

Minimax quantum state estimation under Bregman divergence

Maria Quadeer, Marco Tomamichel, and Christopher Ferrie

Centre for Quantum Software and Information, University of Technology Sydney, Ultimo NSW 2007, Australia
February 28, 2019

We investigate minimax estimators for quantum state tomography under general Bregman divergences. First, generalizing the work of Koyama et al. [Entropy 19, 618 (2017)] for relative entropy, we find that given any estimator for a quantum state, there always exists a sequence of Bayes estimators that asymptotically perform at least as well as the given estimator, on any state. Second, we show that there always exists a sequence of priors for which the corresponding sequence of Bayes estimators is asymptotically minimax (i.e. it minimizes the worst-case risk). Third, by re-formulating Holevo’s theorem for the covariant state estimation problem in terms of estimators, we find that there exists a covariant measurement that is, in fact, minimax (i.e. it minimizes the worst-case risk). Moreover, we find that a measurement that is covariant only under a unitary 2-design is also minimax. Lastly, in an attempt to understand the problem of finding minimax measurements for general state estimation, we study the qubit case in detail and find that every spherical 2-design is a minimax measurement.

1 Introduction

Quantum state tomography [1, 2] refers to the process of determining an unknown quantum state of a physical system by performing quantum measurements. Any information processing task necessarily involves verifying the output of a quantum channel which mandates the study of quantum state tomography, apart from the unavoidable theoretical necessity. With recent developments leading to a transition of quantum computation from theory to practice, the verification of quantum systems and processes is of particular importance.

Given an unknown quantum state with no prior knowledge, it is clear that the measurement must be *informationally complete*, i.e. a measurement with outcome statistics sufficient to fully specify the quantum state [3]. Conventional data-processing techniques such as *direct inversion* and *maximum likelihood estimation*, thus, implicitly assume that the measurement statistics are informationally complete.

In direct inversion, given a fixed informationally complete measurement, one identifies the frequencies of each outcome with the corresponding probabilities. Then, by inverting Born’s rule one obtains a unique *estimator* for the density operator that reproduces the measurement statistics; an estimator is defined as a map on the set of measurement outcomes \mathcal{X} , $\hat{\rho} : \mathcal{X} \mapsto \mathcal{S}(\mathcal{H})$, where $\mathcal{S}(\mathcal{H})$ is the set of density operators on the underlying Hilbert space \mathcal{H} that describes the physical system. However, this strategy suffers from the drawback that such an estimator might not be a *physical* state and would yield negative eigenvalues.

Example 1. *Suppose one measures an unknown quantum state in \mathbb{C}^2 along the x , y and z directions. Assuming that each of the measurements are performed only once, let us suppose that each of the outcome is ‘up’. Thus, $n_x = n_y = n_z = 1$ and $N_x = N_y = N_z = 1$, so that $p_x = n_x/N_x = 1$, etc. Now, an estimator that would yield the same probabilities would be the one with the Bloch vector: $(2p_x - 1, 2p_y - 1, 2p_z - 1) = (1, 1, 1)$. This is an invalid quantum state as it lies outside the Bloch ball, and thus necessarily has negative eigenvalues.*

Maria Quadeer: mariaquadeer@gmail.com

Reference [4] referred to such a shortcoming of direct inversion and proposed an alternative that enforces positivity on the estimator, called *maximum likelihood estimation*.

A likelihood functional $\mathcal{L}[\rho] : \mathcal{S}(\mathcal{H}) \mapsto [0, 1]$ is the probability of observing a data set \mathbb{D} given that the system is in the state ρ :

$$\mathcal{L}[\rho] = p(\mathbb{D}|\rho) \quad (1)$$

The data set \mathbb{D} is characterized by the outcome set of the given measurement $\{E_1, \dots, E_N | E_i \geq 0, \sum_{i=1}^N E_i = \mathbb{I}\}$. Thus, $p(\mathbb{D}|\rho) = \prod_{i=1}^N (\text{Tr}[E_i \rho])^{n_i}$ where n_i is the number of times the i -th outcome is recorded in \mathbb{D} . Maximum likelihood estimation involves maximizing Equation (1) over the space of density operators $\mathcal{S}(\mathcal{H})$, and thus obtaining as the estimator the state that maximizes the likelihood functional.

The problem with MLE is that the estimator $\hat{\rho}_{MLE}$ can be rank-deficient. A rank-deficient estimator is not good, as it would mean that by performing only finite number of measurements we are absolutely certain to rule out many possibilities. This kind of certainty must be bogus, suggesting that there has to be a better estimator. Let us look at Example 1 once again to illustrate this point.

Example 1 (continued). *Given the choice of measurement and the corresponding outcomes, the likelihood functional is $\mathcal{L}[\rho] = (1 + r_x)(1 + r_y)(1 + r_z)/6^3$, which needs to be maximized under the constraint $\|\vec{r}\| \leq 1$ —that characterizes the physical set of states in \mathbb{C}^2 . This implies that $r_x = r_y = r_z = 1/\sqrt{3}$, which corresponds to an estimator that is a pure state.*

This shows that state estimators that are unphysical in direct inversion get mapped to the closest physical states in MLE, that lie on the state space boundary, and are thus rank-deficient. In fact, it can be shown [5] that if there exists a ρ_{DI} obtained via direct inversion over a data set \mathbb{D} which is physical, then, it also maximizes the likelihood functional, i.e. $\hat{\rho}_{DI} = \hat{\rho}_{MLE}$. Although, Example 1 is an instance of an extreme case where probabilities are approximated by frequencies of a single measurement, it suffices to illustrate that in direct inversion as well as MLE, all that one cares about is to obtain an estimate of the true state that reproduces the observed measurement statistics, regardless of the fact that in the light of new data the state's estimate might change completely. Reference [5] gives a detailed critique of both direct inversion and MLE, proposing *Bayesian Mean Estimation* (BME) to be a more plausible estimation technique. Moreover, it has been shown that such an estimation technique is quantitatively better in reference [6].

Generally speaking, in estimation theory [7], the average measure of closeness of an estimator to the actual state is defined as the *risk*,

$$R(\rho, \hat{\rho}) = \mathbb{E}_{X|\rho}[L(\rho, \hat{\rho}(X))], \quad (2)$$

where X is the random variable corresponding to the measurement outcomes and L is a distance-measure between the true state and the estimator. One way of choosing an optimal estimator is to look at the *average risk*—defined as the expectation of risk with respect to a *prior* distribution over $\mathcal{S}(\mathcal{H})$. Then, by minimizing the average risk over the set of all probability distributions over $\mathcal{S}(\mathcal{H})$, one obtains what is called a *Bayes estimator*, $\hat{\rho}_B$ [7, pg. 228]. In fact, it has been shown that the Bayes estimator is the mean if the loss function is the relative entropy [8], while in reference [9] the same was proved for a more general class of distance-measures called *Bregman divergence* (see Definition 3.3), which generalizes two important distance-measures—relative entropy and Hilbert-Schmidt distance, but in the classical setting. We provide a proof for the quantum setting in Appendix A for completion.

Now, the Bayesian mean estimator for a prior distribution $\pi(\rho)$ is given by

$$\hat{\rho}_B(\mathbb{D}) = \int_{\mathcal{S}(\mathcal{H})} p(\rho|\mathbb{D}) \rho \, d\rho, \quad (3)$$

where $p(\rho|\mathbb{D})$ is the posterior probability density given by the Bayes rule:

$$p(\rho|\mathbb{D}) = \frac{p(\mathbb{D}|\rho)\pi(\rho)}{p(\mathbb{D})}, \quad (4)$$

and $p(\mathbb{D}) = \int_{\mathcal{S}(\mathcal{H})} d\pi(\rho)p(\mathbb{D}|\rho)$. However, BME can yield nonsensical estimators if one starts with a *bad* prior, as the following example illustrates.

Example 2. Consider a σ_X measurement on an unknown quantum state ρ in \mathbb{C}^2 . Suppose there exists a prior $\pi(\rho)$ such that it assigns zero measure to all states in \mathbb{C}^2 but $|-\rangle\langle -|$. A single measurement outcome of ‘+’ rules out the outcome ‘−’ and thus annihilates the prior!

In fact, it should be clear from the above example that some priors can be annihilated by a finite number of (independent) measurements. Thus, in general, one needs a *robust* [5] prior that cannot be annihilated in order to prevent rank-deficient estimates. However, the estimator’s knowledge of the true state can *still* be jeopardized in the presence of an adversary who provides her with a wrong prior. Therefore, although BME seems to be the best bet, it remains inherently ambiguous due to its dependence on the choice of priors. A systematic approach towards deriving *optimality* criteria for priors is thus a compelling problem.

The *minimax* approach, complementary to BME, seems to be doing just that. In *classical statistics*, the problem of estimating probability distributions (analogous to state estimation) has been studied using the *minimax approach* [10–13], that offers an alternative characterization of optimality of estimators. In the minimax approach, one looks at the space of all possible estimators defined on \mathcal{X} and, for each estimator $\hat{\rho}$, picks the state ρ for which it has the worst performance or *risk* (quantified in terms of a suitably chosen distance-measure between the estimator and the true state). Then, the *minimax estimator* is the one that has the *best* worst-case risk. Such an estimator necessarily works for all states $\rho \in \mathcal{S}(\mathcal{H})$. It can be shown [10] that such a minimax estimator is a Bayes estimator given a particular choice of ‘non-informative’ prior. Thus, the solution to the minimax problem leads to a natural identification of a prior.

However, as pointed out in reference [5], no such rigorous statements were known for the quantum analogue of the problem until then. Recently, the authors of reference [14] have studied the quantum minimax estimation problem in analogy to the classical problem [13], quantifying the estimator’s risk in terms of relative entropy. To summarize, they find that given an unknown quantum state ρ and some estimator $\hat{\rho}$ of it, there always exists a sequence of Bayes estimators that perform at least as well as $\hat{\rho}$ in the limiting case. Moreover, they show that there always exists a class of priors, called *latent information priors* (although, conventionally, such priors are called *least favourable*, and we shall follow the convention!) for which there is a corresponding sequence of Bayes estimators whose limit is *minimax*. Finally, they define a *minimax POVM* as a POVM that minimizes the minimax risk, see Definition 4.1, and study the qubit (\mathbb{C}^2) case in detail, obtaining the class of the least favourable priors as well as the minimax POVM for \mathbb{C}^2 .

This paper is divided into six sections—we discuss our main results in Section 2, followed by Section 3 that contains the formalism and Section 4 that contains the proofs in detail. Finally, in Section 5, we discuss the state estimation problem for \mathbb{C}^2 . In Section 6, we summarize the results and outline future work.

2 Main Results

Bayesian mean estimation is arguably a more plausible approach towards state estimation as opposed to maximum likelihood estimation or direct inversion. However, the performance of BME is tied to the choice of the prior. A complementary approach towards the state estimation problem—the *minimax* approach provides a window to explore all possible classes of priors, enabling one to narrow down those that are consistent with the requirements of both the Bayesian and the minimax analysis.

We extend the work done in reference [14] on minimax analysis (as discussed earlier) to a more general class of distance-measures called the *Bregman divergence*, see Definition 3.3, that generalizes both relative entropy and Hilbert-Schmidt distance. We also generalize the minimax POVM for \mathbb{C}^2 to Hilbert-Schmidt distance, finding that such a minimax POVM is a spherical 2-design. Moreover, by re-formulating Holevo’s theorem [15, pg. 171] for the *covariant state estimation problem* in terms of estimators, we find that a *covariant* POVM is, in fact, minimax with Bregman divergence as the distance-measure. Let us discuss these results in detail, informally, postponing the formal statements and proofs to Section 4.

Result 2.1. For any estimator $\hat{\rho}$, there always exists a sequence of Bayes estimators such that the

limit of the sequence performs at least as well as $\hat{\rho}$, i.e.

$$R(\rho, \hat{\rho}) \geq R\left(\rho, \lim_{n \rightarrow \infty} \hat{\rho}_B^{\pi_n}\right), \quad \forall \rho \in \mathcal{S}(\mathcal{H}).$$

So, for a given quantum state and some estimator, the above result says that one can always find a sequence of Bayes estimators that asymptotically perform at least as well as the estimator. That there exists such a sequence of Bayes estimators means that there exists a corresponding convergent sequence of priors, see Equation (3). However, in general, the Bayes estimator is uniquely defined only up to the null set of p_π comprised of outcomes that have zero probability under π . This illustrates that one cannot replace any given estimator by a corresponding Bayes estimator. However, one can still find a sequence of Bayes estimators that, in the limiting case, perform at least as well as the given estimator. This is an interesting result as it benchmarks BME against any given estimation technique. A related question is: “Does there exist a Bayes estimator that is also minimax?” Well, if one is given a minimax estimator, then Result 2.1 tells us that there exists a sequence of Bayes estimators that are at least as good as the minimax estimator, which in turns implies that the limit itself is minimax. In fact, the next result gives us a Bayesian procedure to arrive at such a minimax estimator.

Result 2.2. *There always exists a sequence of priors such that the limit of the sequence maximizes the average risk of a Bayes estimator,*

$$r(\pi, \hat{\rho}_B^\pi) = \int_{\mathcal{S}(\mathcal{H})} d\pi(\rho) R(\rho, \hat{\rho}_B^\pi), \quad (5)$$

and the limit of the respective sequence of Bayes estimators minimizes the worst-case risk, i.e., it is minimax.

We find that the prior that maximizes the average risk of the Bayes estimator—referred to as the least favourable prior (as it maximizes the minimum possible average risk) is the limit of a convergent sequence of priors, such that the limit of the corresponding sequence of Bayes estimators is minimax. Note that the average risk with relative entropy as the loss function is the average of the maximal accessible information—Holevo information for the ensemble $\{\pi(\rho|x), \rho\}$ with respect to the total probability of measurement outcomes. So, maximizing this average Holevo information of the given ensemble over all possible priors is like asking “What is the prior for which the posterior yields maximum accessible information for the ensemble $\{\pi(\rho|x), \rho\}$?” However, no such interpretation can be made for general loss functions such as Bregman divergence.

At this point, it must be noted that the underlying measurement in all the analysis done so far is fixed. Thus, a least favourable prior is inherently tied to the measurement. Any construction of such a prior necessarily implies that we also need to find a class of measurements for which it works. Although it is not clear if one would be able to solve this problem in general, one can do so for at least a subset of the estimation problem—the *covariant state estimation* problem.

In covariant state estimation, given a fixed state ρ_0 , one is interested in estimating the states ρ_θ such that $\rho_\theta \in \{V_g \rho_0 V_g^\dagger\}$ where $g \in G$ is a group element acting on the parameter space Θ with $\theta \in \Theta$ and V_g is the projective unitary representation of the parametric group G . By generalizing Holevo’s theorem [15, Theorem 3.1] to Bregman divergences, we obtain a least favourable prior.

Lemma 2.1. *The uniform measure on the parameter space Θ is a least favourable prior for covariant measurements.*

A covariant measurement is a measurement that reflects the transformation of the state under the group action appropriately in the outcome statistics (see Definition 4.2). Now, the question is: “What is the measurement that minimizes the worst-case risk?” An answer to this question is closely related to finding the class of measurements for the least favourable priors. The only additional information that is needed is if such a class of measurements also minimizes the average risk with respect to the least favourable prior. It turns out that this is indeed the case as far as covariant estimation is concerned.

Result 2.3. *There exists a covariant measurement \mathcal{P}_c which is minimax for covariant state estimation. Moreover, if there exists a measurement \mathcal{P}'_c which is covariant under a subgroup H of G*

such that $\{V_h | h \in H\}$ forms a unitary 2-design, where V_h is the projective unitary representation of the subgroup H , and \mathcal{P}_c and \mathcal{P}'_c have the same seed¹, then \mathcal{P}'_c is also minimax.

It is not so straightforward to generalize these results to the general state estimation problem. To better understand the situation for the general case, we look at the simplest system of a single qubit (extending the results of reference [14] to Hilbert-Schmidt distance—details in Section 6. In particular, we find that every spherical 2-design in \mathbb{C}^2 is a minimax POVM.

3 Formalism

Consider a quantum system \mathcal{S} described by a finite-dimensional Hilbert space \mathcal{H} with $\mathcal{S}(\mathcal{H})$ as the set of density operators on \mathcal{H} . Then, consider a quantum measurement to be an experiment in which the quantum system \mathcal{S} is measured and let \mathcal{X} be the corresponding outcome space of the measurement outcomes. Each possible event of the experiment can be identified with a subset $B \subseteq \mathcal{X}$, the event being ‘the measurement outcome x lies in B ’. The probability distribution of the events is thus defined over a Σ -algebra of the measurable subsets $B \subseteq \mathcal{X}$. To be in touch with physical reality, we choose the outcome space to be a Hausdorff space, i.e. a topological space where for any $x_1, x_2 \in \mathcal{X}$ there exist two disjoint open sets $X_1, X_2 \subset \mathcal{X}$ such that $x_1 \in X_1$ and $x_2 \in X_2$. This ensures that the Σ -algebra is a Borel Σ -algebra generated by countable intersections, countable unions and relative complements of open subsets of \mathcal{X} . Let $\mathcal{P}(\mathcal{H})$ be the set of positive operators on \mathcal{H} .

Definition 3.1 (Quantum measurement). *A Positive Operator-Valued Measure (POVM) is a map $P : \Sigma \mapsto \mathcal{P}(\mathcal{H})$, where Σ is the Σ -algebra of all measurable subsets of \mathcal{X} . Thus, a POVM associates an operator $P(B)$ to each $B \in \Sigma$ satisfying the following:*

1. $P(B) \geq 0, \quad \forall B \in \Sigma.$
2. $P(\mathcal{X}) = \mathbb{I}.$
3. $P\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} P(B_i), \quad \text{where } \forall B_i, B_j \text{ s.t. } B_i \cap B_j = \emptyset.$

The set of all POVMs on Σ forms a convex set denoted by \mathcal{P} . A POVM is *informationally complete* [3] if the operators $\{P(B)\}$ span $L(\mathcal{H})$, the space of linear operators on \mathcal{H} . The measurement statistics of such an IC-POVM is sufficient to determine, uniquely, all possible states that the quantum system could be in, in the limit when an infinite number of measurements are performed. Optimization of data-processing deals with the practical aspect of not having infinite resources and minimizing the corresponding statistical error. Reference [16] reviews the theoretical development of optimization techniques in quantum tomography based on informationally complete measurements. However, in this paper, we make no assumptions on the POVM. In fact, we look at an alternative definition for an optimal POVM—to be discussed later in this section.

The following lemma (see Appendix B for proof) provides a convenient way of representing a POVM as an operator-valued density.

Lemma 3.1 (Existence of a POVM density). *Every $P \in \mathcal{P}$ admits a density, i.e. for any POVM P there exists a finite measure $\mu(dx)$ over \mathcal{X} such that $\mu(\mathcal{X}) = 1$ and*

$$P(B) = \int_B d\mu(x) M(x), \tag{6}$$

with $M(x) \geq 0$, and $\text{Tr}[M(x)] = d$ μ -almost everywhere.

The conditional probability of the event ‘the measurement outcome x' lies in B given that the system is in a state ρ ’ is given by Born’s rule as

$$\Pr[x' \in B | \rho] = \text{tr} P(B)\rho = \int_B d\mu(x) \text{tr} M(x)\rho, \tag{7}$$

¹See Lemma E.3.

or, in the differential form as

$$dp(x|\rho) = d\mu(x) \operatorname{tr} M(x)\rho \quad (8)$$

which will come in handy later. Now that we have defined a quantum measurement, we proceed with the formulation.

In estimation theory [7], one typically parametrizes the system \mathcal{S} by a parameter θ . The data set of the measurement outcomes is represented by a *random variable* X . Using this data set one estimates the parameter θ or more generally ρ_θ —the *estimand*. Succinctly, this involves two random variables Θ and X defined as below:

- In the Bayesian model, the quantity θ that parametrizes the system \mathcal{S} is treated as a random variable Θ . This random variable is defined over the parameter space Ω_Θ ² and is distributed according to an a-priori probability distribution $\pi_\Theta \in \mathcal{P}(\Theta)$ (where $\mathcal{P}(\Theta)$ is the set of all probability distributions on Θ).
- X is the random variable associated with the outcomes of the measurement performed on the system \mathcal{S} , defined over the sample space \mathcal{X} . The outcomes of the measurement are conditioned on the random variable Θ . Thus, X is distributed according to the conditional probability $p_X(x|\theta)$, given by Equation (7).

The parameter space Θ is chosen to be a compact metric space. The set of all bounded continuous real-valued function on Θ is denoted by $\mathcal{C}(\Theta, \mathbb{R})$. The set of probability distributions $\mathcal{P}(\Theta)$ on Θ is endowed with a weak topology, which essentially defines the notion of *weak convergence*.

Definition 3.2. A sequence of probability measures $\pi_n \in \mathcal{P}(\Theta)$ weak converges to μ if for every $f \in \mathcal{C}(\Theta, \mathbb{R})$,

$$\int f d\pi_n \rightarrow \int f d\mu, \quad \text{as } n \rightarrow \infty. \quad (9)$$

Then, as Θ is a compact metric space and $\mathcal{P}(\Theta)$ is endowed with a weak topology, by [17, Theorem 6.4], it implies that $\mathcal{P}(\Theta)$ is also a compact metric space.

The central problem in quantum state estimation is to obtain an *estimator* of ρ_θ . We define an *estimator* as the map

$$\hat{\rho} : \mathcal{X} \mapsto \mathcal{S}(\mathcal{H}). \quad (10)$$

The value of $\hat{\rho}(x)$ is the estimate of ρ_θ when the measurement outcome is $X = x$. We want $\hat{\rho}(X)$ to be close to ρ_θ , but $\hat{\rho}(X)$ is a random variable. One way of defining a meaningful measure of closeness is by defining an expectation over the conditional distribution of X , Equation (7). Let $L(\rho_\theta, \hat{\rho}(x))$ be the *loss function* that quantifies the closeness of an estimated state $\hat{\rho}(x)$ to the true state ρ_θ . We assume two things about L :

1. $L(\rho_\theta, \hat{\rho}(x)) \geq 0, \forall \theta \in \Theta, \hat{\rho}$, with equality if and only if $\rho_\theta = \hat{\rho}(x)$.
2. $L(\rho_\theta, \rho_\theta) = 0, \quad \forall \theta \in \Theta$.

The average measure of closeness of $\hat{\rho}(X)$ to ρ_θ is defined as the *risk function*

$$R(\rho_\theta, \hat{\rho}) = \mathbb{E}_{X|\theta}[L(\rho_\theta, \hat{\rho}(X))]. \quad (11)$$

One would like to obtain an estimator that minimizes the risk for all values of θ . Obviously, this problem does not have a solution, i.e. there does not exist an estimator that uniformly minimizes the risk for all values of θ except for the case when ρ_θ is a constant. Instead, one can look at the following two quantities that are a good measure of the risk in a global sense:

²However, we will abuse notation and refer to the parameter space as Θ from now onwards.

1. Average risk:

$$r(\pi, \hat{\rho}) = \int_{\Theta} d\pi(\theta) R(\rho_{\theta}, \hat{\rho}), \quad (12)$$

where $\pi(\theta)$ is an a-priori distribution over the parameter space Θ .

2. Worst-case/minimax risk:

$$\inf_{\hat{\rho}} \sup_{\theta} R(\rho_{\theta}, \hat{\rho}). \quad (13)$$

The estimator that minimizes the average risk is the Bayes estimator $\hat{\rho}_B$ [7, pg.228]. In reference [8] it was shown that the Bayes estimator is the mean if the loss function is the relative entropy $D(\rho_{\theta} || \hat{\rho}(x))$, i.e.,

$$\hat{\rho}_B = x \mapsto \int_{\Theta} d\pi(\theta|x) \rho_{\theta}, \quad (14)$$

where $d\pi(\theta|x)$ is the posterior probability distribution obtained via the Bayes rule,

$$d\pi(\theta|x) = \frac{dp(x|\theta)}{dp_{\pi}(x)} d\pi(\theta),$$

where $p_{\pi}(B) = \int_B \int_{\Theta} dp(x|\theta) d\pi(\theta)$. Note that in the continuous case, the likelihood ratio $\frac{p(x|\theta)}{p_{\pi}(x)}$ is replaced by the corresponding Radon-Nikodym derivative, which is defined uniquely upto the null set of p_{π} . In fact, the same was proved [9] for a more general class of distance-measures called *Bregman divergence* which is the measure we will use in our analysis, but only in the classical setting. We provide a proof for the quantum setting in Appendix A for completion. Let us now define Bregman divergence.

Definition 3.3 (Bregman divergence for density matrices). *Let $f : [0, 1] \mapsto \mathbb{R}$ be a strictly convex continuously-differentiable real-valued function. Then, the Bregman divergence between density matrices ρ, σ is defined as*

$$D_f(\rho, \sigma) = \text{tr} (f(\rho) - f(\sigma) - f'(\sigma)(\rho - \sigma)).$$

Bregman divergence generalizes two important classes of distance-measures: the relative entropy obtained by choosing $f : x \mapsto x \log x$ and the Hilbert-Schmidt distance (Schatten 2-norm) obtained by choosing $f : x \mapsto x^2$.

Let us now look at a few of its important properties. First, Bregman divergence is invariant under unitary transformations of its arguments, i.e. $D_f(U\rho U^{\dagger}, U\sigma U^{\dagger}) = D_f(\rho, \sigma)$. Second, it is not a metric, as it is neither symmetric nor satisfies the triangle inequality, but by the strict convexity of f , $D_f(\rho, \sigma) \geq 0$, with equality if and only if $\rho = \sigma$. Third, the convexity of f implies that $D_f(., .)$ is convex in its first argument; it is jointly convex if f'' is operator convex and numerically non-increasing [18]. Moreover, by generalizing the proof of lower semi-continuity of relative entropy as in reference [19], we obtain the lower semi-continuity of Bregman divergence (see Appendix C for the proof).

We are now ready to state and prove the main results of this paper.

4 Proofs

4.1 Bayesian state estimation

Formally, Result 2.1 is stated as the following theorem.

Theorem 4.1. *Let $\hat{\rho} : \mathcal{X} \mapsto \mathcal{S}(\mathcal{H})$ be an estimator. Then, there exists a convergent sequence of priors such that the corresponding sequence of Bayes estimators $(\hat{\rho}_B^{\pi_n})_n$ converges, with*

$$R(\rho_{\theta}, \hat{\rho}) \geq R\left(\rho_{\theta}, \lim_{n \rightarrow \infty} \hat{\rho}_B^{\pi_n}\right), \quad \forall \theta \in \Theta. \quad (15)$$

Proof. Consider the average distance between the Bayes estimator and a given estimator $\hat{\rho}$ for some prior $\pi \in \mathcal{P}(\Theta)$ as the map

$$g : \pi \mapsto D_{\hat{\rho}}^f(\pi) = \int_{\mathcal{X}} dp_{\pi}(x) D_f(\hat{\rho}_B^{\pi}(x), \hat{\rho}(x)).$$

Now, the Bayes estimator is uniquely defined up to the null set of p_{π} . In fact, it is discontinuous on \mathcal{X} , see Appendix D, and the points of discontinuity belong to the null set of p_{π} . So, unless the null set of p_{π} is empty, the Bayes estimator cannot be defined continuously since there can exist different sequences that converge to the same prior, but the limit of the corresponding sequences of Bayes estimators may not coincide on the null set of p_{π} . To deal with the discontinuity of the Bayes estimator, we consider closed subsets of $\mathcal{P}(\Theta)$ with the defining property that every element of these subsets renders the corresponding Bayes estimator continuous on \mathcal{X} . Then, g is lower semi-continuous on each closed subset as Bregman divergence is lower-semi continuous (Appendix C). Thus, there exists a prior π_n in every subset that minimizes it on that subset. So, we look at the sequence of such priors $(\pi_n)_n$ and find that the corresponding sequence of Bayes estimators $(\hat{\rho}_B^{\pi_n})_n$ converges to a limit that has a risk lower than or equal to that of the given estimator. Let us now proceed with the proof.

We define the closed subsets of $\mathcal{P}(\Theta)$ as

$$\mathcal{P}_{\mu/n} = \left\{ \frac{\mu}{n} + \left(1 - \frac{1}{n}\right) \pi \mid \pi \in \mathcal{P}(\Theta) \right\}, \quad (16)$$

where μ is a measure such that $p_{\mu}(x) > 0$, for all $x \in \mathcal{X}$. The latter condition ensures that the Bayes estimator for a prior that lies in $\mathcal{P}_{\mu/n}$ is continuous on $\mathcal{P}_{\mu/n}$. Then, as a closed subset of a compact set is compact, there exists a prior $\pi_n \in \mathcal{P}_{\mu/n}$ such that $D_{\hat{\rho}}^f(\pi_n) = \inf_{\pi \in \mathcal{P}_{\mu/n}} D_{\hat{\rho}}^f(\pi)$. In fact, as $\mathcal{P}(\Theta)$ is a compact metric space, the sequence of priors $(\pi_n)_n$ has a convergent subsequence. Let us denote this subsequence as $(\pi'_m)_m$. Let n_m be such that $\pi_{n_m} = \pi'_m$.

Then, the idea is to use the fact that each π'_m minimizes $D_{\hat{\rho}}^f$ on the corresponding closed subset \mathcal{P}_{μ/n_m} to obtain a suitable condition. To begin with, we define a prior in the neighbourhood of $\pi'_{m+1} \in \mathcal{P}_{\mu/n_{m+1}}$ by taking a convex sum of it with another element in $\mathcal{P}_{\mu/n_{m+1}}$. Observe that $\left(\frac{n_m}{n_{m+1}}\pi'_{m+1} + \left(1 - \frac{n_m}{n_{m+1}}\right)\delta(\theta - \theta_0)\right)$ lies in $\mathcal{P}_{\mu/n_{m+1}}$, for any $\theta_0 \in \Theta$. So, we define a prior

$$\pi(\theta) = u \left(\frac{n_m}{n_{m+1}}\pi'_{m+1} + \left(1 - \frac{n_m}{n_{m+1}}\right)\delta(\theta - \theta_0) \right) + (1 - u)\pi'_{m+1}, \quad (17)$$

with $0 \leq u \leq 1$. This is like considering a perturbation in the neighbourhood of π'_{m+1} and noting that the derivative of $D_{\hat{\rho}}^f(\pi)$ is positive as one approaches π'_{m+1} as it minimizes $D_{\hat{\rho}}^f(\pi)$ on the set $\mathcal{P}_{\mu/n_{m+1}}$. Thus, we have

$$\begin{aligned} 0 \leq \frac{d}{du} D_{\hat{\rho}}^f(\pi) \Big|_{u=0} &= \frac{d}{du} \int_{\mathcal{X}} dp_{\pi}(x) D_f(\hat{\rho}_B^{\pi}(x), \hat{\rho}(x)) \Big|_{u=0} \\ &= \int_{\mathcal{X}} \frac{dp_{\pi}(x)}{du} \Big|_{u=0} D_f(\hat{\rho}_B^{\pi'_{m+1}}(x), \hat{\rho}(x)) + \int_{\mathcal{X}} dp_{\pi'_{m+1}}(x) \frac{d}{du} D_f(\hat{\rho}_B^{\pi}(x), \hat{\rho}(x)) \Big|_{u=0}. \end{aligned} \quad (18)$$

Let $k = \frac{n_m}{n_{m+1}}$. Then,

$$\begin{aligned} dp_{\pi}(x) &= \int dp(x|\theta) \left(u \left(k\pi'_{m+1}(\theta) + (1 - k)\delta(\theta - \theta_0) \right) + (1 - u)\pi'_{m+1}(\theta) \right) d\theta. \\ \implies \frac{dp_{\pi}(x)}{du} \Big|_{u=0} &= k dp_{\pi'_{m+1}}(x) + (1 - k) dp(x|\theta_0) - dp_{\pi'_{m+1}}(x). \end{aligned}$$

So, the first term (18) is

$$\int_{\mathcal{X}} \left(k dp_{\pi'_m}(x) + (1-k) dp(x|\theta_0) - dp_{\pi'_{m+1}}(x) \right) D_f(\hat{\rho}_B^{\pi'_{m+1}}(x), \hat{\rho}(x)),$$

while the second term, using Lemma E.1, is

$$\begin{aligned} &= \int_{\mathcal{X}} dp_{\pi'_{m+1}}(x) \frac{d}{du} \operatorname{tr} f(\hat{\rho}_B^\pi(x)) - f'(\hat{\rho}(x))(\hat{\rho}_B^\pi(x) - \hat{\rho}(x)) \Big|_{u=0} \\ &= \int_{\mathcal{X}} dp_{\pi'_{m+1}}(x) \operatorname{tr} \left(f'(\hat{\rho}_B^{\pi'_{m+1}}(x)) - f'(\hat{\rho}(x)) \right) \frac{d}{du} \hat{\rho}_B^\pi(x) \Big|_{u=0}. \end{aligned} \quad (19)$$

Let us now calculate the derivative of the Bayes estimator.

$$\begin{aligned} \frac{d}{du} \hat{\rho}_B^\pi(x) \Big|_{u=0} &= \frac{d}{du} \int_{\Theta} \frac{dp(x|\theta)}{dp_\pi(x)} \rho_\theta d\pi(\theta) \Big|_{u=0} \\ &= \frac{d}{du} \int_{\Theta} \frac{1}{\frac{dp_\pi(x)}{dp(x|\theta)}} \rho_\theta d\pi(\theta) \Big|_{u=0} \\ &= \int_{\Theta} \frac{\rho_\theta \left(kd\pi'_m(\theta) + (1-k)\delta(\theta - \theta_0)d\theta - d\pi'_{m+1}(\theta) \right)}{\frac{dp_{\pi'_{m+1}}}{dp(x|\theta)}} \\ &\quad \int_{\Theta} \frac{\rho_\theta d\pi'_{m+1}(\theta)}{\left(\frac{dp_{\pi'_{m+1}}}{dp(x|\theta)} \right)^2} \frac{d}{du} \left(\frac{dp_\pi(x)}{dp(x|\theta)} \right) \Big|_{u=0}. \end{aligned} \quad (20)$$

Now, the derivative with respect to u in the second term can be calculated by exchanging the order of differentiation as $p(x|\theta)$ is independent of u . So, we have

$$\begin{aligned} \frac{d}{du} \left(\frac{dp_\pi(x)}{dp(x|\theta)} \right) \Big|_{u=0} &= \frac{d}{dp(x|\theta)} \left(\frac{dp_\pi(x)}{du} \right) \Big|_{u=0} \\ &= k \frac{dp_{\pi'_m}(x)}{dp(x|\theta)} + (1-k) \frac{dp(x|\theta_0)}{dp(x|\theta)} - \frac{dp_{\pi'_{m+1}}(x)}{dp(x|\theta)}. \end{aligned}$$

Plugging in (20) we obtain,

$$\begin{aligned} \frac{d}{du} \hat{\rho}_B^\pi(x) \Big|_{u=0} &= \int_{\Theta} \frac{\rho_\theta \left(kd\pi'_m(\theta) + (1-k)\delta(\theta - \theta_0)d\theta - d\pi'_{m+1}(\theta) \right)}{\frac{dp_{\pi'_{m+1}}}{dp(x|\theta)}} \\ &\quad \int_{\Theta} \frac{\rho_\theta d\pi'_{m+1}(\theta)}{\left(\frac{dp_{\pi'_{m+1}}}{dp(x|\theta)} \right)^2} \left(k \frac{dp_{\pi'_m}(x)}{dp(x|\theta)} + (1-k) \frac{dp(x|\theta_0)}{dp(x|\theta)} - \frac{dp_{\pi'_{m+1}}(x)}{dp(x|\theta)} \right). \end{aligned}$$

Now, in the limit $m \rightarrow \infty$, the coefficient of k vanishes in the expression above due to weak convergence, while the last term of the first integral cancels with the last term of the second integral. So, we have

$$\lim_{m \rightarrow \infty} \frac{d}{du} \hat{\rho}_B^\pi(x) \Big|_{u=0} = \lim_{m \rightarrow \infty} (1-k) \frac{dp(x|\theta_0)}{dp_{\pi'_{m+1}}(x)} \left(\rho_{\theta_0} - \hat{\rho}_B^{\pi'_{m+1}}(x) \right). \quad (21)$$

Applying the limit and plugging (21) in (19), we find that the second term in (18) is

$$= \lim_{m \rightarrow \infty} \int_{\mathcal{X}} dp_{\pi'_{m+1}}(x) \operatorname{tr} \left(f'(\hat{\rho}_B^{\pi'_{m+1}}(x)) - f'(\hat{\rho}(x)) \right) (1-k) \frac{dp(x|\theta_0)}{dp_{\pi'_{m+1}}(x)} \left(\rho_{\theta_0} - \hat{\rho}_B^{\pi'_{m+1}}(x) \right)$$

$$= \lim_{m \rightarrow \infty} (1-k) \int_{\mathcal{X}} dp(x|\theta_0) \operatorname{tr} \left(f'(\hat{\rho}_B^{\pi_{m+1}'}(x)) - f'(\hat{\rho}(x)) \right) (\rho_{\theta_0} - \hat{\rho}_B^{\pi_{m+1}'}(x)).$$

Finally, combining both the terms of (18) and applying the limit, we find that

$$0 \leq \lim_{m \rightarrow \infty} \frac{d}{du} D_{\hat{\rho}}^f(\pi) \Big|_{u=0} = \lim_{m \rightarrow \infty} \int_{\mathcal{X}} \left(k dp_{\pi_{m+1}'}(x) + (1-k) dp(x|\theta_0) - dp_{\pi_{m+1}'}(x) \right) D_f(\hat{\rho}_B^{\pi_{m+1}'}(x), \hat{\rho}(x)) + (1-k) \int_{\mathcal{X}} dp(x|\theta_0) \operatorname{tr} \left(f'(\hat{\rho}_B^{\pi_{m+1}'}(x)) - f'(\hat{\rho}(x)) \right) (\rho_{\theta_0} - \hat{\rho}_B^{\pi_{m+1}'}(x)).$$

This implies that

$$\lim_{m \rightarrow \infty} (1-k) \int_{\mathcal{X}} dp_{\pi_{m+1}'}(x) D_f(\hat{\rho}_B^{\pi_{m+1}'}(x), \hat{\rho}(x)) \leq \lim_{m \rightarrow \infty} (1-k) \int_{\mathcal{X}} dp(x|\theta_0) \operatorname{tr} \left(f'(\hat{\rho}_B^{\pi_{m+1}'}(x)) - f'(\hat{\rho}(x)) \right) (\rho_{\theta_0} - \hat{\rho}_B^{\pi_{m+1}'}(x)) + \int_{\mathcal{X}} dp(x|\theta_0) D_f(\hat{\rho}_B^{\pi_{m+1}'}(x), \hat{\rho}(x)).$$

The right-hand side of the inequality above can be rearranged to obtain

$$\lim_{m \rightarrow \infty} \underbrace{\int_{\mathcal{X}} dp_{\pi_{m+1}'}(x) D_f(\hat{\rho}_B^{\pi_{m+1}'}(x), \hat{\rho}(x))}_{\geq 0, \text{ due to non-negativity of Bregman divergence}} \leq \lim_{m \rightarrow \infty} \int_{\mathcal{X}} dp(x|\theta_0) D_f(\rho_{\theta_0}, \hat{\rho}(x)) - D_f(\rho_{\theta_0}, \hat{\rho}_B^{\pi_{m+1}'}(x)).$$

This implies that

$$\lim_{m \rightarrow \infty} R(\rho_{\theta_0}, \hat{\rho}_B^{\pi_{m+1}'}(x)) \leq R(\rho_{\theta_0}, \hat{\rho}), \quad \forall \theta_0 \in \Theta.$$

But, as Bregman divergence is lower semi-continuous (Appendix C), we have

$$R(\rho_{\theta_0}, \lim_{m \rightarrow \infty} \hat{\rho}_B^{\pi_{m+1}'}(x)) \leq \lim_{m \rightarrow \infty} R(\rho_{\theta_0}, \hat{\rho}_B^{\pi_{m+1}'}(x)).$$

Therefore, we arrive at our result, i.e.

$$R(\rho_{\theta_0}, \lim_{m \rightarrow \infty} \hat{\rho}_B^{\pi_{m+1}'}(x)) \leq R(\rho_{\theta_0}, \hat{\rho}), \quad \forall \theta_0 \in \Theta.$$

□

4.2 A Bayesian method for minimax state estimation

Formally, Result 2.2 is stated as the following theorem.

Theorem 4.2. *There exists a convergent sequence of priors $(\pi_n)_n$ such that the limit of the sequence maximizes the average risk, Equation (5), of the Bayes estimator. The limit of such a sequence is referred to as a least favourable prior. Moreover, the sequence of Bayes estimators $(\hat{\rho}_B^{\pi_n})_n$ converges such that the limit of the sequence is minimax, i.e*

$$\inf_{\hat{\rho}} \sup_{\theta} R(\rho_{\theta}, \hat{\rho}) = \sup_{\theta} R(\rho_{\theta}, \lim_{n \rightarrow \infty} \hat{\rho}_B^{\pi_n}).$$

Proof. Consider the average risk of the Bayes estimator for a prior $\pi \in \mathcal{P}(\Theta)$ as the map

$$h : \pi \mapsto r(\pi, \hat{\rho}_B^{\pi}) = \int_{\Theta} d\pi(\theta) \int_{\mathcal{X}} dp(x|\theta) D_f(\rho_{\theta}, \hat{\rho}_B^{\pi}(x)).$$

Due to the discontinuity of the Bayes estimator, we follow the same regularization arguments as made earlier and define closed subsets of $\mathcal{P}(\Theta)$, Equation (16), such that the Bayes estimator is continuous on each of these subsets. The map $h : \pi \mapsto r(\pi, \hat{\rho}_B^{\pi})$ is then continuous on each of the subsets. Since these subsets are closed subsets of a compact set they are themselves compact.

Therefore, h attains a maximum on each of the subsets. Then, denoting the maxima in each subset $\mathcal{P}_{\mu/n_{m+1}}$ as π'_{m+1} , we define a prior as done earlier in Equation (17) as a convex sum of π'_{m+1} and another element in $\mathcal{P}_{\mu/n_{m+1}}$. Since the average risk is maximized on $\mathcal{P}_{\mu/n_{m+1}}$, the derivative of $r(\pi, \hat{\rho}_B^\pi)$ is negative as one approaches π'_{m+1} , i.e.

$$\begin{aligned} 0 \geq \frac{d}{du} r(\pi, \hat{\rho}_B^\pi) \Big|_{u=0} &= \frac{d}{du} \int_{\Theta} d\pi(\theta) \int_{\mathcal{X}} dp(x|\theta) D_f(\rho_\theta, \hat{\rho}_B^\pi(x)) \Big|_{u=0} \\ &= \int_{\Theta} \frac{d\pi(\theta)}{du} \Big|_{u=0} \int_{\mathcal{X}} dp(x|\theta) D_f(\rho_\theta, \hat{\rho}_B^{\pi'_{m+1}}(x)) + \\ &\quad \int_{\Theta} d\pi'_{m+1}(\theta) \int_{\mathcal{X}} dp(x|\theta) \frac{d}{du} D_f(\rho_\theta, \hat{\rho}_B^\pi(x)) \Big|_{u=0}. \end{aligned} \quad (22)$$

Evaluating the derivatives using Lemma E.1, the first term in (22) is

$$\int_{\mathcal{X}} dp(x|\theta_0) D_f(\rho_{\theta_0}, \hat{\rho}_B^{\pi'_{m+1}}(x)) - \int_{\Theta} d\pi'_{m+1}(\theta) \int_{\mathcal{X}} dp(x|\theta) D_f(\rho_\theta, \hat{\rho}_B^{\pi'_{m+1}}(x)),$$

while the derivative in the second term of (22) is

$$\begin{aligned} \frac{d}{du} D_f(\rho_\theta, \hat{\rho}_B^\pi(x)) \Big|_{u=0} &= \frac{d}{du} \text{tr} \left[f(\rho_\theta) - f(\hat{\rho}_B^\pi(x)) - f'(\hat{\rho}_B^\pi(x))(\rho_\theta - \hat{\rho}_B^\pi(x)) \right] \Big|_{u=0} \\ &= \text{tr} \left[-f'(\hat{\rho}_B^{\pi'_{m+1}}(x)) \frac{d}{du} \hat{\rho}_B^\pi(x) \Big|_{u=0} + f'(\hat{\rho}_B^{\pi'_{m+1}}(x)) \frac{d}{du} \hat{\rho}_B^\pi(x) \Big|_{u=0} \right. \\ &\quad \left. - (\rho_\theta - \hat{\rho}_B^{\pi'_{m+1}}(x)) \frac{d}{du} f'(\hat{\rho}_B^\pi(x)) \Big|_{u=0} \right] \\ &= \text{tr} \left[(\hat{\rho}_B^{\pi'_{m+1}}(x) - \rho_\theta) \frac{d}{du} f'(\hat{\rho}_B^\pi(x)) \Big|_{u=0} \right]. \end{aligned}$$

So, plugging this in the second term of (22), we have

$$\begin{aligned} &\int_{\Theta} d\pi'_{m+1}(\theta) \int_{\mathcal{X}} dp(x|\theta) \text{tr} \left[(\hat{\rho}_B^{\pi'_{m+1}}(x) - \rho_\theta) \frac{d}{du} f'(\hat{\rho}_B^\pi(x)) \Big|_{u=0} \right] \\ &= \text{tr} \left[\int_{\mathcal{X}} \int_{\Theta} d\pi'_{m+1}(\theta) dp(x|\theta) (\hat{\rho}_B^{\pi'_{m+1}}(x) - \rho_\theta) \frac{d}{du} f'(\hat{\rho}_B^\pi(x)) \Big|_{u=0} \right] \\ &= \text{tr} \left[\int_{\mathcal{X}} \left(dp_{\pi'_{m+1}}(x) \hat{\rho}_B^{\pi'_{m+1}}(x) - \int_{\Theta} d\pi'_{m+1}(\theta) dp(x|\theta) \rho_\theta \right) \frac{d}{du} f'(\hat{\rho}_B^\pi(x)) \Big|_{u=0} \right] = 0. \end{aligned} \quad (23)$$

Thus, applying the limit $m \rightarrow \infty$ we arrive at the following inequality,

$$0 \geq \lim_{m \rightarrow \infty} \int_{\mathcal{X}} dp(x|\theta_0) D_f(\rho_{\theta_0}, \hat{\rho}_B^{\pi'_{m+1}}(x)) - \int_{\Theta} d\pi'_{m+1}(\theta) \int_{\mathcal{X}} dp(x|\theta) D_f(\rho_\theta, \hat{\rho}_B^{\pi'_{m+1}}(x)),$$

which implies that

$$\lim_{m \rightarrow \infty} \int_{\mathcal{X}} dp(x|\theta_0) D_f(\rho_{\theta_0}, \hat{\rho}_B^{\pi'_{m+1}}(x)) \leq \lim_{m \rightarrow \infty} \int_{\Theta} d\pi'_{m+1}(\theta) \int_{\mathcal{X}} dp(x|\theta) D_f(\rho_\theta, \hat{\rho}_B^{\pi'_{m+1}}(x)).$$

So, we have

$$\lim_{m \rightarrow \infty} R(\rho_{\theta_0}, \hat{\rho}_B^{\pi'_{m+1}}) \leq \lim_{m \rightarrow \infty} \int_{\Theta} d\pi'_{m+1}(\theta) R(\rho_\theta, \hat{\rho}_B^{\pi'_{m+1}}), \quad \forall \theta_0 \in \Theta.$$

The lower semi-continuity of Bregman divergence implies that

$$R(\rho_{\theta_0}, \lim_{m \rightarrow \infty} \hat{\rho}_B^{\pi'_{m+1}}) \leq \lim_{m \rightarrow \infty} \int_{\Theta} d\pi'_{m+1}(\theta) R(\rho_\theta, \hat{\rho}_B^{\pi'_{m+1}}).$$

Thus, we have

$$\sup_{\theta_0} R(\rho_{\theta_0}, \lim_{m \rightarrow \infty} \hat{\rho}_B^{\pi_{m+1}'}) \leq \lim_{m \rightarrow \infty} \int_{\Theta} d\pi_{m+1}'(\theta) R(\rho_{\theta}, \hat{\rho}_B^{\pi_{m+1}'}).$$

But, the other direction of the inequality above is true trivially i.e.

$$\sup_{\theta_0} R(\rho_{\theta_0}, \lim_{m \rightarrow \infty} \hat{\rho}_B^{\pi_{m+1}'}) \geq \lim_{m \rightarrow \infty} \int_{\Theta} d\pi_{m+1}'(\theta) R(\rho_{\theta}, \hat{\rho}_B^{\pi_{m+1}'}).$$

Therefore, we obtain

$$\sup_{\theta_0} R(\rho_{\theta_0}, \lim_{m \rightarrow \infty} \hat{\rho}_B^{\pi_{m+1}'}) = \lim_{m \rightarrow \infty} \int_{\Theta} d\pi_{m+1}'(\theta) R(\rho_{\theta}, \hat{\rho}_B^{\pi_{m+1}'}). \quad (24)$$

But,

$$\lim_{m \rightarrow \infty} \int_{\Theta} d\pi_{m+1}'(\theta) R(\rho_{\theta}, \hat{\rho}_B^{\pi_{m+1}'}) = \lim_{m \rightarrow \infty} \sup_{\pi \in \mathcal{P}_{\mu/n_{m+1}}} \int_{\Theta} d\pi(\theta) R(\rho_{\theta}, \hat{\rho}_B^{\pi}).$$

By Lemma E.2, the limit of the suprema over subsets $\mathcal{P}_{\mu/n_{m+1}}$ can be replaced by a supremum over the set $\mathcal{P}(\Theta)$ since the sequence of subsets $\mathcal{P}_{\mu/n_{m+1}}$ is dense in $\mathcal{P}(\Theta)$. Thus, we have

$$\begin{aligned} \lim_{m \rightarrow \infty} \int_{\Theta} d\pi_{m+1}'(\theta) R(\rho_{\theta}, \hat{\rho}_B^{\pi_{m+1}'}) &= \sup_{\pi \in \mathcal{P}(\Theta)} \int_{\Theta} d\pi(\theta) R(\rho_{\theta}, \hat{\rho}_B^{\pi}) \\ &= \sup_{\pi \in \mathcal{P}(\Theta)} \inf_{\hat{\rho}} \int_{\Theta} d\pi(\theta) R(\rho_{\theta}, \hat{\rho}). \end{aligned} \quad (25)$$

Using the minimax theorem for lower semi-continuous and quasi-convex functions [20, Theorem 3.4], we can exchange the infimum and the supremum to obtain

$$\sup_{\pi \in \mathcal{P}(\Theta)} \inf_{\hat{\rho}} \int_{\Theta} d\pi(\theta) R(\rho_{\theta}, \hat{\rho}) = \inf_{\hat{\rho}} \sup_{\pi \in \mathcal{P}(\Theta)} \int_{\Theta} d\pi(\theta) R(\rho_{\theta}, \hat{\rho}) = \inf_{\hat{\rho}} \sup_{\Theta} R(\rho_{\theta}, \hat{\rho}).$$

Thus, by (24) and (25) we have the result

$$\inf_{\hat{\rho}} \sup_{\theta} R(\rho_{\theta}, \hat{\rho}) = \sup_{\theta} R(\rho_{\theta}, \lim_{m \rightarrow \infty} \hat{\rho}_B^{\pi_{m+1}'}).$$

□

Result 2.1 and Result 2.2 are based on the assumption that the underlying POVM is fixed. However, in general, the risk depends on the POVM P , i.e. $R(\rho_{\theta}, \hat{\rho}) \equiv R_P(\rho_{\theta}, \hat{\rho})$. One way of defining an optimal POVM could be to minimize the worst-case risk over \mathcal{P} , the convex set of all POVMs. The POVM that minimizes the worst-case risk is called a *minimax* POVM.

Definition 4.1 (Minimax POVM). *A POVM P^* is minimax if:*

$$\inf_{P \in \mathcal{P}} \inf_{\hat{\rho}} \sup_{\theta} R_P(\rho_{\theta}, \hat{\rho}) = \inf_{\hat{\rho}} \sup_{\theta} R_{P^*}(\rho_{\theta}, \hat{\rho}) \quad (26)$$

where ρ_{θ} is the estimand, $\hat{\rho} : \mathcal{X} \mapsto \mathcal{S}(\mathcal{H})$ is an estimator with risk $R_P(\rho_{\theta}, \hat{\rho})$ which is a function of the POVM P , and \mathcal{P} is the convex set of all POVMs on the measurement outcome space \mathcal{X} .

It remains unclear as to how one could obtain a minimax POVM for general state estimation, but if we restrict ourselves to the case of *covariant state estimation* the situation simplifies.

4.3 Covariant state estimation

In the covariant state estimation problem, as discussed in reference [15], one is given a fixed state ρ_{θ_0} and is interested in estimating all the states ρ_θ that lie in the orbit $\{V_g \rho_{\theta_0} V_g^\dagger\}$, where $g \in G$ is a parametric group of transformations of the parameter space Θ and $g \mapsto V_g$ is a (continuous) projective unitary representation of G . One can think of this as representing the following physical scenario. Given that the parameter θ labels the quantum states of the Hilbert space \mathcal{H} , θ can be assumed to be describing some aspects of the preparation procedure for the state ρ_θ —a transformation g of the parameter θ_0 results in the preparation of the state $\rho_\theta = V_g \rho_{\theta_0} V_g^\dagger$ where $\theta = g\theta_0$. Covariant state estimation thus corresponds to the estimation of the state ρ_θ with the measurement outcome space \mathcal{X} being identical to the parameter space Θ . Let us first define a *covariant* measurement.

Definition 4.2 (Covariant measurement). *Let G be a parametric group of transformations of a set Θ and $g \mapsto V_g$ be a (continuous) projective unitary representation of G in a Hilbert space \mathcal{H} . Let $M(d\hat{\theta})$ be a positive operator-valued measure defined on the σ -algebra $\mathcal{A}(\Theta)$ of Borel subsets of Θ . Then $M(d\hat{\theta})$ is covariant with respect to the representation $g \mapsto V_g$ if*

$$V_g^\dagger M(B) V_g = M(B_{g^{-1}}), \quad g \in G, \quad (27)$$

for any $B \in \mathcal{A}(\Theta)$, where $B_g = \{\theta' : \theta' = g\theta, \theta \in B\}$.

If $M(d\hat{\theta})$ is covariant, then

$$\text{tr } M(B) \rho_\theta = \text{tr } \rho_0 V_g^\dagger M(B) V_g = \text{tr } \rho_0 M(B_{g^{-1}}).$$

Thus,

$$\Pr[\hat{\theta} \in B | g\theta_0] = \Pr[\hat{\theta} \in B_{g^{-1}} | \theta_0],$$

i.e. a covariant measurement preserves the probability distribution under the transformation of the state. We refer the reader to reference [15] for a more detailed discussion on covariant measurements.

Before we start building towards the proof of Result 2.3, let us look at some of the properties of the parametric group G to understand the situation better. First, the group G is chosen to act transitively on Θ . This ensures that the map $g \mapsto g\theta_0$ maps G onto the whole Θ . Second, G is assumed to be unimodular which implies that there exists an invariant measure μ on G . Third, G is assumed to be compact which ensures that the measure $\mu < \infty$. The measure μ is normalized as $\mu(G) = 1$. Now, we are interested in an invariant measure ν on the Σ -algebra $\mathcal{A}(\Theta)$ on Θ such that

$$\nu(B) = \nu(B_g), \quad B \in \mathcal{A}(\Theta),$$

where $B_g = \{\theta' : \theta' = g\theta, \theta \in B\}$. If G_0 , the stationary subgroup of G , is unimodular then such a measure ν exists and if G_0 is compact then ν is finite and can be constructed from μ by demanding that the following relation holds for all integrable functions f on Θ :

$$\int_G f(g\theta_0) d\mu(g) = \int_\Theta f(\theta) d\nu(\theta). \quad (28)$$

We now state Proposition 2.1 from reference [15] as the following lemma that gives a relation between the two measures.

Lemma 4.1. *Let $M(d\theta)$ be a measurement covariant with respect to a projective unitary representation $g \mapsto V_g$ of the parametric group G acting on Θ . For any density operator $\rho \in \mathcal{H}$, and for any Borel set $B \in \mathcal{A}(\Theta)$*

$$\int_G \text{tr}[V_g \rho V_g^\dagger M(B)] d\mu(g) = \nu(B) \quad (29)$$

Let us pause here to look at an example.

Example 3. Let us assume that we are interested in estimating all those states in \mathbb{C}^2 that lie on the Bloch sphere. Thus, the parameter space Θ is \mathbb{S}^2 . This is a covariant estimation problem with the parametric group of transformations on \mathbb{S}^2 being $SO(3)$. Its projective unitary representation is the quotient subgroup $SU(2)/U(1)$. Let us assume that the initial state ρ_{θ_0} is $|0\rangle\langle 0|$. Then, the elements of $SU(2)/U(1)$ generate all the states on the Bloch sphere that are parametrized by a set of two parameters: the latitude θ^3 and azimuth ϕ , with the states being identified as $|\theta, \phi\rangle$. Note that an element of $SU(2)/U(1)$ can be written in terms of these as

$$U_{\theta, \phi} = \begin{bmatrix} \cos \frac{\theta}{2} & -\sin \frac{\theta}{2} e^{-i\phi} \\ \sin \frac{\theta}{2} e^{i\phi} & \cos \frac{\theta}{2} \end{bmatrix}.$$

It can be shown [15, Theorem 2.1] (stated as Lemma E.3 for reference) that starting from a positive operator P_0 that commutes with all the elements of the stationary subgroup of $SU(2)/U(1)$ such that it satisfies Equation (38), setting $M(\theta, \phi) = U_{\theta, \phi} P_0 U_{\theta, \phi}^\dagger$ implies that the measurement defined with $M(\theta, \phi)$ as the operator-valued density with respect to the uniform measure on Θ is covariant. In this case, $P_0 = 2|0\rangle\langle 0|$ and the operator-valued density is

$$M(\theta, \phi) = 2 \begin{bmatrix} \cos \frac{\theta}{2} \\ \sin \frac{\theta}{2} e^{i\phi} \end{bmatrix} \cdot \begin{bmatrix} \cos \frac{\theta}{2} & \sin \frac{\theta}{2} e^{-i\phi} \end{bmatrix} = 2 \begin{bmatrix} \cos^2 \frac{\theta}{2} & \cos \frac{\theta}{2} \sin \frac{\theta}{2} e^{-i\phi} \\ \cos \frac{\theta}{2} \sin \frac{\theta}{2} e^{i\phi} & \sin^2 \frac{\theta}{2} \end{bmatrix}.$$

It is straightforward to verify that $\frac{1}{4\pi} \int_{\mathbb{S}^2} M(\theta, \phi) \sin \theta d\theta d\phi = \mathbb{I}$.

Having defined and illustrated the problem of covariant state estimation, we now recall Holevo's theorem [15, Theorem 3.1], which states that for every loss function that is invariant under the group transformation g , the minimax risk as well as the average risk attain their minima at a covariant measurement. But, the analysis in reference [15] is done for loss functions expressed as functions of the true parameter and the estimator of the parameter. It can be recast in terms of the general framework involving estimators that are functions of the parameter, $\hat{\rho} : \Theta \mapsto \mathcal{D}(\mathcal{H})$ by simply choosing the domain of the loss function to be the set of density matrices $\mathcal{S}(\mathcal{H})$ as opposed to the parameter space Θ . We thus state it as the following lemma which is the main ingredient of the proof of Result 2.3.

Lemma 4.2 (Theorem 3.1, [15]). *In the quantum covariant statistical estimation problem, given an estimator $\hat{\rho}$ of the state, the minima of the average risk $r_P(\nu, \hat{\rho})$ with respect to the uniform Haar measure ν on Θ and the worst case risk $\sup_{\theta} R_P(\rho_{\theta}, \hat{\rho})$ for all Θ -measurements are achieved on a covariant measurement. Moreover, for any covariant measurement P_c , we have*

$$r_{P_c}(\nu, \hat{\rho}) = \sup_{\theta} R_{P_c}(\rho_{\theta}, \hat{\rho}) = R_{P_c}(\rho_{\theta}, \hat{\rho}), \quad \theta \in \Theta \quad (30)$$

Note: A covariant measurement is *not* a unique minimum for either the average or the worst-case risk.

The above theorem implies that for any measurement P , there exists a covariant measurement that minimizes the average risk as well as the worst case risk for a fixed estimator $\hat{\rho}$. But, we know that for a fixed measurement the average risk is minimized by the Bayes estimator $\hat{\rho}_B$. Thus, the Bayes estimator minimizes the average risk for a covariant measurement. However, the invariance of the loss function implies that the Bayes estimator must be covariant under the group transformations as shown below.

Lemma 4.3. *The Bayes estimator is covariant under the group transformations V_g , i.e.*

$$\hat{\rho}_B(B_g) = V_g \hat{\rho}_B(B) V_g^\dagger, \quad B \in \mathcal{A}(\Theta).$$

Proof. Recalling the invariance property of the loss function: $L(\rho_{\theta}, \hat{\rho}(x)) = L(\rho_{g\theta}, \hat{\rho}(gx)) = L(V_g \rho_{\theta} V_g^\dagger, V_g \hat{\rho}(x) V_g^\dagger)$, let us verify for the case of Bayes estimator. Recalling Equation (14), we have

$$\hat{\rho}_B(B_g) = \frac{\int_{\Theta} d\nu(\theta) \operatorname{tr} \rho_{\theta} P(B_g) \rho_{\theta}}{\int_{\Theta} d\nu(\theta) \operatorname{tr} \rho_{\theta} P(B_g)}.$$

³We apologize for the redundancy, but it is best to stick to conventional symbols.

As we are interested in a covariant measurement $P(B_g) = V_g P(B) V_g^\dagger$, therefore

$$\begin{aligned}\hat{\rho}_B(B_g) &= \frac{\int_{\Theta} d\nu(\theta) \rho_\theta \operatorname{tr} \rho_\theta V_g P(B) V_g^\dagger}{\int_{\Theta} d\nu(\theta) \operatorname{tr} \rho_\theta V_g P(B) V_g^\dagger} \\ &= \frac{\int_{\Theta} d\nu(\theta) \rho_\theta \operatorname{tr} V_g^\dagger \rho_\theta V_g P(B)}{\int_{\Theta} d\nu(\theta) \operatorname{tr} V_g^\dagger \rho_\theta V_g P(B)} \\ &= \frac{\int_{\Theta} d\nu(\theta) \rho_\theta \operatorname{tr} \rho_{g^{-1}\theta} P(B)}{\int_{\Theta} d\nu(\theta) \operatorname{tr} \rho_{g^{-1}\theta} P(B)}.\end{aligned}$$

By the invariance of ν and the fact that $\rho_\theta = V_g \rho_{g^{-1}\theta} V_g^\dagger$, we finally obtain

$$\begin{aligned}\hat{\rho}_B(B_g) &= \frac{\int_{\Theta} d\nu(g^{-1}\theta) V_g \rho_{g^{-1}\theta} V_g^\dagger \operatorname{tr} \rho_{g^{-1}\theta} P(B)}{\int_{\Theta} d\nu(g^{-1}\theta) \operatorname{tr} \rho_{g^{-1}\theta} P(B)} \\ &= V_g \left\{ \frac{\int_{\Theta} d\nu(g^{-1}\theta) \rho_{g^{-1}\theta} \operatorname{tr} \rho_{g^{-1}\theta} P(B)}{\int_{\Theta} d\nu(g^{-1}\theta) \operatorname{tr} \rho_{g^{-1}\theta} P(B)} \right\} V_g^\dagger = V_g \hat{\rho}_B(B) V_g^\dagger.\end{aligned}$$

□

Thus, we have established two things about a covariant measurement. First, that the risk of a covariant measurement is independent of the state ρ_θ and second, that it minimizes the average as well as the worst-case risk among all measurements. But, the problem of finding a minimax POVM is closely tied to obtaining a least favourable prior which in turn is tied to the underlying measurement, as we discussed in Section 2. The following lemma gives a least favourable prior and the corresponding measurement in the context of covariant state estimation.

Lemma 2.1. *The uniform measure on the parameter space Θ is a least favourable prior for covariant measurements.*

Proof. As the Bayes estimator $\hat{\rho}_B^\pi$ minimizes the average risk with respect to the prior π ,

$$\sup_{\theta} R_{P_c}(\rho_\theta, \rho_B^\nu) = \sup_{\pi} \int_{\Theta} d\pi(\theta) R_{P_c}(\rho_\theta, \rho_B^\nu) \geq \sup_{\pi} \int_{\Theta} d\pi(\theta) R_{P_c}(\rho_\theta, \rho_B^\pi).$$

However, by Lemma 4.2, we know that for a covariant measurement the risk is independent of the state ρ_θ , i.e.

$$\sup_{\theta} R_{P_c}(\rho_\theta, \rho_B^\nu) = \int_{\Theta} d\nu(\theta) R_{P_c}(\rho_\theta, \rho_B^\nu).$$

This implies that

$$\int_{\Theta} d\nu(\theta) R_{P_c}(\rho_\theta, \rho_B^\nu) \geq \sup_{\pi} \int_{\Theta} d\pi(\theta) R_{P_c}(\rho_\theta, \rho_B^\pi).$$

The other direction of the above inequality holds trivially, i.e.

$$\int_{\Theta} d\nu(\theta) R_{P_c}(\rho_\theta, \rho_B^\nu) \leq \sup_{\pi} \int_{\Theta} d\pi(\theta) R_{P_c}(\rho_\theta, \rho_B^\pi).$$

Therefore,

$$\int_{\Theta} d\nu(\theta) R_{P_c}(\rho_\theta, \rho_B^\nu) = \sup_{\pi} \int_{\Theta} d\pi(\theta) R_{P_c}(\rho_\theta, \rho_B^\pi),$$

and so ν is a least favourable prior for a covariant measurement P_c . □

The only remaining ingredient needed to prove Result 2.3 is the following lemma.

Lemma 4.4. *The Bayes estimator for a covariant measurement P_c is*

$$\hat{\rho}_B(B) = \frac{1}{\nu(B)} \operatorname{tr} \left[\mathbb{I}^R \otimes P_c^{R'}(B) \int_G d\mu(g) (V_g \rho_0 V_g^\dagger)^{\otimes 2} \right], \quad B \in \mathcal{A}(\Theta). \quad (31)$$

Proof. Recalling Equation (14), the Bayes estimator for a covariant measurement P_c is

$$\hat{\rho}_B(B) = \frac{\int_{\Theta} d\nu(\theta) \operatorname{tr}[\rho_\theta P_c(B)] \rho_\theta}{\int_{\Theta} d\nu(\theta) \operatorname{tr} \rho_\theta P_c(B)}, \quad B \in \mathcal{A}(\Theta).$$

Using Lemma 4.1, the denominator in the above expression is

$$\int_{\Theta} d\nu(\theta) \operatorname{tr} \rho_\theta P_c(B) = \nu(B),$$

while the numerator is

$$\begin{aligned} \int_{\Theta} d\nu(\theta) \operatorname{tr}[\rho_\theta P_c(B)] \rho_\theta &= \int_G d\mu(g) \operatorname{tr}[\rho_{g\theta_0} P_c(B)] \rho_{g\theta_0} \\ &= \int_G d\mu(g) \operatorname{tr}[V_g \rho_0 V_g^\dagger P_c(B)] V_g \rho_0 V_g^\dagger \\ &= \operatorname{tr}_{R'} \left[\mathbb{I}^R \otimes P_c^{R'}(B) \int_G d\mu(g) (V_g \rho_0 V_g^\dagger)^R \otimes (V_g \rho_0 V_g^\dagger)^{R'} \right]. \end{aligned}$$

□

Now that we have a class of measurements and the corresponding least favourable prior, in order to show that it is minimax (first part of Result 2.3), we have to show that this class of measurements also minimizes the average risk with respect to such a least favourable prior. Recall Result 2.3. Formally, the first part of Result 2.3 is stated as the following theorem and the second part as a corollary to the theorem.

Theorem 4.3. *There exists a covariant measurement that is minimax for covariant state estimation.*

Proof. Recalling Definition 4.1 of a minimax POVM and the fact that the Bayes estimator minimizes the average risk, we have

$$\begin{aligned} \inf_P \inf_{\hat{\rho}} \sup_{\theta} R_P(\rho_\theta, \hat{\rho}) &= \inf_P \inf_{\hat{\rho}} \sup_{\pi} \int_{\Theta} d\pi(\theta) R_P(\rho_\theta, \hat{\rho}) \\ &\geq \inf_P \sup_{\pi} \int_{\Theta} d\pi(\theta) R_P(\rho_\theta, \rho_B^\pi). \end{aligned}$$

But, as

$$\sup_{\pi} \int_{\Theta} d\pi(\theta) R_P(\rho_\theta, \rho_B^\pi) \geq \int_{\Theta} d\mu(\theta) R_P(\rho_\theta, \rho_B^\mu), \quad \forall \mu \in \mathcal{P}(\Theta),$$

it implies that the same holds for the uniform Haar measure ν as well. Thus,

$$\inf_P \inf_{\hat{\rho}} \sup_{\theta} R_P(\rho_\theta, \hat{\rho}) \geq \inf_P \int_{\Theta} d\nu(\theta) R_P(\rho_\theta, \rho_B^\nu).$$

Now, we know from Lemma 4.2 that

$$\inf_P \int_{\Theta} d\nu(\theta) R_P(\rho_\theta, \rho_B^\nu) = \int_{\Theta} d\nu(\theta) R_{P_c}(\rho_\theta, \rho_B^\nu),$$

where P_c is a covariant measurement that minimizes the average risk. Also, by Lemma 2.1, ν is a least favourable prior which means that $\hat{\rho}_B^\nu$ is a minimax estimator. Therefore, we have

$$\int_{\Theta} d\nu(\theta) R_{P_c}(\rho_\theta, \rho_B^\nu) = \sup_{\theta} R_c(\rho_\theta, \rho_B^\nu) = \inf_{\hat{\rho}} \sup_{\theta} R_{P_c}(\rho_\theta, \hat{\rho}).$$

Thus, we obtain

$$\inf_P \inf_{\hat{\rho}} \sup_{\theta} R_P(\rho_{\theta}, \hat{\rho}) \geq \inf_{\hat{\rho}} \sup_{\theta} R_{P_c}(\rho_{\theta}, \hat{\rho}).$$

The other direction of the inequality above holds trivially, i.e.

$$\inf_P \inf_{\hat{\rho}} \sup_{\theta} R_P(\rho_{\theta}, \hat{\rho}) \leq \inf_{\hat{\rho}} \sup_{\theta} R_{P_c}(\rho_{\theta}, \hat{\rho})$$

Hence, we have proved that P_c is a minimax POVM, i.e.

$$\inf_P \inf_{\hat{\rho}} \sup_{\theta} R_P(\rho_{\theta}, \hat{\rho}) = \inf_{\hat{\rho}} \sup_{\theta} R_{P_c}(\rho_{\theta}, \hat{\rho}).$$

□

As the risk for a covariant measurement is independent of the state ρ_{θ} (by Lemma 4.2) and depends only on the estimator (the Bayes estimator in this case, which is a function of $\int_G d\mu(g)(V_g \rho_0 V_g^\dagger)^{\otimes 2}$), we have the following corollary.

Corollary 4.3.1. *Given a minimax covariant measurement \mathcal{P}_c , if there exists a measurement \mathcal{P}'_c which is covariant under a subgroup H of G such that $\{V_h | h \in H\}$, where V_h is the projective unitary representation of the subgroup H , forms a unitary 2-design, i.e.*

$$\mathcal{P}_c(B) = V_h^\dagger \mathcal{P}_c(B_{g^{-1}}) V_h, \quad B \in \mathcal{A}(\Theta); \quad h \in H,$$

where $B_{g^{-1}} = \{g^{-1}\theta | \theta \in B\}$, and \mathcal{P}_c and \mathcal{P}'_c have the same seed, then \mathcal{P}'_c is also minimax.

In order to understand the above corollary better let us look at what it means for Example 3.

Example 3 (continued). *Now, since any state $|\theta, \phi\rangle$ in \mathbb{S}^2 can be generated by elements of $SU(2)/U(1)$, the following equivalence holds:*

$$\frac{1}{4\pi} \int_{\mathbb{S}^2} |\theta, \phi\rangle\langle\theta, \phi| \sin \theta d\theta d\phi = \int_{SU(2)/U(1)} U_{\theta, \phi} |0\rangle\langle 0| U_{\theta, \phi}^\dagger dU_{\theta, \phi},$$

where $dU_{\theta, \phi}$ is the Haar measure on $SU(2)/U(1)$. In fact, the above implies that

$$\frac{1}{4\pi} \int_{\mathbb{S}^2} (|\theta, \phi\rangle\langle\theta, \phi|)^{\otimes 2} \sin \theta d\theta d\phi = \int_{SU(2)/U(1)} (U_{\theta, \phi} |0\rangle\langle 0| U_{\theta, \phi}^\dagger)^{\otimes 2} dU_{\theta, \phi}.$$

Now, the above equivalence along with [21, Theorem 3.3.1] implies that one can construct a unitary 2-design from a quantum 2-design. Thus, the set of unitary matrices that generate the set of eigenstates of the Pauli matrices $\sigma_x, \sigma_y, \sigma_z$ is a unitary 2-design given as below.

$$\begin{aligned} U_{0,0} &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, U_{\pi,0} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \\ U_{\pi/2,0} &= \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}, U_{\pi/2,\pi} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}, \\ U_{\pi/2,\pi/2} &= \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix}, U_{\pi/2,3\pi/2} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix}. \end{aligned}$$

The corresponding measurement that is covariant under the above unitary 2-design is then obtained via $M_{\theta, \phi} = U_{\theta, \phi} P_0 U_{\theta, \phi}^\dagger$. It is straightforward to see that the corresponding $M_{\theta, \phi}$ are the same as the Pauli measurements apart from a normalization constant. This measurement is thus minimax.

Note: In the above example we obtained a minimax POVM—a covariant measurement for the parameter space \mathbb{S}^2 which describes only pure qubit states. There is no a-priori reason to believe that the same measurement would also be minimax for estimating an arbitrary state of a qubit. However, curiously, it happens to be true for a qubit as we will see in the following section.

5 Example: Minimax measurement for a qubit

We look at the single qubit case as studied in reference [14] wherein the authors obtain such a minimax POVM with relative entropy as the distance-measure. We generalize their results to squared-distance $\|\rho - \sigma\|^2 = \text{tr}(\rho - \sigma)^2$. However, the proof does not follow the generalized treatment in terms of Bregman divergence as done in the previous sections.

To begin with, let us write the most general expression for a POVM on \mathbb{C}^2 . Recalling Lemma 3.1, we can write any POVM, see Definition 3.1, as an operator-valued density, i.e.

$$P(B) = \int_B \mu(dx)M(x),$$

where $B \in \sigma(\mathcal{X})$, $M(x) \geq 0$, $\text{Tr}[M(x)] = 2$ and $\mu(\mathcal{X}) = 1$. Since positive operators can be expanded in the Pauli basis $\{\mathbb{I}, \sigma_x, \sigma_y, \sigma_z\}$ with real coefficients, $M(\vec{x}) = \alpha_0 \mathbb{I} + \vec{x} \cdot \vec{\sigma}$, but the trace 1 condition on $M(\vec{x})$ implies $\alpha_0 = 1/2$. Without loss of generality,

$$M(\vec{x}) = (\mathbb{I} + \vec{x} \cdot \vec{\sigma}).$$

Thus, the most general form of a POVM element on \mathbb{C}^2 is

$$P(B) = \int_B (\mathbb{I} + \vec{x} \cdot \vec{\sigma}) d\mu(\vec{x}). \quad (32)$$

Now that we have obtained the general expression for a POVM on \mathbb{C}^2 , we next evaluate the Bayes estimator which in turn is needed to evaluate the risk $R_P(\rho_\theta, \rho_B^\pi)$. Recalling that the Bayes estimator is given as

$$\rho_B^\pi(x) = \int_\Theta \frac{dp(x|\theta)}{dp_\pi(x)} \rho_\theta d\pi(\theta).$$

Recalling the differential form of Born's rule (8), and using the Bloch sphere notation of ρ_θ , $\rho_\theta = \frac{1}{2}(\mathbb{I} + \vec{\theta} \cdot \vec{\sigma})$, it is a straightforward calculation to obtain

$$\text{tr} M(\vec{x}) \rho_\theta = d\mu(\vec{x})(1 + \vec{x} \cdot \vec{\theta}). \quad (33)$$

However, to be able to further simplify the Bayes estimator, we need to impose some restrictions on the prior π defined on the parameter space Θ (which in this case is \mathbb{R}^3). In particular, we choose a uniform prior π^* supported only on pure states, i.e. $\pi(\theta)$ is zero for all vectors $\vec{\theta}$ with $\|\theta\| < 1$ but is uniformly distributed on the set of unit vectors with $\|\theta\| = 1$. It can be verified that such a prior has the following two properties :

1. $\mathbb{E}_{\pi^*}[\theta_i] = 0, \quad \forall i \in \{x, y, z\}$.
2. $\mathbb{E}_{\pi^*}[\theta_i \theta_j] = \frac{1}{3} \delta_{ij} \quad \forall i, j \in \{x, y, z\}$.

By property (1) of the prior π^* , we have

$$p_\pi(B) = \int_\Theta \int_B d\mu(\vec{x})(1 + \vec{x} \cdot \vec{\theta}) \pi^*(\theta) d\theta = \mu(B). \quad (34)$$

Thus, the Bayes estimator reduces to

$$\begin{aligned} \rho_B^{\pi^*}(\vec{x}) &= \int_\Theta \frac{1}{\frac{dp_{\pi^*}(x)}{dp(x|\theta)}} \rho_\theta d\pi^*(\theta) \\ &= \int_\Theta \frac{1}{\frac{d\mu(\vec{x})}{dp(x|\theta)}} \rho_\theta d\pi^*(\theta) \\ &= \int_\Theta \frac{dp(x|\theta)}{d\mu(\vec{x})} \rho_\theta d\pi^*(\theta) \end{aligned}$$

$$\begin{aligned}
&= \int_{\Theta} \rho_{\theta} d\pi^*(\theta) \operatorname{tr} M(x) \rho_{\theta} \\
&= \frac{1}{2} \int_{\Theta} d\pi^*(\theta) (1 + \vec{x} \cdot \vec{\theta}) (\mathbb{I} + \vec{\theta} \cdot \vec{\sigma}) \\
&= \frac{1}{2} \left(\mathbb{I} + \int_{\Theta} d\pi^*(\theta) (\vec{x} \cdot \vec{\theta}) (\vec{\theta} \cdot \vec{\sigma}) \right) \\
&= \frac{1}{2} \left(\mathbb{I} + \frac{1}{3} \vec{x} \cdot \vec{\sigma} \right).
\end{aligned}$$

Now, recalling that the risk is a function of the POVM P , i.e.

$$R_P(\rho_{\theta}, \rho_B^{\pi^*}) = \int_{\mathcal{X}} d\mu(\vec{x}) \operatorname{tr} M(\vec{x}) \rho_{\theta} D_f(\rho_{\theta}, \rho_B^{\pi^*}(\vec{x})), \quad (35)$$

we evaluate the risk for both relative entropy and Hilbert-Schmidt distance below.

Lemma 5.1. *The risk for relative entropy and Hilbert-Schmidt distance are given as*

$$\begin{aligned}
R_P^{rel}(\rho_{\theta}, \rho_B^{\pi^*}) &= -h\left(\frac{1 + \|\theta\|}{2}\right) + \frac{1}{2} \log \frac{9}{2} - \frac{\log 2}{2} \int_{\mathcal{X}} d\mu(\vec{x}) \sum_{j,k} \theta_j \theta_k x_j x_k, \quad \text{and} \\
R_P^{sq}(\rho_{\theta}, \rho_B^{\pi^*}) &= \frac{1}{2} \left\{ \|\vec{\theta}\|^2 + \frac{1}{9} \int_{\mathcal{X}} d\mu(\vec{x}) \|\vec{x}\|^2 + \frac{1}{9} \int_{\mathcal{X}} d\mu(\vec{x}) \|\vec{x}\|^2 \vec{x} \cdot \vec{\theta} - \frac{2}{3} \int_{\mathcal{X}} d\mu(\vec{x}) \sum_{j,k} \theta_j \theta_k x_j x_k \right\},
\end{aligned}$$

respectively.

Proof. (a) For *relative entropy*, see [14, pg. 11].

(b) For *Hilbert-Schmidt distance*, substituting $f = \|A\|_1^2$ in (35), we get

$$D_{sq}(\rho_{\theta}, \rho_B^{\pi^*}(x)) = \operatorname{tr} (\rho_{\theta} - \rho_B^{\pi^*}(x))^2.$$

Thus,

$$\begin{aligned}
D_{sq}(\rho_{\theta}, \rho_B^{\pi^*}(B)) &= \frac{1}{2} \operatorname{tr} \left[\begin{bmatrix} 1 + \theta_z & \theta_x - i\theta_y \\ \theta_x + i\theta_y & 1 - \theta_z \end{bmatrix} - \begin{bmatrix} 1 + z/3 & (x - iy)/3 \\ (x + iy)/3 & 1 - z/3 \end{bmatrix} \right]^2 \\
&= \frac{1}{2} \operatorname{tr} \left[(\theta_z - z/3)\sigma_z + (\theta_x - x/3)\sigma_x + (\theta_y - y/3)\sigma_y \right]^2 \\
&= \frac{1}{2} \left((\theta_z - z/3)^2 + (\theta_x - x/3)^2 + (\theta_y - y/3)^2 \right) \\
&= \frac{1}{2} \left(\|\vec{\theta}\|^2 + \frac{1}{9} \|\vec{x}\|^2 - \frac{2}{3} \vec{x} \cdot \vec{\theta} \right).
\end{aligned}$$

This implies that the risk is

$$R_P^{sq}(\rho_{\theta}, \rho_B^{\pi^*}) = \frac{1}{2} \int_{\mathcal{X}} d\mu(\vec{x}) (1 + \vec{x} \cdot \vec{\theta}) \left(\|\vec{\theta}\|^2 + \frac{1}{9} \|\vec{x}\|^2 - \frac{2}{3} \vec{x} \cdot \vec{\theta} \right).$$

We can write the above using the property of a general POVM on \mathbb{C}^2 , i.e. $\int_{\mathcal{X}} d\mu(\vec{x}) \vec{x} = 0$ as

$$R_P^{sq}(\rho_{\theta}, \rho_B^{\pi^*}) = \frac{1}{2} \left\{ \|\vec{\theta}\|^2 + \frac{1}{9} \int_{\mathcal{X}} d\mu(\vec{x}) \|\vec{x}\|^2 + \frac{1}{9} \int_{\mathcal{X}} d\mu(\vec{x}) \|\vec{x}\|^2 \vec{x} \cdot \vec{\theta} - \frac{2}{3} \int_{\mathcal{X}} d\mu(\vec{x}) \sum_{j,k} \theta_j \theta_k x_j x_k \right\}.$$

Note: Although in reference [14] it is assumed that the POVM P is rank-1, the above expressions hold in general for any POVM P , i.e. the vector \vec{x} in (32) need not be a unit vector. \square

Lemma 5.2. *For any POVM P , the average risk of the Bayes estimator with respect to the prior π^* satisfies the inequalities:*

$$\begin{aligned}
\int_{\Theta} d\pi^*(\theta) R_P^{rel}(\rho_{\theta}, \rho_B^{\pi^*}) &\geq \frac{1}{2} \log \frac{9}{2} - \frac{1}{6} \log 2, \\
\int_{\Theta} d\pi^*(\theta) R_P^{sq}(\rho_{\theta}, \rho_B^{\pi^*}) &\geq \frac{2}{9},
\end{aligned}$$

for relative entropy and Hilbert-Schmidt distance respectively.

Proof. (a) For *relative entropy* :

From Lemma 5.1 and the properties of the prior π^* we get

$$\begin{aligned} \int_{\Theta} d\pi^*(\theta) R_P^{rel}(\rho_\theta, \rho_B^{\pi^*}) &= \int_{\Theta} d\pi^*(\theta) \left\{ -h\left(\frac{1+\|\theta\|}{2}\right) + \frac{1}{2} \log \frac{9}{2} - \frac{\log 2}{2} \int_{\mathcal{X}} d\mu(\vec{x}) \sum_{j,k} \theta_j \theta_k x_j x_k \right\} \\ &= \frac{1}{2} \log \frac{9}{2} - \frac{\log 2}{2} \int_{\mathcal{X}} d\mu(\vec{x}) \sum_{j,k} \frac{\delta_{jk}}{3} x_j x_k \\ &= \frac{1}{2} \log \frac{9}{2} - \frac{1}{6} \log 2 \sum_j \mathbb{E}_\mu[r_j^2]. \end{aligned}$$

However, since $\sum_j r_j^2 \leq 1$, it implies $\sum_j \mathbb{E}_\mu[r_j^2] \leq 1$, and we obtain the required inequality :

$$\int_{\Theta} d\pi^*(\theta) R_P^{rel}(\rho_\theta, \rho_B^{\pi^*}) \geq \frac{1}{2} \log \frac{9}{2} - \frac{1}{6} \log 2.$$

(b) For *Hilbert-Schmidt distance* :

From Lemma 5.1 and the properties of the prior π^* we get

$$\begin{aligned} \int_{\Theta} d\pi^*(\theta) R_P^{f_2}(\rho_\theta, \rho_B^{\pi^*}) &= \int_{\Theta} d\pi^*(\theta) \frac{1}{2} \left\{ \|\vec{\theta}\|^2 + \frac{1}{9} \int_{\mathcal{X}} d\mu(\vec{x}) \|\vec{x}\|^2 + \frac{1}{9} \int_{\mathcal{X}} d\mu(\vec{x}) \|\vec{x}\|^2 \vec{x} \cdot \vec{\theta} - \right. \\ &\quad \left. \frac{2}{3} \int_{\mathcal{X}} d\mu(\vec{x}) \sum_{j,k} \theta_j \theta_k x_j x_k \right\} \\ &= \frac{1}{2} \left\{ 1 + \frac{1}{9} \sum_j \mathbb{E}_\mu[r_j^2] - \frac{2}{9} \sum_j \mathbb{E}_\mu[r_j^2] \right\} \\ &= \frac{1}{2} \left\{ 1 - \frac{1}{9} \sum_j \mathbb{E}_\mu[r_j^2] \right\}. \end{aligned}$$

Again, as $\sum_j r_j^2 \leq 1$, it implies $\sum_j \mathbb{E}_\mu[r_j^2] \leq 1$, and we obtain the required inequality :

$$\int_{\Theta} d\pi^*(\theta) R_P^{f_2}(\rho_\theta, \rho_B^{\pi^*}) \geq \frac{1}{2} \left(1 - \frac{1}{9}\right) = \frac{4}{9}.$$

□

Lemma 5.3. For any POVM P^* that satisfies $\mathbb{E}_\mu[r_i r_j] = \frac{1}{3} \delta_{ij}$, the average risk of the Bayes estimator coincides with the worst-case risk:

$$\int_{\Theta} d\pi^*(\theta) R_{P^*}(\rho_\theta, \rho_B^{\pi^*}) = \sup_{\theta} R_{P^*}(\rho_\theta, \rho_B^{\pi^*}).$$

Proof. (i) (a) For *relative entropy* : (a) See [14, pg. 11].

(b) For *Hilbert-Schmidt distance*:

As $\mathbb{E}_\mu[r_i r_j] = \frac{1}{3} \delta_{ij}$, it implies $\sum_j \mathbb{E}_\mu[r_j^2] = 1$, and thus by Lemma 5.2,

$$\int_{\Theta} d\pi^*(\theta) R_{P^*}^{f_2}(\rho_\theta, \rho_B^{\pi^*}) = \frac{4}{9}.$$

Now, for such a POVM P^* , $\|\vec{x}\|^2 = 1$, and the risk, see Lemma 5.1, becomes

$$R_{P^*}^{sq}(\rho_\theta, \rho_B^{\pi^*}) = \frac{1}{2} \left(\|\vec{\theta}\|^2 + \frac{1}{9} - \frac{2}{9} \|\vec{\theta}\|^2 \right).$$

Clearly,

$$\max_{\|\vec{\theta}\|} R_{P^*}^{sq}(\rho_\theta, \rho_B^{\pi^*}) = \frac{4}{9}, \text{ at } \|\vec{\theta}\| = 1.$$

This implies that

$$\int_{\Theta} d\pi^*(\theta) R_{P^*}(\rho_{\theta}, \rho_B^{\pi^*}) = \sup_{\theta} R_{P^*}(\rho_{\theta}, \rho_B^{\pi^*}), \text{ for both relative entropy and squared-distance.}$$

□

Lemma 5.4. *The uniform Haar measure on \mathbb{S}^2 is a least favourable prior for spherical 2-designs in \mathbb{C}^2 .*

Proof. Firstly, note that a POVM P^* with $\mathbb{E}_{\mu}[r_i r_j] = \frac{1}{3} \delta_{ij}$ is a spherical 2-design. (See Definition F.2 of spherical t-designs. Examples of spherical 2-designs include the SIC-POVM [22] on \mathbb{C}^2 as well as the POVM defined through the Pauli measurements.) It can be seen from Lemma 5.1 that the risk is a polynomial function of degree 2 in the variables x, y, z . It is straightforward to see that the average of the typical term $x_i x_j$ with respect to the Haar measure on \mathbb{S}^2 is $\frac{\delta_{ij}}{3}$. Thus, any POVM with $\mathbb{E}_{\mu}[x_i x_j] = \frac{1}{3} \delta_{ij}$ is a spherical 2-design. Let us now proceed with the proof of the lemma. As the Bayes estimator $\hat{\rho}_B^{\pi}$ minimizes the average risk with respect to the prior π ,

$$\sup_{\theta} R_{P^*}(\rho_{\theta}, \rho_B^{\pi^*}) = \sup_{\pi} \int_{\Theta} d\pi(\theta) R_{P^*}(\rho_{\theta}, \rho_B^{\pi^*}) \geq \sup_{\pi} \int_{\Theta} d\pi(\theta) R_{P^*}(\rho_{\theta}, \rho_B^{\pi}).$$

However, we just proved in Lemma 5.3 that

$$\begin{aligned} \sup_{\theta} R_{P^*}(\rho_{\theta}, \rho_B^{\pi^*}) &= \int_{\Theta} d\pi^*(\theta) R_{P^*}(\rho_{\theta}, \rho_B^{\pi^*}), \\ \implies \int_{\Theta} d\pi^*(\theta) R_{P^*}(\rho_{\theta}, \rho_B^{\pi^*}) &\geq \sup_{\pi} \int_{\Theta} d\pi(\theta) R_{P^*}(\rho_{\theta}, \rho_B^{\pi}). \end{aligned}$$

The other direction of the above inequality holds trivially, i.e.

$$\begin{aligned} \int_{\Theta} d\pi^*(\theta) R_{P^*}(\rho_{\theta}, \rho_B^{\pi^*}) &\leq \sup_{\pi} \int_{\Theta} d\pi(\theta) R_{P^*}(\rho_{\theta}, \rho_B^{\pi}), \\ \implies \int_{\Theta} d\pi^*(\theta) R_{P^*}(\rho_{\theta}, \rho_B^{\pi^*}) &= \sup_{\pi} \int_{\Theta} d\pi(\theta) R_{P^*}(\rho_{\theta}, \rho_B^{\pi}). \end{aligned}$$

Thus, π^* is a least favourable prior for a spherical 2-design in \mathbb{C}^2 . □

Theorem 5.1. *Any spherical 2-design for \mathbb{C}^2 is a minimax POVM.*

Proof. Recalling Definition F.2 of a minimax POVM and the fact that the Bayes estimator minimizes the average risk, we have:

$$\begin{aligned} \inf_P \inf_{\hat{\rho}} \sup_{\theta} R_P(\rho_{\theta}, \hat{\rho}) &= \inf_P \inf_{\hat{\rho}} \sup_{\pi} \int_{\Theta} d\pi(\theta) R_P(\rho_{\theta}, \hat{\rho}) \\ &\geq \inf_P \sup_{\pi} \int_{\Theta} d\pi(\theta) R_P(\rho_{\theta}, \rho_B^{\pi}) \\ &\geq \inf_P \int_{\Theta} d\pi^*(\theta) R_P(\rho_{\theta}, \rho_B^{\pi^*}). \end{aligned}$$

Now, Lemma 5.2 and Lemma 5.3 together imply that the average risk with respect to π^* is minimized by the POVM P^* , i.e.

$$\inf_P \int_{\Theta} d\pi^*(\theta) R_P(\rho_{\theta}, \rho_B^{\pi^*}) = \int_{\Theta} d\pi^*(\theta) R_{P^*}(\rho_{\theta}, \rho_B^{\pi^*}).$$

Also, by Lemma 5.4, π^* is a least favourable prior which means that $\rho_B^{\pi^*}$ is a minimax estimator. Therefore, we have

$$\int_{\Theta} d\pi^*(\theta) R_{P^*}(\rho_{\theta}, \rho_B^{\pi^*}) = \sup_{\theta} R_{P^*}(\rho_{\theta}, \rho_B^{\pi^*}) = \inf_{\hat{\rho}} \sup_{\theta} R_{P^*}(\rho_{\theta}, \hat{\rho}).$$

Thus, we obtain

$$\inf_P \inf_{\hat{\rho}} \sup_{\theta} R_P(\rho_{\theta}, \hat{\rho}) \geq \inf_{\hat{\rho}} \sup_{\theta} R_{P^*}(\rho_{\theta}, \hat{\rho}).$$

The other direction of the inequality above holds trivially, i.e.

$$\inf_P \inf_{\hat{\rho}} \sup_{\theta} R_P(\rho_{\theta}, \hat{\rho}) \leq \inf_{\hat{\rho}} \sup_{\theta} R_{P^*}(\rho_{\theta}, \hat{\rho})$$

Hence, we have proved that P^* , a spherical 2-design is a minimax POVM, i.e.

$$\inf_P \inf_{\hat{\rho}} \sup_{\theta} R_P(\rho_{\theta}, \hat{\rho}) = \inf_{\hat{\rho}} \sup_{\theta} R_{P^*}(\rho_{\theta}, \hat{\rho}).$$

□

6 Discussion & Future work

To summarize, we extended the work done in reference [14] on minimax analysis to Bregman divergences. Moreover, by re-formulating Holevo's theorem [15, pg. 171] for the *covariant state estimation problem* in terms of estimators, we found that a *covariant* POVM is, in fact, minimax with Bregman divergence as the distance-measure. In addition to that, we found that it suffices that a measurement be covariant only under a subgroup H of G such that the unitary representation of H forms a unitary 2-design for it to be minimax. Finally, in order to understand the problem of finding a minimax POVM for an arbitrary quantum state, we studied the problem for a qubit observing that a spherical 2-design defines a minimax POVM for a qubit.

In the covariant state estimation problem, we assume that the underlying group G is compact. It is natural to ask if these results can be extended to infinite-dimensional systems, or equivalently, non-compact groups. The natural system that comes to mind when one thinks of an infinite-dimensional system is the set of coherent states of a Harmonic oscillator. The underlying group is the translation group \mathcal{T} acting on the complex plane. The projective unitary representation of which is the Weyl-Heisenberg translation operator $\{D(\alpha) \mid \alpha \in \mathbb{C}\}$. Now, the translation group is non-compact. This means that one cannot define a normalisable measure on the group. Our derivation of the main result on covariant state estimation, Theorem 4.3, to obtain a minimax measurement uses a Bayesian approach. Recall that a minimax measurement is the one that minimizes the worst-case risk of a minimax estimator. Thus, we are interested in the following expression :

$$\inf_P \inf_{\hat{\rho}} \sup_{\theta} R_P(\rho_{\theta}, \hat{\rho}).$$

The very first step of the proof involves re-writing the supremum over θ as a supremum over the probability distributions on Θ , i.e.

$$\inf_P \inf_{\hat{\rho}} \sup_{\theta} R_P(\rho_{\theta}, \hat{\rho}) = \inf_P \inf_{\hat{\rho}} \sup_{\pi} \int_{\Theta} d\pi(\theta) R_P(\rho_{\theta}, \hat{\rho}).$$

Obviously, we cannot do so in the case of the translation group \mathcal{T} that acts on the complex plane. So, our approach will not apply to the most general problem of estimating coherent states generated by the Weyl-Heisenberg translation operator $\{D(\alpha) \mid \alpha \in \mathbb{C}\}$. Indeed, a more general theorem for the case of locally compact groups [23, 24] shows that covariant measurements minimize the worst-case risk (average risk cannot be defined for non-compact groups). However, the formalism considered in [23] does not include *estimators*. It would be interesting to extend the same to our setting and, moreover, to come up with an appropriate definition of a Bayesian estimator for such cases.

The next obvious extension of this work is to find minimax POVMs for an arbitrary quantum state. It would be interesting to see if some kind of a t -design comes out as a solution. However, this requires a more generalized approach than mere brute-force calculations which become tedious in higher dimensions. Moreover, one could also generalize this result to arbitrary distance-measures such as Fidelity and Renyi divergences. The authors of reference [25] have derived the Bayes

estimator for distance-measures based on Bhattacharya distance. Partial results [26] are known for fidelity as the distance-measure, but the Bayes estimator remains unknown for a general state with fidelity as the distance-measure. But, these generalizations are not so straight forward either and require a different technique.

Acknowledgements

We acknowledge that reference [24] was brought to our attention by an anonymous reviewer. CF and MT acknowledge Australian Research Council Discovery Early Career Researcher Awards, projects No. DE170100421 and DE160100821, respectively.

References

- [1] U. Fano, [Rev. Mod. Phys.](#) **29**, 74 (1957).
- [2] W. Pauli, *Encyclopedia of Physics* V , 17 (1958).
- [3] J. Watrous, *The Theory of Quantum Information* (Cambridge University Press, 2018).
- [4] Z. Hradil, [Phys. Rev. A](#) **55**, R1561 (1997).
- [5] R. Blume-Kohout, [New Journal of Physics](#) **12**, 043034 (2010), [arXiv:0611080 \[quant-ph\]](#) .
- [6] C. Ferrie and R. Blume-Kohout, *ArXiv e-prints* (2018), [arXiv:1808.01072 \[quant-ph\]](#) .
- [7] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. (Springer-Verlag, New York, NY, USA, 1998).
- [8] F. Tanaka and F. Komaki, [Phys. Rev. A](#) **71**, 052323 (2005).
- [9] A. Banerjee, X. Guo, and H. Wang, [IEEE Transactions on Information Theory](#) **51**, 2664 (2005).
- [10] B. S. Clarke and A. R. Barron, [Journal of Statistical Planning and Inference](#) **41**, 37 (1994).
- [11] N. Merhav and M. Feder, [IEEE Transactions on Information Theory](#) **44**, 2124 (1998).
- [12] Q. Xie and A. R. Barron, [IEEE Transactions on Information Theory](#) **46**, 431 (2000).
- [13] K. F., [Journal of Statistical Planning and Inference](#) **141**, 3705 (2011).
- [14] T. Koyama, T. Matsuda, and F. Komaki, [Entropy](#) **19**, 618 (2017).
- [15] A. S. Holevo, *Probabilistic and Statistical Aspects of Quantum Theory* (Elsevier Science Ltd, Amsterdam ; New York : New York, 1982).
- [16] A. Bisio, G. Chiribella, G. M. D’Ariano, S. Facchini, and P. Perinotti, [IEEE Journal of Selected Topics in Quantum Electronics](#) **15**, 1646 (2009), [arXiv:1702.08751 \[quant-ph\]](#) .
- [17] K. R. Parthasarathy, *Probability measures on Metric Spaces*, Probability and Mathematical Statistics (Academic Press, New York, 1967).
- [18] J. Pitrik and D. Virostek, [Letters in Mathematical Physics](#) **105**, 675 (2015).
- [19] A. Wehrl, [Reviews of Modern Physics](#) **50**, 221 (1978).
- [20] M. Sion, [Pacific J. Math.](#) **8**, 171 (1958).
- [21] M. Derakhshani, *Quantum t-design*, Ph.D. thesis, University of Waterloo (2008).
- [22] J. M. Renes, R. Blume-Kohout, A. J. Scott, and C. M. Caves, [Journal of Mathematical Physics](#) **45**, 2171 (2004), [arXiv:0310075 \[quant-ph\]](#) .
- [23] N. Bogomolov, [Theory of Probability & Its Applications](#) **26**, 787 (1982).
- [24] M. Hayashi, “Group covariance and optimal information processing,” in *A Group Theoretic Approach to Quantum Information* (Springer International Publishing, Cham, 2017) pp. 69–119.
- [25] C. Ferrie and R. Blume-Kohout, *ArXiv e-prints* (2016), [arXiv:1612.07946 \[math.ST\]](#) .
- [26] R. Kueng and C. Ferrie, [New Journal of Physics](#) **17**, 123013 (2015), [arXiv:1503.00677 \[quant-ph\]](#) .
- [27] H. B. Maynard, [Transactions of the American Mathematical Society](#) **173**, 449 (1972).
- [28] D. Prato and C. Tsallis, [Journal of Mathematical Physics](#) **41**, 3278 (2000), [arXiv:9906173 \[cond-mat\]](#) .

A Quantum Bayes estimator for Bregman divergence

Theorem A.1. *If the loss function is Bregman divergence, see Definition 3.3, then*

$$\mathbb{E}_\theta \mathbb{E}_{X|\theta} [D_f(\rho_\theta, \hat{\rho}(X)) - D_f(\rho_\theta, \hat{\rho}_B(X))] \geq 0,$$

for all states $\theta \in \Theta$ and estimators $\hat{\rho}$, where $\hat{\rho}_B$ is the Bayes estimator, see Equation(3).

Proof.

$$\begin{aligned} \mathbb{E}_\theta \mathbb{E}_{X|\theta} [D_f(\rho_\theta, \hat{\rho}(X)) - D_f(\rho_\theta, \hat{\rho}_B(X))] &= \int_\Theta d\pi(\theta) \int_{\mathcal{X}} dp(x|\theta) \operatorname{tr} [f(\rho_\theta) - f(\hat{\rho}(x)) - \\ &\quad f'(\hat{\rho}(x))(\rho_\theta - \hat{\rho}(x)) - f(\rho_\theta) + f(\hat{\rho}_B(x)) + f'(\hat{\rho}_B(x))(\rho_\theta - \hat{\rho}_B(x))] \\ &= \int_\Theta d\pi(\theta) \int_{\mathcal{X}} dp(x|\theta) \operatorname{tr} [f(\hat{\rho}_B(x)) - f(\hat{\rho}(x)) - f'(\hat{\rho}(x))(\rho_\theta - \hat{\rho}(x)) + \\ &\quad f'(\hat{\rho}_B(x))(\rho_\theta - \hat{\rho}_B(x))] \\ &= \int_{\mathcal{X}} dp_\pi(x) \operatorname{tr} [f(\hat{\rho}_B(x)) - f(\hat{\rho}(x)) - f'(\hat{\rho}(x))(\hat{\rho}_B(x) - \hat{\rho}(x))] + \\ &\quad \int_{\mathcal{X}} dp_\pi(x) \operatorname{tr} [f'(\hat{\rho}_B(x))(\hat{\rho}_B(x) - \hat{\rho}(x))] \\ &= \int_{\mathcal{X}} dp_\pi(x) D_f(\hat{\rho}_B(x), \hat{\rho}(x)) \geq 0. \end{aligned}$$

The last inequality follows from the non-negativity of Bregman divergence. \square

Corollary A.1.1. *For all a-priori probability distributions $\pi_\Theta(\theta)$ over the parameter space Ω_Θ ,*

$$r(\pi, \hat{\rho}) \geq r(\pi, \hat{\rho}_B),$$

i.e. the Bayes estimator minimizes the average risk for Bregman divergence.

B Proof of Lemma 3.1

We first present the Radon-Nikodym theorem for operator-valued measures [27], without proof, as stated in [15, pg. 167].

Proposition B.1 (Radon-Nikodym theorem for operator-valued measures). *Let (\mathcal{X}, Σ) be a measurable space and let $\{M(B); B \in \Sigma\}$ be an additive operator-valued function dominated by a measure $\{m(B); B \in \Sigma\}$ in the sense that*

$$|\langle \phi | M(B) | \psi \rangle| \leq m(B) \|\phi\| \|\psi\|, \quad B \in \Sigma,$$

for all $\phi, \psi \in \mathcal{H}$. Then, there exists an operator-valued function $P(\cdot)$ defined uniquely for m -almost all $x \in \mathcal{X}$ (i.e. for all x except for a set of zero m -measure), satisfying $\|P(x)\| \leq 1$ such that

$$\langle \phi | M(B) | \psi \rangle = \int_B \langle \phi | P(x) | \psi \rangle m(dx), \quad B \in \Sigma$$

for all $\phi, \psi \in \mathcal{H}$. If $M(B) \geq 0$ for all $B \in \Sigma$, then $P(x) \geq 0$ for m -almost all $x \in \mathcal{X}$.

Lemma 3.1 (Existence of a POVM density). *Every $P \in \mathcal{P}$ admits a density, i.e. for any POVM P there exists a finite measure $\mu(dx)$ over \mathcal{X} such that $\mu(\mathcal{X}) = 1$ and*

$$P(B) = \int_B d\mu(x) M(x), \quad (6)$$

with $M(x) \geq 0$, and $\operatorname{Tr}[M(x)] = d$ μ -almost everywhere.

Proof. Define $\mu(B) = \text{tr}[P(B)]$, then

$$\langle \phi | P(B) | \phi \rangle \leq \mu(B), \quad \forall |\phi\rangle \in \mathcal{H}.$$

By Cauchy-Schwarz inequality,

$$|\langle \phi | P(B) | \psi \rangle| \leq \mu(B), \quad \forall |\phi\rangle, |\psi\rangle \in \mathcal{H}.$$

Thus, $P(B)$ is dominated by $\mu(B)$. By the Radon-Nikodym theorem, it admits a density $M(x)$ defined uniquely μ - almost everywhere:

$$P(B) = \int_B \mu(dx) M(x).$$

As $P(B) \geq 0, M(x) \geq 0$. Taking trace on both sides of the above equation implies $\text{tr}[M(x)] = d$ μ - almost everywhere. \square

C Lower semi-continuity of Bregman divergence

Before proving the lower semi-continuity of D_f , we state the following lemma which will be used in the proof.

Lemma C.1. *The map $\rho \mapsto \text{tr} f(\rho)$ is continuous on $\Omega = [0, 1]$.*

Proof. Consider a sequence of density operators $(\rho_n)_n$ that weakly converge to a density operator ρ , i.e. $\text{tr} \rho_n a \rightarrow \text{tr} \rho a$ for all $a \in \mathcal{B}(\mathcal{H})$, where \mathcal{H} is a finite dimensional Hilbert space. By choosing $|a\rangle = |i\rangle\langle j|$, where $|i\rangle\langle j|$ is a basis in $\mathcal{B}(\mathcal{H})$, we obtain element-wise convergence of ρ_n to ρ , which in turn implies the convergence of the corresponding eigenvalues.

Now, $\text{tr} f(\rho_n) = \sum_{j=1}^{rk(\rho_n)} f(\lambda_j^n)$ where $\{\lambda_j^n\}_{j=1}^{rk(\rho_n)}$ are the eigenvalues of ρ_n . But, as f is continuous on $[0, 1]$, $f(\lambda_j^n) \rightarrow f(\lambda_j)$, where $\{\lambda_j\}_{j=1}^{rk(\rho)}$ are the eigenvalues of ρ . Thus, the sum $\text{tr} f(\rho_n) = \sum_{j=1}^{rk(\rho_n)} f(\lambda_j^n)$ also converges to $\text{tr} f(\rho) = \sum_{j=1}^{rk(\rho)} f(\lambda_j)$, and this proves the continuity of $\rho \mapsto \text{tr} f(\rho)$ on $[0, 1]$. \square

Theorem C.1. *Bregman divergence $D_f(\cdot, \cdot)$ is lower semi-continuous.*

Proof. We generalize the proof of lower semi-continuity of relative entropy as given in reference [19]. To begin with, we show that there exists a representation of d_f in which the argument of trace does not contain a product of two non-commuting operators. In order to do so, let us define a quantity D_f^λ as

$$D_f^\lambda(\rho, \sigma) = \frac{1}{\lambda} \text{tr} (\lambda f(\rho) + (1 - \lambda) f(\sigma) - f(\lambda \rho + (1 - \lambda) \sigma)),$$

where $\lambda \in (0, 1)$. It is straightforward to verify that the map $\lambda \mapsto \lambda D_f^\lambda$ is concave on the interval $(0, 1)$. Note that $\frac{d}{d\lambda} (\lambda D_f^\lambda) \Big|_{\lambda=0} = D_f$. Thus, as the map $\lambda \mapsto \lambda D_f^\lambda$ is concave on $(0, 1)$, the slope at $\lambda = 0$ will always be greater than the differences $\frac{\lambda D_f^\lambda - 0 \cdot D_f^0}{\lambda}$ for all $\lambda \in (0, 1)$. Assuming that $0 \cdot D_f^0 = 0$, we have

$$\lambda D_f^\lambda \leq \lambda D_f \Leftrightarrow D_f^\lambda \leq D_f, \quad \forall \lambda \in (0, 1).$$

As $\lim_{\lambda \rightarrow 0} D_f^\lambda = D_f$, we have

$$\sup_{\lambda} D_f^\lambda(\rho, \sigma) = D_f(\rho, \sigma).$$

Let $\rho_n \Rightarrow \rho$ and $\sigma_n \Rightarrow \sigma$ be given. Then, by Lemma C.1, the map $(\rho, \sigma) \mapsto D_f^\lambda(\rho, \sigma)$ is continuous. Therefore,

$$D_f(\rho, \sigma) = \sup_{\lambda} D_f^\lambda(\rho, \sigma)$$

$$\begin{aligned}
&= \sup_{\lambda} \lim_{n \rightarrow \infty} D_f^\lambda(\rho_n, \sigma_n) \\
&\leq \liminf_{n \rightarrow \infty} \sup_{\lambda} D_f^\lambda(\rho_n, \sigma_n) \\
&= \liminf_{n \rightarrow \infty} D_f(\rho_n, \sigma_n).
\end{aligned}$$

But, $D_f(\rho, \sigma) \leq \liminf_{n \rightarrow \infty} D_f(\rho_n, \sigma_n)$ defines a lower semi-continuous function. \square

D Why the Bayes estimator is discontinuous.

For the Bayes estimator to be continuous in the prior, it should hold that for any convergent sequence $(\pi_n)_n$,

$$\lim_{n \rightarrow \infty} \hat{\rho}_B^{\pi_n}(x) = \hat{\rho}_B^{\lim_{n \rightarrow \infty} \pi_n}(x), \quad \forall x \in \mathcal{X}.$$

Below is an example that shows that the above is *not* true in general.

Example 4. Let us assume that we are doing a σ_z measurement, thus, $\mathcal{X} = \{0, 1\}$. Now, consider a sequence of priors $(\pi_n)_n$ that converges to $\mu(\theta) = \delta(\theta - \theta_0)$ where θ_0 corresponds to $|0\rangle\langle 0|$ and let each element of the sequence be defined as below:

$$\pi_n(\theta) = \left(1 - \frac{1}{n}\right)\delta(\theta - \theta_0) + \frac{1}{n}\delta(\theta - \theta_1),$$

with θ_1 corresponding to $|1\rangle\langle 1|$. Then, the Bayes estimator, see Equation (14), for each element of the sequence is

$$\hat{\rho}_B^{\pi_n}(x) = \left(1 - \frac{1}{n}\right) \frac{p(x|\theta_0)\rho_{\theta_0}}{\left(1 - \frac{1}{n}\right)p(x|\theta_0) + \frac{1}{n}p(x|\theta_1)} + \frac{1}{n} \frac{p(x|\theta_1)\rho_{\theta_1}}{\left(1 - \frac{1}{n}\right)p(x|\theta_0) + \frac{1}{n}p(x|\theta_1)},$$

which in the limiting case reduces to

$$\lim_{n \rightarrow \infty} \hat{\rho}_B^{\pi_n}(x) = \begin{cases} \rho_{\theta_0} & \text{if } x=0, \\ \rho_{\theta_1} & \text{if } x=1. \end{cases}$$

But, the Bayes estimator for the limit μ of the sequence $(\pi_n)_n$ is

$$\hat{\rho}_B^\mu(x) = \begin{cases} \rho_{\theta_0} & \text{if } x=0, \\ \text{not defined} & \text{if } x=1. \end{cases}$$

Since the Bayes estimator for μ is not defined at $x = 1$, we can define it to be

$$\hat{\rho}_B^\mu(x=1) = \lim_{n \rightarrow \infty} \hat{\rho}_B^{\pi_n}(x=1).$$

However, for the Bayes estimator to be continuous in the prior the above should be true for all sequences of priors that have the same limit point. Let us consider another sequence $(\mu_n)_n$ that converges to μ with each element defined as below:

$$\mu_n(\theta) = \left(1 - \frac{1}{n}\right)\delta(\theta - \theta_0) + \frac{1}{n}\delta(\theta - \theta_+),$$

where θ_+ corresponds to $|+\rangle\langle +|$. Then, the Bayes estimator for μ_n would be

$$\hat{\rho}_B^{\mu_n}(x) = \left(1 - \frac{1}{n}\right) \frac{p(x|\theta_0)\rho_{\theta_0}}{\left(1 - \frac{1}{n}\right)p(x|\theta_0) + \frac{1}{n}p(x|\theta_+)} + \frac{1}{n} \frac{p(x|\theta_+)\rho_{\theta_+}}{\left(1 - \frac{1}{n}\right)p(x|\theta_0) + \frac{1}{n}p(x|\theta_+)},$$

which in the limiting case reduces to

$$\lim_{n \rightarrow \infty} \hat{\rho}_B^{\mu_n}(x) = \begin{cases} \rho_{\theta_0} & \text{if } x=0, \\ \rho_{\theta_+} & \text{if } x=1. \end{cases}$$

Now, if we again chose to define the Bayes estimator for μ at $x = 1$ as $\lim_{n \rightarrow \infty} \hat{\rho}_B^{\mu_n}(x=1)$, we would run into a contradiction since $\lim_{n \rightarrow \infty} \hat{\rho}_B^{\mu_n}(x=1) \neq \lim_{n \rightarrow \infty} \hat{\rho}_B^{\pi_n}(x=1)$.

The above example establishes that the Bayes estimator does not admit a continuous extension on the null set (where it is not defined) of the given prior.

E Additional lemma(s)

Lemma E.1. *Given a self-adjoint operator A parametrized by u , the derivative of a function of the operator with respect to the parameter at $u = u_0$ is given by*

$$\left. \frac{d}{du} \operatorname{tr} [f(A(u))] \right|_{u=u_0} = \operatorname{tr} \left[A'(u) \Big|_{u=u_0} f'(A(u_0)) \right]$$

Proof. See reference [28]. □

Lemma E.2. *Consider a continuous function \mathcal{F} defined on a compact set \mathcal{B} and closed subsets $B_x \subseteq \mathcal{B}$ such that $B_1 \subseteq B_2 \dots \subseteq \mathcal{B}$. Then, assuming that the sequence $(B_x)_x$ is dense in \mathcal{B} , the following holds:*

$$\lim_{x \rightarrow \infty} \sup_{B_x} \mathcal{F} = \sup_{\mathcal{B}} \mathcal{F}.$$

Proof. As $B_1 \subseteq B_2 \dots \subseteq \mathcal{B}$, it implies that

$$\lim_{x \rightarrow \infty} \sup_{B_x} \mathcal{F} \leq \sup_{\mathcal{B}} \mathcal{F}. \quad (36)$$

Note that \mathcal{B} is a compact set and hence

$$\sup_{\mathcal{B}} \mathcal{F} = \max_{\mathcal{B}} \mathcal{F}.$$

Let $b := \arg \max_{\mathcal{B}} \mathcal{F}$. Then, as the sequence of subsets $(B_x)_x$ is dense in \mathcal{B} , it implies that there exists a sequence $(b_x)_x$ that converges to b such that $b_x \in B_x$. Moreover, we know that

$$\sup_{B_x} \mathcal{F} \geq \mathcal{F}(b_x).$$

Thus,

$$\lim_{x \rightarrow \infty} \sup_{B_x} \mathcal{F} \geq \lim_{x \rightarrow \infty} \mathcal{F}(b_x).$$

As \mathcal{F} is continuous, we get

$$\lim_{x \rightarrow \infty} \sup_{B_x} \mathcal{F} \geq \mathcal{F}(\lim_{x \rightarrow \infty} b_x) = \mathcal{F}(b) = \sup_{\mathcal{B}} \mathcal{F}. \quad (37)$$

Equations (36) and (37) imply the result. □

Lemma E.3 ([15], Theorem 2.1). *Let P_0 be a positive operator in the representation space such that $[P_0, V_g] = 0 \ \forall g \in G_0$, where G_0 is the stationary subgroup of G , and satisfying:*

$$\int_G V_g P_0 V_g^\dagger d\mu(g) = \mathbb{I} \quad (38)$$

Then setting $P(g\theta_0) = V_g P_0 V_g^\dagger$, we get an operator-valued function of θ such that:

$$M(B) = \int_B P(\theta) d\nu(\theta), \quad B \in \mathcal{A}(\Theta) \quad (39)$$

is a covariant measurement with respect to $g \mapsto V_g$. Conversely, for any covariant measurement $M(d\theta)$ there is a unique operator P_0 satisfying (38) such that $M(B)$ can be expressed as in (39). P_0 is referred to as the seed of the covariant measurement.

F Aside on t-designs

Definition F.1 (Unitary t-design). *Consider the set of unitary matrices $U(d)$ on a d -dimensional Hilbert space \mathcal{H} . A unitary t -design is a finite subset $\{U_i\}_{i=1}^N \subset U(d)$, such that for all states $\rho \in \mathcal{S}(\mathcal{H})$ the following holds:*

$$\frac{1}{N} \sum_{i=1}^N U_i^{\otimes t} \rho (U_i^\dagger)^{\otimes t} = \int_{U(d)} dU U^{\otimes t} \rho (U^\dagger)^{\otimes t},$$

where ‘ dU ’ is the Haar measure on $U(d)$.

Definition F.2 (Spherical t-design). *Consider the unit sphere \mathbb{S}^{d-1} in the d -dimensional Euclidean space \mathbb{R}^d . A spherical t -design is a finite subset $\mathcal{S} \subset \mathbb{S}^{d-1}$, such that the average value of a polynomial f of degree $\leq t$ on \mathcal{S} equals its average on \mathbb{S}^{d-1} :*

$$\frac{1}{|\mathcal{S}|} \sum_{s_n \in \mathcal{S}} f(s_n) = \int_{\mathbb{S}^{d-1}} ds f(s),$$

where ‘ ds ’ is the Lebesgue measure on \mathbb{S}^{d-1} .