Structural risk minimization for quantum linear classifiers

Casper Gyurik[1], Dyon van Vreumingen[1,2,3], and Vedran Dunjko[1,4]

[1]LIACS, Leiden University, Niels Bohrweg 1, 2333 CA Leiden, Netherlands
[2]QuSoft, Centrum Wiskunde & Informatica (CWI), Science Park 123, 1098 XG Amsterdam, Netherlands
[3]Institute of Physics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, the Netherlands
[4]LION, Leiden University, Niels Bohrweg 2, 2333 CA Leiden, Netherlands
December 4th, 2022

Quantum machine learning (QML) models based on parameterized quantum circuits are often highlighted as candidates for quantum computing's near-term "killer application". However, the understanding of the empirical and generalization performance of these models is still in its infancy. In this paper we study how to balance between training accuracy and generalization performance (also called structural risk minimization) for two prominent QML models introduced by Havlíček et al. [1], and Schuld and Killoran [2]. Firstly, using relationships to well understood classical models, we prove that two model parameters – i.e., the dimension of the sum of the images and the Frobenius norm of the observables used by the model – closely control the models' complexity and therefore its generalization performance. Secondly, using ideas inspired by process tomography, we prove that these model parameters also closely control the models' ability to capture correlations in sets of training examples. In summary, our results give rise to new options for structural risk minimization for QML models.

1 Introduction

After years of efforts the first proof-of-principle quantum computations that arguably surpass what is feasible with classical supercomputers have been realized [3]. As the leap from noisy intermediate-scale quantum (NISQ) devices [4] to full-blown quantum computers may require further decades, finding practically useful NISQ-suitable algorithms is becoming increasingly important. It has been argued that some of the most promising NISQ-suitable algorithms are those that rely on *parameterized quantum circuits* (also called variational quantum circuits) [5, 6]. Such algorithms have been proposed for quantum chemistry [7, 8], for optimization [9], and for machine learning [10]. One of the advantages of parameterized quantum circuits is that restrictions of NISQ devices can be hardwired into the circuit. Moreover, families of parameterized quantum circuits can – under widely believed complexity-theoretic assumptions – realize input-output correlations that are intractable for classical computation [11, 12]. In this paper we discuss the application of parameterized quantum circuits as machine learning models in hybrid quantum-classical methods. The use of NISQ devices in the context of machine learning is particularly appealing as machine learning algorithms may be more tolerant to noise in the quantum hardware [13, 14]. In short, parameterized quantum circuits could yield NISQ-friendly machine learning models that could be used to classify data for which conventional classical machine learning models may struggle.

In machine learning, parameterized quantum circuits can serve as a parameterized family of real-valued functions in a manner similar to neural networks (they are often called quantum neural networks). It has been noted that machine learning models based on parameterized quantum circuit are closely related to *linear classifiers*, which use hyperplanes to separate classes of data embedded in a vector space. This connection was first established by Havlíček et al. [1], and Schuld & Killoran [2], who both defined two machine learning models based on parameterized quantum circuits that efficiently implement certain families of linear classifiers – an illustration of which can be found in Figure 1. In

Casper Gyurik: c.f.s.gyurik@liacs.leidenuniv.nl

this paper we further investigate and exploit this relation between machine learning models based on parameterized quantum circuits and standard linear classifiers to investigate how to perform *structural risk minimization*. More specifically, we study how to tune parameters of quantum machine learning models to optimize their expressivity (i.e., the ability to correctly capture correlations in sets of training examples) while preventing the model from becoming too complex (which could cause it to overfit and generalize poorly to unseen examples).

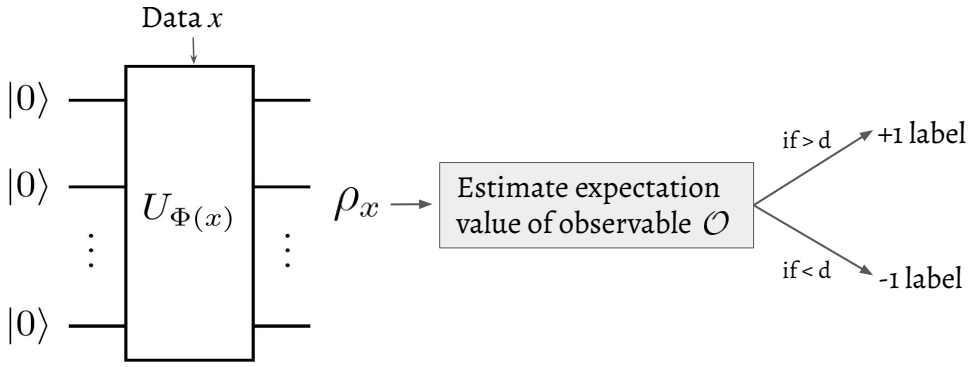

Figure 1: An overview of the quantum machine learning models introduced in [1, 2]. First, a parameterized quantum circuit is used to encode the data into a quantum state ρ_x. Afterwards, an observable $\mathcal{O}$ is measured. If its expectation value lies above d, then we assign the label $+1$, and -1 otherwise. The goal in training is the find the optimal observable $\mathcal{O}$ and threshold d.

Our contributions We identify model parameters that can be used to implement structural risk minimization for a family of quantum machine learning models that includes the models introduced by Havlíček et al. [1], and Schuld & Killoran [2]. Specifically, we theoretically analyze the effect that limiting the rank or Frobenius norm of the observables measured by these models has on the trade-off between 1. generalization performance, and 2. expressivity, and we show that:

1. (a) A measure of complexity called the *VC dimension* can be controlled by limiting the *dimension of the sum of the images* of the observables measured by the quantum model. In particular, we provide explicit analytical bounds on the VC dimension in terms of the dimension of the span of the observables, and the dimension of the sum of the images of the observables. Afterwards, we use this result to devise quantum models for which we can control this VC dimension bound by limiting the ranks of the observables (i.e., they can be regularized by penalizing high-rank observables).

   (b) A measure of complexity called the *fat-shattering dimension* can be controlled by limiting the *Frobenius norm* of the observables measured by the quantum model. In particular, we provide explicit analytical bounds on the fat-shattering dimension in terms of the Frobenius norm of the observable.

   Due to the well-established connection between these complexity measures and upper bounds on the generalization performance [15, 16], our results theoretically quantify the effect that adjusting the dimension of the sum of the images, or Frobenius norm of the observables measured by the quantum model have on its generalization performance. Further we show that:

2. (a) Quantum models that use high-rank observables are strictly more expressive than quantum models that use low-rank observables. In particular, we show that i) any set of examples that can be correctly classified using a low-rank observable can also be correctly classified using a high-rank observable, and ii) there exist sets of examples that can only be correctly classified using an observable of at least a certain rank.

   (b) Quantum models that use observables with large Frobenius norms can achieve larger margins (i.e., empirical quantities measured on a set of training examples that influence certain generalization bounds) compared to quantum models that use observables with small Frobenius

norms. In particular, we show that there exist sets of examples that can only be classified with a given margin using observables of at least a certain Frobenius norm. Since the Frobenius norm controls the fat-shattering dimension, this can actually also have a positive effect on the generalization performance (as discussed in Section 2.2).

To summarize, we show that the rank or Frobenius norm of the observables measured by the quantum model also controls the model's ability to capture correlations in sets of examples.

Additional to the above two points, we also connect quantum machine learning with parameterized quantum circuits to standard structural risk minimization theory and discuss how to use our results to find the best quantum models in practice. In particular, we discuss different types of regularization that are theoretically motivated by our results, which help improve the performance of the quantum models in practice without putting extra requirements on the quantum hardware and are thus NISQ-suitable. Moreover, we find that there exist training methods – i.e., those who penalize high-rank observables – that are theoretically motivated by our results, and for which the resulting quantum model does not necessarily correspond to a kernel method as argued in [17].

Related work The way the observable in Figure 1 is measured typically consists of multiple steps that involve different parts of the quantum model. For instance, a prominent approach consists of first applying a parameterized quantum circuit to the data encoding state ρ_x, and then performing some fixed measurement. Previous works have focused on showing how complexity measures depend on the different parts of the quantum model that implement the observable measurement, such as the quantum circuit ansatz [18, 19, 20], or the level of noise in the model [21]. In this work we study the observable measured by the quantum model as a whole due to the 1-1 correspondence with the normal vectors of separating hyperplanes of linear classifiers. By studying the observable as a whole, our results apply to all quantum models that are of the structure described in Figure 1, independent of how the observable measurement is implemented. Moreover, by being agnostic to how the observable is measured, our results are complementary to results that focus on the specifics of a particular implementation of the observable measurement such as those mentioned above. Other related work has focused on showing that quantum machine learning models are remarkably expressive and satisfy generalization bounds based on different complexity measures [22, 23, 24]. Finally, other related work has studied the generalization performance of quantum machine learning models in order to compare their performance with classical machine learning models [25, 26].

Organization In Section 2, we define the quantum machine learning models studied in this paper, and we provide background on structural risk minimization. In Section 3, we investigate how structural risk minimization can be achieved for the quantum models. First, we determine two capacity measures of the quantum models, which allows us to identify model parameters that control the model's complexity in Subsection 3.1. Afterwards, we investigate the effect of these model parameters on the empirical performance in Subsection 3.2. We end with a discussion of how to implement structural risk minimization of the quantum models in practice in Subsection 3.3.

2 Background and motivation

In this section we provide the necessary background and we motivate our results. First, we introduce the family quantum machine learning models that we will study. Afterwards, we introduce the framework of statistical learning theory, which together with our results will provide an approach to optimally tuning the family of quantum models via so-called *structural risk minimization*.

2.1 Quantum linear classifiers

A fundamental family of classifiers used throughout machine learning are those constructed from *linear functions*. Specifically, these classifiers are constructed from the family of real-valued functions on $\mathbb{R}^\ell$ given by

$$\mathcal{F}_{\text{lin}} = \left\{ f_w(x) = \langle w, x \rangle \;\middle|\; w \in \mathbb{R}^\ell \right\}, \tag{1}$$

where $\langle .,. \rangle$ denotes an inner product on the input space $\mathbb{R}^\ell$. These linear functions are turned into classifiers by adding an offset and taking the sign, i.e., the classifiers are given by

$$\mathcal{C}_{\text{lin}} = \Big\{ c_{w,d}(x) = \text{sign}\big(\langle w, x \rangle - d\big) \;\Big|\; w \in \mathbb{R}^\ell,\ d \in \mathbb{R} \Big\}. \tag{2}$$

These linear classifiers essentially use hyperplanes to separate the different classes in the data.

While this family of classifiers seems relatively limited, it becomes powerful when introducing a *feature map*. Specifically, a feature map $\Phi : \mathbb{R}^\ell \to \mathbb{R}^N$ is used to (non-linearly) map the data to a (much) higher-dimensional space – called the *feature space* – in order to make the data more linearly-separable. We let $\mathcal{C}(\Phi) = \{c \circ \Phi \mid c \in \mathcal{C}\}$ denote the family of classifiers on $\mathbb{R}^\ell$ obtained by combining a family of linear classifiers $\mathcal{C} \subseteq \mathcal{C}_{\text{lin}}$ on $\mathbb{R}^N$ with a feature map Φ. If the feature map is clear from the context we will omit it in the notation and just write $\mathcal{C}$. A well known example of a model based on linear classifiers is the *support vector machine* (SVM), which aims to finds the hyperplane that attains the maximal perpendicular distance to the two classes of points in the two distinct half-spaces (assuming the feature map makes the data linearly separable).

The linear-algebraic nature of linear classifiers makes them particularly well-suited for quantum treatment. In the influential works of Havlíček et al. [1], and Schuld & Killoran [2], the authors propose a model where the space of n-qubit Hermitian operators – denoted $\text{Herm}\big(\mathbb{C}^{2^n}\big)$ – takes the role of the feature space. Specifically, they view $\text{Herm}\big(\mathbb{C}^{2^n}\big)$ as a 4^n-dimensional real vector space equipped with the Frobenius inner product $\langle A, B \rangle = \text{Tr}[A^\dagger B]$. Their feature map maps classical inputs x to n-qubit density matrices $\Phi(x) = \rho_\Phi(x)$ (i.e., positive semi-definite matrices of trace one). Finally, the hyperplanes that separates the states $\rho_{\Phi(x)}$ corresponding to the different classes are given by n-qubit observables. In short, the family of functions their model uses is given by

$$\mathcal{F}_{\text{qlin}} = \Big\{ f_\mathcal{O}(x) = \text{Tr}\big[\mathcal{O}\rho_\Phi(x)\big] \;\Big|\; \mathcal{O} \in \text{Herm}\big(\mathbb{C}^{2^n}\big) \Big\}, \tag{3}$$

and the family of classifiers – which we refer to as *quantum linear classifiers* – is given by

$$\mathcal{C}_{\text{qlin}} = \Big\{ c_{\mathcal{O},d}(x) = \text{sign}\big(\text{Tr}\big[\mathcal{O}\rho_\Phi(x)\big] - d\big) \;\Big|\; \mathcal{O} \in \text{Herm}\big(\mathbb{C}^{2^n}\big),\ d \in \mathbb{R} \Big\}. \tag{4}$$

We can estimate $f_\mathcal{O}(x)$ defined in Equation (3) by preparing the state $\rho_{\Phi(x)}$ and measuring the observable $\mathcal{O}$. In particular, approximating $f_\mathcal{O}(x)$ up to additive error ε requires only $\mathcal{O}\big(1/\varepsilon^2\big)$ samples. While the error creates a fuzzy region around the decision boundary, this turns out to not cause major problems in practical settings [10].

Using parameterized quantum circuits both the preparation of a quantum state that encodes the classical input and the measurement of observables can be done efficiently for certain feature maps and families of observables. We now briefly recap two ways in which parameterized quantum circuits can be used to efficiently implement a family of quantum linear classifiers, as originally proposed by Havlíček et al. [1], and Schuld & Killoran [2]. Both ways use a parameterized quantum circuit to implement the feature map. Specifically, let U_Φ be a parameterized quantum circuit, then we can use it to implement the feature map given by

$$\Phi : x \mapsto \rho_\Phi(x) \coloneqq |\Phi(x)\rangle \langle \Phi(x)|, \tag{5}$$

where $|\Phi(x)\rangle \coloneqq U_\Phi(x)\,|0\rangle^{\otimes n}$. The key difference between the two approaches is which observables they are able to implement (i.e., which separating hyperplanes they can represent) and how the observables are actually measured (i.e., how the functions $f_\mathcal{O}$ are evaluated). An overview of how the two approaches implement quantum linear classifiers can be found in Figure 2, and we discuss the main ideas behind the two approaches below.

Explicit quantum linear classifier[1] The observables measured in this approach are implemented by first applying a parameterized quantum circuit $W(\theta)$, followed by a computational basis measurement

[1] Also called the *quantum variational classifier* [1].

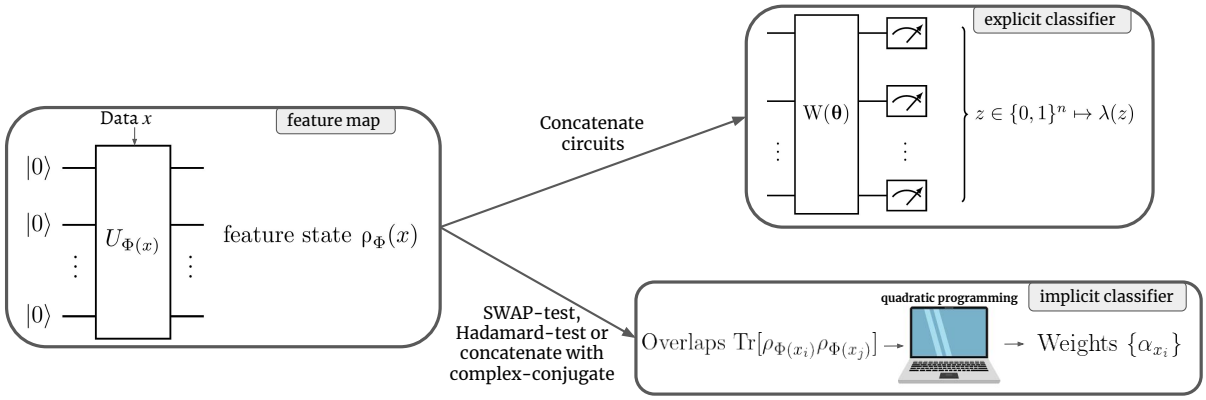

Figure 2: An overview of the implementations of the explicit and implicit quantum linear classifiers defined in Equations (7) and (8), respectively. Note that in the case of the explicit classifier, a universal circuit $W(\theta)$ (specifying the eigenbasis) followed by a computational basis measurement and universal postprocessing λ (specifying the eigenvalues) allows one to measure any observable.

and postprocessing of the measurement outcome $\lambda : [2^n] \to \mathbb{R}$. Upon closer investigation, one can derive that the corresponding observable is given by

$$\mathcal{O}_\theta^\lambda = W^\dagger(\theta) \cdot \mathrm{diag}\Big(\lambda(0), \lambda(1), \ldots, \lambda(2^n - 1)\Big) \cdot W(\theta). \tag{6}$$

Examples of efficiently computable postprocessing functions λ include functions with a polynomially small support (implemented using a lookup table), functions that are efficiently computable from the input bitstring (e.g., the parity of the bitstring, which is equivalent to measuring $Z^{\otimes n}$), or parameterized functions such as neural networks. Note that the postprocessing function λ plays an important role in how the measurement of the observable in Eq. (6) is physically realized. Altogether, this efficiently implements the family of linear classifiers – which we refer to as *explicit quantum linear classifiers* – given by

$$\mathcal{C}_{\mathrm{qlin}}^{\mathrm{explicit}} = \Big\{ c_{\mathcal{O}_\theta^\lambda, d}(x) = \mathrm{sign}\big(\mathrm{Tr}\big[\rho_\Phi(x)\mathcal{O}_\theta^\lambda\big] - d\big) \;\Big|\; \mathcal{O}_\theta^\lambda \text{ as in Equation (6)},\ d \in \mathbb{R} \Big\}. \tag{7}$$

The power of this model lies in the efficient parameterization of the manifold (inside the 4^n-dimensional vector space of Hermitian operators on $\mathbb{C}^{2^n}$) realized by the quantum feature map together with the parameterized separating hyperplanes that can be attained by $W(\theta)$ and λ. Here also lies a restriction of the explicit quantum linear classifier compared to standard linear classifiers, as in the latter all hyperplanes are possible and in the former only the hyperplanes that lie in the manifold parameterized by $W(\theta)$ and λ are possible. Furthermore, explicit quantum linear classifiers can likely not be efficiently evaluated classically, as computing expectation values $\mathrm{Tr}\big[\rho_\Phi(x)\mathcal{O}_\theta^\lambda\big]$ is classically intractable for sufficiently complex feature maps and observables [11, 12].

Implicit quantum linear classifier[2] Another way to implement a linear classifier is by using the so-called *kernel trick* [27]. In short, this trick involves expressing the normal vector of the separating hyperplane, – i.e., the observable $\mathcal{O}$ in the case of quantum linear classifiers – on a set of training examples $\mathcal{D}$ as a linear combination of feature vectors, resulting in the expression

$$\mathcal{O}_\alpha = \sum_{x' \in \mathcal{D}} \alpha_{x'} \rho_{\Phi(x')} = \sum_{x' \in \mathcal{D}} \alpha_{x'} \left|\Phi(x')\right\rangle \left\langle\Phi(x')\right|.$$

Using this expression we can rewrite the corresponding quantum linear classifier as

$$c_{\mathcal{O}_\alpha, d}(x) = \mathrm{sign}\big(\mathrm{Tr}\big[\rho_\Phi(x)\mathcal{O}_\alpha\big] - d\big) = \mathrm{sign}\Big(\sum_{x' \in \mathcal{D}} \alpha_{x'} \mathrm{Tr}\big[\rho_\Phi(x)\rho_\Phi(x')\big] - d\Big).$$

[2] Also called the *quantum kernel estimator* [1].

These type of linear classifiers can also be efficiently realized using parameterized quantum circuits. Using quantum protocols such as the SWAP-test or the Hadamard-test it is possible to efficiently evaluate the overlaps $\mathrm{Tr}[\rho_\Phi(x)\rho_\Phi(x')]$ for the feature map defined in Equation (5). Afterwards, the optimal parameters $\{\alpha_{x'}\}_{x'\in\mathcal{D}}$ are obtained on a classical computer, e.g., by solving a quadratic program. Altogether, this allows us to efficiently implement the family of linear classifiers – which we refer to as *implicit quantum linear classifiers* – given by

$$\mathcal{C}_{\mathrm{qlin}}^{\mathrm{implicit}} = \Big\{ c_{\mathcal{O}_\alpha,d}(x) = \mathrm{sign}\big(\mathrm{Tr}\big[\rho_\Phi(x)\mathcal{O}_\alpha\big] - d\big) \;\Big|\; \mathcal{O}_\alpha = \sum_{x'\in\mathcal{D}} \alpha_{x'}\rho_\Phi(x'),\ \alpha\in\mathbb{R}^{|\mathcal{D}|},\ d\in\mathbb{R} \Big\}. \tag{8}$$

The power of this model comes from the fact that evaluating the overlaps $\mathrm{Tr}\big[\rho_\Phi(x)\rho_\Phi(x')\big]$ is likely classically intractable for sufficiently complex feature maps [1], demonstrating that classical computers can likely neither train nor evaluate this quantum linear classifier efficiently. Moreover, any quantum linear classifier that is the minimizer of a loss functions that includes *regularization* of the Frobenius norm of the observable can be expressed as an implicit quantum linear classifier [17]. However, as we indicate later in Section 3.3, this does not mean that we can forego explicit quantum linear classifiers entirely, as in the explicit approach there are unique types of meaningful regularization for which there is no straightforwards correspondence to the implicit approach.

2.2 Structural risk minimization: generalization bounds and model selection

When looking for the optimal family of classifiers for a given learning problem, it is important to carefully select the family's *complexity* (also known as expressivity or capacity). For instance, in the case of linear classifiers, it is important to select what kind of hyperplanes one allows the classifier to use. Generally, the more complex the family is, the lower the training errors will be. However, if the family becomes overly complex, then it becomes more prone to worse generalization performance (i.e., due to overfitting). Structural risk minimization is a concrete method that balances this trade-off in order to obtain the best possible performance on unseen examples. Specifically, structural risk minimization aims to saturate well-established upper bounds on the expected error of the classifier that consist of the sum of two inversely related terms: a *training error* term, and a *complexity term* penalizing more complex models.

In statistical learning theory it is generally assumed that the data is sampled according to some underlying probability distribution P on $\mathcal{X}\times\{-1,+1\}$. The goal is to find a classifier that minimizes the probability that a random pair sampled according to P is misclassified. That is, the goal is to find a classifier $c_{f,d}(x) = \mathrm{sign}(f(x)-d)$ that minimize the *expected error* given by

$$\mathrm{er}_P(c_{f,d}) = \Pr_{(x,y)\sim P}\big(c_{f,d}(x)\neq y\big). \tag{9}$$

As one generally only has access to training examples $\mathcal{D} = \big\{(x_1,y_1),\ldots,(x_m,y_m)\big\}$ that are sampled according to the distribution P, it is not possible to compute er_P directly. Nonetheless, one can try to approximate Equation (9) using *training errors* such as

$$\widehat{\mathrm{er}}_{\mathcal{D}}(c_{f,d}) = \frac{1}{m}\Big|\big\{i \;\big|\; c_{f,d}(x_i)\neq y_i\big\}\Big|, \tag{10}$$

$$\widehat{\mathrm{er}}^{\gamma}_{\mathcal{D}}(c_{f,d}) = \frac{1}{m}\Big|\big\{i \;\big|\; y_i\cdot\big(f(x_i)-d\big)<\gamma\big\}\Big|,\ \ \gamma\in\mathbb{R}_{\geq 0}. \tag{11}$$

Intuitively, $\widehat{\mathrm{er}}_{\mathcal{D}}$ in Equation (10) represents the frequency of misclassified training examples, and $\widehat{\mathrm{er}}^{\gamma}_{\mathcal{D}}$ in Equation (11) represents the frequency of training examples that are either misclassified or are "within margin γ from being misclassified". In particular, for $\gamma = 0$ both training error estimates are identical (i.e., $\widehat{\mathrm{er}}_{\mathcal{D}} = \widehat{\mathrm{er}}^{0}_{\mathcal{D}}$). When selecting the optimal classifiers from a given model one typically searches for the classifier that minimizes the training error (in practice more elaborate and smooth loss functions are used), which is referred to as *empirical risk minimization*. The problem that structural risk minimization aims to tackle is how to optimally select a model such that one will have some guarantee that the training error will be close to the expected error.

Structural risk minimization uses expected error bounds – two of which we will discuss shortly – that involve a training error term, and a complexity term that penalizes more complex models. This complexity term usually scales with a certain measure of the complexity of the family of classifiers. A well known example of such a complexity measure is the Vapnik-Chervonenkis dimension.

Definition 1 (VC dimension [28]). *Let $\mathcal{C}$ be a family of functions on $\mathcal{X}$ taking values in $\{-1,+1\}$. We say that a set of points $X=\{x_1,\ldots,x_m\}\subset\mathcal{X}$ is shattered by $\mathcal{C}$ if for all $y\in\{-1,+1\}^m$, there exists a classifier $c_y\in\mathcal{C}$ that satisfies $c_y(x_i)=y_i$. The VC dimension of $\mathcal{C}$ defined as*

$$\mathrm{VC}\big(\mathcal{C}\big)=\max\big\{m\mid\exists\{x_1,\ldots,x_m\}\subset\mathcal{X}\text{ that is shattered by }\mathcal{C}\big\}.$$

Besides the VC dimension we also consider a complexity measure called the *fat-shattering dimension*, which can be viewed as a generalization of the VC dimension to real-valued functions. An important difference between the VC dimension and the fat-shattering dimension is that the latter also takes into account the so-called *margins* that the family of classifiers can achieve. Here the margin of a classifier $c_{f,d}(x)=\mathrm{sign}\big(f(x)-d\big)$ on a set of examples $\{x_i\}_{i=1}^m$ is given by $\min_i|f(x_i)-d|$. Throughout the literature, this is often referred to as the functional margin.

Definition 2 (Fat-shattering dimension [29]). *Let $\mathcal{F}$ be a family of real-valued functions on $\mathcal{X}$. We say that a set of points $X=\{x_1,\ldots,x_m\}\subset\mathcal{X}$ is γ-shattered by $\mathcal{F}$ if there exists an $s\in\mathbb{R}^m$ such that for all $y\in\{-1,+1\}^m$, there exists a function $f_y\in\mathcal{F}$ satisfying*

$$f_y(x_i)\begin{cases}\leq s_i-\gamma & \text{if } y_i=-1,\\ \geq s_i+\gamma & \text{if } y_i=+1.\end{cases}$$

The fat-shattering dimension of $\mathcal{F}$ is a function $\mathrm{fat}_{\mathcal{F}}:\mathbb{R}\to\mathbb{Z}_{\geq 0}$ that maps

$$\mathrm{fat}_{\mathcal{F}}(\gamma)=\max\big\{m\;\big|\;\exists\{x_1,\ldots,x_m\}\subset\mathcal{X}\text{ that is }\gamma\text{-shattered by }\mathcal{F}\big\}.$$

We will now state two expected error bounds that can be used to perform structural risk minimization. These error bounds theoretically quantify how an increase in model complexity (i.e., VC dimension or fat-shattering dimension) results in a worse expected error (i.e., due to overfitting). First, we state the expected error bound that involves the VC dimension.

Theorem 1 (Expected error bound using VC dimension [30]). *Consider a set of functions $\mathcal{C}$ on $\mathcal{X}$ taking values in $\{-1,+1\}$. Suppose $\mathcal{D}=\{(x_1,y_1),\ldots,(x_m,y_m)\}$ is sampled using m independent draws from P. Then, with probability at least $1-\delta$, the following holds for all $c\in\mathcal{C}$:*

$$\mathrm{er}_P(c)\leq\widehat{\mathrm{er}}_{\mathcal{D}}(c)+62\sqrt{\frac{k}{m}}+3\sqrt{\frac{\log(2/\delta)}{2m}}\tag{12}$$

where $k=\mathrm{VC}\big(\mathcal{C}\big)$.

Next, we state the expected error bound that involves the fat-shattering dimension. One possible advantage of using the fat-shattering dimension instead of the VC dimension is that it can take into account the margin that the classifier achieves on the training examples. This turns out to be useful since this margin can be used to more precisely fine-tune the expected error bound.

Theorem 2 (Expected error bound using fat-shattering dimension [16]). *Consider a set of real-valued functions $\mathcal{F}$ on $\mathcal{X}$. Suppose $\mathcal{D}=\{(x_1,y_1),\ldots,(x_m,y_m)\}$ is sampled using m independent draws from P. Then, with probability at least $1-\delta$, the following holds for all $c(x)=\mathrm{sign}\big(f(x)-d\big)$ with $f\in\mathcal{F}$ and $d\in\mathbb{R}$:*

$$\mathrm{er}_P(c)\leq\widehat{\mathrm{er}}^{\gamma}_{\mathcal{D}}(c)+\sqrt{\frac{2}{m}\Big(k\log(34em/k)\log_2(578m)+\log(4/\delta)\Big)}.\tag{13}$$

where $k=\mathrm{fat}_{\mathcal{F}}(\gamma/16)$.

Remark(s). *If the classifier can correctly classify all examples in $\mathcal{D}$, then the optimal choice of γ in the above theorem is the margin achieved on the examples in $\mathcal{D}$, i.e., $\gamma=\min_{x_i\in\mathcal{D}}\big|f(x_i)-d\big|$.*

Generally, the more complex a family of classifiers is, the larger its generalization errors are. This correlation between a family's complexity and its generalization errors is theoretically quantified in Theorems 1 and 2. Specifically, the more complex the family is the larger its VC dimension will be, which strictly increases the second term in Equation 12 that corresponds to the generalization error.

Note that for the fat-shattering dimension in Theorem 2 this is not as obvious. In particular, a more complex model could achieve a larger margin γ, which actually decreases the second term in Equation 13 that corresponds to the generalization error.

Theorems 1 and 2 establish that in order to minimize the expected error, we should aim to minimize either of the sums on the right-hand side of Equations (12) or (13) (depending on which complexity measure one wishes to focus on). Note that in both cases the first term corresponds to a training error and the second term corresponds to a complexity term that penalizes more complex models. Crucially, the effect that the complexity measure of the family of classifiers has on these terms is inversely related. Namely, a large complexity measure generally gives rise to smaller training errors, but at the cost of a larger complexity term. Balancing this trade-off is precisely the idea behind structural risk minimization. More precisely, structural risk minimization selects a classifier that minimizes either of the expected error bounds stated in Theorem 1 or 2, by selecting the classifier from a family whose complexity measure is fine-tuned in order to balance both terms on the right-hand side of Equations (12) or (13). Note that limiting the VC dimension and fat-shattering dimension does not achieve the same theoretical guarantees on the generalization error, and it will generally give rise to different performances in practice (as also discussed Section 3.2). An overview of the trade-off in the error bounds stated in Theorems 1 and 2 is depicted in Figure 3.

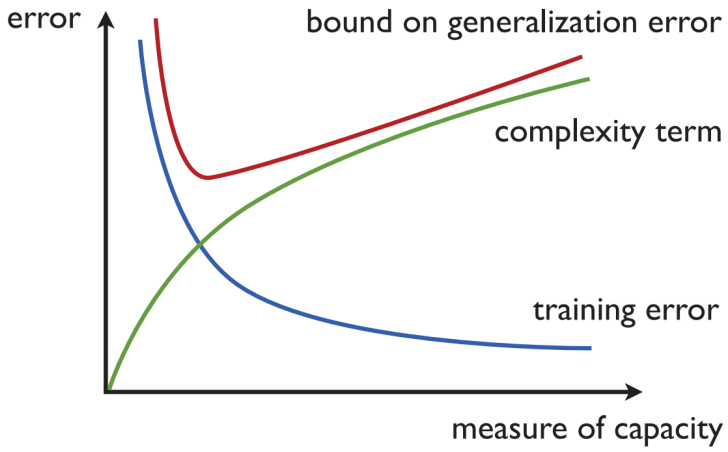

Figure 3: Illustration of structural risk minimization taken from [15]. Increasing the complexity of the classifier family causes the training error (blue) to decrease, while it increases the complexity term (green). Structural risk minimization selects the classifier minimizing the expected error bound in Eqs. (12) and (13) given by the sum of the training error and the complexity term (red).

3 Structural risk minimization for quantum linear classifiers

In this section we theoretically analyze and quantify the influence that model parameters of quantum linear classifiers have on the trade-off in structural risk minimization. We first analyze the effect that model parameters have on the complexity term (i.e., the green line in Figure 3) and afterwards we analyze their effect on the training error (i.e., the blue line in Figure 3). Specifically, in Section 3.1 we analyze the complexity term by establishing analytic upper bounds on complexity measures (i.e., the VC dimension and fat-shattering dimension) of quantum linear classifiers. In Section 3.2 we study the influence that model parameters which influence the established complexity measure bounds have on the training error term. Finally, in Section 3.3, we discuss how to implement structural risk minimization of quantum linear classifiers based on the obtained results.

3.1 Complexity of quantum linear classifiers: fat-shattering and VC dimension

In this section we determine the two complexity measures defined in the previous section – i.e., the fat-shattering dimension and VC dimension – for families of quantum linear classifiers. As a result, we identify model parameters that allow us to control the complexity term in the expected error bounds of

Theorems 1 and 2. In particular, these model parameters can therefore be used to balance the trade-off considered by structural risk minimization, as depicted in Figure 3. Throughout this section we fix the feature map to be the one defined Equation (5) and we allow our separating hyperplanes to come from a family of observables $\mathbb{O} \subseteq \mathrm{Herm}\left(\mathbb{C}^{2^n}\right)$ (e.g., the family of observables implementable using either the explicit or implicit realization of quantum linear classifiers). Our goal is to determine analytical upper bounds on complexity measures of the resulting family of quantum linear classifiers.

First, we show that the VC dimension of a family of quantum linear classifiers is upper bounded by the dimension of the span of the observables that it uses. This in turn is upper bounded by the square of the dimension of the space upon which the observables act nontrivially. We remark that while the VC dimension of quantum linear classifiers also has a clear dependence on the feature map, we chose to focus on the observables because the resulting upper bounds give rise to more explicit guidelines on how to tune the quantum model to perform structural risk minimization (as we discuss in more detail in Section 3.3). We defer the proof to Appendix A.1.

Proposition 3. *Let $\mathbb{O} \subseteq \mathrm{Herm}\left(\mathbb{C}^{2^n}\right)$ be a family of n-qubit observables with $r = \dim\left(\sum_{\mathcal{O}\in\mathbb{O}} \mathrm{Im}\,\mathcal{O}\right)$[3].. Then, the VC dimension of*

$$\mathcal{C}^{\mathbb{O}}_{\mathrm{qlin}} = \left\{ c(x) = \mathrm{sign}\big(\mathrm{Tr}\big[\mathcal{O}\rho_\Phi(x)\big] - d\big) \;\middle|\; \mathcal{O} \in \mathbb{O},\, d \in \mathbb{R} \right\} \tag{14}$$

satisfies

$$\mathrm{VC}\big(\mathcal{C}^{\mathbb{O}}_{\mathrm{qlin}}\big) \leq \dim\left(\mathrm{Span}\big(\mathbb{O}\big)\right) + 1 \leq r^2 + 1. \tag{15}$$

Remark(s). *The quantity r in the above proposition is related to the ranks of the observables. Specifically, note that for any two observables $\mathcal{O}, \mathcal{O}' \in \mathrm{Herm}\big(\mathbb{C}^{2^n}\big)$ we have that*

$$\dim\left(\mathrm{Im}\,\mathcal{O} + \mathrm{Im}\,\mathcal{O}'\right) = \mathrm{rank}\big(\mathcal{O}\big) + \mathrm{rank}\big(\mathcal{O}'\big) - \dim\big(\mathrm{Im}\,\mathcal{O} \cap \mathrm{Im}\,\mathcal{O}'\big).$$

The above proposition implies the (essentially obvious) result that VC dimension of a family of implicit quantum linear classifiers is upper bounded by the number of training examples (i.e., the operators $\{\rho_\Phi(x)\}_{x\in\mathcal{D}}$ span a subspace of dimension at most $\big|\mathcal{D}\big|$). We are however more interested in the application of the above proposition to explicit quantum linear classifiers. In this case, we choose to focus on the upper bound $r^2 + 1$ because it has interpretational advantages as to what parts of the model one has to tune from the perspective of structural risk minimization (i.e., recall from Section 3 that one way to perform structural risk minimization is to tune the VC dimension). Moreover, in the case of explicit quantum linear classifiers, the bound r^2+1 is only quadratically worse than the bound $\dim\left(\mathrm{Span}\big(\mathbb{O}\big)\right)+1$. To see this, we consider a family of explicit quantum linear classifiers with observables $\mathbb{O}_{\mathrm{explicit}} = \left\{\mathcal{O}^\lambda_\theta\right\}$, where

$$\mathcal{O}^\lambda_\theta = W^\dagger(\theta) \cdot \mathrm{diag}\big(\lambda(0), \ldots, \lambda(2^n - 1)\big) \cdot W(\theta)$$

and we denote $W(\theta)\,|i\rangle = |\psi_i(\theta)\rangle$. Next, suppose that $\lambda(j) = 0$ for all $j > L$ and define

$$H = \mathrm{Span}_{\mathbb{C}}\Big\{ |\psi_0(\theta)\rangle, \ldots, |\psi_L(\theta)\rangle \;\;:\;\; \theta \in \mathbb{R}^m \Big\}, \tag{16}$$

$$V = \mathrm{Span}_{\mathbb{R}}\Big\{ \sum_{i=0}^{L} \lambda(i)\, |\psi_i(\theta)\rangle\langle\psi_i(\theta)| \;\;:\;\; \theta \in \mathbb{R}^m \Big\}, \tag{17}$$

Then, Proposition 3 states that

$$\mathrm{VC}\big(\mathcal{C}^{\mathbb{O}_{\mathrm{explicit}}}_{\mathrm{qlin}}\big) \leq \dim\left(V\right) + 1 \leq \dim(H)^2 + 1.$$

Now, by the following lemma, we indeed find that the bound $r^2 + 1$ is only quadratically worse than the bound $\dim\left(\mathrm{Span}\big(\mathbb{O}\big)\right) + 1$. We again defer the proof to Apppendix A.1.

[3]Here $\sum$ denotes the sum of vector spaces and $\mathrm{Im}\,\mathcal{O}$ denotes the image (or column space) of the operator $\mathcal{O}$

Lemma 4. *The vector spaces defined in Eq. (16) and Eq. (17) satisfy*[4]

$$\dim(H) \leq \dim(V) \leq \dim(H)^2.$$

Therefore, if we sufficiently limit $r = \dim(H)$, then this also limits $\dim\big(\mathrm{Span}\big(\mathbb{O}\big)\big) = \dim(V)$. Moreover, even though $\dim\big(\mathrm{Span}\big(\mathbb{O}\big)\big) + 1$ can provide a tighter bound, it can still be advantageous to study the bound $r^2 + 1$ because it might have interpretational advantages. Specifically, it might be easier to construct cases of ansatze where the latter bound allows us to identify a controlable hyperparameter that controls the VC dimension (as we discuss in more detail in Section 3.3).

Note that the quantity r defined in the above proposition, depends on both the structure of the ansatz W as well as the post-processing function λ. One way to potentially limit r is by varying the rank of the final measurement (i.e., the value L defined above). However, for several ansatzes in literature, having either a low-rank or a high-rank final measurement will not make a difference in terms of the VC dimension bound $r^2 + 1$[5]. To see this, consider an ansatz consisting of a single layer of parameterized X-rotations on all qubits, where each rotation is given a separate parameter. Already for this simple ansatz even the first columns $\{\bigotimes_{i=1}^{n} X_i(\theta_i)\,|0\rangle \mid \theta \in [0, 2\pi)^n\}$ span the entire n-qubit Hilbert space. In particular, the above proposition gives the same VC dimension upper bound for the cases where the final measurement is of rank $L = 1$, and where it is of full rank $L = 2^n$ (i.e., we have no guarantee that limiting L limits the VC dimension). This motivates us to design ansatzes for which subsets of columns do not span the entire Hilbert space when varying the variational parameter θ. On the other hand, to exploit the bound $\dim\big(\mathrm{Span}\big(\mathbb{O}\big)\big) + 1$ one needs to consider the span of the projectors onto the first L columns in the vector space of Hermitian operators. This quantity can be slightly less intuitive than the span of the first L columns in the n-qubit Hilbert space, and in Section 3.3 we show that this latter quantity can already be used to affirm the effectiveness of certain regularization techniques. Specifically, in Section 3.3 we discuss examples of ansatzes for which subsets of columns do not span the entire Hilbert space when varying the variational parameter, and we explain how they allow for structural risk minimization by limiting the rank of the final measurement.

Next, we show that the fat-shattering dimension of a family of quantum linear classifiers is related to the Frobenius norm of the observables that it uses. In particular, we show that we can control the fat-shattering dimension of a family of quantum linear classifiers by limiting the Frobenius norm of its observables. We defer the proof to Appendix A.3, where we also discuss the implications of this result in the probably approximately correct (PAC) learning framework.

Proposition 5. *Let $\mathbb{O} \subseteq \mathrm{Herm}\left(\mathbb{C}^{2^n}\right)$ be a family of n-qubit observables with $\eta = \max_{\mathcal{O}\in\mathbb{O}} \|\mathcal{O}\|_F$. Then, the fat-shattering dimension of*

$$\mathcal{F}^{\mathbb{O}}_{\mathrm{qlin}} = \Big\{ f_{\mathcal{O},d}(x) = \mathrm{Tr}\big[\mathcal{O}\rho_\Phi(x)\big] - d \;\Big|\; \mathcal{O} \in \mathbb{O},\ d \in \mathbb{R} \Big\} \tag{18}$$

is upper bounded by

$$\mathrm{fat}_{\mathcal{F}^{\mathbb{O}}_{\mathrm{qlin}}}(\gamma) \leq O\left(\frac{\eta^2}{\gamma^2}\right). \tag{19}$$

Remark(s). *The upper bound in the above proposition matches the result discussed in [26]. This was derived independently by one of the authors of this paper [31], and we include it here for completeness.*

The above proposition shows that the fat-shattering dimension of a family of explicit quantum linear classifiers can be controlled by limiting $||\mathcal{O}^\lambda_\theta||_F = \sqrt{\sum_{i=1}^{2^n} \lambda(i)^2}$. In particular, it shows that the selection of the postprocessing function λ is important when tuning the complexity of the family ofr classifiers. Furthermore, the above proposition shows that the fat-shattering dimension of a family of implicit quantum linear classifiers can be controlled by limiting $||\mathcal{O}_\alpha||_F \leq ||\alpha||_1$. It is important to note that the Frobenius norm itself does not fully characterize the generalization performance of a family of quantum

[4]Note that there exists ansatzes for which the inequalities are strict, i.e., $\dim(H) < \dim(V) < \dim(H)^2$ (e.g., see the first example discussed in Section 3.3).

[5]The relationship between the quantity r and the ranks of the observable can be made explicit by considering the overlaps between the images of the observables. A more detailed explanation of this can be found in Appendix A.2.

linear classifiers. Specifically, plugging Theorem 5 into Proposition 2 we find that the generalization performance bounds depend on both the Frobenius norm as well as the functional margin on training examples[6]. Therefore, to optimize the generalization performance bounds one has to minimize the Frobenius norm, while ensuring the functional margin on training examples stays large. Note that one way to achieve this is by maximizing the so-called geometric margin, which on a set of example $\{x_i\}$ is given by $\min_i \left|\mathrm{Tr}\left[\mathcal{O}\rho_\Phi(x_i)\right] - d\right|/||\mathcal{O}||_F$.

3.2 Expressivity of quantum linear classifiers: model parameters & errors

Having established that the quantity r defined in Proposition 3 and the Frobenius norms of the observables influence the complexity of the family of quantum linear classifiers (i.e., the green line in Figure 3), we will now study the influence of these parameters on the training errors that the classifiers can achieve (i.e., the blue line in Figure 3). First, we study the influence of these model parameters on the ability of the classifiers to correctly classify certain sets of examples. Afterwards, we study the influence of these model parameters on the margins that the classifiers can achieve.

Recall from the previous section that the VC dimension of certain families of quantum linear classifiers depends on the rank of the observables that it uses. For instance, if the observables are such that their images are (largely) overlapping, then the quantity r defined in Proposition 3 can be controlled by limiting the ranks of all observables. In Section 3.3 we use this observation to construct ansatzes for which the VC dimension bound can be tuned by varying the rank of the observable measured on the output of the circuit. Since the VC dimension is only concerned with whether an example is correctly classified (and not what margin it achieves), we choose to investigate the influence of the rank on being able to correctly classify certain sets of examples. In particular, we show that any set of examples that can be correctly classified using a low-rank observable, can also be correctly classified using a high-rank observable. Moreover, we also show that there exist sets of examples that can only be correctly classified using observables of at least a certain rank. We defer the proof to Appendix B.1.

Proposition 6. *Let $\mathcal{C}^{(r)}_{\mathrm{qlin}}$ denote the family of quantum linear classifiers corresponding to observables of exactly rank r, that is,*

$$\mathcal{C}^{(r)}_{\mathrm{qlin}} = \left\{ c(\rho) = \mathrm{sign}\big(\mathrm{Tr}\left[\mathcal{O}\rho\right] - d\big) \;\middle|\; \mathcal{O} \in \mathrm{Herm}\big(\mathbb{C}^{2^n}\big),\, \mathrm{rank}\big(\mathcal{O}\big) = r,\, d \in \mathbb{R} \right\} \tag{20}$$

Then, the following statements hold:

(i) For every finite set of examples $\mathcal{D}$ that is correctly classified by a quantum linear classifier $c \in \mathcal{C}^{(k)}_{\mathrm{qlin}}$ with $0 < k < 2^n$, there exists a quantum linear classifier $c \in \mathcal{C}^{(r)}_{\mathrm{qlin}}$ with $r > k$ that also correctly classifies $\mathcal{D}$.

(ii) There exists a finite set of examples that can be correctly classified by a classifier $c \in \mathcal{C}^{(r)}_{\mathrm{qlin}}$, but which no classifier $c' \in \mathcal{C}^{(k)}_{\mathrm{qlin}}$ with $k < r$ can classify correctly.

Note that in the above proposition we define our classifiers in such a way that high-rank classifiers do not subsume low-rank classifiers. In particular, the family of observables that $\mathcal{C}^{(r)}_{\mathrm{qlin}}$ and $\mathcal{C}^{(k)}_{\mathrm{qlin}}$ use are completely disjoint for $k \neq r$. The construction behind the proof of the above proposition is inspired by tomography of observables. Specifically, we construct a protocol that queries a quantum linear classifier and based on the assigned labels checks whether the underlying observable is approximately equal to a fixed target observable of a certain rank. In particular, we can use this to test whether the underlying observable is really of a given rank, as no low-rank observable can agree with a high-rank observable on the assigned labels during this protocol. Note that if we could query the expectation values of the observable, then tomography would be straightforward. However, the classifier only outputs the sign of the expectation value, which introduces a technical problem that we circumvent. Our protocol could be generalized to a more complete tomographic-protocol which uses queries to a quantum linear classifier in order to find the spectrum of the underlying observable.

Next, we investigate the effect that limitations of the rank of the observables used by a family of quantum linear classifier have on its ability to implement certain families of standard linear classifiers.

[6]Recall that the functional margin of $c_{f,d}(x) = \mathrm{sign}\big(f(x) - d\big)$ on a set of examples $\{x_i\}$ is $\min_i |f(x_i) - d|$.

In particular, assuming that the feature map is bounded (i.e., all feature vectors have finite norm), then the following proposition establishes the following chain of inclusions:

$$\mathcal{C}_{\text{lin}} \text{ on } \mathbb{R}^{2^n} \subseteq \mathcal{C}_{\text{qlin}}^{(\leq 1)} \text{ on } n+1 \text{ qubits} \subseteq \cdots \subseteq \mathcal{C}_{\text{qlin}}^{(\leq r)} \text{ on } n+1 \text{ qubits} \subseteq \cdots \subseteq \mathcal{C}_{\text{lin}} \text{ on } \mathbb{R}^{4^n}, \tag{21}$$

where $\mathcal{C}_{\text{qlin}}^{(\leq r)}$ denotes the family of quantum linear classifiers using observables of rank at most r. Note that $\mathcal{C}_{\text{qlin}}^{(\leq r)} \subsetneq \mathcal{C}_{\text{qlin}}^{(\leq r+1)}$ is strict due to Proposition 6. We defer the proof to Appendix B.2.

Proposition 7. *Let $\mathcal{C}_{\text{lin}}(\Phi)$ denote the family of linear classifiers that is equipped with a feature map Φ. Also, let $\mathcal{C}_{\text{qlin}}^{(\leq r)}(\Phi')$ denote the family of quantum linear classifiers that uses observables of rank at most r and which is equipped with a quantum feature map Φ'. Then, the following statements hold:*

(i) For every feature map $\Phi : \mathbb{R}^\ell \to \mathbb{R}^N$ with $\sup_{x\in\mathbb{R}^\ell} ||\Phi(x)|| = M < \infty$, there exists a feature map $\Phi' : \mathbb{R}^\ell \to \mathbb{R}^{N+1}$ such that $||\Phi'(x)|| = 1$ for all $x \in \mathbb{R}^\ell$ and the families of linear classifiers satisfy $\mathcal{C}_{\text{lin}}(\Phi) \subseteq \mathcal{C}_{\text{lin}}(\Phi')$.

(ii) For every feature map $\Phi : \mathbb{R}^\ell \to \mathbb{R}^N$ with $||\Phi(x)|| = 1$ for all $x \in \mathbb{R}^\ell$, there exists a quantum feature map $\Phi' : \mathbb{R}^\ell \to \text{Herm}\left(\mathbb{C}^{2^n}\right)$ that uses $n = \lceil \log N + 1 \rceil + 1$ qubits such that the families of linear classifiers satisfy $\mathcal{C}_{\text{lin}}(\Phi) \subseteq \mathcal{C}_{\text{qlin}}^{(\leq 1)}(\Phi')$.

(iii) For every quantum feature map $\Phi : \mathbb{R}^\ell \to \text{Herm}\left(\mathbb{C}^{2^n}\right)$, there exists a classical feature map $\Phi' : \mathbb{R}^\ell \to \mathbb{R}^{4^n}$ such that the families of linear classifiers satisfy $\mathcal{C}_{\text{qlin}}(\Phi) = \mathcal{C}_{\text{lin}}(\Phi')$.

Recall from the previous section that the fat-shattering dimension of a family of linear classifiers depends on the Frobenius norm of the observables that is uses. In the following proposition we show that tuning the Frobenius norm changes the margins that the model can achieve, which gives rise to better generalization performance (as discussed in Section 2.2). In particular, we show that there exist sets of examples that can only be classified with a certain margin by a classifier that uses an observable of at least a certain Frobenius norm. We defer the proof to Appendix B.3.

Proposition 8. *Let $\mathcal{C}_{\text{qlin}}^{(\eta)}$ denote the family of quantum linear classifiers corresponding to all n-qubit observables of Frobenius norm η, that is,*

$$\mathcal{C}_{\text{qlin}}^{(\eta)} = \Big\{ c(\rho) = \text{sign}\big(\text{Tr}\big[\mathcal{O}\rho\big] - d\big) \;\Big|\; \mathcal{O} \in \text{Herm}\big(\mathbb{C}^{2^n}\big) \text{ with } ||\mathcal{O}||_F = \eta,\, d \in \mathbb{R} \Big\}. \tag{22}$$

Then, for every $\eta \in \mathbb{R}_{>0}$ and $0 < m \leq 2^n$ there exists a set of m examples consisting of binary labeled n-qubit pure states that satisfies the following two conditions:

(i) There exists a classifier $c \in \mathcal{C}_{\text{qlin}}^{(\eta)}$ that correctly classifies all examples with margin $\eta/\sqrt{m}$.

(ii) No classifier $c' \in \mathcal{C}_{\text{qlin}}^{(\eta')}$ with $\eta' < \eta$ can classify all examples correctly with margin $\geq \eta/\sqrt{m}$.

In conclusion, in Proposition 3 we showed that in certain cases the rank of the observables control the model's complexity (e.g., if the observables have overlapping images), and in Proposition 6 we showed that the rank also controls the model's ability to achieve small training errors. Moreover, in Proposition 8 we similarly showed that the Frobenius norm not only controls the model's complexity (see Proposition 5), but that it also controls the model's ability to achieve large functional margins. However, note that tuning each model parameter achieves a different objective. Namely, increasing the rank of the observable increases the ability to correctly classify sets of examples, whereas increasing the Frobenius norm of the observable increases the margins that it can achieve. For example, one can increase the Frobenius norm of an observable by multiplying it with a positive scalar which increases the margin it achieves, but in order to correctly classify the sets of examples discussed in Proposition 6 one actually has to increase the rank of the observable.

3.3 Structural risk minimization for quantum linear classifiers in practice

Having established how certain model parameters of quantum linear classifiers influence both the model's complexity and its ability to achieve small training errors, we now discuss how to use these results to implement structural risk minimization of quantum linear classifiers in practice. In particular, we will discuss a common approach to structural risk minimization called *regularization*. In short, what regularization entails is instead of minimizing only the training error E_{train}, one simultaneously minimizes an extra term $h(\omega)$, where h is a function that takes larger values for model parameters ω that correspond to more complex models. In this section, we discuss different types of regularization (i.e., different choices of the function h) that can be performed in the context of quantum linear classifiers based on the results of the previous section. These types of regularization help improve the performance of quantum linear classifiers in practice, without putting more stringent requirements on the quantum hardware and are thus NISQ-suitable.

To illustrate how Proposition 3 can be used to implement structural risk minimization in the explicit approach, consider the setting where we have a parameterized quantum circuit $W(\theta)$ (with $\theta \in \mathbb{R}^p$) followed by a fixed measurement that projects onto the first ℓ computational basis states. To use the bound $r^2 + 1$ from Proposition 3 one has to compute the quantity

$$\dim_{\mathbb{C}} \Big(\text{Span}_{\mathbb{C}} \big\{ |\psi_i(\theta)\rangle \;\; : \;\; i = 1, \dots \ell, \;\; \theta \in \mathbb{R}^p \big\} \Big), \tag{23}$$

where $|\psi_i(\theta)\rangle$ denotes the ith column of $W(\theta)$. To use the other bound $\dim\big(\text{Span}(\mathbb{O})\big) + 1$ from Proposition 3 one has to compute the quantity

$$\dim_{\mathbb{R}} \Big(\text{Span}_{\mathbb{R}} \big\{ \sum_{i=1}^{\ell} |\psi_i(\theta)\rangle \langle \psi_i(\theta)| \;\; : \;\; \theta \in \mathbb{R}^p \big\} \Big), \tag{24}$$

Although both are of course possible, in some cases it is slightly easier to see how the quantity in Eq. (23) scales with respect to ℓ. Specifically, utilizing the quantity in Eq. (23) already leads to interesting ansatze that allow for structural risk minimization by limiting ℓ. As discussed below Proposition 3, setting ℓ to be either large or small will not influence the upper bound on the VC dimension independently of the structure of the parameterized quantum circuit ansatz W. The proposition therefore motivates the design of ansatzes whose first ℓ columns define a manifold when varying the variational parameter that is contained in a relatively low-dimensional linear subspace. Specifically, in this case Proposition 3 results in nontrivial bounds on the VC dimension that we aim to control by varying ℓ. We now give three examples of ansatzes that allow one to control the upper bound on the VC dimension by varying ℓ. In particular, these ansatzes allow structural risk minimization to be implemented by regularizing with respect to the rank of the final measurement.

- For the first example, split up the qubits up in a "control register" of size c and a "target register" of size t (i.e., $n = t + c$). Next, let $C{-}U_i(\theta_i)$ denote the controlled gate that applies the t-qubit parameterized unitary $U_i(\theta_i)$ to the target register if the control register is in the state $|i\rangle$. Finally, consider the ansatz

  $$W(\theta) = \big[C{-}U_{2^c}(\theta_{2^c})\big] \cdot \ldots \cdot \big[C{-}U_1(\theta_1)\big].^{7}$$

  Note that the matrix of $W(\theta)$ is given by the block matrix

  $$W(\theta) = \begin{pmatrix} U_1(\theta_1) & & & \\ & U_2(\theta_2) & & \\ & & \ddots & \\ & & & U_{2^c}(\theta_{2^c}) \end{pmatrix}.$$

  For this choice of ansatz, if the final measurement projects onto $\ell = m2^t$ $(m < 2^c)$ computational basis states, then by Proposition 3 the VC dimension is at most $\ell^2 + 1$. Note that t is a controllable hyperparameter that can be used to tune the VC dimension. In particular, we can set it such

[7] We can control the depth of $W(\theta)$ by either limiting the size of the control register or by simply dropping some of the controlled parameterized unitaries (i.e., setting $U_i(\theta_i) = I$).

that the resulting VC dimension is not exponential in n. Let us now consider the other bound $\dim\big(\mathrm{Span}\big(\mathbb{O}\big)\big) + 1$ from Proposition 3. For this choice of ansatz, computing the quantity in Eq. (24) is also straightforward due to the block structure of the unitary. Moreover, for this choice of ansatz the inequalities in Lemma 4 are strict, which shows why being able to compute the quantity in Eq. (23) does not always imply that we can also compute the quantity in Eq. (24) (i.e., one is not simply the square of the other).

- For the second example, consider an ansatz that is composed of parameterized gates of the form $U(\theta) = e^{i\theta P}$ for some Pauli string $P \in \{X, Y, Z, I\}^{\otimes n}$. Specifically, consider the ansatz

$$W(\theta) = e^{i\theta_d P_d} \cdot \ldots \cdot e^{i\theta_1 P_1}.$$

  By the bound $r^2 + 1$ from Proposition 3, for this choice of ansatz if the final measurement projects onto ℓ computational basis states the VC dimension is at most $r^2 + 1$, where $r = \ell \cdot 2^d$. This bound is obtained by computing the quantity in Eq. (23), which can be done by noting that a column of the unitary $U(\theta)$ spans a subspace of dimension at most 2 when varying the variational parameter θ. Moreover, subsequent layers of $U(\theta)$ will only increase the dimension of the span of a column by at most a factor 2. Thus, when applying $U(\theta)$ a total of d times, the dimension of the span of any ℓ columns of $W(\theta)$ is at most $r = \ell \cdot 2^d$. Also in this construction we note that d is a controllable hyperparameter that can be used to tune the VC dimension. In particular, we can set it such that the resulting VC dimension is not exponential in n. For this particular choice of ansatze, computing the quantity in Eq. (24) might also be possible, but it is a bit more involved and not necessary for our main goal of establishing that ℓ controls the VC dimension. In particular, one might be able to compute the quantity in Eq. (24), but the bound $r^2 + 1$ from Proposition 3 already suffices to establish that ℓ is a tunable hyperparameter that controls the VC dimension.

- For the third example, we use symmetry considerations as a tool to control the VC dimension. First, partition the n-qubit register into disjoint subsets $I_1, \ldots, I_k$ of size $|I_j| = m_j$ (i.e., $\sum_j m_j = n$). Next, consider "permutation-symmetry preserving" parameterized unitaries on these partitions, which are defined as

$$S^+_{I_j}(\theta) = e^{i\theta \sum_{i \in I_j} P_i}, \quad \text{and} \quad S^{\otimes}_{I_j}(\theta) = e^{i\theta \prod_{i \in I_j} P_i},$$

  where we have say $P_i = X_i$, $P_i = Y_i$, $P_i = Z_i$ or $P_i = I$ for all $i \in I_j$ (i.e., the same operator acting on all qubits in the partition I_j). Note that if we apply these operators to a permutation invariant state on the m_j-qubits in the jth partition, then it remains permutation invariant (independent of θ). From these symmetric parameterized unitaries we construct parameterized layers $U(\theta_1, \ldots, \theta_k) = \prod_{j=1}^k S^{+/\otimes}_{I_j}(\theta_j)$, from which we construct the ansatz as

$$W(\theta) = U(\theta_1^d, \ldots, \theta_k^d) \cdot \ldots \cdot U(\theta_1^1, \ldots, \theta_k^1), \quad \theta \in [0, 1\pi)^{dk}.$$

  By the bound $r^2 + 1$ from Proposition 3, for this choice of ansatz if the final measurement projects onto ℓ computational basis states the VC dimension is at most $r^2 + 1$, where

$$r = \ell \cdot \prod_{j=1}^{k} (m_j + 1).$$

  This bound is obtained by computing the quantity in Eq. (23), which can be done by noting that if we apply a layer U to an n-qubit state that is invariant under permutations that only permute qubits within each partition, then it remains invariant under these permutations (i.e., independent of the choice of θ). In other words, the first column of $W(\theta)$ is always contained in the space of n-qubit states that are invariant under permutations that only permute qubits within each partition. Next, note that the dimension of the space of n-qubit states that are invariant under permutations that only permute qubits within each partition is equal to $\prod_{j=1}^k (m_j + 1)$. Finally, note that any other column of $W(\theta)$ spans a space whose dimension is at most that of the first column of $W(\theta)$ when varying θ. Thus, any ℓ columns of $W(\theta)$ span a space of dimension is most $r = \ell \cdot \prod_{j=1}^k (m_j + 1)$ when varying θ. Equivalent to the example above, for this particular choice

of ansatze, computing the quantity in Eq. (24) might also be possible, but it is again a bit more involved and not necessary for our main goal of establishing that ℓ controls the VC dimension. In particular, one might be able to compute the quantity in Eq. (24), but the bound $r^2 + 1$ from Proposition 3 again already suffices to establish that ℓ is a tunable hyperparameter that controls the VC dimension.

In all of the above cases we see that we can control the upper bound on the VC dimension by varying the rank of the final measurement ℓ. It is worth noting that in these cases the regularized explicit quantum linear classifiers will generally give rise to a different model then the implicit approach without any theoretical guarantee regarding which will do better, because the standard relationship between the two models [17] will not hold anymore (i.e., the regularized explicit model does not necessarily correspond to a kernel method anymore).

Secondly, recall that by tuning the Frobenius norms of the observables used by a quantum linear classifier, we can balance the trade-off between its fat-shattering dimension and its ability to achieve large margins. In particular, this shows that we can implement structural risk minimization of quantum linear classifiers with respect to the fat-shattering dimension by regularizing the Frobenius norms of the observables. Again, it is important to note that the Frobenius norm itself does not fully characterize the generalization performance, since one also has to take into account the functional margin on training examples. In particular, to optimize the generalization performance one has to minimize the Frobenius norm, while ensuring that the functional margin on training examples stays large. As mentioned earlier, one way to achieve this is by maximizing the geometric margin, which on a set of examples $\{x_i\}$ is given by $\min_i \left|\mathrm{Tr}\left[\mathcal{O}\rho_\Phi(x)\right] - d\right| / ||\mathcal{O}||_F$. As before, for explicit quantum linear classifiers, we can estimate the Frobenius norm by sampling random computational basis states and computing the average of the postprocessing function λ on them in order to estimate $||\mathcal{O}_\theta^\lambda||_F = \sqrt{\sum_{i=1}^{2^n} \lambda(i)^2}$ (note that in some cases the Frobenius norm can be computed more directly). On the other hand, for implicit quantum linear classifiers, we can regularize the Frobenius norm by regularizing $||\alpha||_1$ as $||\mathcal{O}_\alpha||_F \leq ||\alpha||_1$. However, if the weights are obtained by solving the usual quadratic program [1, 2], then the resulting observable is already (optimally) regularized with respect to the Frobenius norm [17].

Besides the types of regularization for which we have established theoretical evidence of the effect on structural risk minimization, there are also other types of regularization that are important to consider. For instance, for explicit quantum linear classifiers, one could regularize the angles of the parameterized quantum circuit [32]. Theoretically analyzing the effect that regularizing the angles of the parameterized quantum circuit has on structural risk minimization would constitute an interesting direction for future research. Another example is regularizing circuit parameters such as depth, width and number of gates for which certain theoretical results are known [19, 18]. Finally, it turns out that one can also regularize quantum linear classifiers by running the circuits under varying levels of noise [21]. For these kinds of regularization the relationships between the regularized explicit and regularized implicit quantum linear classifiers are still to be investigated.

Acknowledgments

The results of this work extend on the MSc thesis of Dyon van Vreumingen [31]. The authors thank Matthias C. Caro, Maria Schuld and Ryan Sweke for giving valuable comments on the manuscript. The authors thank Jordi Tura Brugués for discussions on the permutation-symmetry preserving ansatz. This work was supported by the Dutch Research Council (NWO/OCW), as part of the Quantum Software Consortium programme (project number 024.003.037).

References

[1] Vojtěch Havlíček, Antonio D Córcoles, Kristan Temme, Aram W Harrow, Abhinav Kandala, Jerry M Chow, and Jay M Gambetta. "Supervised learning with quantum-enhanced feature spaces". Nature**567** (2019). arXiv:1804.11326.

[2] Maria Schuld and Nathan Killoran. "Quantum machine learning in feature Hilbert spaces". Physical review letters**122** (2019). arXiv:1803.07128.

[3] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando GSL Brandao, David A Buell, et al. "Quantum supremacy using a programmable superconducting processor". Nature**574** (2019). arXiv:1910.11333.

[4] John Preskill. "Quantum computing in the NISQ era and beyond". Quantum**2** (2018). arXiv:1801.00862.

[5] Marco Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, et al. "Variational quantum algorithms". Nature Reviews Physics**3** (2021). arXiv:2012.09265.

[6] Jarrod R McClean, Jonathan Romero, Ryan Babbush, and Alán Aspuru-Guzik. "The theory of variational hybrid quantum-classical algorithms". New Journal of Physics**18** (2016). arXiv:1509.04279.

[7] Abhinav Kandala, Antonio Mezzacapo, Kristan Temme, Maika Takita, Markus Brink, Jerry M Chow, and Jay M Gambetta. "Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets". Nature**549** (2017). arXiv:1704.05018.

[8] Peter JJ O'Malley, Ryan Babbush, Ian D Kivlichan, Jonathan Romero, Jarrod R McClean, Rami Barends, Julian Kelly, Pedram Roushan, Andrew Tranter, Nan Ding, et al. "Scalable quantum simulation of molecular energies". Physical Review X**6** (2016). arXiv:1512.06860.

[9] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. "A quantum approximate optimization algorithm" (2014). arXiv:1411.4028.

[10] Marcello Benedetti, Erika Lloyd, Stefan Sack, and Mattia Fiorentini. "Parameterized quantum circuits as machine learning models". Quantum Science and Technology**4** (2019). arXiv:1906.07682.

[11] Barbara M Terhal and David P DiVincenzo. "Adaptive quantum computation, constant depth quantum circuits and arthur-merlin games". Quantum Information & Computation**4** (2004). arXiv:quant-ph/0205133.

[12] Michael J Bremner, Richard Jozsa, and Dan J Shepherd. "Classical simulation of commuting quantum computations implies collapse of the polynomial hierarchy". Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences**467** (2011). arXiv:1005.1407.

[13] Edward Grant, Marcello Benedetti, Shuxiang Cao, Andrew Hallam, Joshua Lockhart, Vid Stojevic, Andrew G Green, and Simone Severini. "Hierarchical quantum classifiers". npj Quantum Information**4** (2018). arXiv:1804.03680.

[14] Diego Ristè, Marcus P Da Silva, Colm A Ryan, Andrew W Cross, Antonio D Córcoles, John A Smolin, Jay M Gambetta, Jerry M Chow, and Blake R Johnson. "Demonstration of quantum advantage in machine learning". npj Quantum Information**3** (2017). arXiv:1512.06069.

[15] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. "Foundations of machine learning". MIT press. (2018).

[16] Peter L Bartlett. "The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network". IEEE transactions on Information Theory**44** (1998).

[17] Maria Schuld. "Supervised quantum machine learning models are kernel methods" (2021). arXiv:2101.11020.

[18] Matthias C Caro and Ishaun Datta. "Pseudo-dimension of quantum circuits". Quantum Machine Intelligence**2** (2020). arXiv:2002.01490.

[19] Kaifeng Bu, Dax Enshan Koh, Lu Li, Qingxian Luo, and Yaobo Zhang. "On the statistical complexity of quantum circuits" (2021). arXiv:2101.06154.

[20] Kaifeng Bu, Dax Enshan Koh, Lu Li, Qingxian Luo, and Yaobo Zhang. "Effects of quantum resources on the statistical complexity of quantum circuits" (2021). arXiv:2102.03282.

[21] Kaifeng Bu, Dax Enshan Koh, Lu Li, Qingxian Luo, and Yaobo Zhang. "Rademacher complexity of noisy quantum circuits" (2021). arXiv:2103.03139.

[22] Amira Abbas, David Sutter, Christa Zoufal, Aurélien Lucchi, Alessio Figalli, and Stefan Woerner. "The power of quantum neural networks". Nature Computational Science**1** (2021). arXiv:2011.00027.

[23] Yuxuan Du, Zhuozhuo Tu, Xiao Yuan, and Dacheng Tao. "Efficient measure for the expressivity of variational quantum algorithms". Physical Review Letters**128** (2022). arXiv:2104.09961.

[24] Leonardo Banchi, Jason Pereira, and Stefano Pirandola. "Generalization in quantum machine learning: A quantum information standpoint". PRX Quantum**2** (2021). arXiv:2102.08991.

[25] Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R McClean. "Power of data in quantum machine learning". Nature communications**12** (2021). arXiv:2011.01938.

[26] Yunchao Liu, Srinivasan Arunachalam, and Kristan Temme. "A rigorous and robust quantum speed-up in supervised machine learning". Nature Physics **17**, 1013–1017 (2021). arXiv:2010.02174.

[27] Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. "Learning with kernels: support vector machines, regularization, optimization, and beyond". MIT press. (2002).

[28] Vladimir N Vapnik and A Ya Chervonenkis. "On the uniform convergence of relative frequencies of events to their probabilities". In Measures of complexity. Springer (2015).

[29] Michael J Kearns and Robert E Schapire. "Efficient distribution-free learning of probabilistic concepts". Journal of Computer and System Sciences**48** (1994).

[30] Michael M Wolf. "Mathematical foundations of supervised learning". `https://www-m5.ma.tum.de/foswiki/pub/M5/Allgemeines/MA4801_2020S/ML_notes_main.pdf` (2020).

[31] Dyon van Vreumingen. "Quantum feature space learning: characterisation and possible advantages". Master's thesis. Leiden University. (2020).

[32] Jae-Eun Park, Brian Quanz, Steve Wood, Heather Higgins, and Ray Harishankar. "Practical application improvement to quantum svm: theory to practice" (2020). arXiv:2012.07725.

[33] John Shawe-Taylor, Peter L. Bartlett, Robert C. Williamson, and Martin Anthony. "Structural risk minimization over data-dependent hierarchies". IEEE Transactions on Information Theory (1998).

[34] Martin Anthony and Peter L Bartlett. "Function learning from interpolation". Combinatorics, Probability and Computing**9** (2000).

[35] Peter L Bartlett and Philip M Long. "Prediction, learning, uniform convergence, and scale-sensitive dimensions". Journal of Computer and System Sciences**56** (1998).

[36] Scott Aaronson. "The learnability of quantum states". Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences**463** (2007). arXiv:quant-ph/0608142.

A Proofs of Section 3.1

A.1 Proofs of Proposition 3 and Lemma 4

Proposition 3. *Let $\mathbb{O} \subseteq \mathrm{Herm}\left(\mathbb{C}^{2^n}\right)$ be a family of n-qubit observables with $r = \dim\left(\sum_{\mathcal{O}\in\mathbb{O}} \mathrm{Im}\mathcal{O}\right)$[8].. Then, the VC dimension of*

$$\mathcal{C}^{\mathbb{O}}_{\mathrm{qlin}} = \left\{ c(x) = \mathrm{sign}\big(\mathrm{Tr}\big[\mathcal{O}\rho_\Phi(x)\big] - d\big) \;\middle|\; \mathcal{O} \in \mathbb{O},\, d \in \mathbb{R} \right\} \tag{14}$$

satisfies

$$\mathrm{VC}\big(\mathcal{C}^{\mathbb{O}}_{\mathrm{qlin}}\big) \leq \dim\left(\mathrm{Span}\big(\mathbb{O}\big)\right) + 1 \leq r^2 + 1. \tag{15}$$

Proof. Define $V = \sum_{\mathcal{O}\in\mathbb{O}} \mathrm{Im}\mathcal{O} \subset \mathbb{C}^{2^n}$ and let P_V denote the orthogonal projector onto V. Let $\Phi : \mathcal{X} \to \mathrm{Herm}\big(\mathbb{C}^{2^n}\big)$ denote the feature map of $\mathcal{C}^{\mathbb{O}}_{\mathrm{qlin}}$ and define $\Phi' = P_V \Phi P_V$. Note that $\mathcal{C}^{\mathbb{O}}_{\mathrm{qlin}}(\Phi') = \mathcal{C}^{\mathbb{O}}_{\mathrm{qlin}}(\Phi)$. It is known that the VC dimension of linear classifiers on $\mathbb{R}^\ell$ is $\ell + 1$, and it is clear that $\mathrm{Herm}\big(V\big) \simeq \mathrm{Herm}\big(\mathbb{C}^r\big) \simeq \mathbb{R}^{r^2}$. Also, note that $\mathrm{Span}\big(\mathbb{O}\big)$ is a subspace of $\mathrm{Herm}\big(V\big)$. We therefore conclude that

$$\begin{aligned}\mathrm{VC}\big(\mathcal{C}^{\mathbb{O}}_{\mathrm{qlin}}(\Phi)\big) = \mathrm{VC}\big(\mathcal{C}^{\mathbb{O}}_{\mathrm{qlin}}(\Phi')\big) &\leq \mathrm{VC}\big(\text{linear classifiers on } \mathrm{Span}\big(\mathbb{O}\big)\big) = \dim\left(\mathrm{Span}\big(\mathbb{O}\big)\right) + 1 \\ &\leq \mathrm{VC}\big(\text{linear classifiers on } \mathrm{Herm}\big(V\big) \simeq \mathbb{R}^{r^2}\big) = r^2 + 1.\end{aligned}$$

□

Lemma 4. *The vector spaces defined in Eq. (16) and Eq. (17) satisfy*[9]

$$\dim(H) \leq \dim(V) \leq \dim(H)^2.$$

Proof. First, we note that V is contained in the space of Hermitian operators on H. Since the dimension of the space of Hermitian operators on H is equal to $\dim(H)^2$, this implies that

$$\dim(V) \leq \dim(H)^2.$$

Next, we fix a basis of H which we denote $\{|\psi_k\rangle\}_{k=1}^{\dim(H)}$, where we each $|\psi_k\rangle$ is of the form $|\psi_i(\theta)\rangle$ for some $i \in \{1, \ldots, L\}$ and $\theta \in \mathbb{R}^m$. To show that $\dim(V) \geq \dim(H)$, we will show that the operators $\{|\psi_k\rangle\langle\psi_k|\}_{k=1}^{\dim(H)} \subset V$ are linearly independent. We do so by contradiction, i.e., we assume they are not linearly independent and show that this leads to a contradiction. That is, we assume that there exists a $k' \in \{1, \ldots, \dim(H)\}$ and $\{\alpha_k\}_{k\neq k'} \subset \mathbb{R}$ such that

$$|\psi'_k\rangle\langle\psi'_k| = \sum_{k\neq k'} \alpha_k |\psi_k\rangle\langle\psi_k| .$$

This implies that

$$\begin{aligned}|\psi'_k\rangle &= \left(|\psi'_k\rangle\langle\psi'_k|\right)|\psi'_k\rangle \\ &= \Big(\sum_{k\neq k'} \alpha_k |\psi_k\rangle\langle\psi_k|\Big)|\psi'_k\rangle \\ &= \sum_{k\neq k'} \left(\alpha_k \cdot \langle\psi_k \mid \psi'_k\rangle\right)|\psi_k\rangle ,\end{aligned}$$

which shows that $\{|\psi_k\rangle\}_{k=1}^{\dim(H)}$ are not linearly independent. This clearly contradicts the assumption that $\{|\psi_k\rangle\}_{k=1}^{\dim(H)}$ is basis of H. We therefore conclude that the operators $\{|\psi_k\rangle\langle\psi_k|\}_{k=1}^{\dim(H)} \subset V$ are linearly independent, which shows that $\dim(V) \geq \dim(H)$. □

[8] Here $\sum$ denotes the sum of vector spaces and $\mathrm{Im}\,\mathcal{O}$ denotes the image (or column space) of the operator $\mathcal{O}$

[9] Note that there exists ansatzes for which the inequalities are strict, i.e., $\dim(H) < \dim(V) < \dim(H)^2$ (e.g., see the first example discussed in Section 3.3).

A.2 Relationship between VC dimension bound and ranks of the observables

In this section we discuss one possible way to relate the quantity r in Proposition 3 with the ranks of the observables by considering the overlaps of the images of the observables. Specifically, consider a family of observables $\{\mathcal{O}_i\}_{i=1}^n$, where each observable is of rank R[10]. Next, define the quantities

$$I_i = \dim\left(\operatorname{Im}\mathcal{O}_i \cap [\operatorname{Im}\mathcal{O}_{i+1} + \cdots + \operatorname{Im}\mathcal{O}_n]\right) \tag{25}$$

and

$$O_i = R - I_i. \tag{26}$$

Note that O_i measures the extent to which the image of the observable $\mathcal{O}_i$ overlaps with the images of the observables $\mathcal{O}_{i+1}, \ldots, \mathcal{O}_n$. Specifically, O_i is equal to zero if the images are fully overlapping, and it is equal to R if there is no overlap at all. Now Lemma 9 below provides a way to relate the quantity r in Proposition 3 with the ranks of the observables R and the overlaps of the images O_i. Note that we consider the case where the family of observables is finite, whereas in the case of explicit quantum linear classifiers this family is infinite. However, since all images live in a finite dimensional space, summing only finitely many images is already sufficient. More precisely, for any family of n-qubit observables $\mathbb{O}$ (possibly infinitely large) there exists a $\mathbb{O}' \subseteq \mathbb{O}$ with $|\mathbb{O}'| \leq 2^n$ and

$$\sum_{\mathcal{O}' \in \mathbb{O}'} \operatorname{Im}\mathcal{O}' = \sum_{\mathcal{O} \in \mathbb{O}} \operatorname{Im}\mathcal{O}.$$

In Lemma 9 below we can thus w.l.o.g. consider the case where the family of observables is finite.

Lemma 9. *Consider a family of observables $\mathbb{O} = \{\mathcal{O}_i\}_{i\in I}$, where each observable is of rank R. Then, for r defined in Proposition 3 and $\{O_i\}_{i\in I}$ defined in Eq. (26), we have that*

$$r = R + \sum_{i=1}^{n-1} O_i$$

Proof. The proof is basically a repeated application of the formula

$$\dim\left(\operatorname{Im}\mathcal{O}_1 + \operatorname{Im}\mathcal{O}_2\right) = \dim\left(\operatorname{Im}\mathcal{O}_1\right) + \dim\left(\operatorname{Im}\mathcal{O}_2\right) - \dim\left(\operatorname{Im}\mathcal{O}_1 \cap \operatorname{Im}\mathcal{O}_2\right).$$

Specifically, by repeatedly applying the above formula we find that

$$\begin{aligned}
r = \dim\left(\sum_{i=1}^{n} \operatorname{Im}\mathcal{O}_i\right) &= \dim\left(\operatorname{Im}\mathcal{O}_1\right) + \dim\left(\sum_{i=2}^{n} \operatorname{Im}\mathcal{O}_i\right) - \dim\left(\operatorname{Im}\mathcal{O}_1 \cap \sum_{i=2}^{n} \operatorname{Im}\mathcal{O}_i\right) \\
&= \dim\left(\operatorname{Im}\mathcal{O}_1\right) + \dim\left(\operatorname{Im}\mathcal{O}_2\right) + \dim\left(\sum_{i=3}^{n} \operatorname{Im}\mathcal{O}_i\right) \\
&\quad - \dim\left(\operatorname{Im}\mathcal{O}_1 \cap \sum_{i=2}^{n} \operatorname{Im}\mathcal{O}_i\right) - \dim\left(\operatorname{Im}\mathcal{O}_2 \cap \sum_{i=3}^{n} \operatorname{Im}\mathcal{O}_i\right) \\
&= nR - (I_1 + \cdots + I_{n-1}) \\
&= R - \sum_{i=1}^{n-1}(R - I_i) = R - \sum_{i=1}^{n-1} O_i
\end{aligned}$$

□

[10]The results in this section hold more generally for families with varying ranks, though for simplicity (and to more closely relate it to Proposition 6) we assume all observables have some fixed rank R (from which it should be clear how to adapt it to the case where the observables can have different ranks).

A.3 Proof of Proposition 5

Proposition 5. *Let $\mathbb{O} \subseteq \mathrm{Herm}\left(\mathbb{C}^{2^n}\right)$ be a family of n-qubit observables with $\eta = \max_{\mathcal{O}\in\mathbb{O}} \|\mathcal{O}\|_F$. Then, the fat-shattering dimension of*

$$\mathcal{F}^{\mathbb{O}}_{\mathrm{qlin}} = \left\{ f_{\mathcal{O},d}(x) = \mathrm{Tr}\left[\mathcal{O}\rho_\Phi(x)\right] - d \;\middle|\; \mathcal{O} \in \mathbb{O},\ d \in \mathbb{R} \right\} \tag{18}$$

is upper bounded by

$$\mathrm{fat}_{\mathcal{F}^{\mathbb{O}}_{\mathrm{qlin}}}(\gamma) \leq O\left(\frac{\eta^2}{\gamma^2}\right). \tag{19}$$

Proof. Due to the close relationship to standard linear classifiers, we can utilize previously obtained results in that context. In particular, for our approach we use the following proposition.

Proposition 10 (Fat-shattering dimension of linear functions [33])**.** *Consider the family of real-valued functions on the ball of radius R inside $\mathbb{R}^N$ given by*

$$\mathcal{F}_{\mathrm{lin}} = \left\{ f_{w,d}(x) = \langle w, x\rangle - d \;\middle|\; w \in \mathbb{R}^N \text{ with } ||w|| = 1,\ d \in \mathbb{R} \text{ with } |d| \leq R \right\}.$$

The fat-shattering dimension of $\mathcal{F}_{\mathrm{lin}}$ can be bounded by

$$\mathrm{fat}_{\mathcal{F}_{\mathrm{lin}}}(\gamma) \leq \min\{9R^2/\gamma^2, N+1\} + 1.$$

The context in the above proposition is closely related, yet slightly different than that of quantum linear classifiers. Firstly, n-qubit density matrices lie within the ball of radius $R = 1$ inside $\mathrm{Herm}\left(\mathbb{C}^{2^n}\right)$ equipped with the Frobenius norm. However, as in our case the hyperplanes arise from the family of observables $\mathbb{O}$, whose Frobenius norms are upper bounded by η, we cannot directly apply the above proposition. We therefore adapt the above proposition by exchanging the role of R with the upper bound on the norms of the observables in $\mathbb{O}$, resulting in the following lemma.

Lemma 11. *Consider the family of real-valued functions on the ball of radius $R = 1$ inside $\mathbb{R}^N$ given by*

$$\mathcal{F}^{\leq\eta}_{\mathrm{lin}} = \left\{ f_{w,d}(x) = \langle w, x\rangle - d \;\middle|\; w \in \mathbb{R}^N \text{ with } ||w|| \leq \eta,\ d \in \mathbb{R} \text{ with } |d| \leq \eta \right\}.$$

The fat shattering dimension of $\mathcal{F}^{\leq\eta}_{\mathrm{lin}}$ can be upper bounded by

$$\mathrm{fat}_{\mathcal{F}^{\leq\eta}_{\mathrm{lin}}}(\gamma) \leq \min\{9\eta^2/\gamma^2, N+1\} + 1.$$

Proof. Let us first determine the fat-shattering dimension of the family of linear functions with norm precisely equal to η on points that lie within the ball of radius $R = 1$, i.e.,

$$\mathcal{F}^{=\eta}_{\mathrm{lin}} = \left\{ f_{w,d}(x) = \langle w, x\rangle - d \;\middle|\; w \in \mathbb{R}^N \text{ with } ||w|| = \eta,\ d \in \mathbb{R} \text{ with } |d| \leq \eta \right\}.$$

Suppose $\mathcal{F}^{=\eta}_{\mathrm{lin}}$ can γ-shatter a set of points $\{x_1, \ldots, x_k\}$ that lie within the ball of radius $R = 1$. Because $\langle w, x_i\rangle = \langle w/\eta, \eta x_i\rangle$, we find that $\mathcal{F}^{=1}_{\mathrm{lin}}$ can γ-shatter the set of points $\eta x_1, \ldots, \eta x_k$ that lie within the ball of radius $R = \eta$. By Proposition 10 we have $k \leq \min\{9\eta^2/\gamma^2, N+1\} + 1$. Thus, the fat-shattering dimension of $\mathcal{F}^{=\eta}_{\mathrm{lin}}$ on points within the ball of radius $R = 1$ is upper bounded by

$$\mathrm{fat}_{\mathcal{F}^{=\eta}_{\mathrm{lin}}}(\gamma) \leq \min\{9\eta^2/\gamma^2, N+1\} + 1.$$

To conclude the desired results, note that this bound is monotonically increasing in η, and thus allowing hyperplanes with with norm $\|w\| < \eta$ will not increase the fat-shattering dimension.

□

From the above lemma we can immediately infer an upper bound on the fat-shattering dimension of quantum linear classifiers by identifying that as vector spaces $\mathrm{Herm}\left(\mathbb{C}^{2^n}\right) \simeq \mathbb{R}^{4^n}$.

□

A.3.1 Sample complexity in the PAC-learning framework

Besides being related to generalization performance, the fat-shattering dimension is also related to the so-called *sample complexity* in the probably approximately correct (PAC) learning framework [29]. The sample complexity captures the amount classifier queries required to find another classifier that with high probability agrees with the former classifier on unseen examples.

By plugging the upper bound of Proposition 5 into previously established theorems on the sample complexity of families of classifiers [34, 35], we derive the following corollary, which can be viewed as a dual of the result of [36].

Corollary 12. *Let $\mathbb{O} \subseteq \mathrm{Herm}\left(\mathbb{C}^{2^n}\right)$ be a family of observables with $\eta = \max_{\mathcal{O}\in\mathbb{O}} \|\mathcal{O}\|_F$ and consider the family of real-valued functions $\mathcal{F}^{\mathbb{O}}_{\mathrm{qlin}}$ defined in Eq. (18). Fix an element $F \in \mathcal{F}^{\mathbb{O}}_{\mathrm{qlin}}$ as well as parameters $\varepsilon, \nu, \gamma > 0$ with $\gamma\varepsilon \geq 7\nu$. Suppose we draw m examples $\mathcal{D} = \{\rho_1, \ldots, \rho_m\}$ independently according to a distribution P, and then choose any function $H \in \mathcal{F}^{\mathbb{O}}_{\mathrm{qlin}}$ such that $|H(\rho_i) - F(\rho_i)| \leq \nu$ for all $\rho_i \in \mathcal{D}$. Then, with probability at least $1-\delta$ over P, we have that*

$$\Pr_{\rho\sim P}\big(|H(\rho) - F(\rho)| > \gamma\big) \leq \varepsilon,$$

provided that

$$m \in \Omega\bigg(\frac{1}{\gamma^2\varepsilon^2}\Big(\frac{\eta^2}{\gamma^2\varepsilon^2}\log^2\frac{1}{\gamma\varepsilon} + \log\frac{1}{\delta}\Big)\bigg).$$

Proof. Follows directly from plugging the uppper bound of Proposition 5 into Corollary 2.4 of [36]. □

B Proofs of propositions Section 3.2

B.1 Proof of Proposition 6

Proposition 6. *Let $\mathcal{C}^{(r)}_{\mathrm{qlin}}$ denote the family of quantum linear classifiers corresponding to observables of exactly rank r, that is,*

$$\mathcal{C}^{(r)}_{\mathrm{qlin}} = \Big\{c(\rho) = \mathrm{sign}\big(\mathrm{Tr}\big[\mathcal{O}\rho\big] - d\big) \;\big|\; \mathcal{O} \in \mathrm{Herm}\big(\mathbb{C}^{2^n}\big),\, \mathrm{rank}\big(\mathcal{O}\big) = r,\, d \in \mathbb{R}\Big\} \tag{20}$$

Then, the following statements hold:

(i) For every finite set of examples $\mathcal{D}$ that is correctly classified by a quantum linear classifier $c \in \mathcal{C}^{(k)}_{\mathrm{qlin}}$ with $0 < k < 2^n$, there exists a quantum linear classifier $c \in \mathcal{C}^{(r)}_{\mathrm{qlin}}$ with $r > k$ that also correctly classifies $\mathcal{D}$.

(ii) There exists a finite set of examples that can be correctly classified by a classifier $c \in \mathcal{C}^{(r)}_{\mathrm{qlin}}$, but which no classifier $c' \in \mathcal{C}^{(k)}_{\mathrm{qlin}}$ with $k < r$ can classify correctly.

Proof. (i): Suppose $c_{\mathcal{O},b} \in \mathcal{C}^{(k)}_{\mathrm{qlin}}$ correctly classifies $\mathcal{D}$. Let $\delta = \min_{x\in\mathcal{D}_-}\big|\mathrm{Tr}\big[\mathcal{O}\rho_x\big] - d\big|$, where $\mathcal{D}_-$ is the subset of examples with label -1, and note that since $\mathcal{D}$ is correctly classified we have $\delta > 0$. Fix the basis we work in to be the eigenbasis of $\mathcal{O}$ ordered in such a way that

$$\mathcal{O} = \mathrm{diag}(\lambda_1, \ldots, \lambda_k, 0, \ldots, 0)$$

and define

$$P = \frac{1}{r-k}\mathrm{diag}(\underbrace{0,\ldots,0}_{k \text{ times}}, \underbrace{1,\ldots,1}_{r-k \text{ times}}, \underbrace{0,\ldots,0}_{2^n-r \text{ times}}).$$

For every $0 < \varepsilon < \delta$ we have that $\mathcal{O}' = \mathcal{O} + \varepsilon P$ has $\mathrm{rank}(\mathcal{O}') = r$. What remains to be shown is that $c_{\mathcal{O}',b} \in \mathcal{C}^{(r)}_{\mathrm{qlin}}$ correctly classifies $\mathcal{D}$. To do so, first let $x \in \mathcal{D}_+$ (i.e., labeled $+1$) and note that

$$\mathrm{Tr}\big[\mathcal{O}'\rho_x\big] - b = \underbrace{\big(\mathrm{Tr}\big[\mathcal{O}\rho_x\big] - b\big)}_{\geq 0} + \underbrace{\varepsilon\mathrm{Tr}\big[P\rho_x\big]}_{\geq 0} \geq 0,$$

which shows that indeed $c_{\mathcal{O}',b}(x) = +1$. Next, let $x \in \mathcal{D}_-$ (i.e., labeled -1) and note that

$$\mathrm{Tr}\big[\mathcal{O}'\rho_x\big] - b = \underbrace{\big(\mathrm{Tr}\big[\mathcal{O}\rho_x\big] - b\big)}_{\leq -\delta} + \underbrace{\varepsilon \mathrm{Tr}\big[P\rho_{x^+}\big]}_{<\delta} < 0,$$

which shows that indeed $c_{\mathcal{O}',b}(x) = -1$.

(ii):

We will describe a protocol that queries a classifier $c_{\mathcal{O},b}$ and based on its outcomes checks whether $\mathcal{O}$ is approximately equal to a fixed target observable $\mathcal{T}$ of rank r. We will show that if the queries to $c_{\mathcal{O},b}$ are labeled in a way that agrees with the target classifier that uses the observable $\mathcal{T}$, then the spectrum of $\mathcal{O}$ has to be point-wise within distance ε of the spectrum of $\mathcal{T}$. In particular, this will show that the rank of $\mathcal{O}$ has to be at least r if we make ε small enough. Consequently, if the rank of $\mathcal{O}$ is less than r, then at least one query made during the protocol has to be labeled differently by $c_{\mathcal{O},b}$ than the target classifier. In the end, the queries made to the classifier during the protocol will therefore constitute the set of examples described in the theorem.

Let us start with some definition. For a classifier $c_{\mathcal{O},b}(\rho) = \mathrm{sgn}\big(\mathrm{Tr}\big[\mathcal{O}\rho\big] - b\big)$ we define its effective observable $\mathcal{O}_{\text{eff}} = \mathcal{O} - bI$ which we express in the computational basis as $\mathcal{O}_{\text{eff}} = (O_{ij})$. Next, we define our target classifier to be $c_{\mathcal{T},-1}$ where the observable $\mathcal{T}$ is given by

$$\mathcal{T} = -r \cdot |0\rangle\langle 0| + \sum_{i=1}^{r-1} i \cdot |i\rangle\langle i|,$$

and we define its effective observable $\mathcal{T}_{\text{eff}} = \mathcal{T} + I$ which we express in the computational basis as $\mathcal{T}_{\text{eff}} = (T_{ij})$. Rescaling $\mathcal{O}_{\text{eff}}$ with a positive scalar does not change the output of the corresponding classifier. Therefore, to make the protocol well-defined, we define $\mathcal{O}_{\text{eff}}$ to be the unique effective observable whose first diagonal element is scaled to be equal to $O_{00} = -(r+1)$.

Our approach is as follows. First, we query $c_{\mathcal{O},b}$ in such a way that if the outcomes agree with with the target classifier $c_{\mathcal{T},-1}$, then the absolute values of the off-diagonal entries in the first row and column of $\mathcal{O}_{\text{eff}}$ must be close to zero (i.e., approximately equal to those of $\mathcal{T}_{\text{eff}}$). Afterwards, we again query $c_{\mathcal{O},b}$ but now in such a way that if the outcomes agree with the target classifier $c_{\mathcal{T},-1}$, then the diagonal elements of $\mathcal{O}_{\text{eff}}$ must be approximately equal to those of $\mathcal{T}_{\text{eff}}$. In the end, we query $c_{\mathcal{O},b}$ one final time but this time in such a way that if the outcomes agree with the target classifier $c_{\mathcal{T},-1}$, then the absolute values of the remaining off-diagonal elements of $\mathcal{O}_{\text{eff}}$ must be close to zero (i.e., again approximately equal to those of $\mathcal{T}_{\text{eff}}$). Finally, we use Gershgorin's circle theorem to show that the spectrum of $\mathcal{O}_{\text{eff}}$ has to be point-wise close to the spectrum of $\mathcal{T}_{\text{eff}}$. We remark that this procedure could be generalized to a more complete tomography approach, where one uses queries to the classifier $c_{\mathcal{O},b}$ in order to reconstruct the entire spectrum of $\mathcal{O}_{\text{eff}}$.

First, we query the quantum states $|i\rangle$ for $i = 0, \ldots, 2^n - 1$. Without loss of generality, we can assume that the classifiers $c_{\mathcal{O},b}$ and $c_{\mathcal{T},-1}$ agree on the label, i.e.,

$$c_{\mathcal{O},b}\big(|0\rangle\langle 0|\big) = -1, \text{ and } c_{\mathcal{O},b}\big(|i\rangle\langle i|\big) = +1 \text{ for } i = 1, \ldots, 2^n - 1, \tag{27}$$

as otherwise a set of examples containing just these states would already separate $c_{\mathcal{O},b}$ and $c_{\mathcal{T},-1}$.

In order to show that the absolute value of the off-diagonal elements of the first row and column of $\mathcal{O}_{\text{eff}}$ must be close to zero and that the diagonal elements of $\mathcal{O}_{\text{eff}}$ must be close to those of $\mathcal{T}_{\text{eff}}$, we consider the quantum states given by

$$|\gamma_\theta(\alpha)\rangle = \sqrt{1-\alpha}\,|0\rangle + e^{i\theta}\sqrt{\alpha}\,|j\rangle, \quad \text{with } \alpha \in [0,1] \text{ and } \theta \in [0, 2\pi). \tag{28}$$

Its expectation value with respect to $\mathcal{O}_{\text{eff}}$ is given by

$$\langle\gamma_\theta(\alpha)|\,\mathcal{O}_{\text{eff}}\,|\gamma_\theta(\alpha)\rangle = (1-\alpha)\cdot O_{00} + \alpha \cdot O_{jj} + \sqrt{\alpha(1-\alpha)}\cdot C_\theta, \quad \text{where } C_\theta := \mathrm{Re}\big(e^{i\theta}O_{0j}\big), \tag{29}$$

and its expectation value with respect to $\mathcal{T}_{\text{eff}}$ is given by

$$\langle\gamma_\theta(\alpha)|\,\mathcal{T}_{\text{eff}}\,|\gamma_\theta(\alpha)\rangle = (1-\alpha)\cdot T_{00} + \alpha \cdot T_{jj}. \tag{30}$$

Crucially, by Equation (27) we know that the label of $|\gamma_\theta(\alpha)\rangle$ goes from -1 to $+1$ as α goes $0 \to 1$. Note that the expectation value of $|\gamma_\theta(\alpha)\rangle$ with respect to $\mathcal{T}_{\text{eff}}$ is independent from the phase θ.

To determine that $\left|O_{0j}\right|$ is smaller than $\delta > 0$, we query the classifier $c_{\mathcal{O},b}$ on the states $|\gamma_{\hat{\theta}}(\hat{\alpha})\rangle$ for all $\hat{\theta}$ in a ζ-mesh of $[0, 2\pi)$ and for all $\hat{\alpha}$ in a ξ-mesh of $[0, 1]$ and we suppose they are labeled the same as the target classifier $c_{\mathcal{T},-1}$ would label them. Using these queries we can find estimates $\hat{\alpha}^{\mathcal{O}_{\text{eff}}}_{\text{cross}}(\hat{\theta})$ that are ξ-close to the unique $\alpha^{\mathcal{O}_{\text{eff}}}_{\text{cross}}(\theta) = \alpha'$ that satisfies

$$\langle\gamma_\theta(\alpha')|\,\mathcal{O}_{\text{eff}}\,|\gamma_\theta(\alpha')\rangle = 0, \tag{31}$$

by finding the smallest $\hat{\alpha}$ where the label has gone from -1 to $+1$. We refer to the α' satisfying Equation (31) as the *crossing point at phase* θ. Because the label assigned by $c_{\mathcal{T},-1}$ does not depend on the phase θ, and since all states $|\gamma_{\hat{\theta}}(\hat{\alpha})\rangle$ were assigned the same label by $c_{\mathcal{O},b}$ and $c_{\mathcal{T},-1}$, we find that the crossing point estimate $\hat{\alpha}^{\mathcal{O}_{\text{eff}}}_{\text{cross}}(\hat{\theta})$ is the same for all $\hat{\theta}$. In particular, this implies that the actual crossing points $\alpha^{\mathcal{O}_{\text{eff}}}_{\text{cross}}(\hat{\theta})$ have to be within ξ-distance of each other for all $\hat{\theta}$.

Before we continue, we first show that if $c_{\mathcal{O},b}$ assigns the same labels as $c_{\mathcal{T},-1}$, then O_{jj} is bounded above by a quantity that only depends on n. Fix $\tilde{\theta}$ to be any point inside the ζ-mesh such that $C_{\tilde{\theta}} \leq 0$, and define the function $E(\alpha) = (1-\alpha)\cdot O_{00} + \alpha \cdot O_{jj} + \sqrt{(1-\alpha)\alpha}\cdot C_{\tilde{\theta}}$. By our choice of $\mathcal{T}$, we have that $\alpha^{\mathcal{T}}_{\text{cross}} \in (\frac{r+1}{2r+1}, \frac{r+1}{r+3})$. Therefore, if $c_{\mathcal{O},b}$ and $c_{\mathcal{T},-1}$ agree on the entire ξ-mesh for a small enough ξ, then it must hold that $\alpha^{\mathcal{O}_{\text{eff}}}_{\text{cross}}(\tilde{\theta}) \in (\frac{1}{2}, \frac{2^n+1}{2^n+2})$. By the mean value theorem there exists an $\alpha' \in (\alpha^{\mathcal{O}_{\text{eff}}}_{\text{cross}}(\tilde{\theta}), \frac{2^n+1}{2^n+2})$ such that

$$E'(\alpha') = \frac{E(\frac{2^n+1}{2^n+2}) - E(\alpha^{\mathcal{O}_{\text{eff}}}_{\text{cross}}(\tilde{\theta}))}{\frac{2^n+1}{2^n+2} - \alpha^{\mathcal{O}_{\text{eff}}}_{\text{cross}}(\tilde{\theta})}. \tag{32}$$

After some rewriting, we can indeed conclude from the above equation that O_{jj} is bounded above by a quantity that only depends on n.

Next, write $O_{0j} = \left|O_{0j}\right|e^{i\phi}$ with $\phi \in [0, 2\pi)$, let $\hat{\theta}_{\text{abs}}$ denote the point in the ζ-mesh of $[0, 2\pi)$ that is closest to $2\pi - \phi$, and let $\hat{\theta}_0$ denote the point in the ζ-mesh of $[0, 2\pi)$ that is closest to $\pi/2 - \phi$ modulo 2π. By our previous discussion we know that $\left|\alpha^{\mathcal{O}_{\text{eff}}}_{\text{cross}}(\hat{\theta}_{\text{abs}}) - \alpha^{\mathcal{O}_{\text{eff}}}_{\text{cross}}(\hat{\theta}_0)\right| < \xi$, which together with the previously established bound on O_{jj} implies that

$$\left|C_{\hat{\theta}_{\text{abs}}} - C_{\hat{\theta}_0}\right| < f(\xi), \tag{33}$$

where f is a continuous function (independent from $c_{\mathcal{O},b}$) with $f(\xi) \to 0$ as $\xi \to 0$. Moreover, using the inequality $\cos(\zeta) \geq 1 - \lambda\zeta$, where $\lambda \approx 0.7246$ is a solution of $\lambda\big(\pi - \arcsin(\lambda)\big) = 1 + \sqrt{1-\lambda^2}$, together with the inequality $\cos(\pi/2 - \zeta) \leq \zeta$, we can derive that

$$\begin{aligned}
\left|C_{\hat{\theta}_{\text{abs}}} - C_{\hat{\theta}_0}\right| &= \Big|\left|O_{0j}\right|\cos\left(\hat{\theta}_{\text{abs}} + \phi\right) - \left|O_{0j}\right|\cos\left(\hat{\theta}_0 + \phi\right)\Big| \\
&\geq \left|O_{0j}\right| \cdot \Big|\cos\left(\zeta\right) - \cos\left(\pi/2 - \zeta\right)\Big| \\
&\geq \left|O_{0j}\right| \cdot \Big|1 - \big(\lambda + 1\big)\zeta\Big|.
\end{aligned} \tag{34}$$

Finally, by combining Equation (33) with Equation (34) we can conclude that

$$\left|O_{0j}\right| < \frac{f(\xi)}{1 - (\lambda+1)\zeta},$$

which for ξ and ζ small enough shows that $\left|O_{0j}\right| < \delta$ for any chosen precision $\delta > 0$ (i.e., the fineness of both meshes ξ and ζ will depend on the choice of δ).

To determine that O_{jj} is within distance $\delta' > 0$ of T_{jj} we again query the classifier $c_{\mathcal{O},b}$ but this time on the states $|\gamma_0(\hat{\alpha})\rangle$ for all $\hat{\alpha}$ in a ξ'-mesh of $[0, 1]$ and we suppose they are labeled the same as the target classifier $c_{\mathcal{T},-1}$ would. Using these queries we can find estimates $\hat{\alpha}^{\mathcal{O}_{\text{eff}}}_{\text{cross}}(0)$, $\hat{\alpha}^{\mathcal{T}_{\text{eff}}}_{\text{cross}}(0)$ that are ξ'-close to the corresponding actual crossing point. As we assumed that all queries are labeled the same by $c_{\mathcal{O},b}$ and $c_{\mathcal{T},-1}$, the crossing point estimate $\hat{\alpha}^{\mathcal{O}_{\text{eff}}}_{\text{cross}}(0)$ has to be equal to the crossing point estimate

$\hat{\alpha}^{\mathcal{T}_{\text{eff}}}_{\text{cross}}(0)$. In particular, this implies that the actual crossing points $\alpha^{\mathcal{O}_{\text{eff}}}_{\text{cross}}(0)$ and $\alpha^{\mathcal{T}_{\text{eff}}}_{\text{cross}}(0)$ have to be within ξ'-distance of each other. Next, define $g(\alpha, C)$ to be the unique coefficient $O \in \mathbb{R}_{\geq 0}$ that satisfies

$$(1-\alpha)\cdot O_{00} + \alpha \cdot O + \sqrt{\alpha(1-\alpha)}\cdot C = 0.$$

It is clear that g is a continuous function in α and C that is independent from $c_{\mathcal{O},b}$, and that $T_{jj} = g(\alpha^{\mathcal{T}_{\text{eff}}}_{\text{cross}}(0), 0)$ and $O_{jj} = g(\alpha^{\mathcal{O}_{\text{eff}}}_{\text{cross}}(0), C_0)$. Finally, we let $\delta > 0$ and $\xi' > 0$ be small enough such that if $\left|\alpha^{\mathcal{O}_{\text{eff}}}_{\text{cross}}(0) - \alpha^{\mathcal{T}_{\text{eff}}}_{\text{cross}}(0)\right| < \xi'$ and $\left|C_0\right| < \delta$, then

$$\left|O_{jj} - T_{jj}\right| = \left|g(\alpha^{\mathcal{T}_{\text{eff}}}_{\text{cross}}(0), 0) - g(\alpha^{\mathcal{O}_{\text{eff}}}_{\text{cross}}(0), C_0)\right| < \delta'.$$

In conclusion, to determine that O_{jj} is within distance $\delta' > 0$ of T_{jj} we first do the required queries to determine that $\left|C_0\right| = \left|O_{0j}\right| < \delta$, after which we do the required queries to determine that $\left|\alpha^{\mathcal{O}_{\text{eff}}}_{\text{cross}}(0) - \alpha^{\mathcal{T}_{\text{eff}}}_{\text{cross}}(0)\right| < \xi'$, which together indeed implies that O_{jj} is within distance $\delta' > 0$ of T_{jj}.

In order to show that the absolute value of the remaining off-diagonal elements of $\mathcal{O}_{\text{eff}}$ must be close to zero (i.e., close to those of $\mathcal{T}_{\text{eff}}$) we consider the quantum states given by

$$|\mu_\theta(\alpha)\rangle = \frac{\sqrt{1-\alpha}}{\sqrt{2}}\big(\,|0\rangle + |i\rangle\,\big) + e^{i\theta}\sqrt{\alpha}\,|j\rangle\,, \qquad \text{with } \alpha \in [0,1] \text{ and } \theta \in [0, 2\pi). \tag{35}$$

Its expectation value with respect to $\mathcal{O}_{\text{eff}}$ is given by

$$\langle\mu_\theta(\alpha)|\,\mathcal{O}_{\text{eff}}\,|\mu_\theta(\alpha)\rangle = \big(1-\alpha\big)\cdot\big(O_{00} + O_{ii} + \text{Re}(O_{0i})\big) + \alpha\cdot O_{jj} + \sqrt{2\alpha(1-\alpha)}\cdot C_\theta, \tag{36}$$

where $C_\theta := \text{Re}\big(e^{i\theta}(O_{0j} + O_{ij})\big)$, and its expectation value with respect to $\mathcal{T}_{\text{eff}}$ is given by

$$\langle\mu_\theta(\alpha)|\,\mathcal{T}_{\text{eff}}\,|\mu_\theta(\alpha)\rangle = \big(1-\alpha\big)\cdot\big(T_{00} + T_{ii}\big) + \alpha\cdot T_{jj}. \tag{37}$$

Crucially, by our choice of $\mathcal{T}$ we know that the label of $|\mu_\theta(\alpha)\rangle$ goes from -1 to $+1$ as α goes $0 \to 1$. Note that the expectation value of $|\mu_\theta(\alpha)\rangle$ with respect to $\mathcal{T}_{\text{eff}}$ is independent from the phase θ.

To determine that $\left|O_{ij}\right|$ is smaller than $\delta'' > 0$ for $i, j \geq 1$ and $i \neq j$, we query the classifier $c_{\mathcal{O},b}$ on the states $|\gamma_{\hat{\theta}}(\hat{\alpha})\rangle$ for all $\hat{\theta}$ in a ζ''-mesh of $[0, 2\pi)$ and for all $\hat{\alpha}$ in a ξ''-mesh of $[0, 1]$ and we suppose they are labeled the same as the target classifier $c_{\mathcal{T},-1}$ would. Using these queries we can find estimates $\hat{\alpha}^{\mathcal{O}_{\text{eff}}}_{\text{cross}}(\hat{\theta})$ that are ξ-close to the unique $\alpha^{\mathcal{O}_{\text{eff}}}_{\text{cross}}(\theta) = \alpha'$ that satisfies

$$\langle\mu_\theta(\alpha')|\,\mathcal{O}_{\text{eff}}\,|\mu_\theta(\alpha')\rangle = 0, \tag{38}$$

by finding the smallest $\hat{\alpha}$ where the label has gone from -1 to $+1$. Because the label assigned by $c_{\mathcal{T},-1}$ does not depend on the phase θ, and since all states $|\mu_{\hat{\theta}}(\hat{\alpha})\rangle$ were assigned the same label by $c_{\mathcal{O},b}$ and $c_{\mathcal{T},-1}$, we find that the crossing point estimate $\hat{\alpha}^{\mathcal{O}_{\text{eff}}}_{\text{cross}}(\hat{\theta})$ is the same for all $\hat{\theta}$. In particular, this implies that the actual crossing points $\alpha^{\mathcal{O}_{\text{eff}}}_{\text{cross}}(\hat{\theta})$ have to be within ξ''-distance of each other for all $\hat{\theta}$. Subsequently, write $O_{0j} + O_{ij} = \left|O_{0j} + O_{ij}\right|e^{i\phi}$ with $\phi \in [0, 2\pi)$, let $\hat{\theta}_{\text{abs}}$ denote the point in the ζ''-mesh of $[0, 2\pi)$ that is closest to $2\pi - \phi$, and let $\hat{\theta}_0$ denote the point in the ζ''-mesh of $[0, 2\pi)$ that is closest to $\pi/2 - \phi$ modulo 2π. By our previous discussion we know that $\left|\alpha^{\mathcal{O}_{\text{eff}}}_{\text{cross}}(\hat{\theta}_{\text{abs}}) - \alpha^{\mathcal{O}_{\text{eff}}}_{\text{cross}}(\hat{\theta}_0)\right| < \xi''$, which implies

$$\left|C_{\hat{\theta}_{\text{abs}}} - C_{\hat{\theta}_0}\right| < h(\xi''), \tag{39}$$

where h is a continuous function (independent from $c_{\mathcal{O},b}$ and $c_{\mathcal{T},-1}$) with $h(\xi'') \to 0$ as $\xi'' \to 0$. Moreover, using the inequality $\cos(\zeta'') \geq 1 - \lambda\zeta''$, where $\lambda \approx 0.7246$ is a solution of $\lambda\big(\pi - \arcsin(\lambda)\big) = 1 + \sqrt{1-\lambda^2}$, together with the inequality $\cos(\pi/2 - \zeta'') \leq \zeta''$, we can derive that

$$\begin{aligned}
\left|C_{\hat{\theta}_{\text{abs}}} - C_{\hat{\theta}_0}\right| &= \Big|\left|O_{0j} + O_{ij}\right|\cos\big(\hat{\theta}_{\text{abs}} + \phi\big) - \left|O_{0j} + O_{ij}\right|\cos\big(\hat{\theta}_0 + \phi\big)\Big| \\
&\geq \left|O_{0j} + O_{ij}\right| \cdot \Big|\cos\big(\zeta''\big) - \cos\big(\pi/2 - \zeta''\big)\Big| \\
&\geq \left|O_{0j} + O_{ij}\right| \cdot \Big|1 - \big(\lambda + 1\big)\zeta''\Big|.
\end{aligned} \tag{40}$$

Finally, by combining Equation (39) with Equation (40) we can conclude that

$$\left|O_{0j} + O_{ij}\right| < \frac{h(\xi'')}{1-(\lambda+1)\zeta''},$$

which for ξ'' and ζ'' small enough shows that $\left|O_{0j} + O_{ij}\right| < \delta''/2$ (i.e., the fineness of both meshes ξ'' and ζ'' will depend on the choice of δ''). In conclusion, to determine that $\left|O_{ij}\right|$ is smaller than $\delta'' > 0$ we first do the required queries to determine that $\left|O_{0j}\right| < \delta''/2$, after which we do the required queries to determine that $\left|O_{0j} + O_{ij}\right| < \delta''/2$, which together indeed implies that $\left|O_{ij}\right| < \delta''$.

All in all, we have described a (finite) set of states such that if the label assigned by $c_{\mathcal{O},b}$ agrees with the label assigned by $c_{\mathcal{T},-1}$, then the absolute value of the off-diagonal elements of the first row of $\mathcal{O}_{\text{eff}}$ have to be smaller than δ, the diagonal elements of $\mathcal{O}_{\text{eff}}$ have to be within δ'-distance of those of $\mathcal{T}_{\text{eff}}$, and the remaining off diagonal elements of $\mathcal{O}_{\text{eff}}$ have to be smaller than δ''. Finally, we choose $\delta, \delta', \delta'' = 1/2^{n+1}$ and use the above protocol to establish that for $1 \leq i \leq r-1$ the Gershgorin discs D_i of $\mathcal{O}_{\text{eff}}$ (i.e., with center O_{ii} and radius $\sum_j |O_{ij}|$) have to be contained in the disks $\tilde{D}_i$ with center $i+1$ and radius $1/2$. Moreover, we establish that the Gershgorin disc D_0 has to be contained in the disks $\tilde{D}_0$ with center $-r+1$ and radius $1/2$. Since the disks $\tilde{D}_i$ as disjoint, so are the Gershgorin discs D_i, which implies that $\mathcal{O}_{\text{eff}}$ must have at least r distinct eigenvalues, and thus that $\text{rank}\big(\mathcal{O}\big) \geq r$. Consequently, if $\text{rank}\big(\mathcal{O}\big) < r$, then $c_{\mathcal{O},b}$ must disagree with $c_{\mathcal{T},-1}$ on the label of at least one of the states queried during the protocol.

□

B.2 Proof of Proposition 7

Proposition 7. *Let $\mathcal{C}_{\text{lin}}(\Phi)$ denote the family of linear classifiers that is equipped with a feature map Φ. Also, let $\mathcal{C}_{\text{qlin}}^{(\leq r)}(\Phi')$ denote the family of quantum linear classifiers that uses observables of rank at most r and which is equipped with a quantum feature map Φ'. Then, the following statements hold:*

(i) For every feature map $\Phi : \mathbb{R}^\ell \to \mathbb{R}^N$ with $\sup_{x\in\mathbb{R}^\ell} ||\Phi(x)|| = M < \infty$, there exists a feature map $\Phi' : \mathbb{R}^\ell \to \mathbb{R}^{N+1}$ such that $||\Phi'(x)|| = 1$ for all $x \in \mathbb{R}^\ell$ and the families of linear classifiers satisfy $\mathcal{C}_{\text{lin}}(\Phi) \subseteq \mathcal{C}_{\text{lin}}(\Phi')$.

(ii) For every feature map $\Phi : \mathbb{R}^\ell \to \mathbb{R}^N$ with $||\Phi(x)|| = 1$ for all $x \in \mathbb{R}^\ell$, there exists a quantum feature map $\Phi' : \mathbb{R}^\ell \to \text{Herm}\left(\mathbb{C}^{2^n}\right)$ that uses $n = \lceil \log N + 1 \rceil + 1$ qubits such that the families of linear classifiers satisfy $\mathcal{C}_{\text{lin}}(\Phi) \subseteq \mathcal{C}_{\text{qlin}}^{(\leq 1)}(\Phi')$.

(iii) For every quantum feature map $\Phi : \mathbb{R}^\ell \to \text{Herm}\left(\mathbb{C}^{2^n}\right)$, there exists a classical feature map $\Phi' : \mathbb{R}^\ell \to \mathbb{R}^{4^n}$ such that the families of linear classifiers satisfy $\mathcal{C}_{\text{qlin}}(\Phi) = \mathcal{C}_{\text{lin}}(\Phi')$.

Proof. (i): First, we define the feature map $\Phi' : \mathbb{R}^\ell \to \mathbb{R}^{N+1}$ which maps

$$x \mapsto \frac{\Phi(x)}{M} + \sqrt{1 - \frac{||\Phi(x)||^2}{M^2}} e_{N+1},$$

where e_{N+1} denotes the $(N+1)$-th standard basis vector. Note that this feature map indeed satisfies that $||\Phi'(x)|| = 1$ for all $x \in \mathbb{R}^\ell$. Next, for any classifier $c_{w,b} \in \mathcal{C}_{\text{qlin}}(\Phi)$ we define $w' = w$ and $b' = b/M$ and we note that for any $x \in \mathbb{R}^\ell$ we have

$$c_{w',b'}(\Phi'(x)) = \text{sign}\big(\langle w', \Phi'(x)\rangle - b'\big) = \text{sign}\big(M^{-1}\big[\langle w, \Phi(x)\rangle - b\big]\big) = \text{sign}\big(\langle w, \Phi(x)\rangle - b\big) = c_{w,b}(\Phi(x)).$$

(ii): First, we define the feature map $\widetilde{\Phi} : \mathbb{R}^\ell \to \mathbb{R}^{N+1}$ which maps

$$x \mapsto \Phi(x) + e_{N+1},$$

where e_{N+1} denotes the $(N+1)$-th standard basis vector. Next, for any classifier $c_{w,b} \in \mathcal{C}_{\text{lin}}(\Phi)$ we define $\tilde{w} = w - be_{N+1}$ and we note that for all $x \in \mathbb{R}^\ell$ we have

$$c_{\tilde{w},0}(\widetilde{\Phi}(x)) = \text{sign}\big(\langle \widetilde{\Phi}(x), \widetilde{w}\rangle\big) = \text{sign}\big(\langle \Phi(x), w\rangle - b\big) = c_{w,b}(\Phi(x)).$$

Therefore, it suffices to show that we can implement any linear classifier on $\mathbb{R}^{N+1}$ with $b = 0$ as a quantum linear classifier on $n = \lceil \log N + 1 \rceil + 1$ qubits. To do so, we define the quantum feature map $\Phi' : \mathbb{R}^\ell \to \mathrm{Herm}\big(\mathbb{C}^{2^n}\big)$ which maps

$$x \to \rho_x = \left(\frac{|\Phi(x)\rangle + |0\rangle}{\sqrt{2}}\right)\left(\frac{\langle\Phi(x)| + \langle 0|}{\sqrt{2}}\right),$$

where $|0\rangle$ is a vector that does not lie in the support of Φ (note this vectors exists since we have chosen n large enough). Finally, for any linear classifier $c_{w,0} \in \mathcal{C}_{\mathrm{lin}}(\Phi)$ on $\mathbb{R}^{N+1}$ we define $b' = ||w||^2/2$ and $\mathcal{O} = |w'\rangle\langle w'|$, where $|w'\rangle = |w\rangle + ||w||\,|0\rangle$ and we note that for all $x \in \mathbb{R}$ we have

$$\begin{aligned} c_{\mathcal{O},b'}(\Phi'(x)) &= \mathrm{sign}\big(\mathrm{Tr}\big[\mathcal{O}\rho_x\big] - b'\big) \\ &= \mathrm{sign}\big(\frac{1}{2}\Big|\,\langle w \mid \Phi(x)\rangle + ||w||\,\Big|^2 - \frac{||w||^2}{2}\big) \\ &= \mathrm{sign}\big(\langle w, \Phi(x)\rangle\big) = c_{w,0}(\Phi(x)). \end{aligned}$$

(iii): This follows directly from the fact that $\mathrm{Herm}\left(\mathbb{C}^{2^n}\right) \simeq \mathbb{R}^{4^n}$.

□

B.3 Proof of Proposition 8

Proposition 8. *Let $\mathcal{C}^{(\eta)}_{\mathrm{qlin}}$ denote the family of quantum linear classifiers corresponding to all n-qubit observables of Frobenius norm η, that is,*

$$\mathcal{C}^{(\eta)}_{\mathrm{qlin}} = \Big\{c(\rho) = \mathrm{sign}\big(\mathrm{Tr}\big[\mathcal{O}\rho\big] - d\big) \;\Big|\; \mathcal{O} \in \mathrm{Herm}\big(\mathbb{C}^{2^n}\big) \text{ with } ||\mathcal{O}||_F = \eta,\, d \in \mathbb{R}\Big\}. \tag{22}$$

Then, for every $\eta \in \mathbb{R}_{>0}$ and $0 < m \leq 2^n$ there exists a set of m examples consisting of binary labeled n-qubit pure states that satisfies the following two conditions:

(i) There exists a classifier $c \in \mathcal{C}^{(\eta)}_{\mathrm{qlin}}$ that correctly classifies all examples with margin $\eta/\sqrt{m}$.

(ii) No classifier $c' \in \mathcal{C}^{(\eta')}_{\mathrm{qlin}}$ with $\eta' < \eta$ can classify all examples correctly with margin $\geq \eta/\sqrt{m}$.

Proof. Define $\mathcal{D}_m = \mathcal{D}^+_m \cup \mathcal{D}^-_m$ whose positive examples (i.e., labeled $+1$) are given by

$$\mathcal{D}^+_m = \big\{\,|i\rangle\langle i| \mid i = 1, \ldots, \frac{m}{2}\big\},$$

and whose negative examples (i.e., labeled -1) are given by

$$\mathcal{D}^-_m = \big\{\,|i\rangle\langle i| \mid i = \frac{m}{2} + 1, \ldots, m\big\}.$$

To classify this set of examples we take the classifier $c_{\mathcal{O},0} \in \mathcal{C}^{(\eta)}_{\mathrm{qlin}}$ whose observable is given by

$$\mathcal{O} = \frac{\eta}{\sqrt{m}}\left(\Big(\sum_{i=1}^{m/2}|i\rangle\langle i|\Big) + \Big(\sum_{j=\frac{m}{2}+1}^{m}|j\rangle\langle j|\Big)\right).$$

We remark that $c_{\mathcal{O},0}$ can indeed classify the set of examples $\mathcal{D}_r$ with margin $\eta/\sqrt{m}$.

Now suppose $c_{\mathcal{O}',b'} \in \mathcal{C}^{\eta'}_{\mathrm{qlin}}$ with $\eta' < \eta$ can classify $\mathcal{D}_m$ with margin γ', that is

$$\mathrm{Tr}\big[\mathcal{O}'\,|i\rangle\langle i|\,\big]\begin{cases} \geq b' + \gamma' & \text{if } i = 1, \ldots, \frac{m}{2}, \\ \leq b' - \gamma' & \text{if } i = \frac{m}{2} + 1, \ldots, m. \end{cases} \tag{41}$$

Define $\rho_+ = \sum_{i=1}^{m/2}|i\rangle\langle i|$ and $\rho_- = \sum_{i=\frac{m}{2}+1}^{m}|i\rangle\langle i|$ and note that Equation (41) implies that

$$\mathrm{Tr}\big[\mathcal{O}'\rho_+\big] \geq \frac{m}{2}b' + \frac{m}{2}\gamma'$$

and that

$$\mathrm{Tr}\big[\mathcal{O}'\rho_-\big] \leq \frac{m}{2}b' - \frac{m}{2}\gamma'$$

By combining these two inequalities we find that

$$\mathrm{Tr}\big[\mathcal{O}'(\rho_+ - \rho_-)\big] \geq \frac{m}{2}b' - \frac{m}{2}b' + \frac{m}{2}\gamma' + \frac{m}{2}\gamma' = m\gamma'. \tag{42}$$

Finally, by the Cauchy–Schwarz inequality we find that

$$\mathrm{Tr}\big[\mathcal{O}'(\rho_+ - \rho_-)\big] \leq \underbrace{||\mathcal{O}'||_F}_{<\eta} \cdot \underbrace{||\rho_+ - \rho_-||_F}_{=\sqrt{m}} < \eta\sqrt{m}. \tag{43}$$

Combining Equation (42) and (43) we find that

$$m\gamma' \leq \mathrm{Tr}\big[\mathcal{O}'(\rho_+ - \rho_-)\big] < \eta\sqrt{m}$$

from which we can conclude that $\gamma' < \eta/\sqrt{m}$.

□