Unsupervised Classification of Quantum Data

Gael Sentís,[1] Alex Monràs,[2] Ramon Muñoz-Tapia,[2] John Calsamiglia,[2] and Emilio Bagan[2,3]
[1]*Naturwissenschaftlich-Technische Fakultät, Universität Siegen, 57068 Siegen, Germany*
[2]*Física Teòrica: Informació i Fenòmens Quàntics, Departament de Física, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain*
[3]*Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong*

(Received 19 March 2019; revised manuscript received 4 July 2019; published 8 November 2019)

We introduce the problem of unsupervised classification of quantum data, namely, of systems whose quantum states are unknown. We derive the optimal single-shot protocol for the binary case, where the states in a disordered input array are of two types. Our protocol is universal and able to automatically sort the input under minimal assumptions, yet partially preserves information contained in the states. We quantify analytically its performance for an arbitrary size and dimension of the data. We contrast it with the performance of its classical counterpart, which clusters data that have been sampled from two unknown probability distributions. We find that the quantum protocol fully exploits the dimensionality of the quantum data to achieve a much higher performance, provided the data are at least three dimensional. For the sake of comparison, we discuss the optimal protocol when the classical and quantum states are known.

DOI: 10.1103/PhysRevX.9.041029 Subject Areas: Quantum Information, Statistical Physics

I. INTRODUCTION

Quantum-based communication and computation technologies promise unprecedented applications and unforeseen speed-ups for certain classes of computational problems. In origin, the advantages of quantum computing are exemplary showcased through instances of problems that are hard to solve in a classical computer, such as integer factorization [1], unstructured search [2], discrete optimization [3,4], and simulation of many-body Hamiltonian dynamics [5]. In recent times, the field ventures one step further: Quantum computers are now also envisioned as nodes in a network of quantum devices, where connections are established via quantum channels, and data are quantum systems that flow through the network [6,7]. The design of future quantum networks, in turn, brings up new theoretical challenges, such as devising universal information-processing protocols optimized to work with generic quantum inputs, without the need of human intervention.

Quantum learning algorithms are by design well suited for this class of automated tasks [8]. Generalizing classical machine-learning ideas to operate with quantum data, some algorithms have been devised for quantum template matching [9], quantum anomaly detection [10,11], learning unitary transformations [12] and quantum measurements [13], and classifying quantum states [14–17]. These works fall under the broad category of *supervised* learning [18,19], where the aim is to learn an unknown conditional probability distribution $\Pr(y|x)$ from a number of given samples x_i and associated values or labels y_i, called *training* instances. The performance of a trained learning algorithm is then evaluated by applying the learned function over new data x_i' called *test* instances. In the quantum extension of supervised learning [20], the training instances are quantum—say, copies of the quantum state templates, or a potential anomalous state, or a number of uses of an unknown unitary transformation. The separation between training and testing steps is sometimes not as sharp: In reinforcement learning, training occurs on an instance basis via the interaction of an agent with an environment, and the learning process itself may alter the underlying probability distribution [21].

In contrast, *unsupervised* learning aims at inferring structure in an unknown distribution $\Pr(x)$ given random, unlabeled samples x_i. Typically, this inference is done by grouping the samples in *clusters*, according to a preset definition of similarity. Unsupervised learning is a versatile form of learning, attractive in scenarios where appropriately labeled training data are not available or too costly. But it is also—generically—a much more challenging problem [22,23]. To our knowledge, a quantum extension of unsupervised learning in the sense described above is not yet considered in the literature. In this paper, we take a first step into this branch of quantum learning by introducing the problem of unsupervised binary classification of quantum states. We consider the following scenario: A source

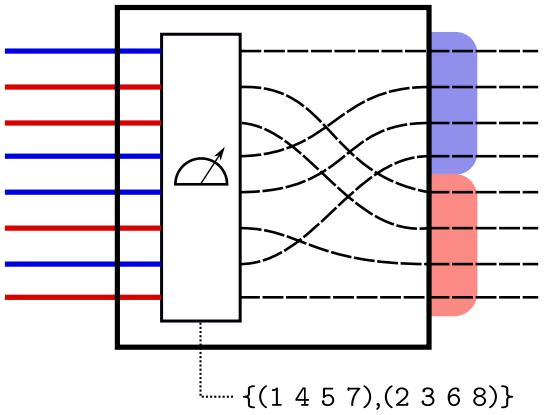


FIG. 1. Pictorial representation of the clustering device for an input of eight quantum states. States of the same type have the same color. States are clustered according to their type by performing a suitable collective measurement, which also provides a classical description of the clustering.

prepares quantum systems in two possible pure states that are completely unknown; after some time, N such systems are produced, and we ask ourselves whether there exists a quantum device that is able to cluster them in two groups according to their states (see Fig. 1). This scenario represents a quantum clustering task in its simplest form, where the single feature defining a cluster of quantum systems is that their states are identical. While clustering classical data under this definition of a cluster—a set of equal data instances—yields a trivial algorithm, merely observing such a simple feature in a finite quantum dataset involves a nontrivial stochastic process and gives rise to a primitive of operational relevance for quantum information. Moreover, in some sense, our scenario actually contains a classical binary clustering problem: If we were to measure each quantum system separately, we would obtain a set of N data points (the measurement outcomes). The points would be effectively sampled from the two probability distributions determined by the quantum states and the choice of measurement. The task would then be to identify which points are sampled from the same distribution. Reciprocally, we can interpret our quantum clustering task as a natural extension of a classical clustering problem with completely unstructured data, where the only single feature that identifies a cluster is that the data points are sampled from a fixed, but arbitrary, categorical probability distribution (i.e., with no order nor metric in the underlying space). The quantum generalization is then to consider (noncommuting) quantum states instead of probability distributions.

We require two important features in our quantum clustering device: (i) It has to be universal—that is, it should be designed to take any possible pair of types of input states—and (ii) it has to provide a classical description of the clustering, that is, which particles belong to each cluster. Feature (i) ensures general-purpose use and versatility of the clustering device, in a similar spirit to programable quantum processors [24]. Feature (ii) allows us to assess the performance of the device purely in terms of the accuracy of the clustering, which, in turn, facilitates the comparison with classical clustering strategies. Also due to (ii), we can justifiably say that the device has not only performed the clustering task but also "learned" that the input is (most likely) partitioned as specified by the output description. Note that relaxing feature (ii), in principle, opens the door to a more general class of *sorting* quantum devices, where the goal could be, e.g., to minimize the distance (under some norm) between the global output state and the state corresponding to perfect clustering of the input. Such devices, however, fall beyond the scope of unsupervised learning.

Requiring the description of the clusters as a classical outcome induces structure in the device. To generate this information, a quantum measurement shall be performed over all N systems with as many outcomes as possible clusterings. Then, the systems will be sorted according to this outcome (see Fig. 1). Depending on the context, e.g., on whether or not the systems will be further used after the clustering, different figures of merit shall be considered in the optimization of the device. In this paper, we focus on the clustering part: Our goal is to find the quantum measurement that maximizes the success probability of a correct clustering.

Features (i) and (ii) allow us to formally regard quantum clustering as a state discrimination task [25–30], albeit with important differences with respect to the standard setting. In quantum state discrimination [25], we want to determine the state of a quantum system among a set of *known* hypotheses (i.e., classical descriptions of quantum states). We can phrase this problem in machine-learning terminology as follows. We have a test state (or several copies of it [29]), and we decide its label based on *infinite training* data. In other words, we have full knowledge about the meaning of the possible labels. Supervised quantum learning algorithms for quantum state classification [14–17] consider the intermediate scenario with *limited training* data. In this case, no description of the states is available. Instead, we are provided with a finite number of copies of systems in each of the possible quantum states, and thus we have only partial classical knowledge about the labels. Extracting the label information from the quantum training data then becomes a key step in the protocol. Following this line of thought, the problem we consider in this paper is a type of unsupervised learning, that is, one with *no training*. There is no information whatsoever about what state each label represents.

We obtain analytical expressions for the performance of the optimal clustering protocol for arbitrary values of the local dimension d of the systems in the cases of a finite number of systems N and in the asymptotic limit of many systems. We show that, in spite of the fact that the number

of possible clusterings grows exponentially with N, the success probability decays only as $O(1/N^2)$. Furthermore, we contrast these results with an optimal clustering algorithm designed for the classical version of the task. We observe a striking phenomenon when analyzing the performance of the two protocols for $d > 2$: Whereas increasing the local dimension has a rapid negative impact in the success probability of the classical protocol (clustering becomes, naturally, harder), it turns out to be beneficial for its quantum counterpart.

We also see, through a numerical analysis, that the quantum measurement that maximizes the success probability is also optimal for a more general class of cost functions that are more natural for clustering problems, including the Hamming distance. In other words, this observation provides evidence that our entire analysis does not depend strongly on the chosen figure of merit but rather on the structure of the problem itself.

Measuring the systems will, in principle, degrade the information encoded in their states; hence, intuitively, there should be a trade-off between how good a clustering is and how much information about the original states is left in the clusters. Remarkably, our analysis reveals that the measurement that clusterizes optimally actually preserves information regarding the type of states that form each cluster. This feature adds to the usability of our device as a universal quantum data sorting processor. It can be regarded as the quantum analog of a sorting network (or sorting memory) [31], used as a fixed network architecture that automatically orders generic inputs coming from an aggregated data pipeline. The details of this second step are, however, left for a subsequent publication.

The paper is organized as follows. In Sec. II, we formalize the problem and derive the optimal clustering protocol and its performance. In Sec. III, we consider a classical clustering protocol and contrast it with the optimal one. We present the proofs of the main results of our work and the necessary theoretical tools to derive them in Sec. IV. We end in Sec. V discussing the features of our quantum clustering device and other cost functions and giving an outlook on future extensions.

II. CLUSTERING QUANTUM STATES

Let us suppose that a source prepares quantum systems randomly in one of two pure d-dimensional states $|\phi_0\rangle$ and $|\phi_1\rangle$ with equal prior probabilities. Given a sequence of N systems produced by the source, and with no knowledge of the states $|\phi_{0/1}\rangle$, we are required to assign labels "0" or "1" to each of the systems. The labeling can be achieved via a generalized quantum measurement that tries to distinguish among all the possible global states of the N systems. Each outcome of the measurement is then associated to a possible label assignment, that is, to a *clustering*.

Consider the case of four systems. All possible clusterings that we may arrange are depicted in Fig. 2 as strings of

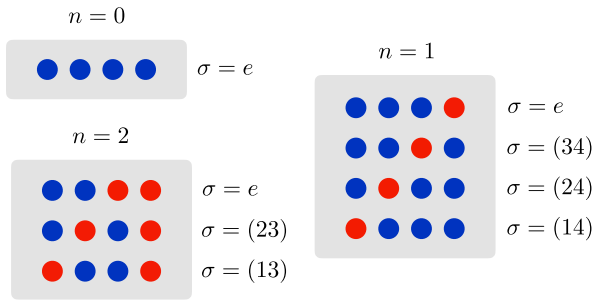


FIG. 2. All possible clusterings of $N = 4$ systems when each can be in one of two possible states, depicted as blue and red. The pair of indices (n, σ) identifies each clustering, where n is the size of the smallest cluster and σ is a permutation of the reference clusterings (those on top of each box), wherein the smallest cluster falls on the right. The symbol e denotes the identity permutation, and (ij) the transposition of systems in positions i and j. Note that the choice of σ is not unique.

red and blue balls. Since the individual states of the systems are unknown, what is labeled as "red" or "blue" is arbitrary; thus, interchanging the labels leads to an equivalent clustering. For arbitrary N, there are 2^{N-1} such clusterings. Figure 2 also illustrates a natural way to label each clustering as (n, σ). The index n counts the number of systems in the smallest cluster. The index σ is a permutation that brings a *reference* clustering, defined as that in which the systems belonging to the smallest cluster fall all on the right, into the desired form. To make this labeling unambiguous, σ is chosen from a restricted set $\mathcal{S}_n \subset S_N$, where S_N stands for the permutation group of N elements and e denotes its unity element. We see that the optimal clustering procedure consists in measuring first the value of n and, depending on the outcome, performing a second measurement that identifies σ among the relevant permutations with a fixed n.

Thus, unsupervised clustering is cast as a multihypothesis discrimination problem, which can be solved for an arbitrary number of systems N with local dimension d. Below, we outline the derivation of our main result: the expression of the maximum average success probability achievable by a quantum clustering protocol. In the limit of large N and for arbitrary d (not necessarily constant with N), we show that this probability behaves as

$$P_s \sim \frac{8(d-1)}{(2d+N)N}. \tag{1}$$

(The symbol $\sim$ stands for "asymptotically equivalent to", as in Ref. [32].). Naturally, P_s goes to zero with N, since the total number of clusterings increases exponentially and it becomes much harder to discriminate among them. What may perhaps come as a surprise is that, despite this exponential growth, the scaling of P_s is only of the order of $O(1/N^2)$. (It is also interesting to see how far one can

improve this result. By letting d scale with N, e.g., by substituting $d \sim sN^{\gamma}$ for some $s > 0$, $\gamma > 1$ in Eq. (1), we obtain the absolute maximum $P_s \sim 4/N$.) Furthermore, increasing the local dimension yields a linear improvement in the asymptotic success probability. As we later see, whereas the asymptotic behavior in N is not an exclusive feature of the optimal quantum protocol—we observe the same scaling in its classical counterpart, albeit only when $d = 2$—the ability to exploit extra dimensions to enhance distinguishability is.

Let us present an outlined derivation of the optimal quantum clustering protocol. Each input can be described by a string of 0's and 1's $\mathbf{x} = (x_1 \ldots x_N)$, so that the global state of the systems entering the device is $|\Phi_{\mathbf{x}}\rangle = |\phi_{x_1}\rangle \otimes |\phi_{x_2}\rangle \otimes \cdots \otimes |\phi_{x_N}\rangle$. The clustering device can generically be defined by a positive operator valued measure (POVM) with elements $\{E_{\mathbf{x}}\}$, fulfilling $E_{\mathbf{x}} \geq 0$ and $\sum_{\mathbf{x}} E_{\mathbf{x}} = \mathbb{1}$, where each operator $E_{\mathbf{x}}$ is associated to the statement "the measured global state corresponds to the string $\mathbf{x}$." We want to find a POVM that maximizes the average success probability $P_s = 2^{1-N} \int d\phi_0 d\phi_1 \sum_{\mathbf{x}} \mathrm{tr}(|\Phi_{\mathbf{x}}\rangle\langle\Phi_{\mathbf{x}}| E_{\mathbf{x}})$, where we assume that each clustering is equally likely at the input and we are averaging over all possible pairs of states $\{|\phi_0\rangle, |\phi_1\rangle\}$ and strings $\mathbf{x}$. Since our goal is to design a universal clustering protocol, the operators $E_{\mathbf{x}}$ cannot depend on $|\phi_{0,1}\rangle$, and we can take the integral inside the trace. The clustering problem can then be regarded as the optimization of a POVM that distinguishes between effective density operators of the form

$$\rho_{\mathbf{x}} = \int d\phi_0 d\phi_1 |\Phi_{\mathbf{x}}\rangle\langle\Phi_{\mathbf{x}}|. \tag{2}$$

It now becomes apparent that $\rho_{\mathbf{x}} = \rho_{\bar{\mathbf{x}}}$, where $\bar{\mathbf{x}}$ is the complementary string of $\mathbf{x}$ (i.e., the values 0 and 1 are exchanged).

The key that reveals the structure of the problem and allows us to deduce the optimal clustering protocol resides in computing the integral in Eq. (2). Averaging over the states leaves out only the information relevant to identify a clustering, that is, n and σ. Certainly, identifying $\mathbf{x} \equiv (n, \sigma)$, we can rewrite $\rho_{\mathbf{x}}$ as

$$\begin{aligned} \rho_{n,\sigma} &= c_n U_\sigma (\mathbb{1}_n^{\mathrm{sym}} \otimes \mathbb{1}_{N-n}^{\mathrm{sym}}) U_\sigma^\dagger \\ &= c_n \bigoplus_{\lambda} \mathbb{1}_{(\lambda)} \otimes \Omega_{\{\lambda\}}^{n,\sigma}. \end{aligned} \tag{3}$$

By applying the Schur lemma, one readily obtains the first line, where $\mathbb{1}_k^{\mathrm{sym}}$ is a projector onto the completely symmetric subspace of k systems, c_n is a normalization factor, and U_σ is a unitary matrix representation of σ. The second line follows from using the Schur basis (see Sec. IV A), in which the states $\rho_{n,\sigma}$ are block diagonal. Here, λ labels the irreducible representations—irreps for short—of the joint action of the groups $\mathrm{SU}(d)$ and S_N over the vector space $(d, \mathbb{C})^{\otimes N}$ and is usually identified with the shape of Young diagrams (or partitions of N). A pair of parentheses () [braces {}], surrounding the subscript λ, e.g., in Eq. (3), are used when λ refers exclusively to irreps of $\mathrm{SU}(d)$ [S_N]; we stick to this convention throughout the paper. Note that averaging over all $\mathrm{SU}(d)$ transformations erases the information contained in the representation subspace (λ). It also follows from Eq. (3) and the rules of the Clebsch-Gordan decomposition that (i) only two-row Young diagrams (partitions of length two) show up in the direct sum above, and (ii) the operators $\Omega_{\{\lambda\}}^{n,\sigma}$ are rank-1 projectors (see Appendix B). They carry all the information relevant for the clustering and are understood to be zero for irreps λ outside the support of $\rho_{n,\sigma}$.

With Eq. (3) at hand, the optimal clustering protocol can be succinctly described as two successive measurements—we state the result here and present an optimality proof in Sec. IV A. The first measurement is a projection onto the irrep subspaces λ, described by the set $\{\mathbb{1}_{(\lambda)} \otimes \mathbb{1}_{\{\lambda\}}\}$. The outcome of this measurement provides an estimate of n, as λ is one to one related to the size of the clusters. More precisely, we have from (i) that $\lambda = (\lambda_1, \lambda_2)$, where λ_1 and λ_2 are non-negative integers such that $\lambda_1 + \lambda_2 = N$ and $\lambda_1 \geq \lambda_2$. Then, given the outcome $\lambda = (\lambda_1, \lambda_2)$ of this first measurement, the optimal guess turns out to be $n = \lambda_2$. Very roughly speaking, the "asymmetry" in the subspace $\lambda = (\lambda_1, \lambda_2)$ increases with λ_2. We recall that $\lambda = (N, 0)$ is the fully symmetric subspace of $(d, \mathbb{C})^N$. Naturally, $\rho_{0,\sigma}$ has support only in this subspace, as all states in the data are of one type. As λ_2 increases from zero, more states of the alternative type are necessary to achieve the increasing asymmetry of $\lambda = (\lambda_1, \lambda_2)$. Hence, for a given λ_2, there is a minimum value of n for which $\rho_{n,\sigma}$ can have support in the subspace $\lambda = (\lambda_1, \lambda_2)$. This minimum n is the optimal guess.

Once we obtain a particular $\lambda = \lambda^*$ as an outcome (and guess n), a second measurement is performed over the subspace $\{\lambda^*\}$ to produce a guess for σ. Since the states $\rho_{n,\sigma}$ are covariant under S_N, the optimal measurement to guess the permutation σ is also covariant, and its seed is the rank-1 operator $\Omega_{\{\lambda^*\}}^{n,e}$, where $\lambda^* = (N-n, n)$. Put together, these two successive measurements yield a joint optimal POVM whose elements take the form

$$E_{n,\sigma} = \xi_{\lambda^*}^n (\mathbb{1}_{(\lambda^*)} \otimes \Omega_{\{\lambda^*\}}^{n,\sigma}), \tag{4}$$

where (n, σ) is the guess for the cluster and $\xi_{\lambda^*}^n$ is some coefficient that guarantees the POVM condition $\sum_{n,\sigma} E_{n,\sigma} = \mathbb{1}$.

The success probability of the optimal protocol can be computed as $P_s = 2^{1-N} \sum_{n,\sigma} \mathrm{tr}(\rho_{n,\sigma} E_{n,\sigma})$ (see Sec. IV A). It reads

$$P_s = 2^{1-N} \sum_{i=0}^{\lfloor N/2 \rfloor} \binom{N}{i} \frac{(d-1)(N-1-2i)^2}{(N-1+d-i)(i+1)^2}, \tag{5}$$

from which the asymptotic limit Eq. (1) follows (see Appendix C).

Before closing this section, we briefly discuss the case when some information about the possible states $|\phi_0\rangle$ and $|\phi_1\rangle$ is available. A clustering device that incorporates this information into its design should succeed with a probability higher than Eq. (5), at the cost of universality. To explore the extent of this performance enhancement, we study the extreme case where we have full knowledge of the states $|\phi_0\rangle$ and $|\phi_1\rangle$. We find that in the large N limit the maximum improvement is by a factor of N. The optimal success probability scales as

$$P_s \sim \frac{4(d-1)}{N} \tag{6}$$

(see Sec. IV B for details).

III. CLUSTERING CLASSICAL STATES

To grasp the significance of our quantum clustering protocol, a comparison with a classical analog is called for. First, in the place of a quantum system whose state is either $|\phi_0\rangle$ or $|\phi_1\rangle$, an input would be an instance of a d-dimensional random variable sampled from either one of two categorical probability distributions: $P = \{p_s\}_{s=1}^d$ and $Q = \{q_s\}_{s=1}^d$. Then, given a string of samples $\mathbf{s} = (s_1 \ldots s_N)$, $s_i \in \{1, \ldots, d\}$, the clustering task would consist in grouping the data points s_i in two clusters so that all points in a cluster have a common underlying probability distribution.

Second, in analogy with the quantum protocol, our goal would be to find the optimal universal (i.e., independent of P and Q) protocol that performs this task. Here, optimality means attaining the maximum average success probability, where the average is over all N-length sequences $\mathbf{x}$ of distributions P and Q from which the string $\mathbf{s}$ is sampled and over all such distributions.

It should be emphasized that this is a very hard classical clustering problem, with absolute minimal assumptions, where there is no metric in the domain of the random variables and, in consequence, no exploitable notion of distance. Therefore, one should expect the optimal algorithm to have a rather low performance and to differ significantly from well-known algorithms for classical unsupervised classification problems.

As a further remark, we note that a choice of prior is required to perform the average over P and Q. We assume that the two are uniformly distributed over the simplex on which they are both defined. This assumption reflects our lack of knowledge about the distributions underlying the string of samples $\mathbf{s}$.

Under all these specifications, the classical clustering problem we just defined naturally connects with the quantum scenario in Sec. II as follows. We can interpret $\mathbf{s}$ as a string of outcomes obtained upon performing the same projective measurement on each individual quantum state $|\phi_{x_i}\rangle$ of our original problem. Furthermore, such local measurements can also be interpreted as a decoherence process affecting the pure quantum states at the input, whereby they decay into classical probability distributions over a fixed basis. We might think of this as the semiclassical analog of our original problem, since quantum resources are not fully exploited.

Let us first lay out the problem in the special case of $d = 2$, where the underlying distributions are Bernoulli, and we can write $P = \{p, 1-p\}$, $Q = \{q, 1-q\}$. Given an N-length string of samples $\mathbf{s}$, our intuition tells us that the best we can do is to assign the same underlying probability distribution to equal values in $\mathbf{s}$. So if, e.g., $\mathbf{s} = (00101\ldots)$, we guess that the underlying sequence of distributions is $\hat{\mathbf{x}} = (PPQPQ\ldots)$ [or, equivalently, the complementary sequence $\hat{\mathbf{x}} = (QQPQP\ldots)$]. Thus, data points are clustered according to their value 0 or 1. The optimality of this guessing rule is a particular case of the result for d-dimensional random variables in Appendix F.

The probability that a string of samples $\mathbf{s}$, with l zeros and $N-l$ ones, arises from the guessed sequence $\hat{\mathbf{x}}$ is given by

$$\Pr(\mathbf{s}|\mathbf{x} = \hat{\mathbf{x}}) = \int_0^1 dp \int_0^1 dq\, p^l q^{N-l} = \frac{1}{(l+1)(N-l+1)}. \tag{7}$$

The average success probability can then be readily computed as $P_s^{\mathrm{cl}} = 2\sum_{\mathbf{x},\mathbf{s}} \delta_{\mathbf{x},\hat{\mathbf{x}}} \Pr(\mathbf{x}) \Pr(\mathbf{s}|\mathbf{x})$ (recall that $\hat{\mathbf{x}}$ depends on $\mathbf{s}$), where $\Pr(\mathbf{x}) = 2^{-N}$ is the prior probability of the sequence $\mathbf{x}$, which we assume to be uniform. The factor 2 takes into account that guessing the complementary sequence leads to the same clustering. It is now quite straightforward to derive the asymptotic expression of P_s^{cl} for large N. In this limit, $\mathbf{x}$ typically has the same number of P and Q distributions, so the guess $\hat{\mathbf{x}}$ is right if $l = N/2$. Then,

$$P_s^{\mathrm{cl}} \sim 2\frac{1}{(N/2+1)^2} \sim \frac{8}{N^2}. \tag{8}$$

This expression coincides with the quantum asymptotic result in Eq. (1) for $d = 2$. As we now see, this coincidence is, however, a particularity of Bernoulli distributions.

The derivation for $d > 2$ is more involved, since the optimal guessing rule is not so obvious (see Appendix F for details). Loosely speaking, we should still assign samples with the same value to the same cluster. By doing so, we obtain up to d preliminary clusters. We next merge them into two clusters in such a way that their final sizes are as balanced as possible. This last step, known as the *partition problem* [33], is weakly NP complete. Namely, its complexity is polynomial in the magnitudes of the data involved (the size of the preliminary clusters, which depends on N)

but nonpolynomial in the input size (the number of such clusters, determined by d). The complexity of this last step implies that the classical and semiclassical protocols cannot be implemented efficiently for arbitrary d. In the asymptotic limit of large N and for arbitrary fixed values of d, we obtain

$$P_s^{\mathrm{cl}} \sim \left(\frac{2}{N}\right)^d \frac{(2d-2)!}{(d-2)!}. \tag{9}$$

There is a huge difference between this result and Eq. (1). Whereas increasing the local dimension provides an asymptotic linear advantage in the optimal quantum clustering protocol—states become more orthogonal—it has the opposite effect in its classical and semiclassical analogs, as it reduces exponentially the success probability.

In the opposite regime, i.e., for d asymptotically large and fixed values of N, the optimal classical and semiclassical strategies provide no improvement over random guessing, and the clustering tasks become exceedingly hard and somewhat uninteresting. This fact follows from observing that the guessing rule relies on grouping repeated data values. In this regime, the typical string of samples $\mathbf{s}$ has no repeated elements; thus, we are left with no alternative but to randomly guess the right clustering of the data and $P_s^{\mathrm{cl}} \sim 2^{1-N}$.

To complete the picture, we end this section by considering known classical probability distributions. Akin to the quantum case, one expects an increase in the success probability of clustering. An immediate consequence of knowing the distributions P and Q is that the rule for assigning a clustering given a string of samples $\mathbf{s}$ becomes trivial. Each symbol $s_i \in \{1, \ldots, d\}$ is assigned to the most likely distribution, that is, to P (Q) if $p_{s_i} > q_{s_i}$ ($p_{s_i} < q_{s_i}$). It is clear that knowing P and Q helps to better classify the data, which becomes apparent by considering the example of two three-dimensional distributions and the data string $\mathbf{s} = (112)$. If the distributions are unknown, such a sequence leads to the guess $\hat{\mathbf{x}} = (PPQ)$ [or, equivalently, to $\hat{\mathbf{x}} = (QQP)$]. In contrast, if P and Q are known and, e.g., $p_1 > q_1$ and $p_2 > q_2$, the same sequence leads to the better guess $\hat{\mathbf{x}} = (PPP)$. The advantage of knowing the distribution, however, vanishes in the large N limit, and the asymptotic performance of the optimal clustering algorithm is shown to be given by Eq. (9). The interested reader can find the details of the proof in Appendix G.

IV. METHODS

Here, we give the full proof of optimality of our quantum clustering protocol and device, which leads to our main result in Eq. (1). The proof relies on the representation theory of the special unitary and the symmetric groups. In particular, the Schur-Weyl duality is used to efficiently represent the structure of the input quantum data and the action of the device. We then leverage this structure to find the optimal POVM and compute the minimum cost. Basic notions of the representation theory that we use in the proof are covered in Appendixes A and B. We close the methods section proving Eq. (6) for the optimal success probability of clustering known quantum states.

A. Clustering quantum states: Unknown input states

In this section, we obtain the optimal POVM for quantum clustering and compute the minimum cost. First, we present a formal optimality proof for an arbitrary cost function $f(\mathbf{x}, \mathbf{x}')$, which specifies the penalty for guessing $\mathbf{x}$ if the input is $\mathbf{x}'$. Second, we particularize to the case of success probability, as discussed in the main text, for which explicit expressions are obtained.

1. Generic cost functions

We say a POVM is optimal if it minimizes the average cost

$$\bar{f} = \int d\phi_0 d\phi_1 \sum_{\mathbf{x},\hat{\mathbf{x}}} \eta_{\mathbf{x}} f(\mathbf{x}, \hat{\mathbf{x}}) \Pr(\hat{\mathbf{x}}|\mathbf{x}), \tag{10}$$

where $\eta_{\mathbf{x}}$ is the prior probability of input string $\mathbf{x}$ and $\Pr(\hat{\mathbf{x}}|\mathbf{x}) = \mathrm{tr}(|\Phi_{\mathbf{x}}\rangle\langle\Phi_{\mathbf{x}}|E_{\hat{\mathbf{x}}})$ is the probability of obtaining measurement outcome (and guess) $\hat{\mathbf{x}}$ given input $\mathbf{x}$; recall that $|\Phi_{\mathbf{x}}\rangle = |\phi_{x_1}\rangle \otimes |\phi_{x_2}\rangle \otimes \ldots \otimes |\phi_{x_N}\rangle$, $x_k = 0, 1$, and an average is taken over all possible pairs of states $\{|\phi_0\rangle, |\phi_1\rangle\}$; hence, $\mathbf{x}$ and its complementary $\bar{\mathbf{x}}$ define the same clustering. A convenient way to identify the different clusterings is by counting the number n, $0 \leq n \leq \lfloor N/2 \rfloor$, of zeros in $\mathbf{x}$ (so, strings with more 0's than 1's are discarded) and giving a unique representative σ of the equivalence class of permutations that turn the reference string $(0^n 1^{\bar{n}})$, $\bar{n} = N - n$, into $\mathbf{x}$. We denote the subset of these representatives by $\mathcal{S}_n \subset S_N$ and the number of elements in each equivalence class by b_n. A simple calculation gives us $b_n = 2(n!)^2$ if $n = \bar{n}$, and $b_n = n!\bar{n}!$ otherwise.

As discussed in the main text, the clustering problem above is equivalent to a multihypothesis discrimination problem, where the hypotheses are given by

$$\begin{aligned}\rho_{\mathbf{x}} &= \int d\phi_0 d\phi_1 |\Phi_{\mathbf{x}}\rangle\langle|\Phi_{\mathbf{x}}| \\ &= c_n U_\sigma (\mathbb{1}_n^{\mathrm{sym}} \otimes \mathbb{1}_{\bar{n}}^{\mathrm{sym}}) U_\sigma^\dagger,\end{aligned} \tag{11}$$

and we have used the Schur lemma to compute the integral. Here, U_σ is a unitary matrix representation of the permutation σ, $\mathbb{1}_k^{\mathrm{sym}}$ is a projector onto the completely symmetric subspace of k systems, and $c_n = 1/(D_n^{\mathrm{sym}} D_{\bar{n}}^{\mathrm{sym}})$, where $D_k^{\mathrm{sym}} = s_{(k,0)}$ [see Eq. (B6)] is the dimension of symmetric subspace of k qudits.

The states (11) are block diagonal in the Schur basis, which decouples the commuting actions of the groups $\mathrm{SU}(d)$ and S_N over product states of the form of $|\Phi_{\mathbf{x}}\rangle$. More

precisely, Schur-Weyl duality states that the representations of the two groups acting on the common space $(d,\mathbb{C})^{\otimes N}$ are each other's commutant. Moreover, it provides a decomposition of this space into decoupled subspaces associated to irreps of both $\mathrm{SU}(d)$ and S_N. We can then express the states $\rho_{\mathbf{x}}$, where $\mathbf{x}$ is specified as (n,σ) [$\mathbf{x}=(n,\sigma)$ for short], in the Schur basis as

$$\rho_{n,\sigma} = c_n \bigoplus_{\lambda} \mathbb{1}_{(\lambda)} \otimes \Omega^{n,\sigma}_{\{\lambda\}}. \tag{12}$$

In this direct sum, λ is a label attached to the irreps of the joint action of $\mathrm{SU}(d)$ and S_N and is usually identified with a partition of N or, equivalently, a Young diagram. As explained in the main text, a pair of parentheses surrounding this type of label, like in (λ), means that it refers specifically to irreps of $\mathrm{SU}(d)$. Likewise, a pair of braces, e.g., $\{\lambda\}$, indicates that the label refers to irreps of S_N. In accordance with this convention, Schur-Weyl duality implies that $\Omega^{n,\sigma}_{\{\lambda\}} = U^{\lambda}_{\sigma}\Omega^{n,e}_{\{\lambda\}}(U^{\lambda}_{\sigma})^{\dagger}$, where U^{λ}_{σ} is the matrix of the irrep λ that represents $\sigma \in S_N$ and e denotes the identity permutation (for simplicity, we omit the index e when no confusion arises). In other words, the family of states $\rho_{n,\sigma}$ is covariant with respect to S_N. One can easily check that $\Omega^{n,\sigma}_{\{\lambda\}}$ is always a rank-1 projector (see Appendix B). In Eq. (12), it is understood that $\Omega^{n,\sigma}_{\{\lambda\}} = 0$ outside of the range of $\rho_{n,\sigma}$.

With no loss of generality, the optimal measurement that discriminates the states $\rho_{n,\sigma}$ can be represented by a POVM whose elements have the form shown in Eq. (12). Moreover, we can assume it to be covariant under S_N [34]. So, such POVM elements can be written as

$$E_{n,\sigma} = \bigoplus_{\lambda} \mathbb{1}_{(\lambda)} \otimes U^{\lambda}_{\sigma}\Xi^{n}_{\{\lambda\}}(U^{\lambda}_{\sigma})^{\dagger}, \tag{13}$$

where $\Xi^{n}_{\{\lambda\}}$ is some positive operator. The resolution of the identity condition imposes constraints on them. The condition reads

$$\begin{aligned}\sum_{n,\sigma} E_{n,\sigma} &= \sum_{n} \frac{1}{b_n} \sum_{\sigma\in S_N} \bigoplus_{\lambda} \mathbb{1}_{(\lambda)} \otimes U^{\lambda}_{\sigma}\Xi^{n}_{\{\lambda\}}(U^{\lambda}_{\sigma})^{\dagger} \\ &= \bigoplus_{\lambda} \mathbb{1}_{(\lambda)} \otimes \mathbb{1}_{\{\lambda\}},\end{aligned} \tag{14}$$

where we use the factor b_n to extend the sum over $\mathcal{S}_n$ to the entire group S_N and apply the Schur lemma. Taking the trace on both sides of the equation, we find the POVM constraint to be

$$\sum_{n} \frac{N!}{b_n} \mathrm{tr}(\Xi^{n}_{\{\lambda\}}) = \nu_{\lambda} \quad \forall\ \lambda, \tag{15}$$

where ν_{λ} is the dimension of $\mathbb{1}_{\{\lambda\}}$ or, equivalently, the multiplicity of the irrep λ of $\mathrm{SU}(d)$ [see Eq. (B5)].

So far, we have analyzed the structure that the symmetries of the problem impose on the states $\rho_{n,\sigma}$ and the measurements. We have learned that, for any choice of operators $\Xi^{n}_{\{\lambda\}}$ that fulfill Eq. (15), the set of operators (13) defines a valid POVM, but it need not be optimal. So, we now proceed to derive optimality conditions for $\Xi^{n}_{\{\lambda\}}$. Those are provided by the Holevo-Yuen-Kennedy-Lax [35,36] necessary and sufficient conditions for minimizing the average cost. For our clustering problem in Eq. (10), they read

$$(W_{\mathbf{x}} - \Gamma)E_{\mathbf{x}} = E_{\mathbf{x}}(W_{\mathbf{x}} - \Gamma) = 0, \tag{16}$$

$$W_{\mathbf{x}} - \Gamma \geq 0. \tag{17}$$

They must hold for all $\mathbf{x}$, where $\Gamma = \sum_{\mathbf{x}} W_{\mathbf{x}}E_{\mathbf{x}} = \sum_{\mathbf{x}} E_{\mathbf{x}}W_{\mathbf{x}}$, and $W_{\mathbf{x}} = \sum_{\mathbf{x}'} f(\mathbf{x},\mathbf{x}')\eta_{\mathbf{x}'}\rho_{\mathbf{x}'}$. We assume that the prior distribution $\eta_{\mathbf{x}}$ is flat and that the cost function is non-negative and covariant with respect to the permutation group, i.e., $f(\mathbf{x},\mathbf{x}') = f(\tau\mathbf{x},\tau\mathbf{x}')$ for all $\tau \in S_N$. Then, $W_{\tau\mathbf{x}} = U_{\tau}W_{\mathbf{x}}U_{\tau}^{\dagger}$, and we need only to ensure that conditions (16) and (17) are met for reference strings, for which $\mathbf{x} = (n,e)$. In the Schur basis, their corresponding operators, which we simply call W_n, and the matrix Γ take the form, respectively,

$$W_n = \bigoplus_{\lambda} \mathbb{1}_{(\lambda)} \otimes \omega^{n}_{\{\lambda\}}, \tag{18}$$

$$\Gamma = \bigoplus_{\lambda} k_{\lambda} \mathbb{1}_{(\lambda)} \otimes \mathbb{1}_{\{\lambda\}}, \tag{19}$$

where we use the Schur lemma to obtain Eq. (19) and define $k_{\lambda} \equiv \sum_n N!\mathrm{tr}(\omega^{n}_{\{\lambda\}}\Xi^{n}_{\{\lambda\}})/(b_n\nu_{\lambda})$. Note that Γ is a diagonal matrix, in spite of the fact that $\omega^{n}_{\{\lambda\}}$ are, at this point, arbitrary full-rank positive operators.

With Eqs. (18) and (19), the optimality conditions (16) and (17) can be made explicit. First, we note that the subspace (λ) is irrelevant in this calculation and that there is an independent condition for each irrep λ. Taking into account these considerations, Eq. (16) now reads

$$\omega^{n}_{\{\lambda\}}\Xi^{n}_{\{\lambda\}} = \Xi^{n}_{\{\lambda\}}\omega^{n}_{\{\lambda\}} = k_{\lambda}\Xi^{n}_{\{\lambda\}} \quad \forall\ n,\lambda. \tag{20}$$

This equation tells us two things: (i) Since the matrices $\omega^{n}_{\{\lambda\}}$ and $\Xi^{n}_{\{\lambda\}}$ commute, they have a common eigenbasis, and (ii) Eq. (20) is a set of eigenvalue equations for $\omega^{n}_{\{\lambda\}}$ with a common eigenvalue k_{λ}, one equation for each eigenvector of $\Xi^{n}_{\{\lambda\}}$. Therefore, the support of $\Xi^{n}_{\{\lambda\}}$ is necessarily restricted to a single eigenspace of $\omega^{n}_{\{\lambda\}}$. Denoting by $\vartheta^{n}_{\lambda,a}$, $a = 1,2,\ldots$, the eigenvalues of $\omega^{n}_{\{\lambda\}}$ sorted in increasing order, we have $k_{\lambda} = \vartheta^{n}_{\lambda,a}$ for some a, which may depend on λ and n, or else $\Xi^{n}_{\{\lambda\}} = 0$.

The second Holevo condition (17), under the same considerations regarding the block-diagonal structure, leads to

$$\omega^n_{\{\lambda\}} \geq k_\lambda \mathbb{1}_{\{\lambda\}} \quad \forall\ n, \lambda. \tag{21}$$

This condition further induces more structure in the POVM. Given λ, Eq. (21) has to hold for *every* value of n. In particular, we must have $\min_{n'} \vartheta^{n'}_{\lambda,1} \geq k_\lambda$. Therefore, $\min_{n'} \vartheta^{n'}_{\lambda,1} \geq \vartheta^n_{\lambda,a}$ for some a, or else $\Xi^n_{\{\lambda\}} = 0$. Since $\Xi^n_{\{\lambda\}}$ cannot vanish for all n because of Eq. (15), we readily see that

$$k_\lambda = \vartheta^{n(\lambda)}_{\lambda,1}, \quad \Xi^n_{\{\lambda\}} = \begin{cases} \xi^n_\lambda \Pi_1(\omega^n_{\{\lambda\}}) & \text{if } n = n(\lambda), \\ 0 & \text{otherwise,} \end{cases} \tag{22}$$

where $n(\lambda) = \mathrm{argmin}_n \vartheta^n_{\lambda,1}$, $\Pi_1(\omega^n_{\{\lambda\}})$ is a projector onto the eigenspace of $\omega^n_{\{\lambda\}}$ (not necessarily the whole subspace) corresponding to the minimum eigenvalue $\vartheta^n_{\lambda,1}$, and ξ^n_λ is a suitable coefficient that can be read off from Eq. (15):

$$\xi^n_\lambda = \frac{\nu_\lambda b_n}{D^n_\lambda N!}, \tag{23}$$

where $D^n_\lambda = \dim[\Pi_1(\omega^n_{\{\lambda\}})]$, which completes the construction of the optimal POVM.

For a generic cost function, we can now write down a closed, implicit formula for the minimum average cost achievable by any quantum clustering protocol. It reads

$$\bar{f} = \mathrm{tr}\Gamma = \sum_\lambda s_\lambda \nu_\lambda \vartheta^{n(\lambda)}_{\lambda,1}, \tag{24}$$

where s_λ is the dimension of $\mathbb{1}_{(\lambda)}$ or, equivalently, the multiplicity of the irrep λ of S_N [see Eq. (B6)]. The only object that remains to be specified is the function $n(\lambda)$, which depends ultimately on the choice of the cost function $f(\mathbf{x}, \mathbf{x}')$.

2. Success probability

We now make Eq. (24) explicit by considering the success probability P_s as a figure of merit; that is, we choose $f(\mathbf{x}, \mathbf{x}') = 1 - \delta_{\mathbf{x},\mathbf{x}'}$ and, hence, $P_s = 1 - \bar{f}$. We also assume that the source that produces the input sequence is equally likely to prepare either state; thus, each string $\mathbf{x}$ has the same prior probability, $\eta_\mathbf{x} = 2^{1-N} \equiv \eta$. In this case, W_n takes the simple form

$$W_n = \bigoplus_\lambda \mathbb{1}_{(\lambda)} \otimes (\mu_\lambda \mathbb{1}_{\{\lambda\}} - \eta c_n \Omega^n_{\{\lambda\}}), \tag{25}$$

where μ_λ are positive coefficients and we recall that the expression in parentheses corresponds to $\omega^n_{\{\lambda\}}$ in Eq. (18). From this expression, one can easily derive the explicit forms of $\vartheta^n_{\lambda,1}$ and $n(\lambda)$. We just need to consider the maximum eigenvalue of the rank-one projector $\Omega^n_{\{\lambda\}}$, which can be either one or zero depending on whether or not the input state $\rho_{n,\sigma}$ has support in the irrep λ space. So, among the values of n for which $\rho_{n,\sigma}$ does have support there, $n(\lambda)$ is one that maximizes c_n. Since c_n is a decreasing function of n in its allowed range (recall that $n \leq \lfloor N/2 \rfloor$), $n(\lambda)$ is the smallest such value.

For the problem at hand, the irreps in the direct sum can be labeled by Young diagrams of at most two rows or, equivalently, by partitions of N of length at most two (see Appendix B); hence, $\lambda = (\lambda_1, \lambda_2)$, where $\lambda_1 + \lambda_2 = N$ and λ_2 runs from 0 to $\lfloor N/2 \rfloor$. Given λ, only states ρ_n with $n = \lambda_2, \ldots, \lfloor N/2 \rfloor$ have support on the irrep λ space, as readily follows from the Clebsch-Gordan decomposition rules. Then,

$$n(\lambda) = \lambda_2, \qquad \vartheta^{n(\lambda)}_{\lambda,1} = \mu_\lambda - \eta c_{n(\lambda)}. \tag{26}$$

Equation (26) gives the optimal guess for the size n of the smallest cluster. The rule is in agreement with our intuition. The irrep $(N, 0)$, i.e., $\lambda_2 = 0$, corresponding to the fully symmetric subspace, is naturally associated with the value $n = 0$, i.e., with all N systems being in the same state or cluster; the irrep with one antisymmetrized index has $\lambda_2 = 1$ and hints at a system being in a different state than the others, i.e., at a cluster of size one, and so on.

We now have all the ingredients to compute the optimal success probability from Eq. (24). It reads

$$\begin{aligned} P_s &= \eta \sum_\lambda c_{n(\lambda)} s_\lambda \nu_\lambda \\ &= \frac{1}{2^{N-1}} \sum_{i=0}^{\lfloor N/2 \rfloor} \binom{N}{i} \frac{(d-1)(N-2i+1)^2}{(d+i-1)(N-i+1)^2}, \end{aligned} \tag{27}$$

where we use the relation $\sum_\lambda s_\lambda \nu_\lambda \mu_\lambda = 1$ that follows from $\mathrm{tr} \sum_\mathbf{x} \eta_\mathbf{x} \rho_\mathbf{x} = 1$ and the expressions of ν_λ and s_λ from Eqs. (B5) and (B6) in Appendix B.

B. Clustering quantum states: Known input states

If the two possible states $|\phi_0\rangle$ and $|\phi_1\rangle$ are known, the optimal clustering protocol must use this information. It is then expected that the average performance will be much higher than for the universal protocol. It is natural in this context not to identify a given string $\mathbf{x}$ with its complementary $\bar{\mathbf{x}}$ (we stick to the notation in the main text), since mistaking one state for the other should clearly count as an error if the two preparations are specified. In this case, then, clustering is equivalent to discriminating the 2^N known pure states $|\Phi_\mathbf{x}\rangle = |\phi_{x_1}\rangle \otimes |\phi_{x_2}\rangle \otimes \cdots \otimes |\phi_{x_N}\rangle$ (hypotheses), where with no loss of generality we can write

$$|\phi_{0/1}\rangle = \sqrt{\frac{1+c}{2}}|0\rangle \pm \sqrt{\frac{1-c}{2}}|1\rangle \tag{28}$$

for a convenient choice of basis. Here, $c = |\langle\phi_0|\phi_1\rangle|$ is the overlap of the two states.

The Gram matrix G encapsulates all the information needed to discriminate the states of the set. It is defined as having elements $G_{\mathbf{x},\mathbf{x}'} = \langle\Phi_{\mathbf{x}}|\Phi_{\mathbf{x}'}\rangle$. It is known that when the diagonal elements of its square root are all equal, i.e., $(\sqrt{G})_{\mathbf{x},\mathbf{x}} \equiv S$ for all $\mathbf{x}$, then the square root measurement is optimal [37,38] and the probability of successful identification reads simply $P_s = S^2$. Notice that we implicitly assume uniformly distributed hypotheses. For the case at hand,

$$\begin{aligned} G_{\mathbf{x},\mathbf{x}'} &= (\langle\phi_{x_1}| \otimes \cdots \otimes \langle\phi_{x_N}|)(|\phi_{x'_1}\rangle \otimes \cdots \otimes |\phi_{x'_N}\rangle) \\ &= \prod_{i=1}^{N} \langle\phi_{x_i}|\phi_{x'_i}\rangle = (\mathcal{G}^{\otimes N})_{\mathbf{x},\mathbf{x}'}, \end{aligned} \tag{29}$$

where

$$\mathcal{G} = \begin{pmatrix} 1 & c \\ c & 1 \end{pmatrix} \tag{30}$$

is the Gram matrix of $\{|\phi_0\rangle, |\phi_1\rangle\}$. Thus, $\sqrt{G} = (\sqrt{\mathcal{G}})^{\otimes N}$, with

$$\sqrt{\mathcal{G}} = \begin{pmatrix} \frac{\sqrt{1+c}+\sqrt{1-c}}{2} & \frac{\sqrt{1+c}-\sqrt{1-c}}{2} \\ \frac{\sqrt{1+c}-\sqrt{1-c}}{2} & \frac{\sqrt{1+c}+\sqrt{1-c}}{2} \end{pmatrix}. \tag{31}$$

As expected, the diagonal terms of $\sqrt{G}$ are all equal, and the success probability is given by

$$P_s(c) = \left(\frac{\sqrt{1+c}+\sqrt{1-c}}{2}\right)^{2N} = \left(\frac{1+\sqrt{1-c^2}}{2}\right)^{N}. \tag{32}$$

We call the reader's attention to the fact that one could attain the very same success probability by performing an individual Helstrom measurement [25], with basis

$$|\psi_{0/1}\rangle = \frac{|0\rangle \pm |1\rangle}{\sqrt{2}}, \tag{33}$$

on each state of the input sequence and guess that the label of that state is the outcome value. In other words, for the problem at hand, global quantum measurements do not provide any improvement over individual fixed measurements.

In order to compare with the results of the main text, we compute the average performance for a uniform distribution of states $|\phi_0\rangle$ and $|\phi_1\rangle$, i.e., the average

$$\begin{aligned} P_s &= \int d\phi_0 d\phi_1 P_s(c) \\ &= \int_0^1 dc^2 P_s(c) \int d\phi_0 d\phi_1 \delta(|\langle\phi_0|\phi_1\rangle|^2 - c^2) \\ &= \int_0^1 dc^2 P_s(c) \int d\phi_1 \delta(|\langle 0|\phi_1\rangle|^2 - c^2) \\ &= \int_0^1 dc^2 \mu(c^2) P_s(c), \end{aligned} \tag{34}$$

where we insert the identity $1 = \int_0^1 dc^2 \delta(a^2 - c^2)$, for $0 < a \equiv |\langle\phi_0|\phi_1\rangle| < 1$, and use the invariance of the measure $d\phi$ under SU(d) transformations. The marginal distribution is $\mu(c^2) = (d-1)(1-c^2)^{d-2}$ (see Appendix E). Using this result, the asymptotic behavior of the last integral is

$$P_s \sim \frac{4(d-1)}{N}. \tag{35}$$

As expected, knowing the two possible states in the input string leads to a better behavior of the success probability: It decreases only linearly in $1/N$, as compared to the best universal quantum clustering protocol, which exhibits a quadratic decrease.

To do a fairer comparison with universal quantum clustering, guessing the complementary string $\bar{\mathbf{x}}$ instead of $\mathbf{x}$ is now counted as a success; that is, now the clusterings are defined by the states

$$\rho_{\mathbf{x}} = \frac{|\Phi_{\mathbf{x}}\rangle\langle\Phi_{\mathbf{x}}| + |\Phi_{\bar{\mathbf{x}}}\rangle\langle\Phi_{\bar{\mathbf{x}}}|}{2}. \tag{36}$$

For this variation of the problem, the optimal measurement is still local and given by a POVM with elements

$$E_{\mathbf{x}} = |\Psi_{\mathbf{x}}\rangle\langle\Psi_{\mathbf{x}}| + |\Psi_{\bar{\mathbf{x}}}\rangle\langle\Psi_{\bar{\mathbf{x}}}|, \tag{37}$$

where $|\Psi_{\mathbf{x}}\rangle = |\psi_{x_1}\rangle \otimes |\psi_{x_2}\rangle \otimes \cdots \otimes |\psi_{x_N}\rangle$ and where we recall that $\{|\psi_0\rangle, |\psi_1\rangle\}$ is the (local) Helstrom measurement basis in Eq. (33). Note that $\{E_{\mathbf{x}}\}$ are orthogonal projectors.

To prove the statement in the last paragraph, we show that the Holevo-Yuen-Kennedy-Lax conditions [Eq. (16)] hold (recall that the Gram matrix technique does not apply to mixed states). For the success probability and assuming equal priors, these conditions take the simpler form

$$\sum_{\mathbf{x}} E_{\mathbf{x}}\rho_{\mathbf{x}} = \sum_{\mathbf{x}} \rho_{\mathbf{x}} E_{\mathbf{x}} \equiv \Gamma, \tag{38}$$

$$\Gamma - \rho_{\mathbf{x}} \geq 0 \quad \forall \ \mathbf{x}, \tag{39}$$

where we drop the irrelevant factor $\eta = 2^{1-N}$. Condition (38) is trivially satisfied. To check that condition (39) also holds, we recall the Weyl inequalities for the eigenvalues of Hermitian $n \times n$ matrices A, B [39]:

$$\vartheta_i(A+B) \leq \vartheta_{i+j}(A) + \vartheta_{n-j}(B), \tag{40}$$

for $j = 0, 1, \ldots, n-i$, where the eigenvalues are labeled in increasing order $\vartheta_1 \leq \vartheta_2 \leq \ldots \leq \vartheta_n$. We use Eq. (40) to write

$$\vartheta_1(\Gamma) \leq \vartheta_3(\Gamma - \rho_{\mathbf{x}}) + \vartheta_{2^N-2}(\rho_{\mathbf{x}}) \tag{41}$$

(note that effectively all these operators act on the 2^N-dimensional subspace spanned by $\{|0\rangle, |1\rangle\}^{\otimes N}$). As proved below, $\Gamma > 0$, which implies that $\vartheta_1(\Gamma) > 0$. We note that $\rho_{\mathbf{x}}$ has rank two; i.e., it has only two strictly positive eigenvalues, so $\vartheta_{2^N-2}(\rho_{\mathbf{x}}) = 0$. Then, Eq. (41) implies

$$\vartheta_3(\Gamma - \rho_{\mathbf{x}}) \geq \vartheta_1(\Gamma) > 0. \tag{42}$$

Finally, notice that $\Gamma - \rho_{\mathbf{x}}$ has two null eigenvalues, with eigenvectors $|\Psi_{\mathbf{x}}\rangle$ and $|\Psi_{\bar{\mathbf{x}}}\rangle$. Hence, $\vartheta_1(\Gamma-\rho_{\mathbf{x}})=\vartheta_2(\Gamma-\rho_{\mathbf{x}})=0$, and it follows from Eq. (42) that condition (39) must hold.

To show the positivity of Γ, which is assumed in the previous paragraph, we use Eqs. (28) and (33) to write

$$\Gamma = \frac{1}{2}\left[\begin{pmatrix} a_1 & 0 \\ 0 & a_2 \end{pmatrix}^{\otimes N} + \begin{pmatrix} b_1 & 0 \\ 0 & b_2 \end{pmatrix}^{\otimes N}\right], \tag{43}$$

where

$$a_{1/2} = \frac{1 \pm c + \sqrt{1-c^2}}{2},$$
$$b_{1/2} = \frac{1 \pm c - \sqrt{1-c^2}}{2}. \tag{44}$$

Notice that $a_1 > b_1$ and $a_2 > |b_2|$. Thus, if $0 \leq c < 1$, we have $\vartheta_k > 0$ for $k = 1, 2, \ldots, 2^N$. The special case $c = 1$ is degenerate. Equation (39) is trivially saturated, rendering $P_s = 2^{1-N}$, as it should be.

The maximum success probability can now be computed by recalling that $P_s(c) = 2^{1-N}\mathrm{tr}\Gamma$. We obtain

$$P_s(c) = \left(\frac{1+\sqrt{1-c^2}}{2}\right)^N + \left(\frac{1-\sqrt{1-c^2}}{2}\right)^N, \tag{45}$$

where the first term corresponds to guessing correctly all the states in the input string, whereas the second one results from guessing the other possible state all along the string. One can easily check that the average over c of the second term vanishes exponentially for large N, and we end up with a success probability given again by Eq. (35).

Finally, we mention that one could consider a simple unambiguous protocol [40–43] whereby each state of the input string is identified with no error with probability $P_s(c) = 1 - c$; i.e., the protocol gives an inconclusive answer with probability $1 - P_s = c$. Therefore, the average unambiguous probability of sorting the data is

$$P_{\mathrm{s}} = 2\int_0^1 dc c\mu(c^2)(1-c)^N \sim \frac{2(d-1)}{N^2}. \tag{46}$$

V. DISCUSSION

Unsupervised learning, which assumes virtually nothing about the distributions underlying the data, is already a hard problem [22,23]. Lifting the notion of classical data to quantum data (i.e., states) factors in additional obstacles, such as the impossibility to repeatedly operate with the quantum data without degrading it. Most prominent classical clustering algorithms heavily rely on the iterative evaluation of a function on the input data (e.g., pairwise distances between points in a feature vector space, as in k-means [44]); hence, they are not equipped to deal with degrading data and expectedly fail in our scenario. The unsupervised quantum classification algorithm we present is thus, by necessity, far away from its classical analogs. In particular, since we are concerned with the optimal quantum strategy, we need to consider the most general collective measurement, which is inherently single shot: It yields a single sample of a stochastic action, namely, a posterior state and an outcome of a quantum measurement, where the latter provides the description of the clustering. The main lesson stemming from our investigation is that, despite these limitations, clustering unknown quantum states is a feasible task. The optimal protocol that solves it showcases some interesting features.

It does not completely erase the information about a given preparation of the input data after clustering.—This result is apparent from Eq. (4), since the action of the POVM on the subspaces (λ) is the identity. After the input data string in the global state $|\Phi_{\mathbf{x}}\rangle$ is measured and outcome λ^* is obtained (recall that λ^* gives us information about the size of the clusters), information relative to the particular states $|\phi_{0/1}\rangle$ remains in the subspace (λ^*) of the global postmeasured state. Therefore, one could potentially use further the posterior (clustered) states down the line as approximations of the two classes of states, which opens the door for our clustering device to be used as an intermediate processor in a quantum network. This notwithstanding, the amount of information that can be retrieved after optimal clustering is currently under investigation.

It outbeats the classical and semiclassical protocols.—If the local dimension of the quantum data is larger than two, the dimensionality of the symmetric subspaces spanned by the global states of the strings of data can be exploited by means of collective measurements with a twofold effect: enhanced distinguishability of states, resulting in improved clustering performance (exemplified by a linear increase in the asymptotic success probability), and information-preserving data handling (to some extent, as discussed above). This effect should be contrasted with the semiclassical protocol, which essentially obliterates the information content of the data (as a von Neumann measurement is performed on each system) and whose success probability vanishes exponentially with the local dimension. In addition, the optimal classical and semiclassical protocols require solving an NP-complete problem, and their

implementation is thus inefficient. In contrast, we observe that the first part of the quantum protocol, which consists in guessing the size of the clusters n, runs efficiently on a quantum computer: This step involves a Schur transform that runs in polynomial time in N and $\log d$ [45,46], followed by a projective measurement with no computational cost. The second part, guessing the permutation σ, requires implementing a group-covariant POVM. The complexity of this step, and, hence, the overall computational complexity of our protocol, is still an open question currently under investigation.

It is optimal for a range of different cost functions.—There are various cost functions that could arguably be better suited to quantum clustering, e.g., the Hamming distance between the guessed and the true clusterings or, likewise, the trace distance or the infidelity between the corresponding effective states $\rho_{n,\sigma}$ and $\rho_{n',\sigma'}$. They are, however, hard to deal with analytically. The question arises as to whether our POVM is still optimal for such cost functions. To answer this question, we formulate an optimality condition that can be checked numerically for problems of a finite size (see Appendix D). Our numerics show that the POVM remains optimal for all these examples. This result is an indication that the optimality of our protocol stems from the structure of the problem, independently of the cost function.

It stands a landmark in multihypothesis state discrimination.—Analytical solutions to multihypothesis state discrimination exist only in a few specific cases [26–28,30,38,47]. Our set of hypotheses arises arguably from the minimal set of assumptions about a pure state source: It produces two states randomly. Variants of this problem with much more restrictive assumptions are considered in Refs. [11,48,49].

Our clustering protocol departs from other notions of quantum unsupervised machine learning that can be found in the literature [50–53]. In these references, data coming from a classical problem are encoded in quantum states that are available on demand via a quantum random access memory [54]. The goal is to surpass classical performance in the number of required operations. In contrast, we deal with unprocessed quantum data as input and aim at performing a task that is genuinely quantum. This scenario is notably harder, where known heuristics for classical algorithms simply cannot work.

Other extensions of this work currently under investigation are clustering systems whose states can be of more than two types, where we expect a similar two-step measurement for the optimal protocol, and clustering of quantum processes, where the aim is to classify instances of unknown processes by letting them run on some input test state of our choice (see Ref. [11] for related work on identifying malfunctioning devices). In this last case, an interesting application arises when considering causal relations as the defining feature of a cluster. A clustering algorithm then aims to identify, within a set of unknown processes, which ones are causally connected. Identifying causal structures has recently attracted attention among the quantum information community [55].

ACKNOWLEDGMENTS

We acknowledge the financial support of the Spanish MINECO, Ref. FIS2016-80681-P (AEI/FEDER, UE), and Generalitat de Catalunya CIRIT, Ref. 2017-SGR-1127. G. S. thanks the support of the Alexander von Humboldt Foundation. E. B. also thanks the hospitality of Computer Science Department of the University of Hong Kong during his stay.

APPENDIX A: PARTITIONS

Partitions play an important role in the representation theory of groups and are central objects in combinatorics. Here, we collect a few definitions and results that are used in the next appendixes, particularly in Appendix B.

A *partition* $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_r, \ldots)$ is a sequence of nonnegative integers in nonincreasing order. The *length* of λ, denoted $l(\lambda)$, is the number of nonzero elements in λ. We denote by $\lambda \vdash N$ a partition λ of the integer N, where $N = \sum_i \lambda_i$. A natural way of ordering partitions is by inverse lexicographic order; i.e., given two partitions λ and λ', we write $\lambda > \lambda'$ iff the first nonzero difference $\lambda_i - \lambda'_i$ is positive.

The total number of partitions of an integer N is denoted by P_N [56], and the number of partitions such that $l(\lambda) \leq r$ by $P_N^{(\leq r)}$. Similarly, the number of partitions of length r is denoted by $P_N^{(r)}$. There exists no closed expression for any of these numbers, but there are widely known results (some of them by Hardy and Ramanujan are very famous [57]) concerning their asymptotic behavior for large N. The one we later use in Appendix F is

$$P_N^{(\leq r)} \sim \frac{N^{r-1}}{r!(r-1)!}, \tag{A1}$$

which gives the dominant contribution for large N. Note that, from the obvious relation $P_N^{(r)} = P_N^{(\leq r)} - P_N^{(\leq r-1)}$, it follows that the same asymptotic expression holds for $P_N^{(r)}$.

Partitions are conveniently represented by *Young diagrams*. The Young diagram associated to the partition $\lambda \vdash N$ is an arrangement of N empty boxes in $l(\lambda)$ rows, with λ_i boxes in the ith row. This association is one to one; hence, λ can be used to label Young diagrams as well. A *Young tableau* of d entries is a Young diagram filled with integers from 1 up to d, one in each box. There are two types of tableaux: A *standard Young tableau* (SYT) of shape $\lambda \vdash N$ is one where $d = N$ and such that the integers in each row increase from left to right and from top to bottom in each column (hence, each integer appears exactly once).

A *semistandard Young tableau* (SSYT) of shape $\lambda \vdash N$ and d entries, $d \geq l(\lambda)$, is one such that integers in each row are nondecreasing from left to right and increasing from top to bottom in each column.

The number of different SYTs of shape $\lambda \vdash N$ is given by the *hook-length* formula

$$\nu_\lambda = \frac{N!}{\prod_{(i,j)\in\lambda} h_{ij}}, \tag{A2}$$

where (i, j) denotes the box located in the ith row and the jth column of the Young diagram and h_{ij} is the hook length of the box (i, j), defined as the number of boxes located beneath or to the right of that box in the Young diagram, counting the box itself. Likewise, the number of SSYTs of shape $\lambda \vdash N$ and d entries is given by the formula

$$s_\lambda = \frac{\Delta(\lambda_1 + d - 1, \lambda_2 + d - 2, \ldots, \lambda_d)}{\Delta(d-1, d-2, \ldots, 0)}, \tag{A3}$$

where $\Delta(x_1, x_2, \ldots, x_d) = \prod_{i<j}(x_i - x_j)$.

APPENDIX B: IRREDUCIBLE REPRESENTATIONS OF SU(d) AND S_N OVER $(d,\mathbb{C})^{\otimes N}$

For the sake of convenience, we recall here some ingredients of representation theory that we use throughout the paper. The results described below can be found in standard textbooks, for instance, in Refs. [58,59].

1. Some results in representation theory

Young diagrams or, equivalently, partitions λ label the irreps of the general linear group $\mathrm{GL}(d)$ and some of its subgroups, e.g., $\mathrm{SU}(d)$, and also the irreps of the symmetric group S_N. The dimension of these irreps are given by s_λ and ν_λ, respectively [Eqs. (A2) and (A3)].

Schur-Weyl duality [59] establishes a connection between irreps of both groups, as follows. Let us consider the transformations $R^{\otimes N}$ and U_σ on the N-fold tensor product space $(d, \mathbb{C})^{\otimes N}$, where $R \in \mathrm{SU}(d)$ and U_σ permutes the N spaces $(d, \mathbb{C})$ of the tensor product according to the permutation $\sigma \in S_N$. Both $R^{\otimes N}$ and U_σ define, respectively, a reducible unitary representation of the groups $\mathrm{SU}(d)$ and S_N on $(d, \mathbb{C})^{\otimes N}$. Moreover, they are each other's commutants. It follows that this reducible representation decomposes into irreps λ, so that their joint action can be expressed as

$$R^{\otimes N} U_\sigma = U_\sigma R^{\otimes N} = \bigoplus_{\lambda \vdash N} R^\lambda \otimes U^\lambda_\sigma, \tag{B1}$$

where R^λ and U^λ_σ are the matrices that represent R and U_σ, respectively, on the irrep λ. To resolve any ambiguity that may arise, we write λ in parentheses, (λ), when it refers to the irreps of $\mathrm{SU}(d)$ or in braces, $\{\lambda\}$, when it refers to those of S_N. Equation (B1) tells us that the dimension of (λ), s_λ, coincides with the *multiplicity* of $\{\lambda\}$, and, conversely, the dimension of $\{\lambda\}$, ν_λ, coincides with the multiplicity of (λ).

This block-diagonal structure provides a decomposition of Hilbert space $\mathcal{H}^{\otimes N} = (d, \mathbb{C})^{\otimes N}$ into subspaces that are invariant under the action of $\mathrm{SU}(d)$ and S_N, as $\mathcal{H}^{\otimes N} = \bigoplus_\lambda H_\lambda$, and, in turn, $H_\lambda = H_{(\lambda)} \otimes H_{\{\lambda\}}$. The basis in which $\mathcal{H}^{\otimes N}$ has this form is known as the *Schur basis*, and the unitary transformation that changes from the computational to the Schur basis is called the *Schur transform*.

To conclude this Appendix, let us recall the rules for reducing the tensor product of two $\mathrm{SU}(d)$ representations as a Clebsch-Gordan series of the form

$$R^\lambda \otimes R^{\lambda'} = \bigoplus_{\lambda''} R^{\lambda''} \otimes \mathbb{1}^{\lambda''} \quad \forall\ R \in \mathrm{SU}(d), \tag{B2}$$

where $\dim(\mathbb{1}^{\lambda''})$ is the multiplicity of irrep λ''. The same rules also apply to the reduction of the outer product of representations of S_n and $S_{n'}$ into irreps of $S_{n''}$, where $n'' = n + n'$. In this case, one has

$$(U^\lambda \otimes U^{\lambda'})_\sigma = \bigoplus_{\lambda''} U^{\lambda''}_\sigma \otimes \mathbb{1}^{\lambda''} \quad \forall\ \sigma \in S_{n''}. \tag{B3}$$

Note the different meanings of $\otimes$ in the last two equations (it is, however, standard notation). The rules are most easily stated in terms of the Young diagrams that label the irreps. They are as follows.

(1) In one of the diagrams that label the irreps on the left-hand side of Eq. (B2) or (B3) (preferably the smallest), write the symbol a in all boxes of the first row, the symbol b in all boxes of the second row, c in all boxes of the third one, and so on.

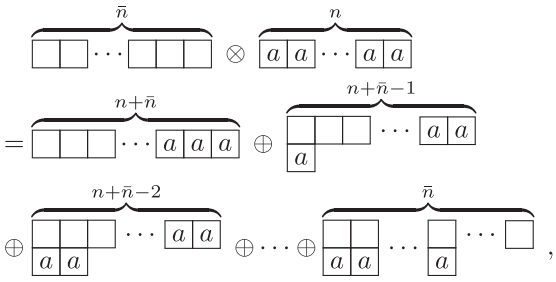
(2) Attach boxes with a to the second Young diagram in all possible ways subjected to the rules that no two a's appear in the same column and that the resulting arrangement of boxes is still a Young diagram. Repeat this process with b's, c's, and so on.

(3) For each Young diagram obtained in step two, read the first row of added symbols from right to left, then the second row in the same order, and so on. The resulting sequence of symbols, e.g., $abaabc\ldots$, must be a lattice permutation; namely, to the left of any point in the sequence, there are not fewer a's than b's, no fewer b's than c's, and so on. Discard all diagrams that do not comply with this rule.

The Young diagrams λ'' that result from this procedure specify the irreps on the right-hand side of Eqs. (B2) and (B3). The same diagram can appear a number M of times, in which case λ'' has multiplicity $\dim(\mathbb{1}^{\lambda''}) = M$.

2. Particularities of quantum clustering

Since the density operators [cf. Eq. (11)] and POVM elements [cf. Eq. (13)] associated to each possible clustering emerge from the joint action of a permutation $\sigma \in S_N$ and a group average over SU(d), it is most convenient to work in the Schur basis, where the mathematical structure is much simpler. A further simplification specific to quantum clustering of two types of states is that the irreps that appear in the block-diagonal decomposition of the states (and, hence, of the POVM elements) have at most length 2; i.e., they are labeled by bipartitions $\lambda = (\lambda_1, \lambda_2)$ and correspond to Young diagrams of at most two rows. We have this simplification because the $\rho_{n,\sigma}$ arise from the tensor product of two *completely symmetric* projectors, $\mathbb{1}_n^{\mathrm{sym}}$ and $\mathbb{1}_{\bar{n}}^{\mathrm{sym}}$, of n and $\bar{n}$ systems [cf. Eq. (11)]. They project into the irrep $\lambda = (n, 0)$ and $\lambda' = (\bar{n}, 0)$ subspaces, respectively. According to the reduction rules above, in the Schur basis the tensor product reduces as



(B4)

which proves our statement.

There is yet another simplification that emerges from Eq. (B4). Note that all the irreps appear only once in the reduction. That is, fixing the indices n, σ, and $\{\lambda\}$ uniquely defines a one-dimensional subspace. Thus, the projectors $\Omega^{n,\sigma}_{\{\lambda\}}$ are rank one.

We conclude by giving explicit expressions for the dimensions of the irreps of S_N and SU(d), in Eqs. (A2) and (A3), for partitions of the form $\lambda = (\lambda_1, \lambda_2)$. These expressions are used to derive Eq. (27) and read

$$\nu_\lambda = \frac{N!(\lambda_1 - \lambda_2 + 1)}{(\lambda_1 + 1)!\lambda_2!}, \tag{B5}$$

$$s_\lambda = \frac{\lambda_1 - \lambda_2 + 1}{\lambda_1 + 1}\binom{\lambda_1 + d - 1}{d - 1}\binom{\lambda_2 + d - 2}{d - 2}. \tag{B6}$$

One can check that Eqs. (B5) and (B6) are consistent with Eq. (B4) by showing that the sum of the dimensions of the irreps on the right-hand side agrees with the product of the two on the left-hand side, namely, by checking that

$$s_{(\bar{n},0)} s_{(n,0)} = \sum_{i=0}^{n} s_{(n+\bar{n}-i,i)}, \tag{B7}$$

$$\nu^{S_{\bar{n}}}_{(\bar{n},0)} \nu^{S_n}_{(n,0)} \binom{n+\bar{n}}{n} = \sum_{i=0}^{n} \nu_{(n+\bar{n}-i,i)}, \tag{B8}$$

where the superscripts remind us that the dimensions on the left-hand side refer to irreps of $S_{\bar{n}}$, S_n. One obviously obtains $\nu^{S_{\bar{n}}}_{(\bar{n},0)} = \nu^{S_n}_{(n,0)} = 1$, since these are the trivial representations of either group. The binomial in Eq. (B8) arises from the definition of outer product representation in Eq. (B3), whereby the action of $S_{n+\bar{n}}$ is defined on basis vectors of the form $\bar{v}_{i_1 i_2 \ldots i_{\bar{n}}} \otimes v_{i_{\bar{n}+1} i_{\bar{n}+2} \ldots i_{\bar{n}+n}}$, with $\bar{v}_{i_1 i_2 \ldots i_{\bar{n}}} \in H^{S_{\bar{n}}}_{\{\lambda\}}$, $v_{i_1 i_2 \ldots i_n} \in H^{S_n}_{\{\lambda'\}}$. There are, naturally, $\binom{\bar{n}+n}{n}$ ways of allocating $\bar{n} + n$ indices in this expression.

APPENDIX C: ASYMPTOTICS OF P_s

We next wish to address the asymptotic behavior of the success probability as the length N of the data string becomes large. Various behaviors are derived, depending on how the local dimension d scales with N.

In the large N limit, it suffices to consider even values of N, which slightly simplifies the derivation of the asymptotic expressions. The success probability in Eq. (27) for $N = 2m$, $m \in \mathbb{N}$, can be written as (just define a new index as $j = m - i$)

$$P_s = \frac{d-1}{2^{2m-1}} \sum_{j=0}^{m} \frac{(2j+1)^2}{(m+1+j)^2(m+d-1-j)} \binom{2m}{m+j}. \tag{C1}$$

For large m, we write $j = mx$ and use

$$\frac{1}{2^{2m-1}} \binom{2m}{m+j} \sim \frac{2e^{-mx^2}}{\sqrt{m\pi}}. \tag{C2}$$

We start by assuming that d scales more slowly than N, e.g., $d \sim N^\gamma$, with $0 \le \gamma < 1$. In this situation, we can neglect d in the denominator of Eq. (C1). Neglecting also other subleading terms in inverse powers of m and using the Euler-Maclaurin formula, we have

$$P_s \sim (d-1) \int_0^\infty dx \frac{4x^2}{(1+x)^2(1-x)} \frac{2e^{-mx^2}}{\sqrt{m\pi}}, \tag{C3}$$

which we can further approximate by substituting 0 for x in the denominator, as the Gaussian factor peaks at $x = 0$ as m becomes larger, so

$$\begin{aligned} P_s &\sim 4(d-1) \int_0^\infty dx x \frac{2xe^{-mx^2}}{\sqrt{m\pi}} \\ &= -4(d-1) \int_0^\infty dx x \frac{d}{dx} \frac{e^{-mx^2}}{m\sqrt{m\pi}}. \end{aligned} \tag{C4}$$

We integrate by parts to obtain

$$P_s \sim \frac{2(d-1)}{m}\int_0^\infty dx \frac{2e^{-mx^2}}{\sqrt{m\pi}} = \frac{2(d-1)}{m^2}. \quad \text{(C5)}$$

Hence, provided that d scales more slowly than N, the probability of success vanishes asymptotically as N^{-2}, or, more precisely, as

$$P_s \sim \frac{8(d-1)}{N^2}. \quad \text{(C6)}$$

Let us next assume that d scales faster than N, e.g., as $d \sim N^\gamma$, with $\gamma > 1$. In this case, d is the leading contribution in the second factor in the denominator of Eq. (C1). Accordingly, we have

$$\begin{aligned} P_s &\sim (d-1)m\int_0^\infty dx \frac{4x^2}{(1+x)^2 d}\frac{2e^{-mx^2}}{\sqrt{m\pi}} \\ &\sim 4m\int_0^\infty dx x\frac{2xe^{-mx^2}}{\sqrt{m\pi}} = \frac{2}{m}, \end{aligned} \quad \text{(C7)}$$

and the asymptotic expression becomes

$$P_s \sim \frac{4}{N}, \quad \text{(C8)}$$

independently of d.

Finally, let us assume that d scales exactly as N and write $d = sN$, $s > 0$. The success probability can be cast as

$$P_s \sim (d-1)\int_0^\infty dx \frac{4x^2}{(1+x)^2(1+2s-x)}\frac{2e^{-mx^2}}{\sqrt{m\pi}}. \quad \text{(C9)}$$

Proceeding as above, we obtain

$$P_s \sim \frac{2(d-1)}{(2s+1)m^2}. \quad \text{(C10)}$$

Thus,

$$P_s \sim \frac{8s}{(2s+1)N}. \quad \text{(C11)}$$

The three expressions, Eqs. (C6), (C8), and (C11), can be combined into a single one as

$$P_s \sim \frac{8(d-1)}{(2d+N)N}. \quad \text{(C12)}$$

APPENDIX D: OPTIMAL POVM FOR GENERAL COST FUNCTIONS

This Appendix deals with the optimization of quantum clustering assuming other cost functions. We introduce a sufficient condition under which the type of POVM we use to maximize the success probability (Sec. IV A) is also optimal for a given generic cost function. We conjecture that the condition holds under reasonable assumptions. We discuss numerical results for the cases of Hamming distance, trace distance, and infidelity.

Recall that Eq. (13) together with Eq. (22) defines the optimal POVM for a generic cost function that preserves covariance under S_N. However, this form is implicit and, thus, not very practical. Particularizing to the success probability, we manage to specify the function $n(\lambda) = \lambda_2$ [cf. Eq. (26)] and the operators $\Xi^n_{\{\lambda\}} = \Omega^n_{\{\lambda\}}\delta_{n,\lambda_2}$. In summary, the POVM is specified solely in terms of the effective states $\rho_{n,\sigma}$ (hypotheses).

Here, we conjecture that the choice $\Xi^n_{\{\lambda\}} = \Omega^n_{\{\lambda\}}\delta_{n,n(\lambda)}$ is still optimal for a large class of cost functions $f(\mathbf{x},\mathbf{x}')$, albeit with varying guessing rules $n(\lambda)$. If the conjecture holds, given $f(\mathbf{x},\mathbf{x}')$, one has only to compute $n(\lambda) = \mathrm{argmin}_n \vartheta^n_{\lambda,1}$ to obtain the optimal POVM. The minimum average cost can then be computed via Eq. (24). We now formulate this conjecture precisely as a testable mathematical condition.

For any cost function (distance) such that $f(\mathbf{x},\mathbf{x}') \geq 0$ and $f(\mathbf{x},\mathbf{x}') = 0$ iff $\mathbf{x} = \mathbf{x}'$, we can always find some constant $t > 0$ such that

$$tf(\mathbf{x},\mathbf{x}') \geq \bar{\delta}_{\mathbf{x},\mathbf{x}'} \equiv 1 - \delta_{\mathbf{x},\mathbf{x}'} \quad \forall\ \mathbf{x},\mathbf{x}'. \quad \text{(D1)}$$

We can then rescale the cost function $f \mapsto t^{-1}f$ and assume with no loss of generality that $f(\mathbf{x},\mathbf{x}') \geq \bar{\delta}_{\mathbf{x},\mathbf{x}'}$. We have

$$W_{\mathbf{x}} = \bar{W}_{\mathbf{x}} + \Delta_{\mathbf{x}}, \quad \text{(D2)}$$

where we use the definition of $W_{\mathbf{x}}$ after Eq. (16) and similarly define $\bar{W}_{\mathbf{x}}$ for the minimal cost $\bar{\delta}_{\mathbf{x},\mathbf{x}'}$. As in Sec. IV A, it suffices to consider $\mathbf{x} = (n,e)$. Then,

$$\Delta_{\mathbf{x}} = \sum_{\mathbf{x}'} \eta_{\mathbf{x}'}[f(\mathbf{x},\mathbf{x}') - \bar{\delta}_{\mathbf{x},\mathbf{x}'}]\rho_{\mathbf{x}'} \geq 0. \quad \text{(D3)}$$

Using the same notation as in Eq. (18), Eq. (D3) is equivalent to

$$\omega^n_{\{\lambda\}} - \bar{\omega}^n_{\{\lambda\}} \geq 0. \quad \text{(D4)}$$

We now recall the meaning of Eqs. (20) and (21): The operators $\Xi^n_{\{\lambda\}}$ must be projectors onto the eigenspace of the minimal eigenvalue of $\omega^n_{\{\lambda\}}$. Then, according to Eq. (22), the choice $\Xi^n_{\{\lambda\}} = \Omega^n_{\{\lambda\}}\delta_{n,n(\lambda)}$ is also optimal for arbitrary cost functions if it holds that

$$\mathrm{supp}(\Omega^n_{\{\lambda\}}) = V_1(\bar{\omega}^n_{\{\lambda\}}) \overset{?}{\subset} V_1(\omega^n_{\{\lambda\}}), \quad \text{(D5)}$$

where $V_1(X)$ is the eigenspace of the minimal eigenvalue of X and the equality follows from Eq. (25).

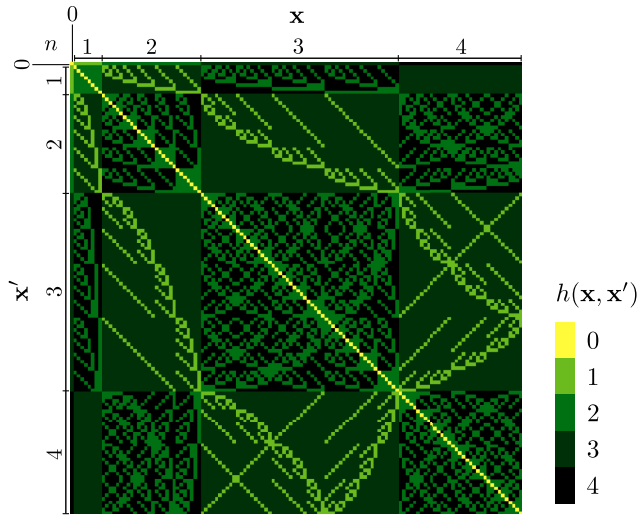


FIG. 3. Heat map of the Hamming distances $h(\mathbf{x},\mathbf{x}')$ between clusterings for $N=8$. The clusterings are grouped by the size of the smallest cluster $n=0,\ 1,\ 2,\ 3,\ 4$. Each group contains all nontrivial permutations σ for a given n. A brighter color means a smaller Hamming distance.

Our conjecture is that Eq. (D5) holds true for the class of "reasonable" cost functions considered in this paper, namely, for those that are non-negative and covariant and satisfy the distance property stated before Eq. (D1). We check its validity for problems of size up to $N=8$, local dimension $d=2$, and uniform prior probabilities for the following cost functions: Hamming distance $h(\mathbf{x},\mathbf{x}')=\min\{|\mathbf{x}-\mathbf{x}'|,|\mathbf{x}-\bar{\mathbf{x}}'|\}$ $(x_i=0,\ 1)$, trace distance $T(\mathbf{x},\mathbf{x}')=||\rho_{\mathbf{x}}-\rho_{\mathbf{x}'}||_1$, and infidelity $I(\mathbf{x},\mathbf{x}')=1-\mathrm{tr}^2[(\sqrt{\rho_{\mathbf{x}}}\rho_{\mathbf{x}'}\sqrt{\rho_{\mathbf{x}}})^{1/2}]$.

The above examples induce a much richer structure in the problem at hand. To illustrate this added complexity, in Fig. 3 we show a heat map of the Hamming distances $h(\mathbf{x},\mathbf{x}')$ between all pairs of clusterings for $N=8$. The figure shows that the largest values of $h(\mathbf{x},\mathbf{x}')$ can occur for two clusterings with equal cluster size n and that $h(\mathbf{x},\mathbf{x}')$ is extremely dependent on the pair of permutations σ,σ'. As a result, the guessing rule $n(\lambda)$ is completely different from the one that maximizes the probability of success P_s. In particular, irreps λ are no longer in one-to-one correspondence with optimal guesses for n. In Table I, we show values of $n(\lambda)$ for our four cost functions and $N=4,\ldots,8$. In contrast to the case of the success probability (the cost function $\bar{\delta}_{\mathbf{x},\mathbf{x}'}$), we note that in some cases it is actually optimal to map several irreps to the same guess while never guessing certain cluster sizes.

Performing the Schur transform is computationally inefficient on a classical computer (In contrast, as mentioned in the main text, there exist efficient quantum circuits able to implement the Schur transform in a quantum computer. A circuit based on the Clebsch-Gordan transform achieves polynomial time in N and d [45]. Recently, an alternative method based on the representation theory of the symmetric group was shown to reduce the dimension scaling to $\mathrm{poly}(\log d)$ [46].), which sets a limit on the size of the data one can test—in our case it is $N=8$. However, it is worth mentioning that this difficulty might actually be overcome. The fundamental objects needed for testing Eq. (D5) are the operators $\Omega^n_{\{\lambda\}}$. Their computation does, in principle, not require the full Schur transform, as they can be expressed in terms of generalized Racah coefficients, which give a direct relation between Schur bases arising from different coupling schemes of the tensor product space. It is indeed possible to calculate generalized Racah coefficients directly without going through a Clebsch-Gordan transform [60], and, should this method be implemented, clustering problems of larger sizes might be tested. However, an extensive numerical analysis is not the aim of this paper.

APPENDIX E: PRIOR DISTRIBUTIONS

In the interest of making this paper self-contained, in this Appendix we include the derivation of some results about the prior distributions used in the paper.

Let $\mathsf{S}_d=\{p_s\geq 0|\sum_{s=1}^d p_s=1\}$ denote the standard $(d-1)$-dimensional (probability) simplex. Every categorical distribution (CD) $P=\{p_s\}_{s=1}^d$ is a point in S_d. The flat distribution of CDs is the volume element divided by the volume of S_d, the latter denoted by V_d. Choosing coordinates $p_1,\ldots,p_{d-1}$, the flat distribution is $\prod_{s=1}^{d-1}dp_s/V_d\equiv dP$.

Let us compute the moments of the flat distribution; as a by-product, we obtain V_d. We have

TABLE I. Values of $n(\lambda)$, i.e., of the optimal guess for the size of the smallest cluster, where $\lambda=(\lambda_1,\lambda_2)$ are the relevant irreps, for data sizes $N=4,5,6,7,8$, and cost functions $\bar{\delta}(\mathbf{x},\mathbf{x}')$ (corresponding to the success probability), Hamming distance $h(\mathbf{x},\mathbf{x}')$, trace distance $T(\mathbf{x},\mathbf{x}')$, and infidelity $I(\mathbf{x},\mathbf{x}')$.

	$N=4$			$N=5$			$N=6$				$N=7$				$N=8$				
λ	(4,0)	(3,1)	(2,2)	(5,0)	(4,1)	(3,2)	(6,0)	(5,1)	(4,2)	(3,3)	(7,0)	(6,1)	(5,2)	(4,3)	(8,0)	(7,1)	(6,2)	(5,3)	(4,4)
$\bar{\delta}$	0	1	2	0	1	2	0	1	2	3	0	1	2	3	0	1	2	3	4
h	0	2	2	0	2	2	0	3	2,3	3	0	3	3	3	0	4	4	4	4
T	1	1	2	1	1	2	1	1	2	3	1	2	2	3	1	2	3	3	4
I	0	1	2	0	1	2	0	1	2	3	0	1	3	3	0	1	3	4	4

$$V_d \int_{\mathsf{S}_d} dP \prod_{s=1}^{d} p_s^{n_s} = \int_0^1 dp_1 \int_0^{1-p_1} dp_2 \ldots \times \int_0^{1-\sum_{s=1}^{d-2} p_s} dp_{d-1} \prod_{s=1}^{d} p_s^{n_s} = \frac{\prod_{s=1}^{d} n_s!}{(d-1+\sum_{s=1}^{d} n_s)!} \tag{E1}$$

[the calculation becomes straightforward by iterating the change of variables $p_r \mapsto x$, where $p_r = (1-\sum_{s=1}^{r-1} p_s)x$, $r = d-2, d-3, \ldots, 2, 1$]. In particular, setting $n_s = 0$ for all s in Eq. (E1), we obtain $V_d = 1/(d-1)!$. Then,

$$\int_{\mathsf{S}_d} dP \prod_{s=1}^{d} p_s^{n_s} = \frac{(d-1)! \prod_{s=1}^{d} n_s!}{(d-1+N)!}, \tag{E2}$$

where $N = \sum_{s=1}^{d} n_s$.

Next, we provide a simple proof that any fixed von Neumann measurement on a uniform distribution of pure states in $(d, \mathbb{C})$ gives rise to CDs whose probability distribution is flat. As a result, the classical and semiclassical strategies discussed in the main text have the same success probability.

Take $|\phi\rangle \in (d, \mathbb{C})$ and let $\{|s\rangle\}_{s=1}^{d}$ be an orthonormal basis of $(d, \mathbb{C})$. By performing the corresponding von Neumann measurement, the probability of an outcome s is $p_s = |\langle s|\phi\rangle|^2$. Thus, any distribution of pure states induces a distribution of CDs $\{p_s = |\langle s|\phi\rangle|^2\}_{s=1}^{d}$ on S_d. Let us compute the moments of the induced distribution, namely,

$$\int d\phi \prod_{s=1}^{d} p_s^{n_s} = \int d\phi \, \mathrm{tr}[\bigotimes_{s=1}^{d} (|s\rangle\langle s|)^{\otimes n_s} (|\phi\rangle\langle\phi|)^{\otimes N}] = \frac{1}{D_N^{\mathrm{sym}}} \mathrm{tr}[\bigotimes_{s=1}^{d} (|s\rangle\langle s|)^{\otimes n_s} \mathbb{1}_N^{\mathrm{sym}}], \tag{E3}$$

where we recall that D_N^{sym} ($\mathbb{1}_N^{\mathrm{sym}}$) is the dimension of (projector on) the symmetric subspace of $(d, \mathbb{C})^{\otimes N}$ and we have used the Schur lemma. A basis of the symmetric subspace is

$$|v_{\mathbf{n}}\rangle = \sqrt{\frac{\prod_{s=1}^{d} n_s!}{N!}} \sum_{\sigma \in S_N} U_\sigma \bigotimes_{s=1}^{d} |s\rangle^{\otimes n_s}, \tag{E4}$$

where $\mathbf{n} = (n_1, n_2, \ldots, n_d)$. Note that there are $\binom{N+d-1}{d-1}$ different strings $\mathbf{n}$ (weak compositions of N in d parts), which agrees with $D_N^{\mathrm{sym}} = s_{(N,0)}$ [recall Eq. (B6)], as it should be. Since $\mathbb{1}_N^{\mathrm{sym}} = \sum_{\mathbf{n}} |v_{\mathbf{n}}\rangle\langle v_{\mathbf{n}}|$, we can easily compute the trace in Eq. (E3) to obtain

$$\int d\phi \prod_{s=1}^{d} p_s^{n_s} = \frac{\prod_{s=1}^{d} n_s!}{N! D_N^{\mathrm{sym}}} = \frac{(d-1)! \prod_{s=1}^{d} n_s!}{(N+d-1)!}. \tag{E5}$$

This equation agrees with Eq. (E2), which means that all the moments of the distribution induced from the uniform distribution of pure states coincide with the moments of a flat distribution of CDs on S_d. Since the moments uniquely determine the distributions with compact support [61] (and S_d is compact), we conclude that they are identical.

As a by-product, we can compute the marginal distribution $\mu(c^2)$, where c is the overlap of $|\phi\rangle$ with a fixed state $|\psi\rangle$. Since we can always find a basis such that $|\psi\rangle$ is its first element, we have $c = |\langle 1|\phi\rangle|$. Because of the results above, the marginal distribution is given by

$$\mu(c^2) = \int_0^{1-p_1} dp_2 \ldots \int_0^{1-\sum_{s=1}^{d-2} p_s} dp_{d-1} \Big|_{p_1 = c^2} = (d-1)(1-c^2)^{d-2}, \tag{E6}$$

in agreement with Ref. [62].

APPENDIX F: OPTIMAL CLUSTERING PROTOCOL FOR UNKNOWN CLASSICAL STATES

In this Appendix, we provide details on the derivation of the optimal protocol for a classical clustering problem, analog to the quantum problem discussed in the main text. The results here also apply to quantum systems when the measurement performed on each of them is restricted to be local, projective, d dimensional, and fixed. We call this type of protocol semiclassical.

Here, we envision a device that takes input strings of N data points $\mathbf{s} = (s_1 s_2 \ldots s_N)$, with the promise that each s_i is a symbol out of an alphabet of d symbols, say, the set $\{1, 2, \ldots, d\}$, and is drawn from either roulette P or from roulette Q, with corresponding categorical probability distributions $P = \{p_s\}_{s=1}^{d}$ and $Q = \{q_s\}_{s=1}^{d}$. To simplify the notation, we use the same symbols for the roulettes and their corresponding probability distributions and for the stochastic variables and their possible outcomes. Also, the range of values of the index s is always understood to be $\{1, 2, \ldots, d\}$, unless specified otherwise. The device's task is to group the data points in two clusters so that all points in either cluster have a common underlying probability distribution (either P or Q). We wish the machine to be universal, meaning that it operates without knowledge on the distributions P and Q. Accordingly, we choose as the figure of merit the probability of correctly classifying *all* data points, averaged over every possible sequence of roulettes $\mathbf{x} = (x_1 x_2 \ldots x_N)$, $x_i \in \{P, Q\}$, and over every possible distribution P and Q. The latter are assumed to be uniformly distributed over the common probability simplex S_d on which they are defined. Formally, this success probability is

$$P_s^{\mathrm{cl}} = \int_{\mathsf{S}_d} dPdQ \sum_{\mathbf{x},\mathbf{s}} \Pr\left(\hat{\mathbf{x}} \in \{\mathbf{x}, \bar{\mathbf{x}}\}, \mathbf{s}, \mathbf{x}; P, Q\right)$$
$$= 2 \int_{\mathsf{S}_d} dPdQ \sum_{\mathbf{x},\mathbf{s}} \delta_{\hat{\mathbf{x}},\mathbf{x}} \Pr\left(\mathbf{s}, \mathbf{x}; P, Q\right), \tag{F1}$$

where $\hat{\mathbf{x}}$ is the guess of $\mathbf{x}$ emitted by the machine, which by the universality requirement, can depend *only* on the data string $\mathbf{s}$. The sums are carried out over all 2^N possible strings $\mathbf{s}$ and sequences of roulettes $\mathbf{x}$. The factor of 2 in the second equality takes into account that P and Q are unknown; hence, identifying the complementary string $\bar{\mathbf{x}}$ leads to the same clustering. By emitting $\hat{\mathbf{x}}$, the device suggests a classification of the N data points s_i in two clusters. In the above equation, we use the notation of Appendix E for the integral over the probability simplex.

An expression for the optimal success probability can be obtained from the trivial upper bound

$$P_s^{\mathrm{cl}} = 2 \sum_{\mathbf{s}} \int dPdQ \Pr\left(\mathbf{s}, \hat{\mathbf{x}}; P, Q\right)$$
$$\leq 2 \sum_{\mathbf{s}} \max_{\mathbf{x}} \int dPdQ \Pr\left(\mathbf{s}, \mathbf{x}; P, Q\right)$$
$$= 2 \sum_{\mathbf{s}} \max_{\mathbf{x}} \Pr\left(\mathbf{s}, \mathbf{x}\right), \tag{F2}$$

where $\Pr(\mathbf{s}, \mathbf{x})$ is the joint marginal distribution of $\mathbf{s}$ and $\mathbf{x}$. This bound is attained by the guessing rule

$$\hat{\mathbf{x}} = \operatorname*{argmax}_{\mathbf{x}} \Pr\left(\mathbf{s}, \mathbf{x}\right). \tag{F3}$$

For two specific distributions P and Q, the probability that a given roulette sequence $\mathbf{x}$ gives rise to a particular data string $\mathbf{s}$ is $\Pr(\mathbf{s}|\mathbf{x}; P, Q) = \prod_s p_s^{n_s} q_s^{m_s}$, where n_s (m_s) is the number of occurrences of symbol s in $\mathbf{s}$ [i.e., how many $s_i \in \mathbf{s}$ satisfy $s_i = s$] arising from roulettes of type P (Q). For later convenience, we define $M_s = n_s + m_s$, which gives the total number of such occurrences. Note that $\{M_s\}$ is independent of $\mathbf{x}$, whereas $\{n_s\}$ and $\{m_s\}$ are not. Performing the integral over P and Q, we have

$$\Pr(\mathbf{s}, \mathbf{x}) = \frac{\Pr(\mathbf{s}|\mathbf{x})}{2^N}$$
$$= \frac{1}{2^N} \int dPdQ \Pr(\mathbf{s}|\mathbf{x}; P, Q)$$
$$= \frac{2^{-N} d_\flat!^2 \prod_s n_s! m_s!}{(d_\flat + \sum_s m_s)!(d_\flat + \sum_s n_s)!}, \tag{F4}$$

where we use Eq. (E2) and in the first equality we assume that the two types of roulette P and Q are equally probable; hence, each possible sequence $\mathbf{x}$ occurs with equal prior probability equal to 2^{-N}. We also introduce the notation $d_\flat \equiv d - 1$ to shorten the expressions throughout this Appendix. Note that all the dependence on $\mathbf{x}$ is through the occurrence numbers m_s and n_s.

According to Eq. (F2), for each string $\mathbf{s}$ we need to maximize the joint probability $\Pr(\mathbf{s}, \mathbf{x})$ in Eq. (F4) over all possible sequences of roulettes $\mathbf{x}$. We first note that, given a total of M_s occurrences of a symbol s in $\mathbf{s}$, $\Pr(\mathbf{s}, \mathbf{x})$ is maximized by a sequence $\mathbf{x}$ whereby all these occurrences come from the same type of roulette, in other words, by a sequence $\mathbf{x}$ such that either $m_s = M_s$ and $n_s = 0$ or else $m_s = 0$ and $n_s = M_s$.

In order to prove the above claim, we single out a particular symbol r that occurs a total number of times $\mu = M_r$ in $\mathbf{s}$. We focus on the dependence of $\Pr(\mathbf{s}, \mathbf{x})$ on the occurrence number $t = m_r$ (so, $n_r = \mu - t$) by writing

$$\Pr(\mathbf{s}, \mathbf{x}) = \frac{a(\mu - t)!t!}{(b + t)!(c - t)!} \equiv f(t), \tag{F5}$$

where the coefficients a, b, and c are defined, respectively, as

$$a = \frac{d_\flat!^2}{2^N} \prod_{s \neq r} n_s! m_s!, \tag{F6}$$

$$b = d_\flat + \sum_{s \neq r} m_s, \tag{F7}$$

$$c = d_\flat + \sum_s n_s + m_r = d_\flat + N - \sum_{s \neq r} m_s \tag{F8}$$

and are independent of t. The function $f(t)$ can be extended to $t \in \mathbb{R}$ using the Euler gamma function and the relation $\Gamma(t + 1) = t!$. This extension enables us to compute the second derivative of $f(t)$ and show that it is a convex function of t in the interval $[0, \mu]$. Indeed,

$$\frac{f''(t)}{f(t)} = [H_1(c - t) - H_1(\mu - t) - H_1(b + t) + H_1(t)]^2$$
$$+ H_2(c - t) - H_2(\mu - t) + H_2(b + t) - H_2(t) \geq 0, \tag{F9}$$

where $H_n(t)$ are the generalized harmonic numbers. For positive integer values of t, they are $H_n(t) = \sum_{j=1}^{t} j^{-n}$. The relation $H_n(t) = \zeta(n) - \sum_{j=1}^{\infty} (t + j)^{-n}$, where $\zeta(n) = \sum_{j=1}^{\infty} j^{-n}$ is the Riemann zeta function, allows one to extend the domain of $H_n(t)$ to real (and complex) values of t.

The positivity of $f''(t)$ follows from the positivity of both $f(t)$ and the two differences of harmonic numbers in the second line of Eq. (F9). Note that $H_2(x)$ is an increasing function of x. Since, obviously, $b + t > t$ and $c - t > \sum_s n_s = \sum_s (M_s - m_s) \geq \mu - t$ [as follows from the definition of c in Eq. (F8)], we see that the two differences are positive.

The convexity of $f(t)$ for $t \in [0, \mu]$ implies that the maximum of $f(t)$ is at either $t = 0$ or $t = \mu$. This implication holds for every value of M_r and every symbol r in the data string, so our claim holds. In summary, the optimal guessing rule must assign the same type of roulette to all the M_s occurrences of a symbol s; i.e., it must group all data points that show the same symbol in the same cluster, which is in full agreement with our own intuition.

The description of the optimal protocol that runs on our device is not yet complete. We need to specify how to reduce the current number of clusters down to two, since at this point we may (and typically will) have up to d clusters, as many as different symbols. The reduction or merging of the d clusters can be based only on their relative sizes, as nothing is known about the underlying probability distributions. This restriction is quite clear: Let P be the subset of symbols (e.g., the subset of $\{1, 2, \ldots, d\}$) for which $n_s = M_s$, and let Q be its complement; i.e., Q contains the symbols for which $m_s = M_s$, and $\mathsf{P} = \bar{\mathsf{Q}}$. The claim we just proved tells us that, in order to find the maximum of $\Pr(\mathbf{s}, \mathbf{x})$, it is enough to consider sequences of roulettes $\mathbf{x}$ that comply with the above conditions on the occurrence numbers (For example, suppose $d = 3$ and $N = 12$. Assuming that $\mathbf{s} = (112321223112)$ is the string of data, the sequence of roulettes $\mathbf{x}$ in the table

i	1	2	3	4	5	6	7	8	9	10	11	12
S	1	1	2	3	2	1	2	2	3	1	1	2
X	*P*	*P*	*Q*	*Q*	*Q*	*P*	*Q*	*Q*	*Q*	*P*	*P*	*Q*

satisfies the conditions $m_s = M_s$ or $n_s = M_s$, since $n_1 = M_1 = 5$, $m_2 = M_2 = 5$, and $m_3 = M_3 = 2$. In this case, $\mathsf{P} = \{1\}$, and $\mathsf{Q} = \{2, 3\}$. The suggested clustering is $\{(1, 2, 6, 10, 11), (3, 4, 5, 7, 8, 9, 12)\}$). For those, the joint probability $\Pr(\mathbf{s}, \mathbf{x})$ can be written as

$$\Pr(\mathbf{s}, \mathbf{x}) = \frac{a}{(d_\flat + \sum_{s\in\mathsf{Q}} M_s)!(d_\flat + \sum_{s\in\mathsf{P}} M_s)!}, \quad \text{(F10)}$$

where a now simplifies to $2^{-N} d_\flat!^2 \Pi_s M_s!$. Thus, it just remains to find the partition $\{\mathsf{P}, \mathsf{Q}\}$ that maximizes this expression. It can also be written as

$$\Pr(\mathbf{s}, \mathbf{x}) = \frac{a}{(d_\flat + x)!(d_\flat + N - x)!}, \quad \text{(F11)}$$

where we define $x = \Sigma_{s\in\mathsf{Q}} M_s$. The maximum of this function is located at $x = N/2$, and one can easily check that it is monotonic on either side of its peak. Note that, depending on the values of the occurrence numbers $\{M_s\}$, the optimal value $x = N/2$ may not be attained. In such cases, the maximum of $\Pr(\mathbf{s}, \mathbf{x})$ is located at $x^* = N/2 \pm \Delta$, where Δ is the bias:

$$\Delta = \frac{1}{2} \min_{\mathsf{Q}} \left| \sum_{s\in\mathsf{Q}} M_s - \sum_{s\in\bar{\mathsf{Q}}} M_s \right|. \quad \text{(F12)}$$

The subset Q that minimizes this expression determines the optimal clustering.

In summary (and not very surprisingly), the optimal guessing rule consists in first partitioning the data $\mathbf{s}$ in up to d groups according to the symbol of the data points and, second, merging those groups (without splitting them) in two clusters in such a way that their sizes are as similar as possible. We have stumbled upon the so-called partition problem [33], which is known to be weakly NP complete. In particular, a large set of distinct occurrence counts $\{M_s\}$ rapidly hinders the efficiency of known algorithms, a situation likely to occur for large d. It follows that the optimal clustering protocol for the classical problem cannot be implemented efficiently in all instances of the problem.

To obtain the maximum success probability P_s^{cl} [Eq. (F2)], we need to sum the maximum joint probability, given by Eq. (F11) with $x = x^*$, over all possible strings $\mathbf{s}$. Those with the same set of occurrence counts $\{M_s\}$ give the same contribution. Moreover, all the dependence on $\{M_s\}$ is through the bias Δ. Therefore, if we define ξ_Δ to be the number of sets $\{M_s\}$ that give rise to a bias Δ, then the corresponding number of data strings is $\xi_\Delta N!/\Pi_s M_s!$. We thus can write

$$P_s^{\mathrm{cl}} = \sum_\Delta \frac{2^{1-N} \xi_\Delta d_\flat!^2 N!}{(d_\flat + \frac{N}{2} + \Delta)!(d_\flat + \frac{N}{2} - \Delta)!}. \quad \text{(F13)}$$

Equation (F13) is as far as we can go, as no explicit formula for the combinatorial factor ξ_Δ is likely to exist for general cases. However, it is possible to work out the asymptotic expression of the maximum success probability for large data sizes N. We first note that a generic term in the sum (F13) can be written as the factor $2^{2d_\flat+1} \xi_\Delta d_\flat!^2 N!/(2d_\flat + N)!$ times a binomial distribution that peaks at $\Delta = 0$ for large N. Hence, the dominant contribution in this limit is

$$P_s^{\mathrm{cl}} \sim \xi_0 \frac{2^{2d_\flat+1} d_\flat!^2 N!}{(2d_\flat + N)!} \sim \xi_0 \frac{2^{2d-1}(d-1)!^2}{N^{2d-2}}. \quad \text{(F14)}$$

From the definition of ξ_Δ, given above Eq. (F13), and that of Δ in Eq. (F12), we readily see that ξ_0 is the number of ordered partitions (i.e., the order matters) of N in d addends or parts [These ordered partitions are known as *weak compositions* of N into d parts in combinatorics, where *weak* means that some addends (or parts) can be zero; in contradistinction, the term *composition* is used when all the parts are strictly positive.] (the occurrence counts M_s) such that a subset of these addends is an ordered partition of $N/2$ as well.

Young diagrams come in handy to compute ξ_0. First, we draw pairs of diagrams $[\lambda, \lambda']$, each of $N/2$ boxes and such that $\lambda \geq \lambda'$ (in lexicographical order; see Appendix A), and $l(\lambda) + l(\lambda') \equiv r + r' \leq d$; i.e., the total number of rows should not exceed d. Next, we fill the boxes with symbols

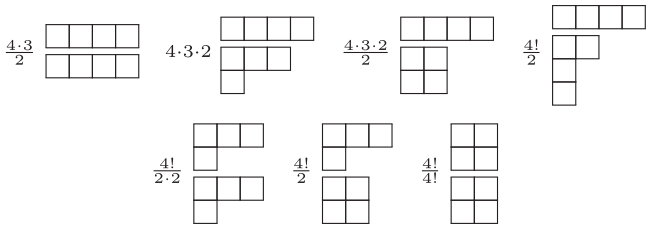


FIG. 4. Use of Young diagrams for computing ξ_0. In the example, $N = 8$ and $d = 4$. The fraction before each pair gives the number of different fillings and hints at how it is computed.

s_i (representing possible data points) so that all the boxes in each row have the same symbol. We readily see that the number of different fillings gives us ξ_0. An example is provided in Fig. 4 for clarity.

Although this pictorial method eases the computation of ξ_0, it becomes unpractical even for relatively small values of N. However, it becomes again very useful in the asymptotic limit, since the number of Young diagrams with at least two rows of equal size becomes negligibly small for large N. (Actually, the number of Young diagrams of a given length with an unequal number of boxes in each row is equal to the number of Young diagrams of $N - r(r-1)/2$ boxes; i.e., it is equal to $P^{(r)}_{N-r(r-1)/2}$. Using the results in Appendix A, we immediately see that for large N one has $P^{(r)}_{N-r(r-1)/2}/P^{(r)}_N \sim 1$, which proves the statement.) The same conclusion applies to the whole pairs $[\lambda, \lambda']$, since, e.g., by reshuffling rows, one could merge the two members into a single diagram of N boxes and length $r + r'$. Thus, we may assume that all pairs of diagrams with a given total length have an unequal number of boxes in each row, which renders the counting of different fillings trivial: There are $d!/(d - r - r' + 1)!$ ways of filling each pair of diagrams. Recalling that there is a one-to-one mapping between partitions and Young diagrams, we can use Eq. (A1) and write

$$\begin{aligned}\xi_0 &\sim \frac{1}{2}\sum_{r=1}^{d-1}\sum_{r'=1}^{d-r} P^{(r)}_{\frac{N}{2}} P^{(r')}_{\frac{N}{2}} \frac{d!}{(d-r-r')!} \\ &\sim \frac{1}{2}\left(\frac{N}{2}\right)^{d-2} \sum_{r=1}^{d} \frac{r(d-r)d!}{r!^2(d-r)!^2} \\ &\sim \frac{1}{2}\left(\frac{N}{2}\right)^{d-2} \frac{(2d-2)!}{(d-2)!(d-1)!^2}. \end{aligned} \tag{F15}$$

This result, together with Eq. (F14), leads us to the desired asymptotic expression for the optimal success probability:

$$P^{\mathrm{cl}}_s \sim \left(\frac{2}{N}\right)^d \frac{(2d-2)!}{(d-2)!}. \tag{F16}$$

APPENDIX G: OPTIMAL CLUSTERING PROTOCOL FOR KNOWN CLASSICAL STATES

In this Appendix, we give a short discussion on clustering classical states under the assumption that the underlying probability distributions are known. In particular, we discuss two low-dimensional cases, $d = 2$, 3, and derive the asymptotic expression of the success probability of clustering for large data string length N and arbitrary data dimension d. We stick to the notation introduced in Appendix F.

If the underlying probability distributions are known, a given data point s is optimally assigned to the probability distribution for which s is most likely. The success probability is thus given by $\max\{p_s, q_s\}/2$ (recall that the data are assumed to be drawn from either P or Q with equal prior probability). The average success probability of clustering over all possible strings of length N then reads

$$P^{\mathrm{cl}}_{s,PQ} = \frac{1}{2^N}\left[\left(\sum_{s=1}^{d}\max\{p_s, q_s\}\right)^N + \left(\sum_{s=1}^{d}\min\{p_s, q_s\}\right)^N\right], \tag{G1}$$

where the second term arises because assigning the wrong probability distribution to *all* data points in $\mathbf{s}$ gives a correct clustering. In order to compare with our results for unknown classical states, we average the success probability over a uniform distribution of categorical probability distributions. This average yields

$$P^{\mathrm{cl}}_s = \int_{\mathsf{S}_d} dP \int_{\mathsf{S}_d} dQ P^{\mathrm{cl}}_{s,PQ}, \tag{G2}$$

where the integration over the simplex S_d, shared by P and Q, is defined in Appendix E.

To perform the integral in Eq. (G2), we need to partition $\mathsf{S}_d \times \mathsf{S}_d$ in different regions according to whether $p_s \le q_s$ or $p_s > q_s$ for the various symbols. By symmetry, the integral can depend only on the number r of symbols for which $p_s \le q_s$ (not on its particular value). Hence, $r = 1, \ldots, d-1$ labels the different types of integral that we need to compute to evaluate P^{cl}_s. Notice that we have the additional symmetry $r \leftrightarrow d - r$, corresponding to exchanging p_s and q_s for all s. Since the value of these integrals does not depend on the specific value of s, we can choose all p_s with $s = 1, 2, \ldots, r$ to satisfy $p_s > q_s$ and all p_s with $s = r+1, r+2, \ldots, d$ to satisfy $p_s \le q_s$. To shorten the expressions below, we define

$$\mathfrak{p}_k := \sum_{s=1}^{k} p_s, \qquad \mathfrak{q}_k := \sum_{s=1}^{k} q_s. \tag{G3}$$

With these definitions, $\mathfrak{p}_d = \mathfrak{q}_d = 1$, $\sum_{s=r+1}^{d} q_s = 1 - \mathfrak{q}_r$, and likewise $\sum_{s=r+1}^{d} p_s = 1 - \mathfrak{p}_r$. The integrals that we need to compute are then

$$I_r^d := \int_{\mathsf{S}_d} dP \frac{1}{V_d} \int_0^{p_1} dq_1 \ldots \int_0^{p_r} dq_r \times \int_{p_{r+1}}^{\mathfrak{p}_{r+1}-\mathfrak{q}_r} dq_{r+1} \ldots \int_{p_{d-1}}^{\mathfrak{p}_{d-1}-\mathfrak{q}_{d-2}} dq_{d-1} \times [(1+\mathfrak{p}_r-\mathfrak{q}_r)^N + (1+\mathfrak{q}_r-\mathfrak{p}_r)^N], \tag{G4}$$

and we note that, as anticipated, $I_r^d = I_{d-r}^d$. The average probability of successful clustering then reads

$$P_s^{\mathrm{cl}} = \frac{1}{2^N} \sum_{r=1}^{d-1} \binom{d}{r} I_r^d, \tag{G5}$$

where the binomial is the number of equivalent integral regions for the given r.

1. Low data dimension

We can now discuss the lowest-dimensional cases, for which explicit closed formulas for I_r^d can be derived. For $d = 2$, one has

$$P_s^{\mathrm{cl}} = \frac{8 - 2^{2-N}}{(N+2)(N+1)}. \tag{G6}$$

This result coincides with that of unknown probability distributions given in Eq. (F13) with $\xi_\Delta = 1$. This result is expected, as the optimal protocol for known and unknown probability distributions is exactly the same: Assign to the same cluster all data points that show the same symbol s. Therefore, knowing the probability distribution does not provide any advantage for $d = 2$.

For $d > 2$, however, knowledge of the distributions P and Q helps classify the data points. If $d = 3$, the success probability (G5) can be computed to be

$$P_s^{\mathrm{cl}} = 6\frac{2^5(N-2) - 2^{2-N}(N^2+7N+18)}{(N+4)(N+3)(N+2)(N+1)}. \tag{G7}$$

In Table II, we compare five values of P_s^{cl} in Eq. (G7), when $N = 2, 3, \ldots, 6$, with those for unknown distributions P and Q given by Eq. (F13). As expected, the success probability is larger if P and Q are known. The source of the increase is illustrated by the string $\mathbf{s} = (112)$, which is labeled as PPQ (or QQP) if P and Q are unknown. However, if they are known and, e.g., $p_1 > q_1$ and $p_2 > q_2$, the string is more appropriately labeled as PPP.

TABLE II. The success probability P_s^{cl} for $d = 3$ and data string lengths $N = 2, \ldots, 6$ in the cases of known and unknown distributions P and Q. For unknown distributions, the values are computed using Eq. (F13) in Appendix F. For known distributions, the values are given by Eq. (G7). The table shows that knowing P and Q increases the success probability of clustering.

N	2	3	4	5	6
Unknown:	7/12	11/30	0.250	0.176	0.130
Known:	3/5	2/5	0.283	0.210	0.160

2. Arbitrary data dimension: Large N limit

For increasing N, however, the advantage of knowing P and Q becomes less significant and vanishes asymptotically, which can be checked explicitly for $d = 2$, 3 by expanding Eqs. (G6) and (G7) in inverse powers of N. In this regime, the average is dominated by distributions for which $\mathfrak{p}_r \approx 1$ and $\mathfrak{q}_r \approx 0$. Since in a typical string approximately half of the data come from the distribution P and the other half from Q, the optimal clustering protocol essentially coincides with that for unknown distributions; i.e., it collects the data points showing the same symbol in the same subcluster and afterwards merges the subclusters into two clusters of approximately the same size. We next prove that this intuition is right for all d.

In the proof, we make repeated use of the trivial observation that, for asymptotically large N and $0 < a < b < c$, one has

$$\int_a^b (c-x)^N dx \sim (c-a)^{N+1}/N. \tag{G8}$$

We also note that the contribution to the success probability coming from the completely wrong assignment of distributions, i.e., $(1+\mathfrak{q}_r-\mathfrak{p}_r)^N$, is exponentially vanishing, since we assume $p_r > q_r$, and thus $\mathfrak{q}_r - \mathfrak{p}_r < 0$ [this fact is well illustrated by the terms proportional to 2^{2-N} in Eqs. (G6) and (G7)].

Because of this last observation, we can drop the second term in the integrand of I_r^d [Eq. (G4)]. The integrals over q_s, $s \le r$, of the remaining term, $(1+\mathfrak{p}_r-\mathfrak{q}_r)^N$, are dominated by the lower limit $q_s = 0$, as this value maximizes $1+\mathfrak{p}_r-\mathfrak{q}_r$. Using Eq. (G8), we get

$$I_r^d \sim \frac{(d-1)!}{N^r} \int_{\mathsf{S}_d} dP \int_{p_{r+1}}^{\mathfrak{p}_{r+1}-\mathfrak{q}_r} dq_{r+1} \ldots \times \int_{p_{d-1}}^{\mathfrak{p}_{d-1}-\mathfrak{q}_{d-2}} dq_{d-1} (1+\mathfrak{p}_r)^{N+r}, \tag{G9}$$

where we recall that the volume of the simplex S_d is $V_d = 1/(d-1)!$. For the remaining integrals over q_s in Eq. (G9), we can take the lower limits to be $p_s \approx 0$, for $s \ge r+1$, since the integrand is maximized by $\mathfrak{p}_r \approx 1$. Therefore, the upper limits become 1, $1-q_{r+1}, \ldots, 1-\sum_{s=r+1}^{d-2} q_s$. We identify these upper and lower limits as those of an integral over a $(d-r-1)$-dimensional probability simplex S_{d-r}. We can thus write

$$I_r^d \sim \frac{(d-1)!}{(d-r-1)!N^r} \int_{\mathsf{S}_d} dP (1+\mathfrak{p}_r)^{N+r}. \tag{G10}$$

The last equation can be cast as

$$I_r^d \sim \frac{(d-1)!}{(d-r-1)!N^r}\int_{\mathsf{S}_d} dP\left(2-\sum_{s=r}^{d-1} p_s\right)^{N+r}, \quad \text{(G11)}$$

where we use again that $\mathfrak{p}_r = 1-\sum_{s=r+1}^{d} p_s$ and note that under the integral sign we are free to relabel the variables p_s. According to the definition of $\int_{\mathsf{S}_d} dP$, we need to perform $d-r$ integrals over the variables p_r, $p_{r+1},\ldots,p_{d-1}$, for which we can use Eq. (G8), which yields a factor $2^{N+d}/N^{d-r}$. The remaining integrals over $p_1, p_2, \ldots, p_{r-1}$ of this constant factor give an additional $1/(r-1)!$, as they effectively correspond to an integral over a $(r-1)$-dimensional simplex. Putting the different pieces together, the asymptotic expression of I_r^d reads

$$I_r^d \simeq \frac{2^{N+d}}{N^d}\frac{[(d-1)!]^2}{(r-1)!(d-r-1)!}. \quad \text{(G12)}$$

We are now in a position to compute the asymptotic success probability. Inserting Eq. (G12) into Eq. (G5), we readily obtain

$$\begin{aligned} P_s^{\mathrm{cl}} &\sim \left(\frac{2}{N}\right)^d (d-1)!(d-1)\sum_{r=1}^{d-1}\binom{d}{r}\binom{d-2}{d-r-1} \\ &= \left(\frac{2}{N}\right)^d \frac{(2d-2)!}{(d-2)!}, \end{aligned} \quad \text{(G13)}$$

where we use the well-known binomial identity $\sum_k \binom{a}{k}\binom{b}{s-k} = \binom{a+b}{s}$ [here, k ranges over all values for which the binomials make sense]. Equation (G13) coincides with the asymptotic expression in the unknown case Eq. (9), as we anticipated.

[1] P. W. Shor, *Fault-Tolerant Quantum Computation*, in *Proceedings of the 37th Conference on Foundations of Computer Science* (IEEE, New York, 1996), pp. 56–65.

[2] L. K. Grover, *Quantum Mechanics Helps in Searching for a Needle in a Haystack*, Phys. Rev. Lett. **79**, 325 (1997).

[3] A. B. Finnila, M. A. Gomez, C. Sebenik, C. Stenson, and J. D. Doll, *Quantum Annealing: A New Method for Minimizing Multidimensional Functions*, Chem. Phys. Lett. **219**, 343 (1994).

[4] T. Kadowaki and H. Nishimori, *Quantum Annealing in the Transverse Ising Model*, Phys. Rev. E **58**, 5355 (1998).

[5] S. Lloyd, *Universal Quantum Simulators*, Science **273**, 1073 (1996).

[6] H. J. Kimble, *The Quantum Internet*, Nature (London) **453**, 1023 (2008).

[7] S. Wehner, D. Elkouss, and R. Hanson, *Quantum Internet: A Vision for the Road Ahead*, Science **362**, eaam9288 (2018).

[8] V. Dunjko and H. J. Briegel, *Machine Learning & Artificial Intelligence in the Quantum Domain: A Review of Recent Progress*, Rep. Prog. Phys. **81**, 074001 (2018).

[9] M. Sasaki and A. Carlini, *Quantum Learning and Universal Quantum Matching Machine*, Phys. Rev. A **66**, 022303 (2002).

[10] N. Liu and P. Rebentrost, *Quantum Machine Learning for Quantum Anomaly Detection*, Phys. Rev. A **97**, 042315 (2018).

[11] M. Skotiniotis, R. Hotz, J. Calsamiglia, and R. Muñoz-Tapia, *Identification of Malfunctioning Quantum Devices*, arXiv:1808.02729.

[12] A. Bisio, G. Chiribella, G. M. D'Ariano, S. Facchini, and P. Perinotti, *Optimal Quantum Learning of a Unitary Transformation*, Phys. Rev. A **81**, 032324 (2010).

[13] A. Bisio, G. M. D'Ariano, P. Perinotti, and M. Sedlák, *Quantum Learning Algorithms for Quantum Measurements*, Phys. Lett. A **375**, 3425 (2011).

[14] M. Guta and W. Kotlowski, *Quantum Learning: Asymptotically Optimal Classification of Qubit States*, New J. Phys. **12**, 123032 (2010).

[15] G. Sentís, J. Calsamiglia, R. Muñoz-Tapia, and E. Bagan, *Quantum Learning without Quantum Memory*, Sci. Rep. **2**, 708 (2012).

[16] G. Sentís, M. Guta, and G. Adesso, *Quantum Learning of Coherent States*, EPJ Quantum Techno. **2**, 17 (2015).

[17] M. Fanizza, A. Mari, and V. Giovannetti, *Optimal Universal Learning Machines for Quantum State Discrimination*, IEEE Trans. Inf. Theory **65**, 5931 (2019).

[18] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Math. Intell. **27**, 83 (2005).

[19] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition* (Springer, New York, 1996).

[20] A. Monràs, G. Sentís, and P. Wittek, *Inductive Supervised Quantum Learning*, Phys. Rev. Lett. **118**, 190503 (2017).

[21] V. Dunjko, J. M. Taylor, and H. J. Briegel, *Quantum-Enhanced Machine Learning*, Phys. Rev. Lett. **117**, 130501 (2016).

[22] D. Aloise, A. Deshpande, P. Hansen, and P. Popat, *NP-Hardness of Euclidean Sum-of-Squares Clustering*, Mach. Learn. **75**, 245 (2009).

[23] S. Ben-David, *Clustering is Easy When....What?*, arXiv:1510.05336.

[24] V. Bužek, M. Hillery, M. Ziman, and M. Roško, *Programmable Quantum Processors*, Quantum Inf. Process. **5**, 313 (2006).

[25] C. W. Helstrom, *Quantum Detection and Estimation Theory* (Academic, New York, 1976).

[26] S. M. Barnett, *Minimum-Error Discrimination between Multiply Symmetric States*, Phys. Rev. A **64**, 030303(R) (2001).

[27] G. Chiribella, G. M. D'Ariano, P. Perinotti, and M. F. Sacchi, *Covariant Quantum Measurements That Maximize the Likelihood*, Phys. Rev. A **70**, 062105 (2004).

[28] G. Chiribella, G. M. D'Ariano, P. Perinotti, and M. F. Sacchi, *Maximum Likelihood Estimation for a Group of Physical Transformations*, Int. J. Quantum. Inform. **04**, 453 (2006).

[29] K. M. R. Audenaert, J. Calsamiglia, R. Munoz-Tapia, E. Bagan, Ll. Masanes, A. Acin, and F. Verstraete,

Discriminating States: The Quantum Chernoff Bound, Phys. Rev. Lett. **98**, 160501 (2007).
[30] H. Krovi, S. Guha, Z. Dutton, and M. P. da Silva, *Optimal Measurements for Symmetric Quantum States with Applications to Optical Communication*, Phys. Rev. A **92**, 062333 (2015).
[31] D. E. Knuth, *Sorting and Searching*, The Art of Computer Programming Vol. 3 (Addison-Wesley, Reading, MA, 1998).
[32] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables* (Dover, New York, 1965).
[33] R. E. Korf, *A Complete Anytime Algorithm for Number Partitioning*, Artif. Intell. **106**, 181 (1998).
[34] A. S. Holevo, *Probabilistic and Statistical Aspects of Quantum Theory* (North-Holland, Amsterdam, 1982).
[35] A. S. Holevo, *Statistical Decision Theory for Quantum Systems*, J. Multivariate Anal. **3**, 337 (1973).
[36] H. Yuen, R. Kennedy, and M. Lax, *Optimum Testing of Multiple Hypotheses in Quantum Detection Theory*, IEEE Trans. Inf. Theory **21**, 125 (1975).
[37] N. D. Pozza and G. Pierobon, *Optimality of Square-Root Measurements in Quantum State Discrimination*, Phys. Rev. A **91**, 042334 (2015).
[38] G. Sentís, E. Bagan, J. Calsamiglia, G. Chiribella, and R. Muñoz-Tapia, *Quantum Change Point*, Phys. Rev. Lett. **117**, 150502 (2016).
[39] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. (Cambridge University Press, Cambridge, England, 2013).
[40] I. D. Ivanovic, *How to Differentiate between Non-Orthogonal States*, Phys. Lett. A **123**, 257 (1987).
[41] D. Dieks, *Overlap and Distinguishability of Quantum States*, Phys. Lett. A **126**, 303 (1988).
[42] A. Peres, *How to Differentiate between Non-Orthogonal States*, Phys. Lett. A **128**, 19 (1988).
[43] A. Chefles and S. M. Barnett, *Optimum Unambiguous Discrimination between Linearly Independent Symmetric States*, Phys. Lett. A **250**, 223 (1998).
[44] S. Lloyd, *Least Squares Quantization in PCM*, IEEE Trans. Inf. Theory **28**, 129 (1982).
[45] A. W. Harrow, *Applications of Coherent Classical Communication and the Schur Transform to Quantum Information Theory*, Ph.D. thesis, Massachusetts Institute of Technology, 2005.
[46] H. Krovi, *An Efficient High Dimensional Quantum Schur Transform*, Quantum **3**, 122 (2019).
[47] G. Sentís, J. Calsamiglia, and R. Muñoz-Tapia, *Exact Identification of a Quantum Change Point*, Phys. Rev. Lett. **119**, 140506 (2017).
[48] J. Von Korff and J. Kempe, *Quantum Advantage in Transmitting a Permutation*, Phys. Rev. Lett. **93**, 260502 (2004).
[49] M. Hillery, E. Andersson, S. M. Barnett, and D. Oi, *Decision Problems with Quantum Black Boxes*, J. Mod. Opt. **57**, 244 (2010).
[50] E. Aïmeur, G. Brassard, and S. Gambs, *Quantum Speed-Up for Unsupervised Learning*, Mach. Learn. **90**, 261 (2013).
[51] S. Lloyd, M. Mohseni, and P. Rebentrost, *Quantum Algorithms for Supervised and Unsupervised Machine Learning*, arXiv:1307.0411.
[52] N. Wiebe, A. Kapoor, and K. Svore, *Quantum Algorithms for Nearest-Neighbor Methods for Supervised and Unsupervised Learning*, Quantum Inf. Comput. **15**, 0318 (2015).
[53] I. Kerenidis, J. Landman, A. Luongo, and A. Prakash, *q-Means: A Quantum Algorithm for Unsupervised Machine Learning*, arXiv:1812.03584.
[54] V. Giovannetti, S. Lloyd, and L. Maccone, *Quantum Random Access Memory*, Phys. Rev. Lett. **100**, 160501 (2008).
[55] G. Chiribella and D. Ebler, *Quantum Speedup in the Identification of Cause-Effect Relations*, Nat. Commun. **10**, 1472 (2019).
[56] P. Flajolet and R. Sedgewick, *Analytic Combinatorics* (Cambridge University Press, Cambridge, England, 2009).
[57] G. E. Andrews, *The Theory of Partitions*, Encyclopedia of Mathematics and Its Applications Vol. 2 (Addison-Wesley, Reading, MA, 1976).
[58] B. E. Sagan, *The Symmetric Group—Representations, Combinatorial Algorithms, and Symmetric Functions* (Springer, New York, 2001).
[59] R. Goodman and N. R. Wallach, in *Symmetry, Representations, and Invariants*, edited by S. Axler and K. A. Ribet, Graduate Texts in Mathematics Vol. 255 (Springer, New York, 2009).
[60] S. Gliske, W. H. Klink, and T. Ton-That, *Algorithms for Computing Generalized U(N) Racah Coefficients*, Acta Appl. Math. **88**, 229 (2005).
[61] N. I. Akhiezer and N. Kemmer, *The Classical Moment Problem: And Some Related Questions in Analysis* (Oliver and Boyd, Edinburgh, 1965), Vol. 5.
[62] L. Alonso and T. Gorin, *Joint Probability Distributions for Projection Probabilities of Random Orthonormal States*, J. Phys. A **49**, 145004 (2016).