Computational speedups and learning separations in quantum machine learning

Gyurik, C.

Chapter 4

Structural risk minimization for quantum linear classifiers

In this chapter 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 2.6) and afterwards we analyze their effect on the training error (i.e., the blue line in Figure 2.6). Specifically, in Section 4.1 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 4.2 we study the influence that model parameters which influence the established complexity measure bounds have on the training error term. Finally, in Section 4.3, we discuss how to implement structural risk minimization of quantum linear classifiers based on the obtained results.

4.1 Complexity of quantum linear classifiers

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. These bounds upper bound the expected error by a sum of a training error and a complexity term that we would like to trade-off to achieve the best possible bound. Using the model parameters that we identify, we can balance this trade-off to construct the best possible model. In short, these model parameters can be used to balance the trade-off considered by structural risk minimization, as depicted in Figure 2.6. Throughout this section we fix the feature map to be the one defined Equation (2.22) 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 4.3). We defer the proof to Appendix B.1.1.

Proposition 12. *Let $\mathbb{O} \subseteq \mathrm{Herm}\left(\mathbb{C}^{2^n}\right)$ be a family of n-qubit observables with $r = \dim\left(\sum_{\mathcal{O}\in\mathbb{O}} \mathrm{Im}\,\mathcal{O}\right)$*[1]*. Then, the VC dimension of*

$$\mathcal{C}_{\text{qlin}}^{\mathbb{O}} = \Big\{c(x) = \mathrm{sign}\big(\mathrm{Tr}\left[\mathcal{O}\rho_\Phi(x)\right] - d\big) \;\big|\; \mathcal{O}\in\mathbb{O},\, d\in\mathbb{R}\Big\} \tag{4.1}$$

satisfies

$$\mathrm{VC}\big(\mathcal{C}_{\text{qlin}}^{\mathbb{O}}\big) \leq \dim\big(\mathrm{Span}\big(\mathbb{O}\big)\big) + 1 \leq r^2 + 1. \tag{4.2}$$

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 2.3 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 $\dim\big(\mathrm{Span}\big(\mathbb{O}\big)\big) + 1$. To see this, we consider a family of explicit quantum linear classifiers with observables $\mathbb{O}_{\text{explicit}} = \left\{\mathcal{O}_\theta^\lambda\right\}$, where

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

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

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

define

$$H = \mathrm{Span}_{\mathbb{C}}\Big\{ \left|\psi_0(\theta)\right\rangle, \ldots, \left|\psi_L(\theta)\right\rangle \ : \ \theta \in \mathbb{R}^m \Big\}, \tag{4.3}$$

$$V = \mathrm{Span}_{\mathbb{R}}\Big\{ \sum_{i=0}^{L} \lambda(i) \left|\psi_i(\theta)\right\rangle \left\langle\psi_i(\theta)\right| \ : \ \theta \in \mathbb{R}^m \Big\}, \tag{4.4}$$

Then, Proposition 12 states that

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

Now, by the following lemma, we indeed find that the bound $r^2 + 1$ is only quadratically worse than the bound $\dim\big(\mathrm{Span}(\mathbb{O})\big) + 1$. We again defer the proof to Appendix B.1.1.

Lemma 13. *The vector spaces defined in Eq.* (4.3) *and Eq.* (4.4) *satisfy*[2]

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

Therefore, if we sufficiently limit $r = \dim(H)$, then this also limits $\dim\big(\mathrm{Span}(\mathbb{O})\big) = \dim(V)$. Moreover, even though $\dim\big(\mathrm{Span}(\mathbb{O})\big) + 1$ 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 4.3).

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 $r^2 + 1$[3]. 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 $\{\bigotimes_{i=1}^n X_i(\theta_i) \left|0\right\rangle \mid \theta \in [0, 2\pi)^n\}$ 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 $\dim\big(\mathrm{Span}(\mathbb{O})\big) + 1$ 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 4.3 we show that this

[2] 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 4.3).

[3] 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 B.1.2.

latter quantity can already be used to affirm the effectiveness of certain regularization techniques. Specifically, in Section 4.3 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 B.1.3, where we also discuss the implications of this result in the probably approximately correct (PAC) learning framework.

Proposition 14. *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*

$$\mathcal{F}^{\mathbb{O}}_{\mathrm{qlin}} = \left\{ f_{\mathcal{O},d}(x) = \mathrm{Tr}\left[\mathcal{O}\rho_\Phi(x)\right] - d \;\middle|\; \mathcal{O} \in \mathbb{O},\ d \in \mathbb{R} \right\} \tag{4.5}$$

is upper bounded by

$$\mathrm{fat}_{\mathcal{F}^{\mathbb{O}}_{\mathrm{qlin}}}(\gamma) \leq O\left(\frac{\eta^2}{\gamma^2}\right). \tag{4.6}$$

Remark(s). *The upper bound in the above proposition matches the result discussed in [127]. This was derived independently by one of the authors of [95] in [193], 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 of 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 linear classifiers. Specifically, plugging Theorem 14 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[4]. 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 $\min_i \left|\mathrm{Tr}\left[\mathcal{O}\rho_\Phi(x_i)\right] - d\right| / ||\mathcal{O}||_F$.

[4] 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 $\min_i |f(x_i) - d|$.

4.2 Expressivity of quantum linear classifiers

Having established that the quantity r defined in Proposition 12 and the Frobenius norms of the observables influence the complexity of the family of quantum linear classifiers (i.e., the green line in Figure 2.6), 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 2.6). 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 12 can be controlled by limiting the ranks of all observables. In Section 4.3 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.2.1.

Proposition 15. *Let $\mathcal{C}_{\text{qlin}}^{(r)}$ denote the family of quantum linear classifiers corresponding to observables of exactly rank r, that is,*

$$\mathcal{C}_{\text{qlin}}^{(r)} = \Big\{ c(\rho) = \text{sign}\big(\text{Tr}\left[\mathcal{O}\rho\right] - d\big) \;\Big|\; \mathcal{O} \in \text{Herm}\big(\mathbb{C}^{2^n}\big),\, \text{rank}\big(\mathcal{O}\big) = r,\, d \in \mathbb{R} \Big\} \tag{4.7}$$

Then, the following statements hold:

(i) For every finite set of examples $\mathcal{D}$ that is correctly classified by a quantum linear classifier $c \in \mathcal{C}_{\text{qlin}}^{(k)}$ with $0 < k < 2^n$, there exists a quantum linear classifier $c \in \mathcal{C}_{\text{qlin}}^{(r)}$ 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 $c \in \mathcal{C}_{\text{qlin}}^{(r)}$, but which no classifier $c' \in \mathcal{C}_{\text{qlin}}^{(k)}$ 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 $\mathcal{C}_{\text{qlin}}^{(r)}$ and $\mathcal{C}_{\text{qlin}}^{(k)}$ 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. 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:

$$\begin{aligned} \mathcal{C}_{\mathrm{lin}} \text{ on } \mathbb{R}^{2^n} &\subseteq \mathcal{C}_{\mathrm{qlin}}^{(\leq 1)} \text{ on } n+1 \text{ qubits} \subseteq \dots && (4.8)\\ &\subseteq \mathcal{C}_{\mathrm{qlin}}^{(\leq r)} \text{ on } n+1 \text{ qubits} \subseteq \dots \subseteq \mathcal{C}_{\mathrm{lin}} \text{ on } \mathbb{R}^{4^n}, && (4.9) \end{aligned}$$

where $\mathcal{C}_{\mathrm{qlin}}^{(\leq r)}$ denotes the family of quantum linear classifiers using observables of rank at most r. Note that $\mathcal{C}_{\mathrm{qlin}}^{(\leq r)} \subsetneq \mathcal{C}_{\mathrm{qlin}}^{(\leq r+1)}$ is strict due to Proposition 15. We defer the proof to Appendix B.2.2.

Proposition 16. *Let $\mathcal{C}_{\mathrm{lin}}(\Phi)$ denote the family of linear classifiers that is equipped with a feature map Φ. Also, let $\mathcal{C}_{\mathrm{qlin}}^{(\leq r)}(\Phi')$ 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 $\mathcal{C}_{\mathrm{lin}}(\Phi) \subseteq \mathcal{C}_{\mathrm{lin}}(\Phi')$.

(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 $\Phi' : \mathbb{R}^\ell \to \mathrm{Herm}\left(\mathbb{C}^{2^n}\right)$ that uses $n = \lceil \log N + 1 \rceil + 1$ qubits such that the families of linear classifiers satisfy $\mathcal{C}_{\mathrm{lin}}(\Phi) \subseteq \mathcal{C}_{\mathrm{qlin}}^{(\leq 1)}(\Phi')$.

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

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 2.3). 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.2.3.

Proposition 17. *Let $\mathcal{C}_{\mathrm{qlin}}^{(\eta)}$ denote the family of quantum linear classifiers corresponding to all n-qubit observables of Frobenius norm η, that is,*

$$\mathcal{C}_{\mathrm{qlin}}^{(\eta)} = \Big\{ c(\rho) = \mathrm{sign}\big(\mathrm{Tr}\,[\mathcal{O}\rho] - d\big) \;\big|\; \mathcal{O} \in \mathrm{Herm}\big(\mathbb{C}^{2^n}\big) \text{ with } ||\mathcal{O}||_F = \eta,\, d \in \mathbb{R} \Big\}. \tag{4.10}$$

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

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

In conclusion, in Proposition 12 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 15 we showed that the rank also controls the model's ability to achieve small training errors. Moreover, in Proposition 17 we similarly showed that the Frobenius norm not only controls the model's complexity (see Proposition 14), 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 15 one actually has to increase the rank of the observable.

4.3 Structural risk minimization 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 12 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 12 one has to compute the quantity

$$\dim_{\mathbb{C}} \Big(\mathrm{Span}_{\mathbb{C}} \big\{ \, |\psi_i(\theta)\rangle \ \ : \ i = 1, \ldots \ell, \ \theta \in \mathbb{R}^p \big\} \Big), \tag{4.11}$$

where $|\psi_i(\theta)\rangle$ denotes the ith column of $W(\theta)$. To use the other bound $\dim\big(\mathrm{Span}\big(\mathbb{O}\big)\big) + 1$ from Proposition 12 one has to compute the quantity

$$\dim_{\mathbb{R}} \Big(\mathrm{Span}_{\mathbb{R}} \big\{ \sum_{i=1}^{\ell} |\psi_i(\theta)\rangle \langle \psi_i(\theta)| \ \ : \ \theta \in \mathbb{R}^p \big\} \Big), \tag{4.12}$$

Although both are of course possible, in some cases it is slightly easier to see how the quantity in Eq. (4.11) scales with respect to ℓ. Specifically, utilizing the quantity in Eq. (4.11) already leads to interesting ansatze that allow for structural risk minimization by limiting ℓ. As discussed below Proposition 12, 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 12 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.

Example 1 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

$$W(\theta) = \big[C–U_{2^c}(\theta_{2^c})\big] \cdot \ldots \cdot \big[C–U_1(\theta_1)\big].^5$$

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 12 the VC dimension is at most $\ell^2 + 1$. Note that t is a controllable hyperparameter that can be used to tune the

[5] 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 $U_i(\theta_i) = I$).

VC dimension. In particular, we can set it such that the resulting VC dimension is not exponential in n. Let us now consider the other bound $\dim\big(\mathrm{Span}(\mathbb{O})\big) + 1$ from Proposition 12. For this choice of ansatz, computing the quantity in Eq. (4.12) is also straightforward due to the block structure of the unitary. Moreover, for this choice of ansatz the inequalities in Lemma 13 are strict, which shows why being able to compute the quantity in Eq. (4.11) does not always imply that we can also compute the quantity in Eq. (4.12) (i.e., one is not simply the square of the other).

Example 2 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 12, 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. (4.11), 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. (4.12) 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. (4.12), but the bound r^2+1 from Proposition 12 already suffices to establish that ℓ is a tunable hyperparameter that controls the VC dimension.

Example 3 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

$$W(\theta) = U(\theta^d_1, \ldots, \theta^d_k) \cdot \cdots \cdot U(\theta^1_1, \ldots, \theta^1_k), \quad \theta \in [0, 1\pi)^{dk}.$$

By the bound r^2+1 from Proposition 12, 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. (4.11), 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. (4.12) 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. (4.12), but the bound r^2+1 from Proposition 12 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 [165] 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 $\min_i \left|\mathrm{Tr}\left[\mathcal{O}\rho_\Phi(x_i)\right] - d\right| / ||\mathcal{O}||_F$. 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 bounding $||\alpha||_1$ as $||\mathcal{O}_\alpha||_F \leq ||\alpha||_1$. However, if the weights are obtained by solving the usual quadratic program [103, 168], then the resulting observable is already (optimally) regularized with respect to the Frobenius norm [165].

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 [153]. 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 [46, 52]. Finally, it turns out that one can also regularize quantum linear classifiers by running the circuits under varying levels of noise [47]. For these kinds of regularization the relationships between the regularized explicit and regularized implicit quantum linear classifiers are still to be investigated.