theta)$, the limiting mean quadratic error matrix W of a sufficiently regular but otherwise arbitrary sequence of estimators must satisfy the asymptotic version of our central inequality $\mathrm{tr}\, F^{-1} W^{-1} \leq d-1$. The existence of conditions on F is very natural. Indeed they imply that θ are smooth parameters in Hilbert space.

VII. ATTAINING THE CRAMÉR-RAO BOUND IN 2 DIMENSIONS

We shall now show that the bounds (26), (28) are sharp in the case of pure states of spin 1/2 systems and of separable measurements in the case of mixed states of spin 1/2 systems. In particular, in the limit of a large number of copies N any target mean quadratic error matrix W that satisfies $\mathrm{tr}\, F^{-1} W^{-1} \leq 1$ can be attained (provided W is non singular). We shall show this by explicitly constructing a measurement strategy that attains the bound. In section VI we have already shown that if $\mathrm{tr}\, F^{-1} W^{-1} > 1$, then it cannot be attained.

A. Attaining the bound at a fixed point θ_0

The first step in the proof is to consider the case of one copy of the unknown state ($N = 1$) and fix a particular point θ_0. Then we show that for any target information matrix $G(\theta_0)$ that satisfies $\mathrm{tr}\, F^{-1}(\theta_0) G(\theta_0) \leq 1$, we can build a measurement $E_\xi = E_\xi^{\theta_0}$, *in general depending on* θ_0, such that $I(E_\xi^{\theta_0}, \theta_0) = G(\theta_0)$. In the next sections we shall show how to use this intermediate result to build a measurement and estimation strategy whose asymptotic mean quadratic error is equal to $W(\theta) = G(\theta)^{-1}$ for *all* θ.

Let us first consider the case of pure states. At θ_0, the state is $|\psi_0\rangle$. We introduce a parameterization θ^1, θ^2 such that in the vicinity of $|\psi_0\rangle$, the unknown state is

$$|\psi_0\rangle = |\psi_0\rangle + (\theta^1 + i\theta^2)|\psi_0^\perp\rangle \ . \tag{85}$$

Thus in this parameterization, the point θ_0 corresponds to $\theta^1 = \theta^2 = 0$. In this parameterization, F is proportional to the identity at $\theta^1 = \eta = 0$: $F_{\theta^1\theta^1}(0) = F_{\theta^2\theta^2}(0) = 1$, $F_{\theta^1\theta^2}(0) = 0$.

We now diagonalize the matrix G. Thus there exist new parameters $\theta'^1 = \cos\lambda\theta^1 + \sin\lambda\theta^2$, $\theta'^2 = -\sin\lambda\theta^1 + \cos\lambda\theta^2$ such that $G_{\theta'^1\theta'^1}(0) = g_1 > 0$, $G_{\theta'^2\theta'^2}(0) = g_2 > 0$, $G_{\theta'^1\theta'^2}(0) = 0$.

In terms of the parameters θ'^1, θ'^2, the unknown state is written

$$|\psi_0\rangle = |\psi_0\rangle + (\theta'^1 + i\theta'^2)|\psi_0^{\perp'}\rangle \tag{86}$$

where $|\psi_0^{\perp'}\rangle = e^{i\lambda}|\psi_0^{\perp}\rangle$.

The POVM $E_\xi^{\theta_0}$ consists of measuring the observable $|\psi_0\rangle\langle\psi_0^{\perp'}| - |\psi_0^{\perp'}\rangle\langle\psi_0^{\perp}|$ with probability g_1, of measuring the observable $i(|\psi_0\rangle\langle\psi_0^{\perp'}| - |\psi_0^{\perp'}\rangle\langle\psi_0^{\perp}|)$ with probability g_2, and of measuring nothing (or measuring the identity) with probability $1-g_1-g_2$. It is straightforward to verify that the Fisher information obtained by carrying out the POVM $E_\xi^{\theta_0}$ is equal to $G(\theta_0)$.

Let us now turn to the case of impure states. We suppose that there are three unknown parameters. We use a parameterization in which $\rho(\theta) = (1/2)(I + \theta\cdot\sigma)$, with $\|\theta\| < 1$. Without loss of generality we can suppose that $\theta_0 = (0,0,n)$, so that $\rho(\theta_0) = (1/2+n/2)\,|1\rangle\,\langle 1| + (1/2-n/2)\,|2\rangle\,\langle 2| = \frac{1}{2}(I+n\sigma_z)$. The tangent space at ρ is spanned by the Pauli matrices $\rho_{,x} = \sigma_x (= \frac{\rho_{,12+}}{2})$, $\rho_{,y} = \sigma_y (= \frac{\rho_{,12-}}{2})$, $\rho_{,z} = \sigma_z (= \rho_{,1}\sqrt{1-n^2})$ where in parenthesis we have given the relation to the basis used in section V D. In this coordinate system $F(\theta_0)$ is diagonal with eigenvalues 1, 1, $1/(1-n^2)$.

Take any symmetric positive matrix G_{ij} satisfying $\operatorname{tr} GF^{-1}(\theta_0) \leq 1$. Define the matrix $H = F^{-\frac{1}{2}}GF^{-\frac{1}{2}} = \sum_i \gamma_i h_i \otimes h_i$, where γ_i and h_i are the eigenvalues and eigenvectors of H. The condition $\operatorname{tr} GF^{-1}(\theta_0) \leq 1$ can then be rewritten $\sum_i \gamma_i \leq 1$. If we define $g_i = F^{\frac{1}{2}} h_i$, then we can write $G = \sum_i \gamma_i g_i \otimes g_i$. Denote $m_i = g_i/\|g_i\|$.

Consider the measurement of the spin along the direction m_i. This is the POVM consisting of the two projectors $P_{+m_i} = \frac{1}{2}(I + m_i.\sigma)$ and $P_{-m_i} = \frac{1}{2}(I - m_i.\sigma)$. The information matrix for this measurement is

$$I(P_{\pm m_i})_{kl} = \sum_{\pm} \frac{\operatorname{tr}(P_{\pm m_i}\sigma_k)\operatorname{tr}(P_{\pm m_i}\sigma_l)}{\operatorname{tr}(P_{\pm m}\rho)} = \frac{m_{ik}m_{il}}{(1-n^2 m_{iz}^2)}. \tag{87}$$

Therefore this information matrix is proportional to $g_i \otimes g_i$. One verifies that it obeys $\operatorname{tr} F^{-1} I(P_{\pm m_i}) = 1$, as it must by our findings in section V since the measurement is exhaustive, $N = 1$, and $p = d^2 - 1$. Therefore

$$I(P_{\pm m_i}) = g_i \otimes g_i\ . \tag{88}$$

We now combine such POVM's to obtain the POVM whose elements are

$$E_\xi = \{\gamma_1 P_{+m^1}, \gamma_1 P_{-m^1}, \gamma_2 P_{+m^2}, \gamma_2 P_{-m^2}, \gamma_3 P_{+m^3}, \gamma_3 P_{-m^3}, (1-\gamma_1-\gamma_2-\gamma_3)I\}\ . \tag{89}$$

The information matrix for this measurement is just the sum $I(E_\xi) = \gamma_1 I(P_{\pm m^1}) + \gamma_2 I(P_{\pm m^2}) + \gamma_3 I(P_{\pm m^3}) = \sum_i \gamma_i g_i \otimes g_i = G$. Thus the POVM E_ξ we have constructed attains the target information G at the point θ_0.

B. Attaining the bound for every θ and arbitrary N by separable measurements

We now prove Theorem VII that states that we can attain the bound (28) for every θ. Give yourself a continuous matrix $W(\theta)$, the target mean quadratic error matrix, satisfying (28) for every θ. Define $G(\theta) = W(\theta)^{-1}$, the target information matrix, which satisfies therefore (26). We will show that there exists a separable measurement and an estimation strategy on N copies of the state $\rho(\theta)$ such that the mean quadratic error matrix of the estimator satisfies

$$\mathrm{mqe}_{\hat\theta}(\theta)_{ij} = \mathrm{E}_\theta((\hat\theta^i - \theta^i)(\hat\theta^j - \theta^j)) = \frac{W_{ij}(\theta)}{N} + o\left(\frac{1}{N}\right) \tag{90}$$

for all θ. In fact this holds uniformly in θ in a sufficiently small neighborhood of any given point. This is proven by constructing explicitly a measurement and estimation strategy that satisfies (90), following the lines of [5].

The measurement and estimation strategy we propose is the following: first take a fraction $N_0 = O(N^a)$ of the states, for some fixed $0 < a < 1$, and on 1/3 of them measure σ_x, on one third σ_y and on one third σ_z. One obtains from each measurement of σ_x the outcome ± 1 with probabilities $\frac{1}{2}(1 \pm \theta_x)$, and similarly for σ_y, σ_z. Using this data we make a first estimate of θ, call it $\tilde\theta$, for instance by equating the observed relative frequencies of ± 1 in the three kinds of measurement to their theoretical values. If the state is pure this determines a first estimate of the direction of polarization. If the state is mixed it is possible that the initial estimate suggests that the Bloch vector lies outside the unit sphere. This only occurs with exponentially small probability (in N_0) and if this is the case the measurement is discarded. As discussed below this only affects the mean quadratic error by $o(1/N)$.

On the remaining $N' = N - N_0$ states we carry out the measurement $E_\xi = E_\xi^{\tilde\theta}$ such that $I(E_\xi, \tilde\theta) = G(\tilde\theta)$ which we have just shown how to construct. Note that $I(E_\xi, \tilde\theta) = G(\tilde\theta)$ only when the true value of θ is precisely equal to $\tilde\theta$.

Write $I(E_\xi, \theta; \tilde\theta)$ for the Fisher information about θ, based on the measurement $E_\xi = E_\xi^{\tilde\theta}$ optimal at $\tilde\theta$, while the true value of the parameter is actually θ. Given $\tilde\theta$, each of the N' second stage measurements represents one draw from the probability distribution $p(\xi|\theta;\tilde\theta) = \mathrm{tr}\, E_\xi^{\tilde\theta}\rho(\theta)$. We use the classical m.l.e. based on this data only (with $\tilde\theta$ fixed at its observed value) to estimate what is the value of θ. Call this estimated value $\hat\theta$.

Let $\epsilon > 0$ be fixed, arbitrarily small. Let θ_0 denote the true value of θ. For given $\delta > 0$ let $B(\theta_0, \delta)$ denote the ball of radius δ about θ_0. Fix a convenient matrix norm $\|\cdot\|$. We have the exponential bound

$$\Pr\{\tilde\theta \in B(\theta_0, \delta) \geq 1 - Ce^{-DN_0\delta_0^2} \tag{91}$$

for some positive numbers C and D (depending on δ). The reason we take N_0 proportional to N^a for some $0 < a < 1$ is that this ensures that $1 - Ce^{-DN_0} = o(1/N)$.

Modern results [14] on the m.l.e. $\hat\theta$ state that, under certain regularity conditions, conditional on $\tilde\theta$ the mean quadratic error matrix $\mathrm{mqe}_{\hat\theta}(\theta;\tilde\theta)_{ij} = \mathrm{E}_\theta(\,(\hat\theta^i - \theta^i)(\hat\theta^j - \theta^j)\,|\,\tilde\theta\,)$ satisfies

$$\mathrm{mqe}_{\hat\theta}(\theta_0;\tilde\theta) = \frac{I(E_\xi,\theta_0;\tilde\theta)^{-1}}{N'} + o\left(\frac{1}{N'}\right) \tag{92}$$

uniformly in θ_0. We need however for the next step in our argument that this same result is true uniformly in $\tilde\theta$ for given θ_0. This could be verified by careful reworking of the proof in [14]. Rather than doing that, we will explicitly calculate in subsection VII C the mean quadratic error matrix of our estimator and show that conditional on $\tilde\theta$ it satisfies (92) uniformly in $\tilde\theta$ in a small enough neighborhood $B(\theta_0, \delta)$ of θ_0. The 'little o' in (92) refers to the chosen matrix norm.

We will also need that $I(E_\xi, \theta_0; \tilde\theta)^{-1}$ is continuous in $\tilde\theta$ at $\tilde\theta = \theta_0$, at which point it equals by our construction the target mean quadratic error $W(\theta_0)$. This is also established in subsection VII C. Therefore, replacing if necessary δ by a smaller value, we can guarantee that $I(E_\xi, \theta_0; \tilde\theta)^{-1}$ is within ϵ of $I(E_\xi, \theta_0; \theta_0)^{-1} = W(\theta_0)$ for all $\tilde\theta \in B(\theta_0, \delta)$.

If $\tilde\theta$ is outside the domain $B(\theta_0, \delta)$, then the norm of $\mathrm{mqe}_{\hat\theta}(\theta;\tilde\theta)$ is bounded by a constant A since θ belongs to a compact domain.

Putting everything together we find that

$$\begin{aligned}
\|N'\mathrm{mqe}_{\hat\theta}(\theta_0) - W(\theta_0)\| &= \left\| \int \left(N'\mathrm{mqe}_{\hat\theta}(\theta_0;\tilde\theta) - W(\theta_0) \right) dP(\tilde\theta) \right\| \\
&\leq \int_{B(\theta_0,\delta)} \|N'\mathrm{mqe}_{\hat\theta}(\theta_0;\tilde\theta) - W(\theta_0)\| dP(\tilde\theta) + AN'C'e^{-DN_0} \\
&= \int_{B(\theta_0,\delta)} \|I(E_\xi,\theta_0;\tilde\theta)^{-1} + o(1) - W(\theta_0)\| dP(\tilde\theta) + o(1) \\
&\leq \epsilon + o(1) + o(1).
\end{aligned}$$

It follows since $N'/N \to 1$ as $N \to \infty$ that

$$\limsup_{N\to\infty} \|N\mathrm{mqe}_{\hat\theta}(\theta_0) - W(\theta_0)\| \leq \epsilon.$$

Since ϵ was arbitrary, we obtain (90).

C. Analysis of the conditional mean quadratic error

We first consider the case of impure states, with the parameterization

$$\rho = \frac{1}{2}(I + \theta.\sigma), \quad \text{with } \sum(\theta^i)^2 < 1. \tag{93}$$

where we have imposed that the state is never be pure. This case turns out to allow the most explicit and straightforward analysis because the relation between the frequency of the outcomes and the parameters θ is linear. For other cases the analysis is more delicate and is discussed in the next subsection. In general, smoothness assumptions will have to be made on the parameterization $\rho = \rho(\theta)$.

We suppose that $W(\theta)$ is non-singular and continuous in θ. Consequently the γ_i (defined in section VII A) depend continuously on θ and are all strictly positive at the true value θ_0 of θ.

Given the initial estimate, the second stage measurement can be implemented as follows: for each of the $N' = N - N_0$ observations, independently of one another, with probability γ_i measure the projectors $P_{\pm m^i}$, in other words, measure the spin observable $m^i.\sigma$. With probability $1 - \sum \gamma_i$ do nothing.

We emphasize that the γ_i and m^i all depend on the initial estimate $\tilde{\theta}$ through $W(\tilde{\theta})$ and $F(\tilde{\theta})$. In the following, all probability calculations are conditional on a given value of $\tilde{\theta}$.

For simplicity we will modify the procedure in the following two ways: firstly, rather than taking a random number of each of the three types of measurement, we will take the fixed (expected) numbers $\lfloor \gamma_i N' \rfloor$ (and neglect the difference between $\lfloor \gamma_i N' \rfloor$ and $\gamma_i N'$). Secondly, we will ignore the constraint $\sum (\theta^i)^2 \leq 1$. These two modifications make the maximum likelihood estimator easier to analyze, but do not change its asymptotic mean quadratic error. Later we will sketch how to extend the calculations to the original *constrained* maximum likelihood estimator based on *random* numbers of measurements of each observable.

Now measuring $m^i.\sigma$ produces the values ± 1 with probabilities $p_{\pm i} = \frac{1}{2}(1 \pm \theta.m^i)$. Since our data consists of three binomially distributed counts and we have three parameters $\theta^1, \theta^2, \theta^3$ the maximum likelihood estimator can be described, using the invariance of maximum likelihood estimators under 1–1 reparameterization, as follows: set the theoretical values $p_{\pm i}$ equal to their empirical counterparts (relative frequencies of ± 1 in the $\gamma_i N'$ observations of the i'th spin) and solve the resulting three equations in three unknowns. Define $\eta_i = 2p_{+i} - 1 = \theta.m^i$ and let $\hat{\eta}_i$ be its empirical counterpart. Recall that $m^i = g^i/\|g^i\|$, $g^i = F^{1/2}h^i$, where the h^i are the orthonormal eigenvectors of $F^{-1/2}GF^{-1/2}$, and where F and G are $F(\tilde{\theta})$, $G(\tilde{\theta})$, and $\tilde{\theta}$ is the preliminary estimate of θ.

Then we can rewrite

$$\eta_i = \theta.m^i = \theta.g^i/\|g^i\| = \theta.F^{1/2}h^i/\|F^{1/2}h^i\| = (F^{1/2}\theta).h^i/\|F^{1/2}h^i\|$$

from which we obtain

$$(F^{1/2}\theta).h^i = \|F^{1/2}h^i\|\eta_i$$

and hence

$$\theta = F^{-1/2}\sum_i \|F^{1/2}h^i\|\eta_i h^i.$$

The same relation holds between $\hat{\theta}$ and $\hat{\eta}$. The $\hat{\eta}_i$ are independent with variance $4p_{+m_i}p_{-m_i}/(\gamma_i N') = (1 - (\theta.m^i)^2)/(\gamma_i N')$. Thus the mean quadratic error matrix of $\hat{\theta}$, conditional on the preliminary estimate $\tilde{\theta}$, is

$$\mathrm{mqe}_{\hat{\theta}}(\theta_0;\tilde{\theta}) = \frac{1}{N'}\sum_i \frac{1}{\gamma_i}\left(1 - \frac{(\theta_0.F^{1/2}h^i)^2}{\|F^{1/2}h^i\|^2}\right)\|F^{1/2}h^i\|^2 F^{-1/2}(h^i \otimes h^i)F^{-1/2}. \tag{94}$$

There is no $o(1/N')$ term here so we do not have to check uniform convergence: the limiting value is attained exactly. Actually we cheated by replacing $\lfloor \gamma_i N' \rfloor$ by $\gamma_i N'$. This does introduce a $o(1/N')$ error into (94) uniformly in a neighborhood of θ_0 in which the γ_i, which depend on $\tilde{\theta}$, are bounded away from zero, and F and its inverse are bounded.

One may verify that (94) reduces to $W(\theta_0)/N'$ at $\tilde{\theta} = \theta_0$ (indeed at $\theta_0 = \tilde{\theta}$, $(\tilde{\theta}.F^{1/2}h^i)^2 = \frac{n^2 h_z^{i2}}{1-n^2}$ and $\|F^{1/2}h^i\|^2 = \frac{1-n^2+n^2h_z^{i2}}{1-n^2}$). But this computation is really superfluous since at this point, we are computing the mean quadratic error of the maximum likelihood estimator based on a measurement with, by our construction, Fisher information equal to the inverse of $W(\theta_0)$. (The modifications to our procedure will not alter the Fisher information). The two quantities must be equal by the classical large sample results for the maximum likelihood estimator.

We finally need to show the continuity in $\tilde{\theta}$ at $\tilde{\theta} = \theta_0$ of N' times the quantity in (94). This is evident if the γ_i are all different at θ_0. Both the eigenvalues and the eigenvectors of $F^{-\frac{1}{2}}GF^{-\frac{1}{2}}$ are then continuous functions of $\tilde{\theta}$ at θ_0. There is a potential difficulty however if some γ_i are equal to one another at $\tilde{\theta} = \theta_0$. In this case, the eigenvectors h^i are not continuous functions of $\tilde{\theta}$ at this point, and not even uniquely defined there. We argue as follows that this does not destroy continuity of the mean quadratic error. Consider a sequence of points $\tilde{\theta}_n$ approaching θ_0. This generates a sequence of eigenvectors h_n^i and eigenvalues γ_{in}. The eigenvalues converge to the γ_i but the eigenvectors need not converge at all. However by compactness of the set of unit vectors in R^3, there is a subsequence along which the eigenvectors h_n^i converge; and they must converge to a possible choice of eigenvectors at θ_0. Thus along this subsequence the mean quadratic error (94) does converge to a limit given by the same formula evaluated at the limiting h_i etc. But this limit is equal by construction to the inverse of the target information matrix $G(\theta)$. A standard argument now shows that the limiting mean quadratic error is continuous at $\tilde{\theta} = \theta_0$.

The mean quadratic error of $\hat{\theta}$ given $\tilde{\theta}$ (times N') therefore converges uniformly in a sufficiently small neighborhood of θ_0 to a limit continuous at that point and equal to $W(\theta_0)$ there.

In our derivation of (90) we required the parameter and its estimator to be bounded. By dropping the constraint on the length of θ we have inadvertently lost this property. Suppose we replace our modified estimator $\hat{\theta}$ by the actual maximum likelihood estimator respecting the constraint. The two only differ when the unconstrained estimator lies outside the unit sphere; but this event only occurs with an exponentially small probability, uniformly in $\tilde{\theta}$, provided the γ_i are uniformly bounded away from 0 in the given neighborhood of θ_0. From this it can be shown that the mean quadratic error is altered by an amount $o(1/N')$ uniformly in $\tilde{\theta}$.

If we had worked with random numbers of measurements of each spin variable, when computing the mean quadratic error we would first have copied the computation above conditional on the numbers of measurements, say X_i, of each spin m^i. These numbers are binomially distributed with parameter N' and γ_i. The conditional mean quadratic error would be the same as the expression above but with $1/(\gamma_i N')$ replaced by $1/X_i$ (and special provision taken for the possible outcome $X_i = 0$). So to complete the argument we must show that $\mathrm{E}(1/X_i) = 1/(\gamma_i N') + o(1/N')$ uniformly in $\tilde{\theta}$. This can also be shown to be true, using the fact that X_i/N' only differs from its mean by more than a fixed amount with exponentially small probability as $N' \to \infty$ and we restrict attention to $\tilde{\theta}$ in a neighborhood of θ_0 where the γ_i are bounded away from zero.

Inspection of our argument shows that the convergence of the mean quadratic error is uniform in θ_0 as long as we keep away from the boundary of the parameter space.

By the convergence of the normalized binomial distribution to the normal distribution, the representation of the estimator we gave above also shows that it is asymptotically normally distributed with asymptotic covariance matrix equal to the target covariance matrix W. Moreover, if X has the binomial(n,p) distribution, then $n^{\frac{1}{2}}(X/n - p)$ converges in distribution to the normal with mean zero and variance $p(1-p)$, uniformly in p. Thus the convergence in distribution is also uniform in θ_0 as long as we keep away from the boundary of the parameter space.

D. Conditional mean quadratic error for other models

The preceding subsection gave a complete analysis of the mean quadratic error, given the preliminary estimate $\tilde{\theta}$ for the 3 unknown parameters θ^j of the parameterization (93). We shall first analyze the mean quadratic error when the unknown parameters are functions $\phi^i(\theta^j)$ of the parameters θ^j. We shall then consider the important case when the state is pure and depends on two unknown parameters, and finally the case when the state is pure or mixed and depends on one unknown parameter, or is mixed and depends on two unknown parameters.

Our first result is that if the change of parameters $\phi^i(\theta^j)$ is locally C^1, then the m.q.e. matrix of the ϕ^i is obtained from the m.q.e. of the θ^j by the Jacobian $\partial\phi^i/\partial\theta^j$ except eventually at isolated points. This follows from the fact that under a smooth (locally C^1) parameterization, the delta method (first order Taylor expansion) allows us to conclude uniform convergence of the probability distribution of $\sqrt{N}(\widehat{\phi}^N - \phi)$ to a normal limit with the target mean quadratic error. If the ϕ^i and their derivatives $\partial\phi^i/\partial\theta^j$ are bounded then this proves our claim. If there are points where the ϕ^i or their derivatives $\partial\phi^i/\partial\theta^j$ are infinite, then convergence in distribution does not necessarily imply convergence of moments. However a truncation device allows one to modify the estimate $\widehat{\phi}$, replacing it by 0 if any component is larger than cN^a for given c and a (use the method of [14], Lemma II.8.2 together with the exponential inequality (91) for the multinomial distribution). With this minor modification one can show (uniform in ϕ in a neighborhood of ϕ_0) convergence of the moments of the corresponding $\sqrt{N}(\widehat{\phi} - \phi)$ to the moments of its limiting distribution, hence achievement of the bound in the sense of Theorem IV. In particular if the parameter ϕ is bounded then the truncation is superfluous.

Now turn to the pure state analog of model (93). Obtain a preliminary estimate of the location of ρ on the surface of the Poincaré sphere using the same method as in the mixed case, but always projecting onto the surface of the sphere. Next, after rotation to transform the preliminary estimate into 'spin up', reparameterize to $\rho = \frac{1}{2}(1 + \phi \cdot \sigma)$ where the parameters to be estimated are $(\phi_1, \phi_2) = (\theta'^1, \theta'^2)$ of the parameterization (86) while $\phi_3 = \sqrt{(1 - \phi_1^2 - \phi_2^2)}$. The preliminary estimate is at $\phi_1 = \phi_2 = 0$. The optimal measurement at this point according to Section VII A consists of measurements of the spins σ_1 and σ_2 on specified proportions of the remaining copies. The resulting estimator of the parameter (ϕ_1, ϕ_2) is a linear function of binomial counts and hence its mean quadratic error can be studied exactly as in section VII C. Then we must transfer back to the originally specified parameterization, for instance polar coordinates. This is done as in the preceeding paragraph. If the transformation is locally C_1 then uniform convergence in distribution to the normal law also transfers back; convergence of mean quadratic error too if the original parameter space is bounded. Otherwise a truncation might be necessary. In any case, we can exhibit a procedure optimal in the sense of Theorem IV.

It remains to consider one- and two-dimensional sub-models of the full mixed model, and one-dimensional sub-models of the full pure model. We suppose that the model specifies a smooth curve or surface in the interior of the Poincaré sphere, or a smooth curve on its surface; smoothly parameterized by a one- or two-dimensional parameter as appropriate. The first stage of the procedure is just as before, finishing in projection of an estimated density matrix into the model. Then we reparameterize locally, augmenting the dimension of the parameter to convert the model into a full mixed or pure model respectively. The target information for the extra parameters is zero. Compute as before the optimal measurement at this point. Because of the zero values in the target information matrix, there will be zero eigenvalues γ_i in the computation of section VII A. Thus the optimal measurement will involve specified fractions of measurement of spin in the same number of directions as the dimension of the model. Compute the maximum likelihood estimator of the original parameters based on this data. If the parameterization is smooth enough the estimator will yet again achieve the bound of Theorem IV.

VIII. CONCLUSIONS AND OPEN QUESTIONS

In this paper we have solved some of the problems that arise when trying to estimate the state of a quantum system of which one possesses a large number of copies. This constitutes a preliminary step towards solving the question with which Helstrom concluded his book [2]: "(...) mathematical statisticians are often concerned with asymptotic properties of decision strategies and estimators. (...) When the parameters of a quantum density operator are estimated on the basis of many observations, how does the accuracy of the estimates depend on the number of observations as that number grows very large? Under what conditions have the estimates asymptotic normal distributions? Problems such as these, and still others that doubtless will occur to physicists and mathematicians, remain to be solved within the framework of the quantum-mechanical theory.'

In the case of pure states of spin 1/2 particles the problem has been completely solved. In the limit of large N the variance of the estimate is bounded by (27), and the bound can be attained by separate Von Neumann measurements on each particle.

In the case of mixed states of spin 1/2 particles the state estimation problem for large N has been solved if one restricts oneself to separable measurements. However if one considers non separable measurements, then one can improve the quality of the estimate, which shows that the Fisher information, which in classical statistics is additive, is no longer so for quantum state estimation.

For the case of mixed states of spin 1/2 particles, or for higher spins we do not know what the "outer" boundary of the set of (rescaled) achievable Fisher information matrices based on arbitrary (non separable) measurements of N systems looks like. We have some indications about the shape of this set (see section V G) and we know that it is convex and compact.

Acknowledgements.
S.M. thanks Utrecht university, Netherlands, where part of this work was carried out. He is a "chercheur qualifié du FNRS". R.D.G. thanks the generous hospitality of the Department of Mathematics and Statistics, University of Western Australia.

---

[1] A. S. Holevo, *Probabilistic and Statistical Aspects of Quantum Theory*, North Holland, Amsterdam, 1982.
[2] C. W. Helstrom, *Quantum Detection and Estimation Theory*, Academic, New York, 1976.
[3] W. K. Wootters, Phys. Rev. D **23** (1981) 357.
[4] S. L. Braunstein and C. M. Caves, Phys. Rev. Lett. **72** (1994) 3439.
[5] O. E. Barndorff-Nielsen and R. D. Gill, *an example of non attainability of expected quantum information*, quant-ph/9808009.
[6] H. P. H. Yuen and M. Lax, IEEE Trans. IT-19(6) (1973) 740.
[7] A. Peres and W. K. Wootters, Phys. Rev. Lett. **66**, 1119 (1991)
[8] S. Massar and S. Popescu, Phys. Rev. Lett. **74** (1995) 1259.
[9] C. H. Bennett, et. al., *Quantum nonlocality without entanglement*, quant-ph/9804053

[10] H. Cramér, *Mathematical methods of Statistics*, Princeton University, Princeton, NJ, 1946
[11] H. L. van Trees, *Detection, Estimation and Modulation Theory*, Part 1, New York: Wiley, 1968.
[12] R. D. Gill and B. Y. Levit, Bernouilli **1(1/2)** (1995) 59.
[13] A. W. van der Vaart, ch. 27 pp. 397–410 in *Festschrift for Lucien Le Cam*, ed. D. Pollard, E. Torgersen, G.L. Yang; Springer, 1997.
[14] I.A. Ibragimov, R.Z Has'minskii, *Statistical estimation—asymptotic theory*, Springer, 1981. Theorem III.1.1 (p. 174) is the theorem on uniform convergence of m.l.e.'s.
[15] A.P.Dempster, N.M. Laird and D.R. Rubin, J. Roy. Statist. Soc. (B) 39 (1977), 1-38