Fisher Information in Noisy Intermediate-Scale Quantum Applications

Johannes Jakob Meyer[1,2]

[1]Dahlem Center for Complex Quantum Systems, Freie Universität Berlin, 14195 Berlin, Germany
[2]QMATH, Department of Mathematical Sciences, Københavns Universitet, 2100 København Ø, Denmark

26-06-2021

The recent advent of noisy intermediate-scale quantum devices, especially near-term quantum computers, has sparked extensive research efforts concerned with their possible applications. At the forefront of the considered approaches are variational methods that use parametrized quantum circuits. The classical and quantum Fisher information are firmly rooted in the field of quantum sensing and have proven to be versatile tools to study such parametrized quantum systems. Their utility in the study of other applications of noisy intermediate-scale quantum devices, however, has only been discovered recently. Hoping to stimulate more such applications, this article aims to further popularize classical and quantum Fisher information as useful tools for near-term applications beyond quantum sensing. We start with a tutorial that builds an intuitive understanding of classical and quantum Fisher information and outlines how both quantities can be calculated on near-term devices. We also elucidate their relationship and how they are influenced by noise processes. Next, we give an overview of the core results of the quantum sensing literature and proceed to a comprehensive review of recent applications in variational quantum algorithms and quantum machine learning.

Progress in science and engineering as well as considerable investments have increased our capabilities to precisely control quantum systems, leading to the development of quantum devices that are capable of impressive feats. One prominent example of this new generation of devices are quantum computers. They have low numbers of qubits and are still plagued by noise and relatively low coherence times and cannot yet fulfill the promises of fault-tolerant quantum computation, but it was recently shown that they can outperform classical computers [1] – however only in certain contrived tasks with unknown practical relevance. Yet, given these positive results and the speed of improvement of these devices, it is no surprise that the search for practically relevant applications of these *noisy intermediate-scale quantum (NISQ)* [2] computers has become a very active field of research in quantum information science. Particularly prominent candidates to make use of near-term quantum computers are *variational quantum algorithms* [3, 4] where parametrized quantum states are combined with a classical computer into a hybrid quantum-classical algorithm [5]. Another much researched direction is to use these quantum devices as *quantum machine learning models* [6], again employing parametrized quantum states.

Along with the development of these new techniques, there is a need for a better *understanding* of parametrized quantum systems and consequently the approaches that employ them. In the field of quantum sensing, a particular type of parametrized quantum system – namely parametrized by the parameters that are sensed – has long been studied. The principal tools employed in this regard are the *classical* and *quantum Fisher information.*

Intuitively, the quantum Fisher information is a measure of how much a parametrized quantum state changes under a change of a parameter. This parameter could be the underlying magnetic field that a quantum system is designed to sense, but it can also be a control knob turned by an experimenter or the angle of a rotation gate that is changed by a quantum computer programmer. The classical Fisher information in turn captures how much a change of the underlying parameter affects the probabilities with which different outcomes of a specific measurement that is performed on the parametrized quantum state are observed.

As parametrized quantum states appear in many approaches in NISQ applications beyond quantum sensing, it is no surprise that the first works have explored the use of classical and quantum Fisher information to study them. We do, however, believe that there are many more fruitful applications waiting to be discovered. This article thus intends to further popularize classical and quantum Fisher information as versatile tools to understand NISQ applications.

While much has been written about the Fisher information, in classical and quantum settings alike, the barriers to penetrate the literature can be quite substantial. This article aims to complement the excellent reviews provided in Refs. [7–9] with a tutorial that provides a very gentle and intuitive introduction to both the classical and quantum Fisher information that does require only little prerequisites. Since many of the results related to these quantities were developed in the context of quantum sensing, we will also review the results from this field that are most relevant to understanding applications of Fisher information. However, we will not explore specific applications, as it is not the goal of this article to provide a comprehensive introduction to quantum sensing. We conclude this work with a comprehensive review of the different uses of Fisher information that have emerged in quantum machine learning and variational quantum algorithms.

The principal target audience of this paper are people that are interested in learning about Fisher information and its use in the context of NISQ applications. It thus only requires familiarity with the absolute basics of NISQ applications. The intuitive approach to the subject should also be beneficial to a wider class of readers, *e.g.* people that are taking their first steps in quantum sensing. After reading this paper, the reader will have an intuitive understanding of the origins and applications of both classical and quantum Fisher information and will know how they are related to each other. We will discuss how these quantities can be calculated in the context of NISQ applications, how the ever-present device noise enters the picture and how they are applied in quantum machine learning and optimization of variational quantum algorithms. Along the way, we will review important results from quantum sensing and try to demystify technical terms usually encountered when reading about Fisher information like "metric" or "pullback" and quantum-specific jargon like "Heisenberg scaling".

This work is organized as follows: In Sec. 1, we will build intuition about parametrized quantum states and how we can properly assign distances to pairs of parameters. Sec. 2 outlines how we can extract information about parametrized quantum systems in the form of information matrices. We follow up on this with the introduction of the classical Fisher information matrix in Sec. 3 and the quantum Fisher information matrix in Sec. 4. We will put a particular focus on how we can actually compute these quantities in a NISQ context. The subsequent Sec. 5 elucidates the relationship between the quantum and the classical Fisher information, whereas Sec. 6 treats the role of noise, a very important impediment and namesake of NISQ devices. In Sec. 7 we will highlight the different areas in which the classical and quantum Fisher information have been applied in the context of NISQ devices, rounding up with a short outlook on fascinating applications beyond the scope of NISQ devices in Sec. 8. The work concludes with a look to the future in Sec. 9. To make this work as self-contained and explanatory as possible, most derivations are found in the Appendix, alongside with proofs of the properties of classical and quantum Fisher information. In the course of this work we will give recommendations on literature for further study.

1 Parametrized Quantum States

The study of NISQ devices places a huge emphasis on the study of *parametrized quantum states*, *i.e.* quantum states that depend continuously on a vector of parameters $\boldsymbol{\theta} \in \mathbb{R}^d$. We will denote them with $|\psi(\boldsymbol{\theta})\rangle$ for pure states and $\rho(\boldsymbol{\theta})$ for mixed states. Parametrized quantum states arise in many settings: In NISQ computing, *parametrized quantum circuits* are used as an *ansatz* for variational quantum algorithms [3]. In quantum optimal control, the parameters of the radio frequency pulses determine the executed gate or the created quantum state. In quantum metrology, a quantity that needs to be measured, for example a magnetic field, is imprinted on a quantum state via an interaction, thus parametrizing it.

It is of course of tremendous interest to understand what happens if the parameters $\boldsymbol{\theta}$ are changed. We can attempt to quantify this by measuring the *distance* between the parameters themselves – which can be done via the regular Euclidean distance. But only in the rarest cases all parameters have an equal influence on the underlying state. So a smarter way to measure distances between parameters would actually be to measure the distance of the associated *states*. If we are given a distance measure d between quantum states, we can define – by a slight abuse of notation – a new distance measure between the associated parameters:

$$d(\boldsymbol{\theta}, \boldsymbol{\theta}') = d(\rho(\boldsymbol{\theta}), \rho(\boldsymbol{\theta}')). \tag{1}$$

This strategy of measuring the distance in the space of quantum states instead of the parameter space is known as a *pullback*, because we pull back the distance measure to the space of parameters.

In the following we will be concerned with distance measures that are *monotonic*. This means that the distance between two quantum states can only decrease if both are subject to the same quantum operation. A quantum operation can take many forms, including unitary evolution, noisy evolution or even measurements – but we will formalize this more later on. Monotonicity is a very desirable property for a distance measure. Performing a quantum operation cannot add additional information, so it should not be easier to distinguish two states after such an operation is performed. Moreover, if the operation is noisy, it will actually *destroy* information, which should be echoed in a decrease of the associated distance as adding noise cannot make two states more distinguishable [10].

We have already argued that it is more sensible to measure distances between parameters of a quantum

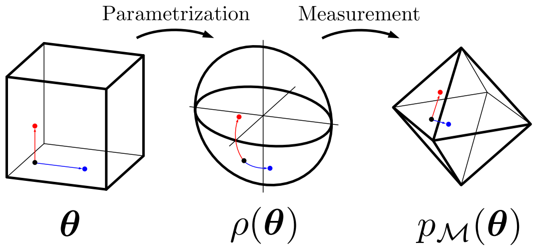

Parametrization
Measurement
$\boldsymbol{\theta}$
$\rho(\boldsymbol{\theta})$
$p_{\mathcal{M}}(\boldsymbol{\theta})$

Figure 1: Large distances between parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ need not correspond to large distances of the corresponding quantum states $\rho(\boldsymbol{\theta})$ and $\rho(\boldsymbol{\theta}')$. Equally, large distances between quantum states need not correspond to large distances between the output probability distributions after the measurement, $p_{\mathcal{M}}(\boldsymbol{\theta})$ and $p_{\mathcal{M}}(\boldsymbol{\theta}')$. Measuring the distance between two parameters by measuring the distance between the corresponding quantum states or output probability distributions is therefore a sensible approach known as a *pullback*.

state by measuring the distances in the space of quantum states. But as we are mere classical observers, any NISQ application must involve a measurement of the underlying quantum state. A measurement will inevitably collapse the quantum state into a classical probability distribution over the possible measurement outcomes. We can formalize this by defining a measurement $\mathcal{M} = \{\Pi_l\}$, where the operator Π_l identifies the l-th outcome of the experiment [10]. In NISQ applications, the operators $\{\Pi_l\}$ are usually just the projectors onto the basis states. The probabilities of observing the different outcomes are then given by

$$p_l(\boldsymbol{\theta}) = \operatorname{Tr}\{\rho(\boldsymbol{\theta})\Pi_l\}. \tag{2}$$

We will use $p_{\mathcal{M}}(\boldsymbol{\theta})$ to refer to the full output distribution for a specific measurement $\mathcal{M}$. We will drop the subscript $\mathcal{M}$ if we talk about generic probability distributions.

If we fix a certain measurement $\mathcal{M}$, we also fix a certain way of collapsing a parametrized quantum state into a probability distribution which is now also dependent on the parameter $\boldsymbol{\theta}$. We thus have a *parametrized probability distribution*. But now the same argument that we used to motivate measuring the distance between parameters in the space of quantum states applies here, too. A large change in the underlying quantum state might not correspond to a large change in the output probability distribution that is observed. If we fix a measurement, we should therefore consider the pullback of a distance d between two probability distributions

$$d_{\mathcal{M}}(\boldsymbol{\theta}, \boldsymbol{\theta}') = d(p_{\mathcal{M}}(\boldsymbol{\theta}), p_{\mathcal{M}}(\boldsymbol{\theta}')). \tag{3}$$

In the following, we require some very natural properties from the distance measures between quantum states or probability distributions we employ. First, that it is always positive $d(\boldsymbol{\theta}, \boldsymbol{\theta}') \geq 0$ and second, that the distance between identical objects is zero $d(\boldsymbol{\theta}, \boldsymbol{\theta}) = 0$. This means that the distance measure has to satisfy the axiomatic definition of a *divergence* as employed in statistics, which is less strict than the definition of distance encountered in other areas of mathematics.

In summary, we have now set the arena: the parameters $\boldsymbol{\theta}$ define a quantum state $\rho(\boldsymbol{\theta})$ which then undergoes the measurement $\mathcal{M}$. We can measure distances between parameters by going to the space of quantum states or to the space of probability distributions over the measurement outcomes, dependent on the question we seek to answer. Fig. 1 illustrates these relations.

2 Information Matrices

We are often confronted with scenarios where a quantum system is in a state associated to a particular parameter $\boldsymbol{\theta}$ but where it is important to understand how much a change of the parameter $\boldsymbol{\theta}$ in a particular direction results in a change of the underlying quantum state or the output probability distribution.

To gain that understanding, we will look how a slight perturbation of the parameter $\boldsymbol{\theta} + \boldsymbol{\delta}$ reflects in the chosen distance by analyzing $d(\boldsymbol{\theta}, \boldsymbol{\theta}+\boldsymbol{\delta})$. If the distance measure d is *differentiable*, we can develop this into a Taylor series around $\boldsymbol{\delta} = 0$. Because d is a distance measure, we can assume that it is both positive and vanishes for identical parameters, *i.e.* $d(\boldsymbol{\theta}, \boldsymbol{\theta}) = 0$ is a minimum. But we know that the first order contributions vanish around minima and that the second order is therefore the first contribution of the Taylor series that does not vanish. To write down the Taylor expansion, we first define the matrix M with entries

$$M(\boldsymbol{\theta})_{ij} = \frac{\partial^2}{\partial \delta_i \partial \delta_j} d(\boldsymbol{\theta}, \boldsymbol{\theta} + \boldsymbol{\delta}) \bigg|_{\boldsymbol{\delta}=0}. \tag{4}$$

In more mathematical terms, this is the matrix of second order derivatives – also known as the *Hessian* – of the function

$$g_{\boldsymbol{\theta}}(\boldsymbol{\delta}) = d(\boldsymbol{\theta}, \boldsymbol{\theta} + \boldsymbol{\delta}), \tag{5}$$

at $\boldsymbol{\delta} = 0$. With this we can express the Taylor expansion as

$$\begin{aligned} d(\boldsymbol{\theta}, \boldsymbol{\theta} + \boldsymbol{\delta}) &= \frac{1}{2} \sum_{i,j=1}^{d} \delta_i \delta_j M(\boldsymbol{\theta})_{ij} + O(\|\boldsymbol{\delta}\|^3) & (6) \\ &= \frac{1}{2} \boldsymbol{\delta}^T M(\boldsymbol{\theta}) \boldsymbol{\delta} + O(\|\boldsymbol{\delta}\|^3). & (7) \end{aligned}$$

The matrix $M(\boldsymbol{\theta})$ therefore captures all we need to know about the local vicinity of $\boldsymbol{\theta}$ in parameter space, but measured by the distance d in the underlying space of either quantum states or output probability distributions! Intuitively, large entries of $M(\boldsymbol{\theta})$ indicate that a change in the corresponding parameters results in a large change of the underlying object. In the following, we will not denote the dependence on $\boldsymbol{\theta}$ explicitly and simply use the notation M.

The matrix M is an example of a *metric*. To elucidate the meaning of this, we first need to remind ourselves how measuring distances and angles in Euclidean space works through the standard scalar product. It is defined as

$$\langle \boldsymbol{\delta}, \boldsymbol{\delta}' \rangle = \boldsymbol{\delta}^T \boldsymbol{\delta}' = \sum_{i=1}^{d} \delta_i \delta_i'. \tag{8}$$

The standard scalar product acts as an "umbrella" of sorts as it allows us to measure lengths

$$\|\boldsymbol{\delta}\| = \sqrt{\langle \boldsymbol{\delta}, \boldsymbol{\delta} \rangle}, \tag{9}$$

distances

$$d(\boldsymbol{\delta}, \boldsymbol{\delta}') = \sqrt{\langle \boldsymbol{\delta} - \boldsymbol{\delta}', \boldsymbol{\delta} - \boldsymbol{\delta}' \rangle}, \tag{10}$$

and angles between vectors

$$\sphericalangle(\boldsymbol{\delta}, \boldsymbol{\delta}') = \arccos \frac{\langle \boldsymbol{\delta}, \boldsymbol{\delta}' \rangle}{\sqrt{\langle \boldsymbol{\delta}, \boldsymbol{\delta} \rangle \langle \boldsymbol{\delta}', \boldsymbol{\delta}' \rangle}}. \tag{11}$$

It is quite suggestive that $\boldsymbol{\delta}^T \boldsymbol{\delta}' = \boldsymbol{\delta}^T \mathbb{I} \boldsymbol{\delta}'$ resembles the expression in Eq. (7) with M replaced by the identity matrix $\mathbb{I}$. And indeed, we can use the matrix M to define a new scalar product

$$\langle \boldsymbol{\delta}, \boldsymbol{\delta}' \rangle_M = \boldsymbol{\delta}^T M \boldsymbol{\delta}' \tag{12}$$

that now includes information about the local environment of $\rho(\boldsymbol{\theta})$ or $p_{\mathcal{M}}(\boldsymbol{\theta})$! And this is what a metric does in a nutshell – it allows to measure distances and angles between vectors in parameter space, but with the local structure of the underlying quantum states or probability distribution parametrized by those vectors taken into account.

Because these matrices contain information about the underlying quantum states and probability distributions – which are by nature objects that have an information theoretic meaning – we will call them *information matrices*.

3 The Classical Fisher Information

After having kept the derivations general, we now turn our attention to the Fisher information itself and begin with the classical case. As the name suggests, the classical Fisher information is defined for parametrized probability distributions. In the context of NISQ devices, this means on the probability distributions of measurement outcomes. To make use of the machinery we developed in the last section, we need a distance measure between probability distributions. There exists numerous ways to measure distances between probability distributions, but one of the most popular is certainly the *Kullback-Leibler (KL) divergence*, also known as the *relative entropy*. It is defined as[1]

$$d_{\mathrm{KL}}(p_{\mathcal{M}}(\boldsymbol{\theta}), p_{\mathcal{M}}(\boldsymbol{\theta}')) = \sum_{l \in \mathcal{M}} p_l(\boldsymbol{\theta}) \log \frac{p_l(\boldsymbol{\theta})}{p_l(\boldsymbol{\theta}')}. \tag{13}$$

The intuition behind the KL divergence is not obvious at first sight, but Nielsen and Chuang give some accounts in Ref. [10]. Imagine we are given an unknown probability distribution that can be either $p_{\mathcal{M}}(\boldsymbol{\theta})$ or $p_{\mathcal{M}}(\boldsymbol{\theta}')$ and we are tasked with deciding which of the two distributions it is. Then, the KL divergence essentially captures how fast the "false negative" error decreases with the number of repetitions we are allowed when there is a constraint on the "false positive" probability.

Now, we need to perform the second order expansion to get a formula for the corresponding information matrix. For conciseness of the main text, you find the complete derivation in App. A. The result of the calculation is the following formula for the information matrix associated to the KL divergence:

$$\begin{aligned}[M_{\mathrm{KL}}]_{ij} &= \sum_{l \in \mathcal{M}} p_l(\boldsymbol{\theta}) \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p_l(\boldsymbol{\theta}) && (14)\\ &= \sum_{l \in \mathcal{M}} \frac{1}{p_l(\boldsymbol{\theta})} \frac{\partial p_l(\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial p_l(\boldsymbol{\theta})}{\partial \theta_j}. && (15)\end{aligned}$$

The matrix M_{KL} is nothing else than the *(classical) Fisher information matrix (CFIM)* and we will in the following denote it as $I = M_{\mathrm{KL}}$. A resource with substantial information about it is the textbook by Lehmann and Casella [11]. For the reader's convenience, we list and prove the most important properties of the classical Fisher information matrix in App. B.

Uniqueness. We derived the classical Fisher information from the Kullback-Leibler divergence. But what happens if we repeat the same procedure for another distance measure? It turns out that the derivation will *always* yield a constant multiple of the classical Fisher information if the distance measure is monotonic.

We shallowly introduced the idea of monotonicity in Sec. 1, but we will formalize it now. To this end, we first need another concept, namely that of a *stochastic map*. A stochastic map is a linear operation that takes in probability distributions and always outputs probability distributions. This is a very broad definition and includes every admissible thing we can do with probability distributions. We formally call a distance measure between two probability distributions p and q monotonic if the distance cannot increase under any

[1] Note that there are edge cases where the KL divergence is not properly defined, for example if there exists a $p_l(\boldsymbol{\theta}') = 0$ where at the same time $p_l(\boldsymbol{\theta}) \neq 0$. We will exclude this cases in this work.

stochastic map T:

$$d(T[p], T[q]) \leq d(p, q). \tag{16}$$

The fact that we always end up with a constant multiple of the Fisher information if we start from a monotonic distance measure is known as the *uniqueness* of the classical Fisher information which is a celebrated result due to Morozova and Chentsov [12]. A particularly important and widely applied monotonic distance measure between probability distributions is given by the *total variation distance*

$$d_{\mathrm{TV}}(p, q) = \frac{1}{2} \sum_l |p_l - q_l|. \tag{17}$$

It measures the maximum difference in probability that p and q can assign to the same event. It would be tempting to ask which information matrix is related to this distance measure. But we have to remember that we required the distance measure to be differentiable in order to define an information matrix – and as the total variation distance includes an absolute value function, it is not differentiable at $p = q$ as would be required for our construction. The total variation distance therefore does not induce an information matrix.

Calculation. To work with the classical Fisher information matrix in a NISQ context we need to actually calculate it. If we take a closer look at Eq. (14), we see that we need two ingredients to do so: First, the probabilities of the different measurement outcomes $p_l(\boldsymbol{\theta})$ and second their derivatives with respect to the parameters $\boldsymbol{\theta}$, $\partial p_l(\boldsymbol{\theta})/\partial\theta_i$.

First, let us talk about the output probabilities. Each run of a NISQ device with a fixed measurement setting will add a data point, and with sufficiently many data points, the output probabilities $p_l(\boldsymbol{\theta})$ can be estimated. This is also how one approaches the estimation of expectation values: a unitary transformation changes the measurement basis to the eigenbasis of the desired operator $H = \sum_l h_l |h_l\rangle\langle h_l|$. We then perform multiple repetitions of the experiment, estimate the output probabilities $p_l(\boldsymbol{\theta})$ and then calculate the expectation value as

$$\langle H(\boldsymbol{\theta})\rangle = \sum_l p_l(\boldsymbol{\theta}) h_l. \tag{18}$$

This means that an estimation of the output probabilities is actually a prerequisite for the calculation of expectation values, the most ubiquitous primitive in NISQ settings. Note however that the output probability distribution contains more information than the expectation value – which means that a faithful estimate of the probability distribution will usually require more runs of the experiment than an estimation of an expectation value.

In fact, in the worst case the number of samples required to estimate the probability distribution is proportional to the number of different measurement outcomes which usually is exponential in the number of qubits [13]. This problem is ameliorated if many of the output probabilities are very small – in this case we need fewer samples to guarantee a good estimate. This can be understood intuitively: If we do not observe a particular measurement outcome l, we estimate $p_l = 0$. But as we know that p_l is small, this is already a good estimate.

As the estimation of the output probability distribution is a very important task in NISQ settings, there also exist more sophisticated methods than just estimating the output probabilities from the number of times they were observed in a limited number of test runs. One possibility is a Bayesian approach, where a prior estimate of the output distribution is updated using new samples. Techniques based on machine learning have also been shown to be very effective tools to capture information about quantum states and therefore also the associated output probability distributions [14].

We now turn to the derivatives. In many cases, gradient based schemes are deployed to optimize Variational Quantum Algorithms. We will now argue that these schemes already entail the calculation of the derivatives of the output probabilities necessary for the calculation of the Fisher information matrix. These methods use either finite differences or the *parameter-shift rule* [15, 16]. In all cases, the derivatives of expectation values are calculated by evaluating expectation values at different parameter settings. We will make this exemplary by considering two-sided finite differences, where the derivative is approximated via a small perturbation ϵ:

$$\frac{\partial}{\partial\theta_i}\langle H(\boldsymbol{\theta})\rangle \approx \frac{\langle H(\boldsymbol{\theta} + \epsilon \boldsymbol{e}_i) - \langle H(\boldsymbol{\theta} - \epsilon \boldsymbol{e}_i)\rangle}{2\epsilon}, \tag{19}$$

where $\boldsymbol{e}_i$ denotes the i-th unit vector, or, equivalently, that we only perturb θ_i. The following argument, however, equally works for other means of estimating derivatives.

In order to perform the finite-difference approximation, we have to compute the expectation values $\langle H(\boldsymbol{\theta} \pm \boldsymbol{e}_i\epsilon)\rangle$, which proceeds via the estimation of the output probability distribution as explained above. But this means that we actually compute the derivative of the output probability distribution:

$$\frac{\partial}{\partial\theta_i}\langle H(\boldsymbol{\theta})\rangle \approx \frac{\sum_l p_l(\boldsymbol{\theta} + \epsilon\boldsymbol{e}_i)h_l - \sum_l p_l(\boldsymbol{\theta} - \epsilon\boldsymbol{e}_i)h_l}{2\epsilon} \tag{20}$$

$$= \sum_l \frac{p_l(\boldsymbol{\theta} + \epsilon\boldsymbol{e}_i) - p_l(\boldsymbol{\theta} - \epsilon\boldsymbol{e}_i)}{2\epsilon} h_l \tag{21}$$

$$\approx \sum_l \frac{\partial p_l(\boldsymbol{\theta})}{\partial\theta_i} h_l. \tag{22}$$

This means that the same process we use to estimate derivatives of expectation values will also yield the derivatives of the output probabilities we need to calculate the classical Fisher information matrix. Again a word of caution is advised, as the faithful estimation of the derivatives of the whole output probability distribution will usually require more runs of the experiment than used for estimating the derivative of an expectation value.

In summary, we learned that the processes that are already used to compute expectation values and their gradients implicitly give us the data necessary to estimate both the output probabilities $p_l(\boldsymbol{\theta})$ and their derivatives $\partial p_l(\boldsymbol{\theta})/\partial\theta_i$ and thus the whole classical Fisher information matrix. This brings us to the conclusion that the calculation of the Fisher information matrix is not harder than the calculation of expectation values and their derivatives which are routine tasks in the context of NISQ computing.

We should also note that – especially for large quantum systems – one usually does not have sufficiently many samples to accurately estimate the whole distribution, which means that many elements of the probability distribution p_l are estimated as zero. In this case, the corresponding terms in the formula for the classical Fisher information of Eq. (14) would diverge. This is usually mitigated by the fact that the associated derivatives are also zero in which case we can just remove the corresponding terms from the sum. At points where p_l vanishes but a derivative $\partial p_l(\boldsymbol{\theta})/\partial\theta_i$ does not we actually have a discontinuity of the Fisher information [17], which means that it is not properly defined at these points, a property that is inherited from the Kullback-Leibler divergence. Another way around this problem is to use a Bayesian approach for the estimation of the output probability distribution with a prior that does not contain zero probabilities.

4 The Quantum Fisher Information

We always strive to find quantum generalizations of classical concepts, and the Fisher information is no exception. We know that any classical probability distribution can be expressed by some quantum state with a diagonal density matrix, which means that classical probability distributions are actually a subset of all quantum states. Because of the uniqueness of the classical Fisher information, any "quantum Fisher information" should reduce to the classical Fisher information when looking on classical states.

We have furthermore learned that the classical Fisher information is associated with *monotonic* distance measures. To properly define monotonicity in the quantum setting we first need to introduce the quantum generalization of stochastic maps, the *quantum channels*. Stochastic maps are linear operations that map probability distributions to probability distributions. Likewise, quantum channels are defined as linear operations that take density matrices in and output density matrices, even when ancillary systems are included. Like stochastic maps, quantum channels are a very broad concept, including not only unitary and noisy evolution but also measurements! This can actually be seen quite easily: measurements turn a quantum state into a classical probability distribution over the measurement outcomes. But as we just learned these are also a subset of all quantum states, which means that measurements match the requirements of a quantum channel. A distance measure d between quantum states is monotonic if it cannot increase under any quantum channel Φ:

$$d(\Phi[\rho], \Phi[\sigma]) \leq d(\rho, \sigma). \tag{23}$$

As quantum channels encompass basically everything we can do in quantum information processing, they are heavily studied. You find introductions to quantum channels in the excellent book by Wilde [18], the classic book by Nielsen and Chuang [10] and the more mathematical lecture notes by Wolf [19].

To find a quantum generalization of the classical Fisher information, we will again use the machinery developed in Sec. 2. We will limit ourselves to *pure* quantum states in this section to keep the developments simple, but return to the general case in Sec. 6.

To start the machinery, we need to choose an appropriate distance measure between quantum states. From all possible contenders, the *fidelity* stands out due to its beautiful operational interpretation. The fidelity between two pure quantum states is given by

$$f(|\psi(\boldsymbol{\theta})\rangle, |\psi(\boldsymbol{\theta}')\rangle) = |\langle\psi(\boldsymbol{\theta})|\psi(\boldsymbol{\theta}')\rangle|^2. \tag{24}$$

The fidelity is so important because the probability with which we can distinguish the two states $|\psi(\boldsymbol{\theta})\rangle$ and $|\psi(\boldsymbol{\theta}')\rangle$ when using the optimal measurement is

$$d_f(|\psi(\boldsymbol{\theta})\rangle, |\psi(\boldsymbol{\theta}')\rangle) = 1 - f(|\psi(\boldsymbol{\theta})\rangle, |\psi(\boldsymbol{\theta}')\rangle). \tag{25}$$

As we need a measure of distance that is 0 for indistinguishable states, we will use d_f in our calculations.[2]

We will again leave the complete derivation to App. D and directly go to the formula for the information matrix associated with the fidelity:

$$\begin{aligned}[M_f]_{ij} = 2\,\mathrm{Re}\big[&\langle\partial_i\psi(\boldsymbol{\theta})|\partial_j\psi(\boldsymbol{\theta})\rangle \\ &- \langle\partial_i\psi(\boldsymbol{\theta})|\psi(\boldsymbol{\theta})\rangle\langle\psi(\boldsymbol{\theta})|\partial_j\psi(\boldsymbol{\theta})\rangle\big].\end{aligned} \tag{26}$$

In this formula, we have used the shorthand $\partial_i = \partial/\partial\theta_i$. If one checks the consistency of this information matrix with the classical Fisher information, however one realizes that the prefactor is wrong. As this is only an artifact of how defined our distance we can simply correct this by multiplying with a constant. Doing so, we obtain the formula for the *quan-*

[2] We could also have used a different convention for our calculations, where the square root of the fidelity is used instead of the fidelity. App. C contains an explanation why this will only incur a constant prefactor.

tum Fisher information matrix (QFIM) which is associated with the distance $2d_f$:

$$\begin{aligned}\mathcal{F}_{ij} = 4\,\mathrm{Re}\big[&\langle \partial_i \psi(\boldsymbol{\theta}) | \partial_j \psi(\boldsymbol{\theta})\rangle \\ &- \langle \partial_i \psi(\boldsymbol{\theta}) | \psi(\boldsymbol{\theta})\rangle \langle \psi(\boldsymbol{\theta}) | \partial_j \psi(\boldsymbol{\theta})\rangle\big].\end{aligned} \tag{27}$$

A great review about the quantum Fisher information matrix, various ways to calculate it and many applications was recently written by Liu *et al.* [7]. For the reader's convenience, however, the important properties of the quantum Fisher information matrix are summarized and proven in App. E.

Non-Uniqueness. In the classical case, it does not matter from which monotonic distance measure between probability distributions we start our derivation, we will always end up with a constant multiple of the classical Fisher information. But this is actually not the case in the quantum setting! Indeed, as was proven by Pétz [20], there are infinitely many monotonic metrics in the quantum setting, and multiple ones have found application in quantum information theory. To distinguish the quantum Fisher information that arises from the fidelity from the other ones it is also often referred to as *SLD quantum Fisher information*. Here, SLD stands for *symmetric logarithmic derivative*. This is related to a different way in which we can define the (SLD) quantum Fisher information. It is

$$\mathcal{F}_{ij} = \frac{1}{2}\,\mathrm{Tr}\{\rho(L_i L_j + L_j L_i)\} \tag{28}$$

where L_i is the *symmetric logarithmic derivative (SLD) operator* corresponding to the coordinate θ_i. It is implicitly defined by

$$\frac{\partial \rho}{\partial \theta_i} = \frac{1}{2}(L_i \rho + \rho L_i). \tag{29}$$

You can think of this operator as a way to rewrite derivatives of the quantum state ρ. We chose to not introduce the quantum Fisher information via this approach because it is not only unintuitive but also somewhat unwieldy. This, however, is the way in which the first results on the quantum Fisher information were obtained [21].

Pétz also showed that the (SLD) quantum Fisher information that we just derived from the fidelity between quantum states is special because it is the *smallest* monotone metric in a certain sense [20]. One can furthermore make a case that it is the "most natural" in certain respects. If your are interested in this, have a look at Ref. [22] which contains a pedagogical exposition relating it to the "natural" geometry of the Hilbert space of quantum states. Another useful property of the SLD quantum Fisher information that sets it apart from the competition is that it is actually defined for pure states – many other possible generalizations of the classical Fisher information become infinite in this case.

Calculation. The quantum Fisher information is a much more peculiar object to work with than its classical counterpart. We will now outline the techniques that have been developed to tackle the calculation of the quantum Fisher information. We will keep our focus on the pure state case and come back to the practically important noisy case later.

In many NISQ applications, especially in NISQ computing, the parameters of the quantum state are usually rotation angles of gates with a certain generator, *e.g.* $U(\theta_i) = e^{-i\theta_i G_i}$. In this case, our job will be a lot easier. If we execute the circuit that prepares our state until the point where the gate in question is applied and call the state before the gate happens $|\psi_0\rangle$ we can express the derivative in terms of the generator:

$$|\partial_i \psi(\boldsymbol{\theta})\rangle = \partial_i e^{-i\theta_i G_i} |\psi_0\rangle = -iG_i |\psi(\boldsymbol{\theta})\rangle . \tag{30}$$

Putting this into the formula for the quantum Fisher information of Eq. (27), we get a simple formula for the diagonal elements:

$$\mathcal{F}_{ii} = 4\left(\langle \psi_0 | G_i^2 | \psi_0\rangle - \langle \psi_0 | G_i | \psi_0\rangle^2\right). \tag{31}$$

This is nothing but the fourfold variance of the generator G_i with respect to the state $|\psi_0\rangle$. Note that we dropped the real part from the formula as the variance is already a real number. This means that we can get the diagonal elements of the quantum Fisher information matrix by executing the circuit in question until our gate happens and then evaluating the expectation values of the two observables G_i^2 and G_i.

If we have multiple gates happening in parallel, we can also use the same approach to compute the off-diagonal elements of the quantum Fisher information matrix related to these gates. If we evaluate the formula for the quantum Fisher information in this case, we get

$$\begin{aligned}\mathcal{F}_{ij} = 4\Big(&\langle \psi_0 | \frac{\{G_i, G_j\}}{2} | \psi_0\rangle \\ &- \langle \psi_0 | G_i | \psi_0\rangle \langle \psi_0 | G_j | \psi_0\rangle\Big),\end{aligned} \tag{32}$$

where $\{G_i, G_j\}/2 = (G_i G_j + G_j G_i)/2$ can be understood as the "real part" of the product $G_i G_j$. The quantity above is nothing else but the fourfold *covariance* of the generators G_i and G_j with respect to the state $|\psi_0\rangle$. This means that we can evaluate all "blocks" of the quantum Fisher information matrix corresponding to gates executed in parallel by evaluating the aforementioned observables on the state right before the layer of parallel gates is executed, $|\psi_0\rangle$ [23].

Elements of the quantum Fisher information matrix that correspond to gates that are not executed in parallel are harder to deal with, because the observables that need to be evaluated now also depend on the intermediary circuit elements between the gates.

But we can still evaluate the elements of the quantum Fisher information in this case using more sophisticated techniques.

If the quantum gates that shall be differentiated support a *parameter-shift rule* we can do this across layers. A parameter-shift rule [16] states that the expectation value of any operator O evaluated on a state $|\psi(\boldsymbol{\theta})\rangle$, $\langle O(\boldsymbol{\theta})\rangle = \langle\psi(\boldsymbol{\theta})|O|\psi(\boldsymbol{\theta})\rangle$ can be differentiated as

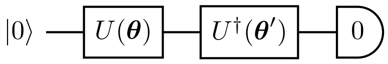
$$\frac{\partial}{\partial\theta_i}\langle O(\boldsymbol{\theta})\rangle = r\left(\langle O(\boldsymbol{\theta}+\boldsymbol{e}_i\frac{\pi}{4r})\rangle - \langle O(\boldsymbol{\theta}-\boldsymbol{e}_i\frac{\pi}{4r})\rangle\right) \tag{33}$$

where the constant r depends on the nature of the gate in question. We see that the derivative can be evaluated by running the same circuit but "shifting" the parameter in question in both directions. Luckily, most quantum gates available on near-term quantum devices support such a parameter-shift rule.

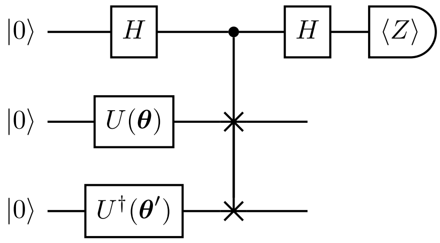
Using the parameter-shift rule and the fact that the quantum Fisher information matrix can be expressed via the second derivatives of the fidelity, the authors of Ref. [24] derived a formula for the quantum Fisher information that reads

$$\begin{aligned}\mathcal{F}_{ij} = -\frac{1}{2}\Big(&|\langle\psi(\boldsymbol{\theta})|\psi(\boldsymbol{\theta}+(\boldsymbol{e}_i+\boldsymbol{e}_j)\frac{\pi}{2})\rangle|^2 \\ &-|\langle\psi(\boldsymbol{\theta})|\psi(\boldsymbol{\theta}+(\boldsymbol{e}_i-\boldsymbol{e}_j)\frac{\pi}{2})\rangle|^2 \\ &-|\langle\psi(\boldsymbol{\theta})|\psi(\boldsymbol{\theta}-(\boldsymbol{e}_i-\boldsymbol{e}_j)\frac{\pi}{2})\rangle|^2 \\ &+|\langle\psi(\boldsymbol{\theta})|\psi(\boldsymbol{\theta}-(\boldsymbol{e}_i+\boldsymbol{e}_j)\frac{\pi}{2})\rangle|^2\Big).\end{aligned} \tag{34}$$

This formula contains the fidelities between different parametrizations of the same state. To evaluate those via quantum circuits, we have two principal ways: if we denote the circuit preparing the quantum state $|\psi(\boldsymbol{\theta})\rangle$ as $U(\boldsymbol{\theta})$, then the overlap $|\langle\psi(\boldsymbol{\theta})|\psi(\boldsymbol{\theta}')\rangle|^2$ can be evaluated via first applying the unitary $U(\boldsymbol{\theta})$ and then the inverse of $U(\boldsymbol{\theta}')$. After this circuit, the overlap is exactly the probability of observing the outcome 0 [25]:



Alternatively, we can make use of the *SWAP test* which achieves the same thing:



Both approaches have their downsides. The compute and reverse approach has twice the depth of the original state preparation and the SWAP test requires twice as many qubits, an additional ancilla and controlled SWAP operations but retains the same depth up to some constant number of gates. We should also not forget that these approaches only work for pure states. As soon as the operations preparing the state become noisy, the quantities calculated via these approaches do not coincide with the fidelity but instead with the state overlap which does not give rise to the quantum Fisher information.

The property that the quantum Fisher information is associated to the second derivatives of the fidelity can also be used to get an approximation to the quantum Fisher information matrix via a finite differences approach. We have introduced the quantum Fisher information matrix as the second derivative of twice the fidelity distance. With finite differences, we can approximate the second derivative of any function $g(t)$ as

$$\partial_t^2 g(t) \approx \frac{g(t+\epsilon)-2g(t)+g(t-\epsilon)}{\epsilon^2}. \tag{35}$$

If we now set

$$g(t) = 2d_f(\boldsymbol{\theta}, \boldsymbol{\theta}+t\boldsymbol{v}) \tag{36}$$

for some arbitrary unit vector $\boldsymbol{v}$, we see that $g(0)=0$ because d_f is a distance and $g(t+\epsilon)=g(t-\epsilon)$ for small ϵ, because the second order is the first non-vanishing order of the Taylor expansion of d_f. Putting all of this together allows us to compute the projection of the quantum Fisher information matrix in a particular direction:

$$\boldsymbol{v}^T\mathcal{F}\boldsymbol{v} \approx \frac{4d_f(\boldsymbol{\theta}, \boldsymbol{\theta}+\epsilon\boldsymbol{v})}{\epsilon^2}, \tag{37}$$

where $\boldsymbol{v}$ is an arbitrary vector of unit length and ϵ is small. We can therefore use quantum circuits that calculate the overlap between two pure states along with small perturbations to approximate the quantum Fisher information matrix. Note that you might also find other formulas that contain the square root of the fidelity. As argued in App. C, these formulas are equally valid and stem from a different convention for the fidelity distance.

Recently, Ref. [26] followed this spirit of approximation to generalize the *simultaneous perturbation stochastic approximation (SPSA)* method for the stochastic approximation of gradients to the calculation of the quantum Fisher information matrix. The original SPSA method computes estimates of the gradient of a function $f(\boldsymbol{\theta})$ using the finite differences approximation with small *random* perturbations of the parameters. The same strategy can be applied to the finite differences expression for the second derivative, where two distinct random perturbations are used to obtain an estimate of the Hessian matrix. Applied to the estimation of the quantum Fisher information matrix, the technique proceeds by first selecting two random vectors of unit length, $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. Then, for a small ϵ, the following quantity is computed:

$$\begin{aligned}\delta\mathcal{F} = f_{\boldsymbol{\theta}}(\epsilon\boldsymbol{v}_1 + \epsilon\boldsymbol{v}_2) - f_{\boldsymbol{\theta}}(-\epsilon\boldsymbol{v}_1) \\ - f_{\boldsymbol{\theta}}(-\epsilon\boldsymbol{v}_1 + \epsilon\boldsymbol{v}_2) + f_{\boldsymbol{\theta}}(+\epsilon\boldsymbol{v}_1).\end{aligned} \tag{38}$$

The shorthand $f_{\boldsymbol{\theta}}(\boldsymbol{\delta})$ denotes the fidelity of the system state with the state whose parameters are perturbed by $\boldsymbol{\delta}$,

$$f_{\boldsymbol{\theta}}(\boldsymbol{\delta}) = |\langle\psi(\boldsymbol{\theta})|\psi(\boldsymbol{\theta}+\boldsymbol{\delta})\rangle|^2. \tag{39}$$

From this quantity, we can use the outer products of the perturbation vectors to generate an approximation of the quantum Fisher information matrix as

$$\hat{\mathcal{F}} = -\frac{\delta\mathcal{F}}{2\epsilon^2}(\boldsymbol{v}_1\boldsymbol{v}_2^T + \boldsymbol{v}_2\boldsymbol{v}_1^T). \tag{40}$$

Note that the authors of Ref. [26] use a different convention for the quantum Fisher information matrix that causes a difference in the prefactor. This approximation is not enough to faithfully estimate the whole quantum Fisher information matrix because it has a maximal rank of 2. But we can average multiple such estimates to get an approximation of the quantum Fisher information matrix with increasing quality.

5 Relation of Classical and Quantum Fisher Information

To clarify the relation between classical and quantum Fisher information, we first return to the notion of *monotonicity* that we required for the distance measures we look at. Intuitively, we would expect that the associated information measure also "decreases" if a quantum channel is applied to the underlying quantum system.

And indeed, the monotonicity of the distance measure carries over to the associated information matrix. To elucidate *how* this happens, we will now use the notation of quantum channels and quantum states, because it includes the case of stochastic maps on classical probability distributions. The following reasoning is therefore valid for both the classical and the quantum Fisher information matrix. Recall that the information matrix arises at the second order approximation when we perturb the parameters of a quantum state slightly:

$$d(\rho(\boldsymbol{\theta}), \rho(\boldsymbol{\theta}+\boldsymbol{\delta})) = \frac{1}{2}\boldsymbol{\delta}^T M[\rho(\boldsymbol{\theta})]\boldsymbol{\delta} + O(\|\boldsymbol{\delta}\|^3). \tag{41}$$

The condition of monotonicity implies that the distance measure decreases if we apply some sort of quantum channel Φ, that can represent a variety of different operations:

$$d(\Phi[\rho(\boldsymbol{\theta})], \Phi[\rho(\boldsymbol{\theta}+\boldsymbol{\delta})]) \leq d(\rho(\boldsymbol{\theta}), \rho(\boldsymbol{\theta}+\boldsymbol{\delta})). \tag{42}$$

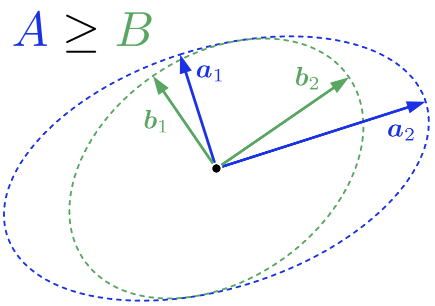


Figure 2: A visual explanation of the matrix inequality $A \geq B$ in the case of 2×2 matrices. To every positive matrix, we can assign an ellipse by considering its action on the unit sphere – in the 2D case the unit circle. The axes of the ellipse are then given by the eigenvectors scaled to the length of the associated eigenvector. The relation $A \geq B$ means that the ellipsis of B lies inside the ellipsis of A. If the ellipses do not touch, we have the stronger relation $A > B$.

But this also must hold in the limit $\|\boldsymbol{\delta}\| \to 0$, where we can drop higher order terms and only use the second order approximation of the distance. This means that

$$\frac{1}{2}\boldsymbol{\delta}^T M[\Phi[\rho(\boldsymbol{\theta})]]\boldsymbol{\delta} \leq \frac{1}{2}\boldsymbol{\delta}^T M[\rho(\boldsymbol{\theta})]\boldsymbol{\delta}. \tag{43}$$

We assumed that $\boldsymbol{\delta}$ was a very short vector to derive this inequality, but now we can also rescale it again to arbitrary length. This means that the above inequality holds for *any* $\boldsymbol{\delta}$ and that it implies the *matrix inequality*

$$M[\Phi[\rho(\boldsymbol{\theta})]] \leq M[\rho(\boldsymbol{\theta})]. \tag{44}$$

You might be a bit confused, as we are looking at matrices here and not numbers. The matrix inequality $A \geq B$ implies that the matrix $A - B$ has only non-negative eigenvalues, a statement that is equivalent to $\boldsymbol{\delta}^T A\boldsymbol{\delta} \geq \boldsymbol{\delta}^T B\boldsymbol{\delta}$ for any vector $\boldsymbol{\delta}$. A more visual explanation of this is given in Fig. 2.

We have shown that the monotonicity of the distance measure means that the associated information matrix also has a monotonicity property as it can also only decrease under quantum operations. This fact helps us to shine light on the relation between the quantum Fisher information and the classical Fisher information. Remember that we already learned that we can also model *measurements* as quantum channels. Together with the monotonicity of information matrices we just derived we know that applying a measurement $\mathcal{M}$ need necessarily make the information matrix associated to any monotonic distance smaller:

$$M[\mathcal{M}[\rho(\boldsymbol{\theta})] \leq M[\rho(\boldsymbol{\theta})]. \tag{45}$$

But the outcome of a measurement will always be a classical probability distribution over the measurement outcomes, ergo $\mathcal{M}[\rho(\boldsymbol{\theta})]$ is a classical probability distribution. And due to the uniqueness of the classical Fisher information we therefore know that no matter what kind of distance we used to derive M, after the measurement it will be a constant multiple of the classical Fisher information matrix associated with the probability distribution after the measurement:

$$M[\mathcal{M}[\rho(\boldsymbol{\theta})] = \alpha I[\mathcal{M}[\rho(\boldsymbol{\theta})]]. \tag{46}$$

As we are free to rescale the distance measure we use and therefore also the associated information matrix we can always make it consistent with the classical Fisher information when evaluated on classical probability distributions, which means that we can choose $\alpha = 1$, which we will also assume in the following. The relation is especially true for the quantum and the classical Fisher information, where we have that

$$I[\mathcal{M}[\rho(\boldsymbol{\theta})]] \leq \mathcal{F}[\rho(\boldsymbol{\theta})] \text{ for all } \mathcal{M}. \tag{47}$$

Let us quickly take a step back and marvel at the feat we just accomplished: We have used the simple requirement of monotonicity of the distance measure to show that any quantum information matrix is an upper bound to the classical Fisher information matrix associated with *any* probability distribution that results of a measurement of the state. We have initially motivated the derivation of this quantity via a geometric intuition, and it also pops up here: Indeed, a quantum information measure like the quantum Fisher information will measure how much the underlying quantum state $\rho(\boldsymbol{\theta})$ will change if we change the parameters of the state slightly. The relation we just derived tells us now that the change of the underlying quantum state directly gives us an upper bound on how much of this change we can make visible when performing a measurement.

This also gives us as a hint on what kinds of applications these two quantities enable. The quantum Fisher information really captures information about the underlying quantum state and therefore also phenomena that are quantum mechanical in nature. But the importance of the classical Fisher information should not be discounted, as we are always forced to perform measurements to extract information about the quantum system, which means that at the end of the day the classical Fisher information will always be the measure that quantifies the objects we can actually observe. This means that both quantities are very important for near-term applications, as the classical Fisher information captures the outputs of our experiments, but the quantum Fisher information can inform us about the quantum phenomena happening and also quantifies the ultimate limits of our approaches.

One might wonder if there always exists a measurement that achieves equality of the classical and the quantum Fisher information matrix. In the single-parameter case this is actually always possible [27], and consequently we can always find a measurement that achieves the quantum Fisher information for every *individual* parameter θ_i. The optimal measurement can be found by computing the SLD operator L_i that realizes the derivative of the underlying quantum state with respect to the parameter θ_i as in Eq. (29) [7] and can depend on the actual value of $\boldsymbol{\theta}$ [28]. But for multiple parameters we cannot necessarily find a measurement that achieves equality of classical and quantum Fisher information matrix, as the optimal measurements for the individual parameters need not be compatible with each other. A detailed discussion of the question of optimal measurements is found in Refs. [29, 30]. More detailed studies quantifying the (in)compatibility of different measurements can be found in Refs. [31–33].

6 The Role of Noise

As is already evident from the name *noisy* intermediate-scale quantum devices, noise is one of the principal impediments and a defining factor for NISQ devices. It is therefore imperative for us to understand how noise influences the classical and quantum Fisher information. We already got to know the concept of a quantum channel and learned that it can also be used to model any noisy quantum evolution.

But up to now, we have only treated the quantum Fisher information for pure quantum states, resting on the particularly simple formula for the fidelity for pure states. If we wish to extend it to mixed quantum states, we have to find a quantity that reproduces the known fidelity formula for pure states but also stays valid for mixed quantum states. This generalization is given by the *Bures fidelity*, also known as *Uhlmann's fidelity*[3]

$$f_B(\rho, \sigma) = \operatorname{Tr}\{(\rho^{1/2}\sigma\rho^{1/2})^{1/2}\}^2. \tag{48}$$

The notation $\rho^{1/2}$ denotes the unique positive *matrix square root* of ρ, which is the only positive semidefinite matrix that fulfills $(\rho^{1/2})^2 = \rho$. If we look at the eigendecomposition of ρ,

$$\rho = \sum_i \lambda_i |\lambda_i\rangle\langle\lambda_i|, \tag{49}$$

we can easily see that it is given by

$$\rho^{1/2} = \sum_i \sqrt{\lambda_i} |\lambda_i\rangle\langle\lambda_i|. \tag{50}$$

Constructions of the form $\rho^{1/2}\sigma\rho^{1/2}$ often appear in the context of quantum information theory because this is a way to multiply ρ and σ that – contrary to just using $\rho\sigma$ – yields a Hermitian matrix.

[3] Again, there exist different conventions if the square should be included or not.

As in the pure case, we need to transform the fidelity to get the proper distance measure associated to it, namely the *Bures distance*

$$d_B(\rho, \sigma) = 2 - 2 f_B(\rho, \sigma). \tag{51}$$

The prefactor 2 ensures that the the associated information matrix is consistent with the classical Fisher information matrix, as in Sec. 4.

We won't reproduce the whole derivation of the quantum Fisher information matrix for the Bures distance, as it is quite laborious. The interested reader can refer to Ref. [34] for a detailed derivation. The resulting formula is:

$$\begin{aligned}\mathcal{F}_{ij} &= \sum_{\substack{kl \\ \lambda_k + \lambda_l \neq 0}} \frac{2\operatorname{Re}(\langle \lambda_k | \partial_i \rho | \lambda_l \rangle \langle \lambda_l | \partial_j \rho | \lambda_k \rangle)}{\lambda_k + \lambda_l} && (52)\\ &= \sum_{\substack{k \\ \lambda_k \neq 0}} \frac{(\partial_i \lambda_k)(\partial_j \lambda_k)}{\lambda_k} + 4\lambda_k \operatorname{Re}(\langle \partial_i \lambda_k | \partial_j \lambda_k \rangle) \\ &\quad - \sum_{\substack{kl \\ \lambda_k, \lambda_l \neq 0}} \frac{8\lambda_k \lambda_l}{\lambda_k + \lambda_l} \operatorname{Re}(\langle \partial_i \lambda_k | \lambda_l \rangle \langle \lambda_l | \partial_j \lambda_k \rangle). && (53)\end{aligned}$$

Let us analyze this behemoth. First, we see that the equations contain a division by $\lambda_k + \lambda_l$. As ρ is a positive semidefinite matrix, the eigenvalues λ_k cannot be negative, which means that the case excluded in the sum can only occur if both λ_k and λ_l are zero. It is an advantage of the second equation that the sums now only run over λ_k and λ_l that are non-zero.

Now let's go over the two parts of the second equation. The first term,

$$\sum_{\substack{k \\ \lambda_k \neq 0}} \frac{(\partial_i \lambda_k)(\partial_j \lambda_k)}{\lambda_k}, \tag{54}$$

looks familiar. Recall that in the eigenbasis, ρ is diagonal and therefore represents a *classical* probability distribution over the basis states with probabilities λ_k – this means that the term above is the "classical" part of the quantum Fisher information and quantifies how the eigenvalues themselves change. Note that here we only sum over $\lambda_k \neq 0$ – which means that we effectively exclude the possibility of an λ_k being zero but $\partial_i \lambda_k$ being nonzero. In such a case, the rank of the density matrix would change which causes the quantum Fisher information to be undefined [17, 35]. This, however, is more of a technicality as there always exists a full-rank state arbitrarily close to any state we could be interested.

The next term now captures how much the *eigenstates* themselves change under the parameters θ_i and θ_j:

$$\begin{aligned}&\sum_{\substack{k \\ \lambda_k \neq 0}} 4\lambda_k \operatorname{Re}(\langle \partial_i \lambda_k | \partial_j \lambda_k \rangle) \\ &- \sum_{\substack{kl \\ \lambda_k, \lambda_l \neq 0}} \frac{8\lambda_k \lambda_l}{\lambda_k + \lambda_l} \operatorname{Re}(\langle \partial_i \lambda_k | \lambda_l \rangle \langle \lambda_l | \partial_j \lambda_k \rangle).\end{aligned} \tag{55}$$

This constitutes the quantum part of the quantum Fisher information. The changing of the eigenvectors is a non-classical phenomenon as they are always fixed and identified with the different measurement outcomes for classical states.

Calculation. We have seen that the noisy quantum Fisher information is much more complicated than the quantum Fisher information for pure states. To evaluate it exactly one usually has to perform a full tomography of the underlying state, an operation that is too costly for near-term applications because the number of samples is exponential in the number of qubits [36]. For quantum states that are nearly pure approximations can be used, but they also break down at certain noise levels [37, 38].

Recently, variational approaches for the computation of the Bures fidelity f_B have been suggested. They can be used to compute the quantum Fisher information for noisy states in conjunction with the perturbation techniques described in Sec. 4. Ref. [39] proposes to alleviate the resource requirements for tomography through the use of a variational quantum autoencoder. A variational autoencoder compresses an N-qubit input state into K qubits [40]. The authors of Ref. [39] show that the compressed state of a perfectly trained autoencoder has the same spectrum as the original state. They propose to estimate the Bures fidelity between two states ρ and σ by first performing tomography of the compressed state of ρ and then running separate SWAP tests that involve the eigenstates of ρ and the state σ. The fidelity is then computed in classical post-processing and guarantees on the estimation precision are given based on how well the autoencoder was trained. The authors provide numerical evidence that the proposed strategy works well for low-rank states if the variational circuit for the autoencoder is suitably chosen.

Ref. [41], on the other hand, proposed to exploit Uhlmann's theorem that states that the Bures fidelity of two states ρ and σ is equal to the maximum fidelity over all possible purifications $|\psi_\rho\rangle$ and $|\psi_\sigma\rangle$:

$$f_B(\rho, \sigma) = \max_{|\psi_\rho\rangle, |\psi_\sigma\rangle} f(|\psi_\rho\rangle, |\psi_\sigma\rangle). \tag{56}$$

The authors suggest to use a variational circuits to learn purifications of the input states ρ and σ and then perform another optimization to extract the maximum in Eq. (56). Another approach for the calculation of the Bures fidelity was put forward in Ref. [42]. It is based on a different subroutine, *purity minimization*, that can be used to calculate the matrix square roots in Eq. (48). The two latter approaches require multiple copies of the input states and therefore incur an overhead that is large for NISQ applications. They furthermore rely on the success of variational subroutines which is closely tied to the successful selection of a circuit ansatz. These approaches are therefore only suitable to calculate the quantum Fisher information for small systems.

Ref. [43] introduced the *truncated quantum Fisher information (TQFI)* as a way to approximate the quantum Fisher information by only including the quantum state's m eigenvectors with the largest eigenvalues in the computation of the quantum Fisher information. For increasing m, the approximations obtained via the truncated approach get closer to the true quantum Fisher information. Recent efforts showed that the TQFI and other upper and lower bounds on the quantum Fisher information can be effectively evaluated on NISQ devices via variational procedures [44].

The authors of Ref. [45] suggest another strategy in the same vein. They propose a hierarchy of lower bounds $\mathcal{F}_n$ based on a series expansion of the quantum Fisher information that can be computed using randomized measurements [46]. The first level, $\mathcal{F}_0$, already appeared in Refs. [47, 48] and represents a tighter bound than the sub-quantum Fisher information of Ref. [44]. With increasing n, the complexity of computing the lower bound increases, but the distance of the bound to the true quantum Fisher information decreases exponentially in n.

7 NISQ Applications

We will now survey the variety of different contexts in which Fisher information popped up related to NISQ devices. The fact that these application are widely different shows the value of these techniques.

7.1 Quantum Sensing

Quantum sensing, also known as quantum metrology or quantum parameter estimation, is one of the main pillars of near-term quantum technologies. It is the area of application in which the classical and quantum Fisher information have been explored the most. We will explore quantum sensing to the degree necessary to understand the role of Fisher information, but, as the focus of this work is on near-term applications *beyond* quantum sensing, we will not delve deeper into applications. But before we can do that, we first have to set the stage.

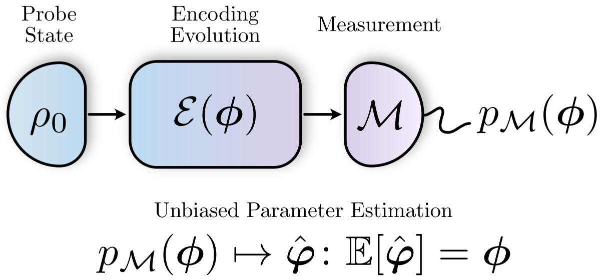

Probe State
Encoding Evolution
Measurement
ρ_0
$\mathcal{E}(\boldsymbol{\phi})$
$\mathcal{M}$
$p_{\mathcal{M}}(\boldsymbol{\phi})$
Unbiased Parameter Estimation
$p_{\mathcal{M}}(\boldsymbol{\phi}) \mapsto \hat{\boldsymbol{\varphi}} \colon \mathbb{E}[\hat{\boldsymbol{\varphi}}] = \boldsymbol{\phi}$

Figure 3: Mathematical representation of a quantum sensing experiment. A probe state ρ_0 undergoes an interaction with the environment that imprints the physical parameters $\boldsymbol{\phi}$ we want to sense onto the state. To extract information about the parameters we perform a measurement $\mathcal{M}$ of the final state, yielding a probability distribution $p_{\mathcal{M}}(\boldsymbol{\phi})$ that depends on both the chosen measurement and the physical parameters. From this probability distribution, an estimator of the physical parameters, $\hat{\boldsymbol{\varphi}}$ is constructed.

Introduction. The object of quantum sensing is to measure some physical quantity, say a magnetic field, pressure or temperature, which we will denote as $\boldsymbol{\phi}$. We consider a vector-valued quantity because we often want to measure multiple things, *e.g.* all components of the magnetic field. The measurement proceeds by preparing a physical system that interacts with its environment so that the physical parameter is imprinted on it. Consider the example of a magnetic field: to find out what the components are we can use one or multiple spins that undergo precession depending on the strength of the magnetic field and the alignment of the spin relative to the magnetic field.

In the end, we can model this whole process by taking a quantum system in an initial state ρ_0 which then undergoes a quantum channel $\mathcal{E}(\boldsymbol{\phi})$ that depends on the physical parameters. The result will again be a parametrized state:

$$\rho(\boldsymbol{\phi}) = \mathcal{E}(\boldsymbol{\phi})[\rho_0]. \tag{57}$$

As the state ρ_0 "probes" the environment, it is called the *probe state*. Note also that we are actually back in the framework of parametrized quantum states with which we started our discussion in Sec. 1, only that the parameters this time are the physical parameters we want to measure instead of some tunable parameters of the state preparation.

The story is of course not over here. We want to use the obtained quantum state $\rho(\boldsymbol{\phi})$ to learn as much as we can about the underlying parameters $\boldsymbol{\phi}$. To this end, we have to perform some measurement $\mathcal{M}$. After the measurement, we are left with a probability distribution over measurement outcomes that depends on the physical parameters and the chosen measurement, $p_{\mathcal{M}}(\boldsymbol{\phi})$. If we make an estimate of the underlying parameters from the observed measurement statistics we formally construct an *estimator*, denoted as $\hat{\boldsymbol{\varphi}}$. As the measurement results are random, our estimate will necessarily also be and the estimator $\hat{\boldsymbol{\varphi}}$ is therefore a *random vector*. A property of an estimator that we can aim for is that it is *unbiased*, *i.e.* that the predictions are correct in expectation: $\mathbb{E}[\hat{\boldsymbol{\varphi}}] = \boldsymbol{\phi}$. There are also other notions of "unbiasedness" that are less restrictive: An estimator can be *locally unbiased*, which means it is only unbiased in the neighborhood of a certain parameter value $\boldsymbol{\phi}_0$. This property is of interest if one already has prior knowledge about the underlying parameter, *e.g.* from previous measurements. Furthermore, an estimator can be *asymptotically un-*

biased, *i.e.* it is unbiased in the limit of infinitely many samples. The whole formalization of a sensing experiment is depicted in Fig. 3.

Quantifying Sensing Performance. The classical Fisher information matrix comes into play when we want to quantify how good an estimator can be if we perform our experiment n times. This manifests itself in the *Cramér-Rao bound*, which is central to the field of quantum sensing and quantifies the best attainable performance of an unbiased estimator:

$$\operatorname{Cov}[\hat{\boldsymbol{\varphi}}] \geq \frac{1}{n} I_{\mathcal{M}}(\boldsymbol{\phi})^{-1}. \tag{58}$$

This is again a matrix inequality, as explained in Fig. 2. Let us decode this inequality. On the left hand side there is the *covariance matrix* of our estimator, whose entries are the covariances of the separate components

$$\operatorname{Cov}[\hat{\boldsymbol{\varphi}}]_{ij} = \mathbb{E}[\hat{\varphi}_i \hat{\varphi}_j] - \mathbb{E}[\hat{\varphi}_i]\mathbb{E}[\hat{\varphi}_j]. \tag{59}$$

As we want to be sure of our estimates, we want our estimator to vary as little as possible, which means that the covariances should be as small as possible. We can also connect the covariance matrix to the *mean-squared error* of our estimate

$$\operatorname{MSE}[\hat{\boldsymbol{\varphi}}] = \mathbb{E}[\|\hat{\boldsymbol{\varphi}} - \boldsymbol{\phi}\|^2] = \operatorname{Tr}\{\operatorname{Cov}[\hat{\boldsymbol{\varphi}}]\}, \tag{60}$$

which can be more easily interpreted as a performance measure. Note that this identity only holds for unbiased estimators.

The right hand side of the Cramér-Rao bound Eq. (58) imposes a fundamental limit on how small the covariances can get. And here we find the inverse of the classical Fisher information matrix $I_{\mathcal{M}}(\boldsymbol{\phi})$. Note that we denoted the chosen measurement explicitly, because the classical Fisher information matrices for different measurements usually do not coincide.

The appearance of the Fisher information has an intuitive explanation: Remember that large entries of the Fisher information indicate that a change of the associated parameters results in a large change of the underlying probability distribution. This would be desirable for the purpose of estimation because this means that a probability distribution we observe can be associated with a certain parameter with higher confidence – if the parameter was different the probability distribution would also be notably different. As the Cramér-Rao bound is concerned with the covariance, a quantity that we want to be as small as possible, we need the *inverse* of the Fisher information matrix to account for that.

A property that makes the scalar Cramér-Rao bound very useful is that it can *always* be saturated in the limit of infinitely many samples. In this limit, a process called maximum likelihood estimation is guaranteed to give an optimal estimate. This is appealing, because it means that we can stop worrying about how to actually construct a good estimator because we know there must be one that achieves the Cramér-Rao bound.

Optimal Sensing. We see that a sensing procedure has two levers we can pull to optimize it: we want to choose a probe state that is "maximally susceptible" to the evolution $\mathcal{E}(\boldsymbol{\phi})$ and we want to find a measurement $\mathcal{M}$ that extracts as much of this information from the quantum state.

We learned in Sec. 5 that the quantum Fisher information matrix gives an upper bound to the classical Fisher information matrices that can be obtained by measuring the underlying quantum system. This upper bound on the classical Fisher information matrix directly implies a more fundamental lower bound on the attainable precision, the so-called *quantum Cramér-Rao bound* [49, 50]

$$\operatorname{Cov}[\hat{\boldsymbol{\varphi}}] \geq \frac{1}{n} I_{\mathcal{M}}(\boldsymbol{\phi})^{-1} \geq \frac{1}{n} \mathcal{F}(\boldsymbol{\phi})^{-1}. \tag{61}$$

The quantum Cramér-Rao bound has the advantage that it is independent of the chosen measurement and only depends on the probe state. We can therefore formulate an intuitive recipe to find an optimal quantum sensing scheme: first try to find the probe state with the largest quantum Fisher information matrix and then optimize the POVM to bring the classical Fisher information matrix as close to the quantum one as possible.

As pointed out in Sec. 5, we are often unable to construct a measurement for which $I_{\mathcal{M}} = \mathcal{F}$ due to the possible incompatibility of the parameters ϕ_i. But how can we then decide which possible measurement is "best"? We have just learned that we can construct an estimator that achieves the classical Cramér-Rao bound for any measurement. But we can usually not use the covariance matrix to quantitatively compare the performance of two estimators $\hat{\boldsymbol{\varphi}}_1$ and $\hat{\boldsymbol{\varphi}}_2$ corresponding to different measurements. This is because it can be that neither $\operatorname{Cov}[\hat{\boldsymbol{\varphi}}_1] \geq \operatorname{Cov}[\hat{\boldsymbol{\varphi}}_2]$ nor $\operatorname{Cov}[\hat{\boldsymbol{\varphi}}_2] \geq \operatorname{Cov}[\hat{\boldsymbol{\varphi}}_1]$. This can be intuitively understood by looking at Fig. 2: if neither ellipse corresponding to the two covariance matrices is contained in the other, we cannot make a statement of one being "larger" than the other.

To still be able to make a comparison we need a scalar quantity. To get one, we can generalize the idea of Eq. (60) and perform a trace of the covariance matrix with the addition of a positive semidefinite *weight matrix* W. As the name suggests, we can use it to put additional emphasis on certain parameters or just use the identity matrix to perform an equal weighting. In this case, we get a scalar quantity that quantifies the performance of an estimator and that fulfills a scalar (quantum) Cramér-Rao bound:

$$\begin{aligned}\mathrm{Tr}\{W\,\mathrm{Cov}[\hat{\boldsymbol{\varphi}}]\} &\geq \frac{1}{n}\,\mathrm{Tr}\{W I_{\mathcal{M}}(\boldsymbol{\phi})^{-1}\} \\ &\geq \frac{1}{n}\,\mathrm{Tr}\{W \mathcal{F}(\boldsymbol{\phi})^{-1}\}.\end{aligned} \tag{62}$$

We should note that the quantum Cramér-Rao bound is not the end of the story. If we consider a weighting of the estimator's covariance matrix as in Eq. (62), the *Holevo Cramér-Rao bound (HCRB)* gives the ultimate limit to estimation precision and is attainable in the limit of infinitely many samples [51]. It is computed by solving an optimization problem and does not involve the quantum Fisher information matrix. It can, however, not exceed the quantum Cramér-Rao by more than a factor of 2 [52] and is equal to it when the underlying parameters can be estimated simultaneously.

For a review on multi-parameter quantum sensing, have a look at Refs. [7, 8, 53]. Ref. [8] that provides an exhaustive review with a particular focus on the underlying geometry.

Exploiting Quantum Effects. The (quantum) Cramér-Rao bound contains a factor $1/n$ which accounts for the fact that we can simply decrease the variance of our estimates by averaging over n independent repetitions of the same experiment. The scaling

$$\mathrm{Cov}[\hat{\boldsymbol{\varphi}}] \propto \frac{1}{n} \tag{63}$$

is called the *standard quantum limit (SQL)* or *shot-noise limit*.

But the approach of just performing independent repetitions of the same experiment does not make use of one of the most crucial properties of quantum mechanics, namely *entanglement*. To elucidate how we can make use of it, we have a look at a simple sensing task where we want to measure the rate Δ at which a qubit acquires a phase. In our experiment every single-qubit probe acquires a phase $\phi = \Delta t$ between the $|0\rangle$ and the $|1\rangle$ state. If we perform n repetitions separately, we obtain the scaling of the standard quantum limit. But we can also entangle the n probes into a generalized GHZ state:

$$|\mathrm{GHZ}_n\rangle = \frac{1}{\sqrt{2}}\,|\underbrace{00\ldots0}_{n\text{ times}}\rangle + \frac{1}{\sqrt{2}}\,|\underbrace{11\ldots1}_{n\text{ times}}\rangle \tag{64}$$

$$= \frac{1}{\sqrt{2}}\,|0_n\rangle + \frac{1}{\sqrt{2}}\,|1_n\rangle\,. \tag{65}$$

Under the sensing interaction, the state evolves to:

$$\frac{1}{\sqrt{2}}\,|0_n\rangle + e^{-in\Delta t}\frac{1}{\sqrt{2}}\,|1_n\rangle\,. \tag{66}$$

Effectively, we have enhanced the signal by a factor of n compared to a single probe.

This factor of n enhancement lies at the heart of quantum advantage in sensing. You might rightfully ask why this is the case, because we also got a factor n enhancement from just performing n separate repetitions. But it turns out that the fact that we enhanced the *signal* by a factor of n actually gives us a factor n^2 improvement in the Fisher information! Recall that the classical Fisher information with respect to the parameter Δ is given by

$$I(\Delta) = \sum_l \frac{(\partial_\Delta p_l)^2}{p_l}. \tag{67}$$

By enhancing the signal by a factor of n we actually changed the rate with which it changes – the derivative with respect to Δ – by a factor of n compared to a single repetition. Because the Fisher information contains the *square* of the derivative, this effectively gives us the enhancement by a factor of n^2.

The scaling of this entangled approach,

$$\mathrm{Cov}[\hat{\boldsymbol{\varphi}}] \propto \frac{1}{n^2}, \tag{68}$$

is actually the most fundamental limit attainable when exploiting quantum mechanical effects and is called the *Heisenberg limit* [54].

But there is a catch. If we consider realistic sensing problems with noise we see that the generalized GHZ state we just used to achieve Heisenberg scaling is actually no better than the simple single-qubit strategy [55], even if there is only an infinitesimal amount of noise. This is due to the fact that not only the signal but also the noise itself gets "amplified", therefore canceling out the advantage from quantum entanglement. But not all hope is lost as Ref. [55] also showed that other, less entangled states, can still give an advantage in the noisy regime. More recently, it was shown in Ref. [56] that one can only hope to get a constant factor improvement to the standard quantum limit for many noise models. To regain the quantum advantage, one needs to combine quantum error correction with quantum sensing. Indeed, if the noise acts sufficiently different than the signal, we can retain the Heisenberg scaling by using a *metrological code* [57, 58]. For certain noise models, scalings between standard and Heisenberg scaling can be achieved [59].

Estimation of Expectation Values. Estimating expectation values is a key subroutine of variational quantum algorithms. Recently, the authors of Refs. [60, 61] used reasoning from quantum sensing to find a way to reduce the number of evaluations of a NISQ experiment necessary to faithfully evaluate an expectation values at the cost of increased circuit depth.

The approach is based on the observation that the estimation quality of an expectation value of a Pauli observable $\langle P\rangle$ is also governed by the Cramér-Rao bound. We can use the same constructions as we have already outlined but identify the unknown parameter with the expectation value: $\phi = \langle P \rangle$. If we want to achieve a fixed precision we can reduce the number of experimental repetitions necessary to achieve it by increasing the classical Fisher information with respect to the parameter $\langle P \rangle$. The construction of Refs. [60, 61] is enabled by the use of a Bayesian approach, where an initial estimate about the distribution of $\langle P \rangle$ is updated with new information from experiments. The classical Fisher information is increased by a subroutine inspired from Grover's famous algorithm. The authors call this approach *Engineered Likelihood Functions*, as the subroutines inspired by Grover's algorithm have parameters that are adjusted to yield the highest possible Fisher information for the specific expectation value that shall be estimated.

Variational Algorithms for Sensing. As metrology is such a fundamental application, it is important to construct optimal protocols. This is rather difficult because noise and device limitations have to be taken into account. But recently the idea of using NISQ devices themselves for the optimization of quantum sensing schemes has gained some attention and a number of variational algorithms for that purpose have been put forward.

In the above exposition of quantum sensing, we looked only at the physical parameters $\boldsymbol{\phi}$ as parameters of the state. But nothing holds us back from parametrizing the probe state as well, maybe in a simulation on a NISQ computer or by providing a tunable state preparation in an experiment. The same also holds for the measurement – we can fix a certain detection scheme but precede it by a unitary, again with parametrized gates. If we denote the parameters of the probe state preparation with $\boldsymbol{\theta}$ and the parameters of the measurement with $\boldsymbol{\mu}$, we work with a parametrized state $\rho(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\mu})$.

To perform a variational algorithm we also need a *cost function* that measures how well we do with our estimation. The first works in this direction considered the case of estimating only a single parameter. This line of work was started in Ref. [62], which considered the probe state's *spin squeezing* as a cost function, as it acts as surrogate for the sensing precision. Subsequently, the authors of Ref. [63] used the change of the fidelity under small perturbation of the sensing parameter to estimate and optimize the quantum Fisher information, a method outlined in detail in Sec. 4. Note that looking at the quantum Fisher information is interesting because it captures the best achievable performance of the sensing scheme, but it also removes the performed measurement from the equation, as the quantum Fisher information is independent of $\boldsymbol{\mu}$. The authors of Ref. [64] argue that this will not be desirable in realistic applications where measurement capabilities are limited and the optimal measurement cannot be realized. Instead, they propose to use a cost function based on the classical Fisher information and extend the proposal to multi-parameter estimation. This technique was further improved in Ref. [65] by combining it with an additional optimization over the parametrizations of the probe state and the measurement.

Another strategy to tackle this problem was put forward in Ref. [44], where efficient schemes to evaluate different bounds on the quantum Fisher information were proposed and applied to the optimization of quantum sensing protocols. In a follow-up work, it was proven that the global optima of a specific bound on the quantum Fisher, the *sub-quantum Fisher information*, first derived in Ref. [66], coincide with that of the quantum Fisher information itself [67], underscoring its quality as a surrogate for the quantum Fisher information.

7.2 Quantum Natural Gradient Descent

Another intriguing application of Fisher information can be found in the field of variational quantum algorithms. In these applications, we desire to minimize a cost function C that depends on some expectation values evaluated on a parametrized quantum state $|\psi(\boldsymbol{\theta})\rangle$, rendering the cost function a function of the circuit parameters, $C = C(\boldsymbol{\theta})$.

A popular way to minimize the cost function is by using *gradient descent*. In this scheme, we start from an initial guess of parameters $\boldsymbol{\theta}^{(0)}$ and perform multiple optimization steps where we update $\boldsymbol{\theta}$ according to the rule

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta \nabla C(\boldsymbol{\theta}^{(t)}). \tag{69}$$

The gradient of $C(\boldsymbol{\theta})$ always points in the direction of the steepest increase of the cost function. Subtracting it therefore ensures we go into a direction where the cost function decreases. The parameter η is the *step size* that controls how far we step.

The authors of Ref. [23] proposed a way to make use of the knowledge about the underlying quantum states encoded in the quantum Fisher information matrix $\mathcal{F}$ in the context of optimization, namely *quantum natural gradient descent*. In this approach, the gradient descent update rule is modified by the use of the inverse of the quantum Fisher information matrix:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta \mathcal{F}(\boldsymbol{\theta})^{-1} \nabla C(\boldsymbol{\theta}^{(t)}). \tag{70}$$

Let us gather some intuitive understanding of this approach. As we have seen in Eq. (4), the fidelity distance is given by

$$d_f(\boldsymbol{\theta}, \boldsymbol{\theta} + \boldsymbol{\delta}) = \frac{1}{4} \boldsymbol{\delta}^T \mathcal{F} \boldsymbol{\delta} \tag{71}$$

in leading order, see also Sec. 4. This means that large entries of the quantum Fisher information matrix correspond to large changes in the distance and small entries to small changes. By taking the inverse of the quantum Fisher information matrix we thus *normalize* the gradient step we take: directions that correspond to large changes of the underlying quantum state are scaled down while directions that correspond to small changes are scaled up.

But we can also give a beautiful mathematical justification of this approach. This can be seen by analyzing the origin of the original gradient descent rule. We can recast the next set of parameters $\boldsymbol{\theta}^{(t+1)}$ as the solution to the following optimization problem:

$$\operatorname*{argmin}_{\boldsymbol{\theta}} \left\{ \langle \boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}, \nabla C(\boldsymbol{\theta}^{(t)}) \rangle + \frac{1}{2\eta} \| \boldsymbol{\theta} - \boldsymbol{\theta}^{(t)} \|_2^2 \right\}. \tag{72}$$

This includes a term that ensures we go in the opposite direction of the gradient and a so-called *regularization term* that ensures that we do not step too far, thereby controlling the step size.

But we have already learned that we can use the quantum Fisher information matrix to measure lengths of vectors, thereby including our knowledge about the underlying quantum state. We thus replace the 2-norm in the regularization term with the norm induced by $\mathcal{F}$, $\|\boldsymbol{\theta}\|_{\mathcal{F}}^2 = \boldsymbol{\theta}^T \mathcal{F} \boldsymbol{\theta}$ – which is for small step sizes approximately equal to the expected fidelity distance d_f up to a constant factor. This means that we do no longer penalize large distances in parameter space, but instead in the space of quantum states! The solution to the new optimization problem

$$\operatorname*{argmin}_{\boldsymbol{\theta}} \left\{ \langle \boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}, \nabla C(\boldsymbol{\theta}^{(t)}) \rangle + \frac{1}{2\eta} \| \boldsymbol{\theta} - \boldsymbol{\theta}^{(t)} \|_{\mathcal{F}}^2 \right\} \tag{73}$$

is nothing but the update step of quantum natural gradient descent from Eq. (69).

It was shown in Ref. [68] that this method actually provides an advantage in optimizing realistic quantum systems when large systems are concerned due to it taking different optimization paths than optimization strategies that do not relate to the underlying quantum states.

A weak point of quantum natural gradient descent is that the quantum Fisher information matrix has to be calculated at each step, which can be a very costly endeavor, as also outlined in Sec. 4. The authors of Ref. [69], however, argue that the additional cost of estimating the quantum Fisher information matrix is negligible in gradient-descent applications. This is because the number of shots necessary to faithfully evaluate the gradient of the cost function itself increases as one approaches a minimum whereas the cost of estimating the quantum Fisher information does not.

To reduce the complexity one can also resort to stochastic approximations of the quantum Fisher information matrix as proposed in Ref. [26]. The authors use a variation of SPSA to approximate the quantum Fisher information matrix, as was already introduced in Sec. 4. The authors also use the assumption that the quantum Fisher information matrix only changes slowly between iterations to perform a smooth approximation of the quantum Fisher information matrix. In this approach, the approximation of the matrix is held in memory and the new stochastic approximations obtained at the current optimization step are used to update this approximation. This approach has the downside that it is not guaranteed that this approximation of the quantum Fisher information matrix has only positive eigenvalues, but the authors introduce measures to mitigate this issue.

Another issue with quantum natural gradient descent arises when we consider noisy quantum devices. As we already learned in Sec. 6, the true quantum Fisher information is very hard to calculate for a noisy quantum state. The authors of Ref. [38] argue that at least in the case of states that are not too noisy, approximations to the quantum Fisher information matrix are sufficient.

7.3 Analyzing Quantum Learning Models

Quantum Machine Learning (QML) [70, 71] is a research field with increasingly growing traction. The field searches both for applications of quantum computers to improve machine learning techniques, as well as applications of machine learning to analyze quantum systems. A special focus lies on variational learning techniques useful in NISQ applications.

The classical Fisher information has already been successfully used in classical machine learning to analyze learning models like neural networks. A particular example is the *Fisher-Rao norm* [72], which is nothing else than the norm induced by the classical Fisher information matrix that we already encountered:

$$\|\boldsymbol{\delta}\|_{\mathrm{fr}}^2 = \boldsymbol{\delta}^T I \boldsymbol{\delta}. \tag{74}$$

The Fisher-Rao norm is treated as a measure that correlates with the *capacity* of a learning model, which measures how complicated the relationships a learning model can express are.

The authors of Ref. [73] have looked at the classical Fisher information of a parametrized quantum circuit to quantify its capacity [73]. They define a new capacity measure they call the *effective dimension* which can be used to bound how well a variational quantum learning model can generalize on unseen data. The effective dimension also takes into account how many datapoints are available to train a learning model. In the limit of infinitely many datapoints, the effective dimension is equal to the number of non-zero eigenvalues – the rank – of the classical Fisher information matrix. A higher effective dimension is associated with a "flatter" eigenvalue spectrum of the classical Fisher information matrix. The authors also provide numerics that suggest that this "flatness" is a generic feature of certain quantum learning models.

Recently, the authors of Ref. [74] introduced the *effective quantum dimension* as a measure of how much of the underlying quantum state space a quantum learning model can explore. The effective quantum dimension as proposed in this work is equal to the rank of the quantum Fisher information matrix. It does not take into account the number of available datapoints and should therefore not be confused with the quantum generalization of the effective dimension of Ref. [73]. It is still an intuitive measure as the rank directly captures in how many directions a varying of the parameters will also result in a varying of the underlying quantum state. If the effective quantum dimension is lower than the number of parameters, then some of the parameters are redundant and the model is therefore overparametrized. The numerical investigations of the authors confirm that the specific choice of the circuit parametrization leads to different effective quantum dimensions of the resulting quantum learning models.

8 Beyond NISQ

The focus of this work is to showcase and explore the application of classical and quantum Fisher information in the NISQ context. But these tools have of course very interesting applications beyond that. We want to give two notable mentions that display how versatile these tools are:

In Ref. [75], techniques from quantum sensing have been used to prove how large "quantum programs" need to be to allow a quantum computer to perform a unitary with given target precision. They construct an optimal quantum program interpreter that basically estimates the unitary that should be performed from the quantum program, therefore putting it in the realm where the tools from quantum metrology can be used to prove that the approach matches a lower bound.

Another very interesting application arose recently when other techniques from quantum metrology were used to to provide a new and simple proof for the approximate Eastin-Knill theorem in quantum error correction [76]. The theorem shows that certain classes of error correcting codes with very favorable properties cannot exist. The new proof uses upper bounds on the quantum Fisher information to show that an error correcting code that violates the Eastin-Knill theorem would allow for too large Fisher information.

Recently, Ref. [77] showed that the classical and quantum Fisher information can be used in the framework of quantum resource theories [78]. The authors propose a general way to construct a parameter estimation task for which the presence of a resource gives an advantage. As this construction is generic, it also allows for the converse reasoning: every quantum resource is useful for metrology, because there exists a parameter estimation task for which it provides an advantage.

It is also worth noting that the review by Liu *et al.* [7] contains a chapter of other applications of the quantum Fisher information in the contexts of quantum thermodynamics, quantum speed limits and the study of non-Markovianity.

9 Outlook

Both the classical and the quantum Fisher information capture important information about the parametrized quantum systems that lie at the heart of many applications of near-term quantum devices. They can therefore further our understanding of the capabilities of these techniques and guide us in the development of new approaches. We hope that the present work will motivate some of its readers to incorporate the classical and quantum Fisher information into their practical and theoretical toolboxes and that it can inspire new uses of these tools in the context of NISQ devices.

As discussed in the main text, there have been many developments aiming to simplify the calculation of both the classical and quantum Fisher information on NISQ hardware. To further increase the applicability, it will, however, be necessary to further improve on these techniques.

While the classical Fisher information is in principle easily accessible from the output probability distributions, there is still a need for rigorous analyses of different estimation strategies. A promising avenue, especially for variational applications, would be to consider Bayesian techniques, as the classical Fisher information is expected not to change too rapidly if circuit parameters are updated incrementally. Another important development would be rigorous bounds on the number of samples needed to faithfully estimate the classical Fisher information that could guide the computation in practical applications.

The quantum Fisher information on the other hand suffers from problems related to its higher complexity. An especially impactful development would be a technique that allows for the efficient calculation of the quantum Fisher information for noisy quantum states that are encountered in relevant settings. Machine learning approaches that have recently been developed to capture the properties of quantum states [14] could possibly be useful in this regard.

As classical and quantum Fisher information are so intimately tied to the structure of parametrized quantum states, we expect them to be useful in many more applications than discussed in this paper.

An interesting direction is to further understand the properties of learning models that are constructed from parametrized quantum circuits through the lens of quantum Fisher information. Especially generalization bounds that are "truly quantum" would be of high interest, as they could capture the ultimate limits of quantum-enhanced machine learning models and lead to a better understanding of the influence of noise. It would be especially important in that regard to derive bounds that take into account the dichotomy between circuit parameters that represent data inputs and circuit parameters that are trainable.

It would furthermore be intriguing to see if more tools that were developed in the context of quantum sensing, *e.g.* measures of parameter incompatibility, could be applied to the analysis of near-term applications. As variational quantum algorithms usually perform measurements in many bases to estimate expectation values of relevant observables, one possible directions would be to analyze the difference between classical and quantum Fisher information as this quantifies the information loss associated with specific measurements.

The fact that both classical and quantum Fisher information are susceptible to noise also suggests that they could be applicable to near-term quantum error correction and error mitigation, both in practical applications as well as theoretical tools to prove rigorous mathematical statements.

Another alluring avenue of research is to develop applications of monotone metrics beyond the quantum Fisher information, like the Wigner-Yanase information or the Kubo-Mori information. While these quantities do not provide tighter bounds in the quantum Cramér-Rao bound, they can give tighter bounds in other contexts [79]. It is natural to ask if generalizations of the approaches already developed with the quantum Fisher information in mind, *e.g.* the quantum natural gradient descent, also perform well when it is replaced with a different monotone metric. An important prerequisite for such applications would be to find efficient ways to calculate those other metrics in practical applications.

Acknowledgments

The author would like to thank Jacob Beckey, Francesco Albarelli, Daniel Stilck-França, Nathan Killoran and Josh Izaac for providing feedback that greatly improved the quality of the manuscript.

The author acknowledges funding from the German Federal Ministry for Economic Affairs and Energy under the PlanQK initiative.

References

[1] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. S. L. Brandao, D. A. Buell, B. Burkett, Y. Chen, Z. Chen, B. Chiaro, R. Collins, W. Courtney, A. Dunsworth, E. Farhi, B. Foxen, A. Fowler, C. Gidney, M. Giustina, R. Graff, K. Guerin, S. Habegger, M. P. Harrigan, M. J. Hartmann, A. Ho, M. Hoffmann, T. Huang, T. S. Humble, S. V. Isakov, E. Jeffrey, Z. Jiang, D. Kafri, K. Kechedzhi, J. Kelly, P. V. Klimov, S. Knysh, A. Korotkov, F. Kostritsa, D. Landhuis, M. Lindmark, E. Lucero, D. Lyakh, S. Mandrà, J. R. McClean, M. McEwen, A. Megrant, X. Mi, K. Michielsen, M. Mohseni, J. Mutus, O. Naaman, M. Neeley, C. Neill, M. Y. Niu, E. Ostby, A. Petukhov, J. C. Platt, C. Quintana, E. G. Rieffel, P. Roushan, N. C. Rubin, D. Sank, K. J. Satzinger, V. Smelyanskiy, K. J. Sung, M. D. Trevithick, A. Vainsencher, B. Villalonga, T. White, Z. J. Yao, P. Yeh, A. Zalcman, H. Neven, and J. M. Martinis, "Quantum supremacy using a programmable superconducting processor," Nature **574**, 505–510 (2019) DOI: 10.1038/s41586-019-1666-5.

[2] J. Preskill, "Quantum computing in the NISQ era and beyond," Quantum **2**, 79 (2018) DOI: 10.22331/q-2018-08-06-79.

[3] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, and P. J. Coles, "Variational quantum algorithms," arXiv:2012.09265 (2020).

[4] K. Bharti, A. Cervera-Lierta, T. H. Kyaw, T. Haug, S. Alperin-Lea, A. Anand, M. Degroote, H. Heimonen, J. S. Kottmann, T. Menke, W.-K. Mok, S. Sim, L.-C. Kwek, and A. Aspuru-Guzik, "Noisy intermediate-scale quantum (NISQ) algorithms," arXiv:2101.08448 (2021).

[5] J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, "The theory of variational hybrid quantum-classical algorithms," New Journal of Physics **18**, 023023 (2016) DOI: 10.1088/1367-2630/18/2/023023.

[6] M. Benedetti, E. Lloyd, S. Sack, and M. Fiorentini, "Parameterized quantum circuits as machine learning models," Quantum Science and Technology **4**, 043001 (2019) DOI: 10.1088/2058-9565/ab4eb5.

[7] J. Liu, H. Yuan, X.-M. Lu, and X. Wang, "Quantum fisher information matrix and multiparameter estimation," Journal of Physics A: Mathematical and Theoretical **53**, 023001 (2020) DOI: 10.1088/1751-8121/ab5d4d.

[8] J. S. Sidhu and P. Kok, "Geometric perspective on quantum parameter estimation," AVS Quantum Science **2**, 014701 (2020) DOI: 10.1116/1.5119961.

[9] V. Katariya and M. M. Wilde, "Geometric distinguishability measures limit quantum channel estimation and discrimination," Quantum Information Processing **20**, 78 (2021) DOI: 10.1007/s11128-021-02992-7.

[10] M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information*, 10th anniversary ed (Cambridge University Press, Cambridge ; New York, 2010), 676 pp.

[11] E. L Lehmann and G. Casella, *Theory of point estimation* (Springer London, s.l., 1998).

[12] E. A. Morozova and N. N. Chentsov, "Markov invariant geometry on manifolds of states," Journal of Soviet Mathematics **56**, 2648–2669 (1991) DOI: 10.1007/BF01095975.

[13] C. L. Canonne, "A short note on learning discrete distributions," arXiv:2002.11457 (2020).

[14] G. Torlai and R. G. Melko, "Machine-learning quantum states in the NISQ era," Annual Review of Condensed Matter Physics **11**, 325–344 (2020) DOI: 10.1146/annurev-conmatphys-031119-050651.

[15] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, "Quantum circuit learning," Physical Review A **98**, 032309 (2018) DOI: 10.1103/PhysRevA.98.032309.

[16] M. Schuld, V. Bergholm, C. Gogolin, J. Izaac, and N. Killoran, "Evaluating analytic gradients on quantum hardware," Physical Review A **99**, 032331 (2019) DOI: 10.1103/PhysRevA.99.032331.

[17] L. Seveso, F. Albarelli, M. G. Genoni, and M. G. A. Paris, "On the discontinuity of the quantum fisher information for quantum statistical models with parameter dependent rank," Journal of Physics A: Mathematical and Theoretical **53**, 02LT01 (2020) DOI: 10.1088/1751-8121/ab599b.

[18] M. M. Wilde, "From classical to quantum shannon theory," arXiv:1106.1445, DOI: 10.1017/9781316809976.001 (2017) DOI: 10.1017/9781316809976.001.

[19] M. M. Wolf, *Quantum channels & operations guided tour*, 2014.

[20] D. Petz, "Monotone metrics on matrix spaces," Linear Algebra and its Applications **244**, 81–96 (1996) DOI: 10.1016/0024-3795(94)00211-8.

[21] C. Helstrom, "Minimum mean-squared error of estimates in quantum statistics," Physics Letters A **25**, 101–102 (1967) DOI: 10.1016/0375-9601(67)90366-0.

[22] R. Cheng, "Quantum geometric tensor (fubini-study metric) in simple quantum system: a pedagogical introduction," arXiv:1012.1337 (2013).

[23] J. Stokes, J. Izaac, N. Killoran, and G. Carleo, "Quantum natural gradient," Quantum **4**, 269 (2020) DOI: 10.22331/q-2020-05-25-269.

[24] A. Mari, T. R. Bromley, and N. Killoran, "Estimating the gradient and higher-order derivatives on quantum hardware," Physical Review A **103**, 012405 (2021) DOI: 10.1103/PhysRevA.103.012405.

[25] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, "Supervised learning with quantum-enhanced feature spaces," Nature **567**, 209–212 (2019) DOI: 10.1038/s41586-019-0980-2.

[26] J. Gacon, C. Zoufal, G. Carleo, and S. Woerner, "Simultaneous perturbation stochastic approximation of the quantum fisher information," arXiv:2103.09232 (2021).

[27] S. L. Braunstein and C. M. Caves, "Statistical distance and the geometry of quantum states," Physical Review Letters **72**, 3439–3443 (1994) DOI: 10.1103/PhysRevLett.72.3439.

[28] O. E. Barndorff-Nielsen and R. D. Gill, "Fisher information in quantum statistics," Journal of Physics A: Mathematical and General **33**, 4481 (2000) DOI: 10.1088/0305-4470/33/24/306.

[29] L. Pezzè, M. A. Ciampini, N. Spagnolo, P. C. Humphreys, A. Datta, I. A. Walmsley, M. Barbieri, F. Sciarrino, and A. Smerzi, "Optimal measurements for simultaneous quantum estimation of multiple phases," Physical Review Letters **119**, DOI: 10.1103/PhysRevLett.119.130504 (2017) DOI: 10.1103/PhysRevLett.119.130504.

[30] J. Yang, S. Pang, Y. Zhou, and A. N. Jordan, "Optimal measurements for quantum multiparameter estimation with general states," Physical Review A **100**, 032104 (2019) DOI: 10.1103/PhysRevA.100.032104.

[31] S. Ragy, M. Jarzyna, and R. Demkowicz-Dobrzański, "Compatibility in multiparameter quantum metrology," Physical Review A **94**, 052108 (2016) DOI: 10.1103/PhysRevA.94.052108.

[32] X.-M. Lu and X. Wang, "Incorporating heisenberg's uncertainty principle into quantum multiparameter estimation," Physical Review Letters **126**, 120503 (2021) DOI: 10.1103/PhysRevLett.126.120503.

[33] F. Belliardo and V. Giovannetti, "Incompatibility in quantum parameter estimation," New Journal of Physics **23**, 063055 (2021) DOI: 10.1088/1367-2630/ac04ca.

[34] J. Liu, H.-N. Xiong, F. Song, and X. Wang, "Fidelity susceptibility and quantum fisher information for density operators with arbitrary ranks," Physica A: Statistical Mechanics and its Applications **410**, 167–173 (2014) DOI: 10.1016/j.physa.2014.05.028.

[35] D. Šafránek, "Discontinuities of the quantum fisher information and the bures metric," Physical Review A **95**, 052320 (2017) DOI: 10.1103/PhysRevA.95.052320.

[36] T. Baumgratz, A. Nüßeler, M. Cramer, and M. B. Plenio, "A scalable maximum likelihood method for quantum state tomography," New Journal of Physics **15**, 125004 (2013) DOI: 10.1088/1367-2630/15/12/125004.

[37] G. Toth, "Lower bounds on the quantum fisher information based on the variance and various types of entropies," arXiv:1701.07461 (2018).

[38] B. Koczor and S. C. Benjamin, "Quantum natural gradient generalised to non-unitary circuits," arXiv:1912.08660 (2020).

[39] Y. Du and D. Tao, "On exploring practical potentials of quantum auto-encoder with advantages," arXiv:2106.15432 (2021).

[40] J. Romero, J. P. Olson, and A. Aspuru-Guzik, "Quantum autoencoders for efficient compression of quantum data," Quantum Science and Technology **2**, 045001 (2017) DOI: 10.1088/2058-9565/aa8072.

[41] R. Chen, Z. Song, X. Zhao, and X. Wang, "Variational quantum algorithms for trace distance and fidelity estimation," arXiv:2012.05768 (2020).

[42] K. C. Tan and T. Volkoff, "Variational quantum algorithms to estimate rank, quantum entropies, fidelity and fisher information via purity minimization," arXiv:2103.15956 (2021).

[43] A. Sone, M. Cerezo, J. L. Beckey, and P. J. Coles, "A generalized measure of quantum fisher information," arXiv:2010.02904 (2021).

[44] J. L. Beckey, M. Cerezo, A. Sone, and P. J. Coles, "Variational quantum algorithm for estimating the quantum fisher information," arXiv:2010.10488 (2020).

[45] A. Rath, C. Branciard, A. Minguzzi, and B. Vermersch, "Quantum fisher information from randomized measurements," arXiv:2105.13164 (2021).

[46] H.-Y. Huang, R. Kueng, and J. Preskill, "Predicting many properties of a quantum system from very few measurements," Nature Physics **16**, 1050–1057 (2020) DOI: 10.1038/s41567-020-0932-7.

[47] C. Zhang, B. Yadin, Z.-B. Hou, H. Cao, B.-H. Liu, Y.-F. Huang, R. Maity, V. Vedral, C.-F. Li, G.-C. Guo, and D. Girolami, "Detecting metrologically useful asymmetry and entanglement by a few local measurements," Physical Review A **96**, 042327 (2017) DOI: 10.1103/PhysRevA.96.042327.

[48] D. Girolami and B. Yadin, "Witnessing multipartite entanglement by detecting asymmetry," Entropy **19**, 124 (2017) DOI: 10.3390/e19030124.

[49] C. W. Helstrom, "Quantum detection and estimation theory," Journal of Statistical Physics **1**, 231–252 (1969) DOI: 10.1007/BF01007479.

[50] A. S. Holevo, *Probabilistic and statistical aspects of quantum theory* (2011).

[51] R. Demkowicz-Dobrzański, W. Górecki, and M. Guţă, "Multi-parameter estimation beyond quantum fisher information," Journal of Physics A: Mathematical and Theoretical **53**, 363001 (2020) DOI: 10.1088/1751-8121/ab8ef3.

[52] M. Tsang, F. Albarelli, and A. Datta, "Quantum semiparametric estimation," Physical Review X **10**, 031023 (2020) DOI: 10.1103/PhysRevX.10.031023.

[53] F. Albarelli, M. Barbieri, M. G. Genoni, and I. Gianani, "A perspective on multiparameter quantum metrology: from theoretical tools to applications in quantum imaging," Physics Letters A **384**, 126311 (2020) DOI: 10.1016/j.physleta.2020.126311.

[54] J. J. Bollinger, W. M. Itano, D. J. Wineland, and D. J. Heinzen, "Optimal frequency measurements with maximally correlated states," Physical Review A **54**, R4649–R4652 (1996) DOI: 10.1103/PhysRevA.54.R4649.

[55] S. F. Huelga, C. Macchiavello, T. Pellizzari, A. K. Ekert, M. B. Plenio, and J. I. Cirac, "Improvement of frequency standards with quantum entanglement," Physical Review Letters **79**, 3865–3868 (1997) DOI: 10.1103/PhysRevLett.79.3865.

[56] R. Demkowicz-Dobrzański, J. Kołodyński, and M. Guţă, "The elusive heisenberg limit in quantum-enhanced metrology," Nature Communications **3**, 1063 (2012) DOI: 10.1038/ncomms2067.

[57] S. Zhou, M. Zhang, J. Preskill, and L. Jiang, "Achieving the heisenberg limit in quantum metrology using quantum error correction," Nature Communications **9**, 78 (2018) DOI: 10.1038/s41467-017-02510-3.

[58] W. Górecki, S. Zhou, L. Jiang, and R. Demkowicz-Dobrzański, "Optimal probes and error-correction schemes in multi-parameter quantum metrology," Quantum **4**, 288 (2020) DOI: 10.22331/q-2020-07-02-288.

[59] J. F. Haase, A. Smirne, S. F. Huelga, J. Kołodynski, and R. Demkowicz-Dobrzanski, "Precision limits in quantum metrology with open quantum systems," Quantum Measurements and Quantum Metrology **5**, 13–39 (2016) DOI: 10.1515/qmetro-2018-0002.

[60] G. Wang, D. E. Koh, P. D. Johnson, and Y. Cao, "Minimizing estimation runtime on noisy quantum computers," PRX Quantum **2**, 010346 (2021) DOI: 10.1103/PRXQuantum.2.010346.

[61] D. E. Koh, G. Wang, P. D. Johnson, and Y. Cao, "A framework for engineering quantum likelihood functions for expectation estimation," arXiv:2006.09349 (2020).

[62] R. Kaubruegger, P. Silvi, C. Kokail, R. van Bijnen, A. M. Rey, J. Ye, A. M. Kaufman, and P. Zoller, "Variational spin-squeezing algorithms on programmable quantum sensors," Physical Review Letters **123**, 260505 (2019) DOI: 10.1103/PhysRevLett.123.260505.

[63] B. Koczor, S. Endo, T. Jones, Y. Matsuzaki, and S. C. Benjamin, "Variational-state quantum metrology," New Journal of Physics **22**, 083038 (2020) DOI: 10.1088/1367-2630/ab965e.

[64] J. J. Meyer, J. Borregaard, and J. Eisert, "A variational toolbox for quantum multiparameter estimation," npj Quantum Information **7**, 1–5 (2021) DOI: 10.1038/s41534-021-00425-y.

[65] Z. Ma, P. Gokhale, T.-X. Zheng, S. Zhou, X. Yu, L. Jiang, P. Maurer, and F. T. Chong, "Adaptive circuit learning for quantum metrology," arXiv:2010.08702 (2020).

[66] M. Gärttner, P. Hauke, and A. M. Rey, "Relating out-of-time-order correlations to entanglement via multiple-quantum coherences," Physical Review Letters **120**, 040402 (2018) DOI: 10.1103/PhysRevLett.120.040402.

[67] M. Cerezo, A. Sone, J. L. Beckey, and P. J. Coles, "Sub-quantum fisher information," Quantum Science and Technology **6**, 035008 (2021) DOI: 10.1088/2058-9565/abfbef.

[68] D. Wierichs, C. Gogolin, and M. Kastoryano, "Avoiding local minima in variational quantum eigensolvers with the natural gradient optimizer," Physical Review Research **2**, 043246 (2020) DOI: 10.1103/PhysRevResearch.2.043246.

[69] B. van Straaten and B. Koczor, "Measurement cost of metric-aware variational quantum algorithms," PRX Quantum **2**, 030324 (2021) DOI: 10.1103/PRXQuantum.2.030324.

[70] M. Schuld, I. Sinayskiy, and F. Petruccione, "An introduction to quantum machine learning," Contemporary Physics **56**, 172–185 (2015) DOI: 10.1080/00107514.2014.964942.

[71] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, "Quantum machine learning," Nature **549**, 195–202 (2017) DOI: 10.1038/nature23474.

[72] T. Liang, T. Poggio, A. Rakhlin, and J. Stokes, "Fisher-rao metric, geometry, and complexity of neural networks," in The 22nd international conference on artificial intelligence and statistics (2019), p. 9.

[73] A. Abbas, D. Sutter, C. Zoufal, A. Lucchi, A. Figalli, and S. Woerner, "The power of quantum neural networks," Nature Computational Science **1**, 403–409 (2021) DOI: 10.1038/s43588-021-00084-1.

[74] T. Haug, K. Bharti, and M. S. Kim, "Capacity and quantum geometry of parametrized quantum circuits," arXiv:2102.01659 (2021).

[75] Y. Yang, R. Renner, and G. Chiribella, "Optimal universal programming of unitary gates," Physical Review Letters **125**, 210501 (2020) DOI: 10.1103/PhysRevLett.125.210501.

[76] A. Kubica and R. Demkowicz-Dobrzański, "Using quantum metrological bounds in quantum error correction: a simple proof of the approximate eastin-knill theorem," Physical Review Letters **126**, 150503 (2021) DOI: 10.1103/PhysRevLett.126.150503.

[77] K. C. Tan, V. Narasimhachar, and B. Regula, "Fisher information universally identifies quantum resources," arXiv:2104.01763 (2021).

[78] E. Chitambar and G. Gour, "Quantum resource theories," Reviews of Modern Physics **91**, 025001 (2019) DOI: 10.1103/RevModPhys.91.025001.

[79] D. P. Pires, M. Cianciaruso, L. C. Céleri, G. Adesso, and D. O. Soares-Pinto, "Generalized geometric quantum speed limits," Physical Review X **6**, 021031 (2016) DOI: 10.1103/PhysRevX.6.021031.

[80] D. Spehner, F. Illuminati, M. Orszag, and W. Roga, "Geometric measures of quantum correlations with bures and hellinger distances," in *Lectures on general quantum correlations and their applications*, edited by F. F. Fanchini, D. d. O. Soares Pinto, and G. Adesso (Springer International Publishing, Cham, 2017), pp. 105–157, DOI: 10.1007/978-3-319-53412-1˙6.

A Derivation of the Classical Fisher Information

We start out derivation of the classical Fisher information matrix from the Kullback-Leibler- or KL-divergence between a probability distribution and its perturbed counterpart

$$d_{\mathrm{KL}}(p_{\mathcal{M}}(\boldsymbol{\theta}), p_{\mathcal{M}}(\boldsymbol{\theta}+\boldsymbol{\delta})) = \sum_{l\in\mathcal{M}} p_l(\boldsymbol{\theta}) \log \frac{p_l(\boldsymbol{\theta})}{p_l(\boldsymbol{\theta}+\boldsymbol{\delta})}. \tag{75}$$

We can exploit the fact that $\log(a/b) = \log a - \log b$ and rewrite the perturbed KL divergence as

$$d_{\mathrm{KL}}(p_{\mathcal{M}}(\boldsymbol{\theta}), p_{\mathcal{M}}(\boldsymbol{\theta}+\boldsymbol{\delta})) = \sum_{l\in\mathcal{M}} p_l(\boldsymbol{\theta})(\log p_l(\boldsymbol{\theta}) - \log p_l(\boldsymbol{\theta}+\boldsymbol{\delta})). \tag{76}$$

We will now compute the metric by performing a second order expansion around $\boldsymbol{\delta} = 0$. To do so, we take the second derivatives with respect to the components of $\boldsymbol{\delta}$, so it is immediately clear that only the second term of Eq. (76) will contribute. We thus have

$$\begin{aligned}
[M_{\mathrm{KL}}]_{ij} &= -\frac{\partial^2}{\partial\delta_i\partial\delta_j} \sum_{l\in\mathcal{M}} p_l(\boldsymbol{\theta}) \log p_l(\boldsymbol{\theta}+\boldsymbol{\delta})\bigg|_{\boldsymbol{\delta}=0} && (77)\\
&= -\sum_{l\in\mathcal{M}} p_l(\boldsymbol{\theta})\, \frac{\partial^2}{\partial\delta_i\partial\delta_j} \log p_l(\boldsymbol{\theta}+\boldsymbol{\delta})\bigg|_{\boldsymbol{\delta}=0} && (78)\\
&= -\mathbb{E}\left\{ \frac{\partial^2}{\partial\delta_i\partial\delta_j} \log p_l(\boldsymbol{\theta}+\boldsymbol{\delta})\bigg|_{\boldsymbol{\delta}=0} \right\}. && (79)
\end{aligned}$$

We can now substitute $\boldsymbol{\xi} = \boldsymbol{\theta}+\boldsymbol{\delta}$, in which case the derivatives transform as

$$\frac{\partial}{\partial\delta_i} = \frac{\partial}{\partial\xi_i}\frac{\partial\xi_i}{\partial\delta_i} = \frac{\partial}{\partial\xi_i} \tag{80}$$

to obtain

$$[M_{\mathrm{KL}}]_{ij} = \mathbb{E}\left\{ -\frac{\partial^2}{\partial\xi_i\partial\xi_j} \log p_l(\boldsymbol{\xi})\bigg|_{\boldsymbol{\xi}=\boldsymbol{\theta}} \right\}. \tag{81}$$

Now we can rename $\boldsymbol{\xi}$ to $\boldsymbol{\theta}$ again to get the more familiar looking

$$[M_{\mathrm{KL}}]_{ij} = \mathbb{E}\left\{ -\frac{\partial^2}{\partial\theta_i\partial\theta_j} \log p_l(\boldsymbol{\theta}) \right\}. \tag{82}$$

We can also continue this derivation a bit more to get a second form of this expression. Consider that

$$-\frac{\partial^2}{\partial\theta_i\partial\theta_j} \log p_l(\boldsymbol{\theta}) = -\frac{\partial}{\partial\theta_j}\frac{1}{p_l(\boldsymbol{\theta})}\frac{\partial p_l(\boldsymbol{\theta})}{\partial\theta_i} = \frac{1}{p_l^2(\boldsymbol{\theta})}\frac{\partial p_l(\boldsymbol{\theta})}{\partial\theta_j}\frac{\partial p_l(\boldsymbol{\theta})}{\partial\theta_i} - \frac{1}{p_l(\boldsymbol{\theta})}\frac{\partial^2 p_l(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}. \tag{83}$$

To get to the expression for $[M_{\mathrm{KL}}]_{ij}$, we have to take the expectation value of this expression over the whole probability distribution. In the process of doing so, the second term will actually disappear:

$$\begin{aligned}
\sum_{l\in\mathcal{M}} p_l(\boldsymbol{\theta})\frac{1}{p_l(\boldsymbol{\theta})}\frac{\partial^2 p_l(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j} &= \sum_{l\in\mathcal{M}} \frac{\partial^2 p_l(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j} && (84)\\
&= \frac{\partial^2}{\partial\theta_i\partial\theta_j}\sum_{l\in\mathcal{M}} p_l(\boldsymbol{\theta}) && (85)\\
&= \frac{\partial^2}{\partial\theta_i\partial\theta_j}1 && (86)\\
&= 0. && (87)
\end{aligned}$$

To summarize, we arrive at the following equivalent formulas:

$$\begin{aligned}
[M_{\mathrm{KL}}]_{ij} &= \sum_{l\in\mathcal{M}} p_l(\boldsymbol{\theta})\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p_l(\boldsymbol{\theta}) && (88)\\
&= \sum_{l\in\mathcal{M}} \frac{1}{p_l(\boldsymbol{\theta})}\frac{\partial p_l(\boldsymbol{\theta})}{\partial\theta_j}\frac{\partial p_l(\boldsymbol{\theta})}{\partial\theta_i}. && (89)
\end{aligned}$$

B Properties of the Classical Fisher Information Matrix

The classical Fisher Information Matrix I of a probability distribution $p(\boldsymbol{\theta})$ with respect to a set of d parameters $\boldsymbol{\theta} \in \mathbb{R}^d$ is given by

$$I_{ij} = \sum_l p_l(\boldsymbol{\theta}) \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p_l(\boldsymbol{\theta}) = \sum_l \frac{1}{p_l(\boldsymbol{\theta})} \frac{\partial p_l(\boldsymbol{\theta})}{\partial\theta_i} \frac{\partial p_l(\boldsymbol{\theta})}{\partial\theta_j} \tag{90}$$

where the index l iterates over all elements of $p(\boldsymbol{\theta})$. It has the following properties:

(i) I is a real symmetric $d \times d$ matrix.

$$I \in \mathbb{R}^{d\times d} \qquad I_{ij} = I_{ji} \tag{91}$$

(ii) I is positive semidefinite, *i.e.* it only has non-negative eigenvalues. We write this as

$$I \geq 0. \tag{92}$$

(iii) Convexity. I is convex, which means that for any two probability distributions $p(\boldsymbol{\theta})$ and $q(\boldsymbol{\theta})$,

$$I[\lambda p(\boldsymbol{\theta}) + (1-\lambda) q(\boldsymbol{\theta})] \leq \lambda I[p(\boldsymbol{\theta})] + (1-\lambda) I[q(\boldsymbol{\theta})] \tag{93}$$

for $0 \leq \lambda \leq 1$.

(iv) I is additive under direct sum, in the sense that if we look at a two probability distributions $p(\boldsymbol{\theta})$ and $q(\boldsymbol{\theta})$, then

$$I[\lambda p(\boldsymbol{\theta}) \oplus (1-\lambda) q(\boldsymbol{\theta})] = \lambda I[p(\boldsymbol{\theta})] + (1-\lambda) I[q(\boldsymbol{\theta})], \tag{94}$$

where the parameter $0 \leq \lambda \leq 1$ ensures that the resulting object is a proper probability distribution.

(v) I is non-increasing under stochastic maps. Let T be a stochastic map (a map between probability distributions), then

$$I[T[p(\boldsymbol{\theta})]] \leq I[p(\boldsymbol{\theta})] \tag{95}$$

which reads as the matrix $I[p(\boldsymbol{\theta})] - I[T[p(\boldsymbol{\theta})]]$ being positive semidefinite.

(vi) Transformation rule. Suppose we have a reparametrization of the parameters $\boldsymbol{\theta}$ given by a multivariate function $\boldsymbol{f}(\boldsymbol{\theta})$. The classical Fisher information matrix with respect to the new parameters is given by

$$I[p(\boldsymbol{f}(\boldsymbol{\theta}))] = J(\boldsymbol{\theta}) I_{\boldsymbol{\theta}}[p(\boldsymbol{\theta})] J^T(\boldsymbol{\theta}), \tag{96}$$

where $J_{ij} = \partial\theta_j / \partial f_i$ is the inverse of the Jacobian of the mapping $\boldsymbol{f}(\boldsymbol{\theta})$.

(vii) Uniqueness as monotone metric. The second order expansion of any monotonic distance measure, *i.e.* a distance measure decreasing under stochastic maps, will yield a constant multiple of the classical Fisher information matrix:

$$d(p(\boldsymbol{\theta}), p(\boldsymbol{\theta} + \boldsymbol{\delta})) \propto \boldsymbol{\delta}^T I \boldsymbol{\delta} + O(\|\boldsymbol{\delta}\|^3). \tag{97}$$

Proof. Recall that the classical Fisher information matrix arises in the second order expansion of the KL divergence,

$$d_{\mathrm{KL}}(p(\boldsymbol{\theta}), p(\boldsymbol{\theta} + \boldsymbol{\delta})) = \frac{1}{2} \boldsymbol{\delta}^T I \boldsymbol{\delta} + O(\|\boldsymbol{\delta}\|^3), \tag{98}$$

where

$$I_{ij} = \frac{\partial^2}{\partial\delta_i \partial\delta_j} d_{\mathrm{KL}}(p(\boldsymbol{\theta}), p(\boldsymbol{\theta} + \boldsymbol{\delta})) \bigg|_{\boldsymbol{\delta}=0} \tag{99}$$

(i) Symmetry follows from the fact that the derivatives in Eq. (99) can be exchanged. Reality is a consequence of the fact that d_{KL} is a real-valued function.

(ii) Positive-semidefiniteness follows from the fact that d_{KL} is a distance measure, meaning that $d_{\mathrm{KL}}(p, q) \geq 0$ and $d_{\mathrm{KL}}(p(\boldsymbol{\theta}), p(\boldsymbol{\theta})) = 0$. This implies that we evaluate the second derivative in Eq. (99) at a minimum, implying that the curvature in no direction can be negative.

(iii) Convexity follows from the joint convexity of d_{KL}, *i.e.* the fact that

$$d_{\mathrm{KL}}(\lambda p_1 + (1-\lambda)q_1, \lambda p_2 + (1-\lambda)q_2) \leq \lambda d_{\mathrm{KL}}(p_1, p_2) + (1-\lambda) d_{\mathrm{KL}}(q_1, q_2). \tag{100}$$

In our case, we can identify $p_1 = p(\boldsymbol{\theta})$, $q_1 = q(\boldsymbol{\theta})$, $p_2 = p(\boldsymbol{\theta}+\boldsymbol{\delta})$ and $q_2 = q(\boldsymbol{\theta}+\boldsymbol{\delta})$ and have

$$\begin{aligned} d_{\mathrm{KL}}(&\lambda p(\boldsymbol{\theta}) + (1-\lambda)q(\boldsymbol{\theta}), \lambda p(\boldsymbol{\theta}+\boldsymbol{\delta}) + (1-\lambda)q(\boldsymbol{\theta}+\boldsymbol{\delta})) \\ &\leq \lambda d_{\mathrm{KL}}(p(\boldsymbol{\theta}), p(\boldsymbol{\theta}+\boldsymbol{\delta})) + (1-\lambda) d_{\mathrm{KL}}(q(\boldsymbol{\theta}), q(\boldsymbol{\theta}+\boldsymbol{\delta})). \end{aligned} \tag{101}$$

The convexity of the classical Fisher information matrix immediately follows by linearity when taking the second derivatives with respect to the components of $\boldsymbol{\delta}$ evaluated at $\boldsymbol{\delta} = 0$.

(iv) Due to the direct sum structure, the terms relating to $p(\boldsymbol{\theta})$ and $q(\boldsymbol{\theta})$ in Eq. (90) are independent of each other, yielding the desired result.

(v) Monotonicity of the KL divergence implies that under the action of a stochastic map T

$$d_{\mathrm{KL}}(T[p(\boldsymbol{\theta})], T[p(\boldsymbol{\theta}+\boldsymbol{\delta})]) \leq d_{\mathrm{KL}}(p(\boldsymbol{\theta}), p(\boldsymbol{\theta}+\boldsymbol{\delta})), \tag{102}$$

which holds independently of $\boldsymbol{\delta}$. In the limit $\|\boldsymbol{\delta}\| \to 0$, we have

$$d_{\mathrm{KL}}(p(\boldsymbol{\theta}), p(\boldsymbol{\theta}+\boldsymbol{\delta})) \approx \frac{1}{2}\boldsymbol{\delta}^{T} I[p(\boldsymbol{\theta})]\boldsymbol{\delta}, \tag{103}$$

and therefore

$$\boldsymbol{\delta}^{T} I[T[p(\boldsymbol{\theta})]]\boldsymbol{\delta} \leq \boldsymbol{\delta}^{T} I[p(\boldsymbol{\theta})]\boldsymbol{\delta}. \tag{104}$$

As this needs to hold for any $\boldsymbol{\delta}$, it implies the sought matrix inequality

$$I[T[p(\boldsymbol{\theta})]] \leq I[p(\boldsymbol{\theta})]. \tag{105}$$

(vi) To prove the transformation rule, we make use of Faà di Bruno's formula for the Hessian. If we have a look at a function $g(\boldsymbol{\theta})$ and seek its second derivatives with respect to coordinates $\boldsymbol{f}(\boldsymbol{\theta})$, we have that

$$\frac{\partial^2 g(\boldsymbol{\theta})}{\partial f_i \partial f_j} = \sum_k \frac{\partial g(\boldsymbol{\theta})}{\partial \theta_k} \frac{\partial^2 u_k}{\partial f_i \partial f_j} + \sum_{kl} \frac{\partial^2 g(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_l} \frac{\partial \theta_k}{\partial f_i} \frac{\partial \theta_l}{\partial f_j}. \tag{106}$$

If $g(\boldsymbol{\theta})$ is a local optimum (as will be the case for us), the gradient terms vanish and we can rewrite the above using the Jacobian of the coordinate transform $J_{ik} = \partial \theta_k / \partial f_i$ as

$$\frac{\partial^2 g(\boldsymbol{\theta})}{\partial f_i \partial f_j} = \sum_{kl} \frac{\partial^2 g(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_l} J_{ik} J_{jl}. \tag{107}$$

If we denote $H^{f}_{ij} = \partial^2 g(\boldsymbol{\theta}) / \partial f_i \partial f_j$ and $H^{\theta}_{ij} = \partial^2 g(\boldsymbol{\theta}) / \partial \theta_i \partial \theta_j$ we can deduce the transformation rule

$$H^{f} = J H^{\theta} J^{T}. \tag{108}$$

The transformation rule of the classical Fisher information matrix follows from applying this general result to the definition of the classical Fisher information matrix as the Hessian of the underlying distance function.

(vii) For this result see the original work in Ref. [12].

□

C Modified Distance Functions

In this section, we show that applying a post-processing function to a distance function will only change the associated information matrix by a scalar prefactor.

To do so, we will first have a look at the second derivatives of functions of the form $h(g(\boldsymbol{\delta}))$:

$$\begin{aligned}\frac{\partial^2}{\partial\delta_i\partial\delta_j}h(g(\boldsymbol{\delta})) &= \frac{\partial}{\partial\delta_j}\frac{\partial h(g(\boldsymbol{\delta}))}{\partial\delta_i} &(109)\\ &= \frac{\partial}{\partial\delta_j}\left(f'(g(\boldsymbol{\delta}))\frac{\partial g(\boldsymbol{\delta})}{\partial\delta_i}\right) &(110)\\ &= f'(g(\boldsymbol{\delta}))\frac{\partial^2 g(\boldsymbol{\delta})}{\partial\delta_i\partial\delta_j} + f''(g(\boldsymbol{\delta}))\left(\frac{\partial g(\boldsymbol{\delta})}{\partial\delta_i}\right)\left(\frac{\partial g(\boldsymbol{\delta})}{\partial\delta_j}\right). &(111)\end{aligned}$$

In the case of distance functions $g(\boldsymbol{\delta}) = d(\boldsymbol{\theta}, \boldsymbol{\theta}+\boldsymbol{\delta})$, we want to evaluate them at $\boldsymbol{\delta} = 0$ where they are extremal. This means that the first derivatives need to vanish at this point. Therefore, at any extremal point $\boldsymbol{\delta}^*$, we have that

$$\left.\frac{\partial^2}{\partial\delta_i\partial\delta_j}h(g(\boldsymbol{\delta}))\right|_{\boldsymbol{\delta}=\boldsymbol{\delta}^*} = f'(g(\boldsymbol{\delta}^*))\left.\frac{\partial^2 g(\boldsymbol{\delta})}{\partial\delta_i\partial\delta_j}\right|_{\boldsymbol{\delta}=\boldsymbol{\delta}^*}. \tag{112}$$

This means that the resulting matrix for $h(g(\boldsymbol{\delta}))$ differs from the one for $g(\boldsymbol{\delta})$ only by a factor of $f'(g(\boldsymbol{\delta}^*))$.

The distance measures $d(\boldsymbol{\theta}, \boldsymbol{\theta}+\boldsymbol{\delta})$ we are treating in the main part of this work fulfill the extremality property at $\boldsymbol{\delta} = 0$ where $d(\boldsymbol{\theta}, \boldsymbol{\theta}) = 0$. We can therefore say something about the information matrix associated with $h(d)$ compared to d:

$$[M_{h(d)}]_{ij} = \left.\frac{\partial^2}{\partial\delta_i\partial\delta_j}h(d(\boldsymbol{\theta}, \boldsymbol{\theta}+\boldsymbol{\delta}))\right|_{\boldsymbol{\delta}=0} = f'(0)\left.\frac{\partial^2 d(\boldsymbol{\theta}, \boldsymbol{\theta}+\boldsymbol{\delta})}{\partial\delta_i\partial\delta_j}\right|_{\boldsymbol{\delta}=0} = f'(0)[M_d]_{ij} \tag{113}$$

From this formula we immediately see that we need to have that $f'(0) > 0$ to make for a sensible transformation of our information matrix.

The reasoning outlined above also extends beyond functions of the distance. For example, consider the fidelity distance we used to introduce the quantum Fisher information matrix:

$$d_f(\boldsymbol{\theta}, \boldsymbol{\theta}+\boldsymbol{\delta}) = 1 - |\langle\psi(\boldsymbol{\theta})|\psi(\boldsymbol{\theta}+\boldsymbol{\delta})\rangle|^2. \tag{114}$$

We know that the fidelity $f(\boldsymbol{\theta}, \boldsymbol{\theta}+\boldsymbol{\delta}) = |\langle\psi(\boldsymbol{\theta})|\psi(\boldsymbol{\theta}+\boldsymbol{\delta})\rangle|^2$ is extremal for $\boldsymbol{\delta} = 0$ where it takes the value 1. We could also define the fidelity distance using the square root of the fidelity – and some authors do that – to get

$$d_{\sqrt{f}}(\boldsymbol{\theta}, \boldsymbol{\theta}+\boldsymbol{\delta}) = 1 - |\langle\psi(\boldsymbol{\theta})|\psi(\boldsymbol{\theta}+\boldsymbol{\delta})\rangle|. \tag{115}$$

With our knowledge from above and with seeing that $h(x) = \sqrt{x}$, we immediately see that

$$M_{\sqrt{f}} = h'(f(\boldsymbol{\theta}, \boldsymbol{\theta}))M_f = \frac{1}{2}M_f. \tag{116}$$

This shows that the information matrices of the two conventions only differ by a factor of 1/2.

D Derivation of the Quantum Fisher Information

We start the derivation of the quantum Fisher information for pure states from the fidelity distance

$$d_f(|\psi(\boldsymbol{\theta})\rangle, |\psi(\boldsymbol{\theta}')\rangle) = 1 - f(|\psi(\boldsymbol{\theta})\rangle, |\psi(\boldsymbol{\theta}')\rangle) = 1 - |\langle\psi(\boldsymbol{\theta})|\psi(\boldsymbol{\theta}')\rangle|^2. \tag{117}$$

For small displacements, we have

$$|\psi(\boldsymbol{\theta}+\boldsymbol{\delta})\rangle = |\psi(\boldsymbol{\theta})\rangle + \sum_i \delta_i\,|\partial_i\psi(\boldsymbol{\theta})\rangle\,. \tag{118}$$

Therefore,

$$f(|\psi(\boldsymbol{\theta})\rangle, |\psi(\boldsymbol{\theta}+\boldsymbol{\delta})\rangle) = |\langle\psi(\boldsymbol{\theta})|\psi(\boldsymbol{\theta})\rangle + \sum_i \delta_i\langle\psi(\boldsymbol{\theta})|\partial_i\psi(\boldsymbol{\theta})\rangle|^2 = |1 + \sum_i \delta_i\langle\psi(\boldsymbol{\theta})|\partial_i\psi(\boldsymbol{\theta})\rangle|^2. \tag{119}$$

Rewriting the absolute value yields

$$f(\left|\psi(\boldsymbol{\theta})\right\rangle, \left|\psi(\boldsymbol{\theta}+\boldsymbol{\delta})\right\rangle) = 1 + \sum_i \delta_i(\langle\psi(\boldsymbol{\theta})|\partial_i\psi(\boldsymbol{\theta})\rangle + \langle\partial_i\psi(\boldsymbol{\theta})|\psi(\boldsymbol{\theta})\rangle) + \sum_{ij}\delta_i\delta_j\langle\partial_i\psi(\boldsymbol{\theta})|\psi(\boldsymbol{\theta})\rangle\langle\psi(\boldsymbol{\theta})|\partial_j\psi(\boldsymbol{\theta})\rangle. \tag{120}$$

We are only interested in second order terms for the computation of our metric, so it would appear that only the last term is of interest to us. But there is a hidden second order dependence in the second term. It arises because $\left|\psi(\boldsymbol{\theta}+\boldsymbol{\delta})\right\rangle = \left|\psi(\boldsymbol{\theta})\right\rangle + \sum_i \delta_i \left|\partial_i\psi(\boldsymbol{\theta})\right\rangle$ needs to be a pure state, and therefore

$$1 = f(\left|\psi(\boldsymbol{\theta}+\boldsymbol{\delta})\right\rangle, \left|\psi(\boldsymbol{\theta}+\boldsymbol{\delta})\right\rangle) \tag{121}$$

$$= \langle\psi(\boldsymbol{\theta})|\psi(\boldsymbol{\theta})\rangle + \sum_i \delta_i(\langle\psi(\boldsymbol{\theta})|\partial_i\psi(\boldsymbol{\theta})\rangle + \langle\partial_i\psi(\boldsymbol{\theta})|\psi(\boldsymbol{\theta})\rangle) + \sum_{ij}\delta_i\delta_j\langle\partial_i\psi(\boldsymbol{\theta})|\partial_j\psi(\boldsymbol{\theta})\rangle, \tag{122}$$

which enforces

$$\sum_i \delta_i(\langle\psi(\boldsymbol{\theta})|\partial_i\psi(\boldsymbol{\theta})\rangle + \langle\partial_i\psi(\boldsymbol{\theta})|\psi(\boldsymbol{\theta})\rangle) = -\sum_{ij}\delta_i\delta_j\langle\partial_i\psi(\boldsymbol{\theta})|\partial_j\psi(\boldsymbol{\theta})\rangle. \tag{123}$$

With these results at hand, we see that

$$[M_f]_{ij} = \frac{\partial^2}{\partial\delta_i\partial\delta_j} d_f(\left|\psi(\boldsymbol{\theta})\right\rangle, \left|\psi(\boldsymbol{\theta}+\boldsymbol{\delta})\right\rangle) \tag{124}$$

$$= 2\,\mathrm{Re}[\langle\partial_i\psi(\boldsymbol{\theta})|\partial_j\psi(\boldsymbol{\theta})\rangle - \langle\partial_i\psi(\boldsymbol{\theta})|\psi(\boldsymbol{\theta})\rangle\langle\psi(\boldsymbol{\theta})|\partial_j\psi(\boldsymbol{\theta})\rangle]. \tag{125}$$

The real part and the factor 2 account for the fact that $\delta_i\delta_j$ appears in the sum of Eq. (120) twice, but with conjugated terms that follow it – so that we can use that for any complex number z

$$z + z^* = 2\,\mathrm{Re}[z]. \tag{126}$$

Now, we are left with a final step – we have to ensure that our information matrix is consistent with the classical case. This means that for a "classical" state with classical measurements, we should recover the classical Fisher information. To this end, let us consider the state

$$\left|\psi(\boldsymbol{\theta})\right\rangle = \sum_l \sqrt{p_l(\boldsymbol{\theta})}\left|l\right\rangle, \tag{127}$$

where the sum is over the computational basis states $\left|l\right\rangle$. Together with measurements in the computational basis, this state represents the classical probability distribution $p(\boldsymbol{\theta})$. If we put this into the formula we just derived, we see that the first term already resembles the classical Fisher information:

$$\langle\partial_i\psi(\boldsymbol{\theta})|\partial_j\psi(\boldsymbol{\theta})\rangle = \sum_l(\partial_i\sqrt{p_l(\boldsymbol{\theta})})(\partial_j\sqrt{p_l(\boldsymbol{\theta})}) \tag{128}$$

$$= \frac{1}{4}\sum_l \frac{\partial_i p_l(\boldsymbol{\theta})}{\sqrt{p_l(\boldsymbol{\theta})}}\frac{\partial_j p_l(\boldsymbol{\theta})}{\sqrt{p_l(\boldsymbol{\theta})}} \tag{129}$$

$$= \frac{1}{4}\sum_l \frac{(\partial_i p_l(\boldsymbol{\theta}))(\partial_j p_l(\boldsymbol{\theta}))}{p_l(\boldsymbol{\theta})}. \tag{130}$$

And luckily for us, the second term won't contribute, because

$$\langle\partial_i\psi(\boldsymbol{\theta})|\psi(\boldsymbol{\theta})\rangle = \sum_l(\partial_i\sqrt{p_l(\boldsymbol{\theta})})\sqrt{p_l(\boldsymbol{\theta})} \tag{131}$$

$$= \frac{1}{2}\sum_l \frac{\partial_i p_l(\boldsymbol{\theta})}{\sqrt{p_l(\boldsymbol{\theta})}}\sqrt{p_l(\boldsymbol{\theta})} \tag{132}$$

$$= \frac{1}{2}\sum_l \partial_i p_l(\boldsymbol{\theta}) \tag{133}$$

$$= \frac{1}{2}\partial_i\sum_l p_l(\boldsymbol{\theta}) \tag{134}$$

$$= \frac{1}{2}\partial_i 1 \tag{135}$$

$$= 0. \tag{136}$$

Combining this with Eq. (125), we see that – for the classical state we currently consider – we have

$$M_f = \frac{1}{2} I. \tag{137}$$

If we correct for the factor $\frac{1}{2}$ we arrive at the definition of the quantum Fisher information, which is now consistent with the classical limit:

$$\mathcal{F}_{ij} = 2[M_f]_{ij} = 4 \operatorname{Re}[\langle \partial_i \psi(\boldsymbol{\theta}) | \partial_j \psi(\boldsymbol{\theta}) \rangle - \langle \partial_i \psi(\boldsymbol{\theta}) | \psi(\boldsymbol{\theta}) \rangle \langle \psi(\boldsymbol{\theta}) | \partial_j \psi(\boldsymbol{\theta}) \rangle]. \tag{138}$$

E Properties of the Quantum Fisher Information Matrix

The quantum Fisher Information Matrix $\mathcal{F}$ of a quantum state $\rho(\boldsymbol{\theta}) = \sum_i \lambda_i |\lambda_i\rangle\langle\lambda_i|$ with respect to a set of d parameters $\boldsymbol{\theta}$ is given by [7]

$$\begin{aligned} \mathcal{F}_{ij} &= \sum_{\substack{kl \\ \lambda_k + \lambda_l \neq 0}} \frac{2 \operatorname{Re}(\langle \lambda_k | \partial_i \rho | \lambda_l \rangle\langle \lambda_l | \partial_j \rho | \lambda_k \rangle)}{\lambda_k + \lambda_l} & (139) \\ &= \sum_{\substack{k \\ \lambda_k \neq 0}} \frac{(\partial_i \lambda_k)(\partial_j \lambda_k)}{\lambda_k} + 4\lambda_k \operatorname{Re}(\langle \partial_i \lambda_k | \partial_j \lambda_k \rangle) - \sum_{\substack{kl \\ \lambda_k, \lambda_l \neq 0}} \frac{8 \lambda_k \lambda_l}{\lambda_k + \lambda_l} \operatorname{Re}(\langle \partial_i \lambda_k | \lambda_l \rangle\langle \lambda_l | \partial_j \lambda_k \rangle) & (140) \end{aligned}$$

In the case of of a pure state $\rho(\boldsymbol{\theta}) = |\psi(\boldsymbol{\theta})\rangle\langle\psi(\boldsymbol{\theta})|$ this simplifies to

$$\mathcal{F}_{ij} = 4 \operatorname{Re}[\langle \partial_i \psi | \partial_j \psi \rangle - \langle \partial_i \psi | \psi \rangle\langle \psi | \partial_j \psi \rangle]. \tag{141}$$

The properties of the quantum Fisher information matrix are listed in the excellent review by Liu *et al.* [7], but will be given here again for the sake of completeness, along with proofs to foster understanding:

(i) $\mathcal{F}$ is a real symmetric $d \times d$ matrix.

$$\mathcal{F} \in \mathbb{R}^{d \times d} \qquad \mathcal{F} = \mathcal{F}^T. \tag{142}$$

(ii) $\mathcal{F}$ is positive semidefinite, *i.e.* it only has non-negative eigenvalues. We write this as

$$\mathcal{F} \geq 0. \tag{143}$$

(iii) Convexity. $\mathcal{F}$ is convex, which means that for any two quantum states $\rho(\boldsymbol{\theta})$ and $\sigma(\boldsymbol{\theta})$,

$$\mathcal{F}[\lambda \rho(\boldsymbol{\theta}) + (1 - \lambda)\sigma(\boldsymbol{\theta})] \leq \lambda \mathcal{F}[\rho(\boldsymbol{\theta})] + (1 - \lambda)\mathcal{F}[\sigma(\boldsymbol{\theta})] \tag{144}$$

for $0 \leq \lambda \leq 1$.

(iv) Invariance under unitary transformations. For any unitary transformation U, we have that

$$\mathcal{F}[U \rho(\boldsymbol{\theta}) U^\dagger] = \mathcal{F}[\rho(\boldsymbol{\theta})]. \tag{145}$$

(v) $\mathcal{F}$ is additive under direct sum, in the sense that if we look at a two quantum states $\rho(\boldsymbol{\theta})$ and $\sigma(\boldsymbol{\theta})$, then

$$\mathcal{F}[\lambda \rho(\boldsymbol{\theta}) \oplus (1 - \lambda)\sigma(\boldsymbol{\theta})] = \lambda \mathcal{F}[\rho(\boldsymbol{\theta})] + (1 - \lambda)\mathcal{F}[\sigma(\boldsymbol{\theta})], \tag{146}$$

where the parameter $0 \leq \lambda \leq 1$ ensures that the resulting object is a proper quantum state.

(vi) $\mathcal{F}$ is additive under tensor products, in the sense that if we look at a two quantum states $\rho(\boldsymbol{\theta})$ and $\sigma(\boldsymbol{\theta})$, then

$$\mathcal{F}[\rho(\boldsymbol{\theta}) \otimes \sigma(\boldsymbol{\theta})] = \mathcal{F}[\rho(\boldsymbol{\theta})] + \mathcal{F}[\sigma(\boldsymbol{\theta})]. \tag{147}$$

(vii) $\mathcal{F}$ is non-increasing under quantum channels. Let Φ be a quantum channel (a map between density matrices), then

$$\mathcal{F}[\Phi[\rho(\boldsymbol{\theta})]] \leq \mathcal{F}[\rho(\boldsymbol{\theta})] \tag{148}$$

which reads as the matrix $\mathcal{F}[\rho(\boldsymbol{\theta})] - \mathcal{F}[\Phi[\rho(\boldsymbol{\theta})]]$ being positive semidefinite.

(viii) Transformation rule. Suppose we have a reparametrization of the parameters $\boldsymbol{\theta}$ given by a multivariate function $\boldsymbol{f}(\boldsymbol{\theta})$. The quantum Fisher information matrix with respect to the new parameters is given by

$$\mathcal{F}[\rho(\boldsymbol{f}(\boldsymbol{\theta}))] = J(\boldsymbol{\theta})\mathcal{F}_{\boldsymbol{\theta}}[\rho(\boldsymbol{\theta})]J^{T}(\boldsymbol{\theta}), \tag{149}$$

where $J_{ij} = \partial\theta_j/\partial f_i$ is the inverse of the Jacobian of the mapping $\boldsymbol{f}(\boldsymbol{\theta})$.

Proof. In the mixed state case, the quantum Fisher information matrix arises in the second order expansion of the Bures distance. If we have a state $\rho = \sum_i \lambda_i |\lambda_i\rangle\langle\lambda_i|$ we denote its unique positive square root as $\rho^{1/2} = \sum_i \sqrt{\lambda_i}|\lambda_i\rangle\langle\lambda_i|$. The Bures distance is then given by

$$d_B(\rho, \sigma) = 2 - 2\operatorname{Tr}\{(\rho^{1/2}\sigma\rho^{1/2})^{1/2}\}^2 \tag{150}$$

as

$$d_B(\rho(\boldsymbol{\theta}), \rho(\boldsymbol{\theta}+\boldsymbol{\delta})) = \frac{1}{2}\boldsymbol{\delta}^{T}\mathcal{F}\boldsymbol{\delta} + O(\|\boldsymbol{\delta}\|^3), \tag{151}$$

where

$$\mathcal{F}_{ij} = \frac{\partial^2}{\partial\delta_i\partial\delta_j} d_B(\rho(\boldsymbol{\theta}), \rho(\boldsymbol{\theta}+\boldsymbol{\delta}))\bigg|_{\boldsymbol{\delta}=0}. \tag{152}$$

(i) Symmetry follows from the fact that the derivatives in Eq. (152) can be exchanged. Reality is a consequence of the fact that d_B is a real-valued function.

(ii) Positive-semidefiniteness follows from the fact that d_B is a distance measure, meaning that $d_B(\rho, \sigma) \geq 0$ and $d_B(\rho(\boldsymbol{\theta}), \rho(\boldsymbol{\theta})) = 0$. This implies that we evaluate the second derivative in Eq. (152) at a minimum, implying that the curvature in no direction can be negative.

(iii) Convexity follows from the joint convexity of d_B, *i.e.* the fact that [80]

$$d_B(\lambda\rho_1 + (1-\lambda)\sigma_1, \lambda\rho_2 + (1-\lambda)\sigma_2) \leq \lambda d_B(\rho_1, \rho_2) + (1-\lambda)d_B(\sigma_1, \sigma_2). \tag{153}$$

In our case, we can identify $\rho_1 = \rho(\boldsymbol{\theta})$, $\sigma_1 = \sigma(\boldsymbol{\theta})$, $\rho_2 = \rho(\boldsymbol{\theta}+\boldsymbol{\delta})$ and $\sigma_2 = \sigma(\boldsymbol{\theta}+\boldsymbol{\delta})$ and have

$$\begin{aligned} &d_B(\lambda\rho(\boldsymbol{\theta}) + (1-\lambda)\sigma(\boldsymbol{\theta}), \lambda\rho(\boldsymbol{\theta}+\boldsymbol{\delta}) + (1-\lambda)\sigma(\boldsymbol{\theta}+\boldsymbol{\delta})) \\ &\qquad \leq \lambda d_B(\rho(\boldsymbol{\theta}), \rho(\boldsymbol{\theta}+\boldsymbol{\delta})) + (1-\lambda)d_B(\sigma(\boldsymbol{\theta}), \sigma(\boldsymbol{\theta}+\boldsymbol{\delta})). \end{aligned} \tag{154}$$

The convexity of the quantum Fisher information matrix immediately follows by linearity when taking the second derivatives with respect to the components of $\boldsymbol{\delta}$ evaluated at $\boldsymbol{\delta} = 0$.

(iv) The unitary invariance directly follows from the unitary invariance of the Bures distance. To see that the Bures distance is indeed unitarily invariant we first note that

$$\begin{aligned} (U\rho U^\dagger)^{1/2} &= \left(\sum_i \lambda_i U|\lambda_i\rangle\langle\lambda_i|U^\dagger\right)^{1/2} && (155)\\ &= \sum_i \sqrt{\lambda_i} U|\lambda_i\rangle\langle\lambda_i|U^\dagger && (156)\\ &= U\left(\sum_i \sqrt{\lambda_i}|\lambda_i\rangle\langle\lambda_i|\right)U^\dagger && (157)\\ &= U\rho^{1/2}U^\dagger. && (158) \end{aligned}$$

We thus have

$$\begin{aligned} d_B(U\rho U^\dagger, U\sigma U^\dagger) &= 2 - 2\operatorname{Tr}\{((U\rho U^\dagger)^{1/2}U\sigma U^\dagger(U\rho U^\dagger)^{1/2})^{1/2}\}^2 && (159)\\ &= 2 - 2\operatorname{Tr}\{(U\rho^{1/2}U^\dagger U\sigma U^\dagger U\rho^{1/2}U^\dagger)^{1/2}\}^2 && (160)\\ &= 2 - 2\operatorname{Tr}\{U(\rho^{1/2}\sigma\rho^{1/2})^{1/2}U^\dagger\}^2 && (161)\\ &= 2 - 2\operatorname{Tr}\{(\rho^{1/2}\sigma\rho^{1/2})^{1/2}U^\dagger U\}^2 && (162)\\ &= 2 - 2\operatorname{Tr}\{(\rho^{1/2}\sigma\rho^{1/2})^{1/2}\}^2 && (163)\\ &= d_B(\rho, \sigma). && (164) \end{aligned}$$

The unitary invariance of the quantum Fisher information matrix then follows directly from its definition:

$$\mathcal{F}[U\rho(\boldsymbol{\theta})U^\dagger]_{ij} = \frac{\partial^2}{\partial\delta_i\partial\delta_j} d_B(U\rho(\boldsymbol{\theta})U^\dagger, U\rho(\boldsymbol{\theta}+\boldsymbol{\delta})U^\dagger)\bigg|_{\boldsymbol{\delta}=0} \tag{165}$$

$$= \frac{\partial^2}{\partial\delta_i\partial\delta_j} d_B(\rho(\boldsymbol{\theta}), \rho(\boldsymbol{\theta}+\boldsymbol{\delta}))\bigg|_{\boldsymbol{\delta}=0} \tag{166}$$

$$= \mathcal{F}[\rho(\boldsymbol{\theta})]_{ij}. \tag{167}$$

(v) Due to the direct sum structure, the terms relating to $\rho(\boldsymbol{\theta})$ and $\sigma(\boldsymbol{\theta})$ in Eq. (139) fall into different parts of the sum, yielding the desired result.

(vi) The tensorization property is a consequence of the tensorization of the Bures fidelity $f_B(\rho,\sigma) = \operatorname{Tr}\{(\rho^{1/2}\sigma\rho^{1/2})^{1/2}\}^2$. Note that the matrix square root of a tensor product is the tensor product of the matrix square roots:

$$(\rho\otimes\sigma)^{1/2} = (\rho^{1/2}\otimes\sigma^{1/2}). \tag{168}$$

This implies that the Bures fidelity is multiplicative with respect to tensor products:

$$f_B(\rho_1\otimes\rho_2, \sigma_1\otimes\sigma_2) = \operatorname{Tr}\{([\rho_1\otimes\rho_2]^{1/2}[\sigma_1\otimes\sigma_2][\rho_1\otimes\rho_2]^{1/2})^{1/2}\}^2 \tag{169}$$

$$= \operatorname{Tr}\{(\rho_1^{1/2}\sigma_1\rho_1^{1/2})^{1/2}\otimes(\rho_2^{1/2}\sigma_2\rho_2^{1/2})^{1/2}\}^2 \tag{170}$$

$$= \operatorname{Tr}\{(\rho_1^{1/2}\sigma_1\rho_1^{1/2})^{1/2}\}^2 \operatorname{Tr}\{(\rho_2^{1/2}\sigma_2\rho_2^{1/2})^{1/2}\}^2 \tag{171}$$

$$= f_B(\rho_1,\sigma_1)f_B(\rho_2,\sigma_2). \tag{172}$$

The quantum Fisher information is proportional to the matrix of second derivatives of the Bures fidelity under small perturbation. We can use this to prove the tensorization property:

$$\mathcal{F}_{ij}[\rho(\boldsymbol{\theta})\otimes\sigma(\boldsymbol{\theta})] = -2\,\frac{\partial^2}{\partial\delta_i\partial\delta_j} f_B(\rho(\boldsymbol{\theta})\otimes\sigma(\boldsymbol{\theta}), \rho(\boldsymbol{\theta}+\boldsymbol{\delta})\otimes\sigma(\boldsymbol{\theta}+\boldsymbol{\delta}))\bigg|_{\boldsymbol{\delta}=0} \tag{173}$$

$$= -2\,\frac{\partial^2}{\partial\delta_i\partial\delta_j} f_B(\rho(\boldsymbol{\theta}), \rho(\boldsymbol{\theta}+\boldsymbol{\delta}))f_B(\sigma(\boldsymbol{\theta}), \sigma(\boldsymbol{\theta}+\boldsymbol{\delta}))\bigg|_{\boldsymbol{\delta}=0} \tag{174}$$

$$= -2\Bigg[\frac{\partial^2 f_B(\rho(\boldsymbol{\theta}), \rho(\boldsymbol{\theta}+\boldsymbol{\delta}))}{\partial\delta_i\partial\delta_j} f_B(\sigma(\boldsymbol{\theta}), \sigma(\boldsymbol{\theta}+\boldsymbol{\delta})) + f_B(\rho(\boldsymbol{\theta}), \rho(\boldsymbol{\theta}+\boldsymbol{\delta}))\frac{\partial^2 f_B(\sigma(\boldsymbol{\theta}), \sigma(\boldsymbol{\theta}+\boldsymbol{\delta}))}{\partial\delta_i\partial\delta_j} \tag{175}$$

$$+ \frac{\partial f_B(\rho(\boldsymbol{\theta}), \rho(\boldsymbol{\theta}+\boldsymbol{\delta}))}{\partial\delta_i}\frac{\partial f_B(\sigma(\boldsymbol{\theta}), \sigma(\boldsymbol{\theta}+\boldsymbol{\delta}))}{\partial\delta_j} + \frac{\partial f_B(\rho(\boldsymbol{\theta}), \rho(\boldsymbol{\theta}+\boldsymbol{\delta}))}{\partial\delta_j}\frac{\partial f_B(\sigma(\boldsymbol{\theta}), \sigma(\boldsymbol{\theta}+\boldsymbol{\delta}))}{\partial\delta_i}\Bigg]_{\boldsymbol{\delta}=0}. \tag{176}$$

We can now exploit the fact that we expand around a minimum, which means that first derivatives vanish. Furthermore, $f_B(\rho,\rho) = 1$ for any ρ, which leaves us with only two terms that we can identify with the quantum Fisher information associated to the individual states:

$$\mathcal{F}_{ij}[\rho(\boldsymbol{\theta})\otimes\sigma(\boldsymbol{\theta})] = -2\left[\frac{\partial^2 f_B(\rho(\boldsymbol{\theta}), \rho(\boldsymbol{\theta}+\boldsymbol{\delta}))}{\partial\delta_i\partial\delta_j} + \frac{\partial^2 f_B(\sigma(\boldsymbol{\theta}), \sigma(\boldsymbol{\theta}+\boldsymbol{\delta}))}{\partial\delta_i\partial\delta_j}\right]_{\boldsymbol{\delta}=0} \tag{177}$$

$$= \mathcal{F}_{ij}[\rho(\boldsymbol{\theta})] + \mathcal{F}_{ij}[\sigma(\boldsymbol{\theta})], \tag{178}$$

which concludes the proof.

(vii) Monotonicity of the Bures divergence implies that under the action of a quantum channel Φ

$$d_B(\Phi[\rho(\boldsymbol{\theta})], \Phi[\rho(\boldsymbol{\theta}+\boldsymbol{\delta})]) \leq d_B(\rho(\boldsymbol{\theta}), \rho(\boldsymbol{\theta}+\boldsymbol{\delta})), \tag{179}$$

which holds independently of $\boldsymbol{\delta}$. In the limit $\|\boldsymbol{\delta}\|\to 0$, we have

$$d_B(\rho(\boldsymbol{\theta}), \rho(\boldsymbol{\theta}+\boldsymbol{\delta})) \approx \frac{1}{2}\boldsymbol{\delta}^T\mathcal{F}[\rho(\boldsymbol{\theta})]\boldsymbol{\delta}, \tag{180}$$

and therefore

$$\boldsymbol{\delta}^{\scriptscriptstyle T}\mathcal{F}[\Phi[\rho(\boldsymbol{\theta})]]\boldsymbol{\delta} \leq \boldsymbol{\delta}^{\scriptscriptstyle T}\mathcal{F}[\rho(\boldsymbol{\theta})]\boldsymbol{\delta}. \tag{181}$$

As this needs to hold for any $\boldsymbol{\delta}$, it implies the sought matrix inequality

$$\mathcal{F}[\Phi[\rho(\boldsymbol{\theta})]] \leq \mathcal{F}[\rho(\boldsymbol{\theta})]. \tag{182}$$

(viii) To prove the transformation rule, we make use of Faà di Bruno's formula for the Hessian. If we have a look at a function $g(\boldsymbol{\theta})$ and seek its second derivatives with respect to coordinates $\boldsymbol{f}(\boldsymbol{\theta})$, we have that

$$\frac{\partial^2 g(\boldsymbol{\theta})}{\partial f_i \partial f_j} = \sum_k \frac{\partial g(\boldsymbol{\theta})}{\partial \theta_k}\frac{\partial^2 u_k}{\partial f_i \partial f_j} + \sum_{kl} \frac{\partial^2 g(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_l}\frac{\partial \theta_k}{\partial f_i}\frac{\partial \theta_l}{\partial f_j}. \tag{183}$$

If the $g(\boldsymbol{\theta})$ is a local optimum (as will be the case for us), the gradient terms vanish and we can rewrite the above using the Jacobian of the coordinate transform $J_{ik} = \partial\theta_k/\partial f_i$ as

$$\frac{\partial^2 g(\boldsymbol{\theta})}{\partial f_i \partial f_j} = \sum_{kl} \frac{\partial^2 g(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_l} J_{ik} J_{jl}. \tag{184}$$

If we denote $H^f_{ij} = \partial^2 g(\boldsymbol{\theta})/\partial f_i \partial f_j$ and $H^\theta_{ij} = \partial^2 g(\boldsymbol{\theta})/\partial \theta_i \partial \theta_j$ we can deduce the transformation rule

$$H^f = J H^\theta J^{\scriptscriptstyle T}. \tag{185}$$

The transformation rule of the quantum Fisher information matrix follows from applying this general result to its definition as the Hessian of the Bures distance.

□