Regularization of Riemannian optimization: Application to process tomography and quantum machine learning

Felix Soest[1], Konstantin Beyer[2,1], and Walter T. Strunz[1]

[1]Institute of Theoretical Physics, TUD Dresden University of Technology, 01062, Dresden, Germany
[2]Department of Physics, Stevens Institute of Technology, Hoboken, New Jersey 07030, USA

Gradient descent algorithms on Riemannian manifolds have been used recently for the optimization of quantum channels. In this contribution, we investigate the influence of various regularization terms added to the cost function of these gradient descent approaches. Motivated by Lasso regularization, we apply penalties for large ranks of the quantum channel, favoring solutions that can be represented by as few Kraus operators as possible. We apply the method to quantum process tomography and a quantum machine learning problem. Suitably regularized models show faster convergence of the optimization as well as better fidelities in the case of process tomography. Applied to quantum classification scenarios, the regularization terms can simplify the classifying quantum channel without degrading the accuracy of the classification, thereby revealing the minimum channel rank needed for the given input data.

1 Introduction

Quantum channels or completely positive and trace preserving (CPT) maps describe all physically valid transformations of an arbitrary input quantum state to a corresponding output quantum state. In this sense, quantum channels are the most general framework of how information can be processed. Quantum computers, for example, implement specifically tailored state transformations to make use of quantum effects for the efficient processing of classical data encoded in the input state. Noise influences the overall channel in a way which is usually detrimental for useful information processing. Similar problems arise in quantum communication schemes, where the transferred message is directly affected by imperfections of the transmission channel. Currently, many research efforts focus on the question of how to cope with this noise and how to make applications more resistant against it [1–6]. In order to do so, it is often instrumental to identify the influence of the noise, that is, to analyze how the channel which transforms the quantum can be characterized [7–9]. Such a task is usually called quantum channel or process tomography and aims at finding a numerical representation of a channel that reproduces the experimentally obtained data [10–12]. Various methods have been proposed, ranging from linear inversion and maximum-likelihood methods [13–17] over convex optimization [18] and projection techniques [19, 20] to machine learning approaches [21, 22]. Experimentally, process tomography has been implemented, for example, in superconducting qubits [23–28], optical setups [29, 30], trapped ions [31], and nuclear spins [32].

Recently, a method based on Riemannian gradient descent has been proposed for quantum process tomography [33]. This approach makes use of the fact that Kraus representations of quantum channels form a Stiefel manifold for which efficient optimization techniques exist [34, 35]. The parametrization in terms of Kraus operators leads to a valid quantum channel by construction, thereby circumventing problems arising in other tomography methods, where the resulting maps

have to be projected to the closest valid quantum state transformation after the optimization [33].

In general, a full-rank representation of a quantum channel scales exponentially with system size. However, often quantum channels can be well approximated by low-rank channels, especially if the channel emerges from a quantum circuit with limited connectivity and depth. Methods based on matrix product state representations of the Choi matrix [22] or compression approaches [36–39] can be used in such a case, to enable tomography of higher-dimensional systems. For parametrization on the Stiefel manifold, the maximal rank of the channel can easily be constrained by limiting the number of Kraus operators in the representation. However, if the actual target channel has a rank smaller than the number of Kraus operators, the gradient descent method does not necessarily converge to a solution with a minimal number of nonzero Kraus operators, due to the non-uniqueness of Kraus representations.

In this paper, we investigate the influence of various regularization terms added to the cost function of the Riemannian optimization with the aim of lowering the number of relevant Kraus operators in the representation. More specifically, we analyze the performance of three different regularization schemes based on the Hilbert-Schmidt norm of the involved Kraus operators, the purity of the Choi matrix of the channel, and an L_1-norm of the Stiefel vector that represents the Kraus decomposition, respectively. The first two terms are directly motivated by the fact that they penalize channels with high rank and large numbers of nontrivial Kraus operators, respectively. The L_1-norm regularization has been considered before in Ref. [33]. In that paper, however, it appeared only as a side remark without detailed motivation and examination of its influence, so we include it here for comparison. We will see that all three terms – to different extent – support a convergence of the optimization toward simple representations of the channel under tomography.

Quantum process tomography is certainly the prime example for numerical channel optimization. However, the framework is applicable to more general settings that do not necessarily require learning the full representation of a quantum channel [40, 41]. Quantum machine learning (QML) problems can be of this form, and we will consider them as a second test example for regularized Riemannian gradient descent algorithms.

Assume we have a classification problem where classical input data is to be discriminated into various classes. Today, such a problem is typically solved by classical algorithms. However, it has been shown that quantum circuits can also be utilized for classification tasks [42]. There, the input information is encoded in a quantum state, which is then mapped by a quantum transformation to an output state that can be measured to determine the class the input data belongs to. The task is then to optimize the mapping such that it correctly classifies the input data.

It is an ongoing debate under which conditions quantum machine learning approaches can actually provide an advantage over classical methods. Theoretical improvements have been reported [43–45]. Still, it is often questioned whether QML will eventually have a practical impact [46, 47]. Note that the quest for a quantum advantage is not our concern here. Instead, we show how the regularized Riemannian optimization can yield insight into the properties of a quantum classification problem. In particular, the regularization terms considered here can help to understand which effective channel rank is necessary to classify a certain data set.

The paper is organized as follows. We start with a detailed description of the general setting. We then give a brief overview of Riemannian optimization on the Stiefel manifold and introduce the regularization terms we are going to investigate. In Sec. 5 and Sec. 6 we apply the optimization with regularization to quantum process tomography and a quantum machine learning problem, showing the influence of the terms in illustrative examples. In Sec. 7, we conclude.

2 Setting

We consider the following general optimization scenario for a quantum channel or CPT map. The aim is to optimize a channel T that maps an input state ρ_α to a state $T[\rho_\alpha]$. A subsequent measurement by a positive operator valued measure (POVM) with elements $\{M_\beta\}$ would yield outcome β with probability

$$p_T(\beta|\alpha) = \mathrm{tr}[M_\beta T[\rho_\alpha]], \tag{1}$$

conditioned on the choice of input state ρ_α. The probability distribution $p_T(\beta|\alpha)$ will be subject to a cost function $\mathcal{L}_p$ which specifies the desired properties for the channel. $\mathcal{L}_p$ depends on T through Eq. (1). The channel T is then optimized to minimize the cost.

The choice of suitable input states ρ_α and measurements M_β, as well as the specific form of the cost function $\mathcal{L}$, depend on the problem to be solved.

3 Riemannian optimization of Kraus maps

In order to numerically optimize the channel T, it needs to be represented in a suitable form. In this paper, we use Kraus decompositions. Kraus channels with a fixed number of Kraus operators form a Stiefel manifold, and the optimization of the channel can be done by Riemannian gradient descent on that manifold [40]. We will review the framework in the following. More extensive presentations on that matter can be found in Refs. [34, 48].

Consider a channel T on a d-dimensional quantum system. The channel is represented by m Kraus operators κ_k that satisfy the condition

$$\sum_{k=1}^{m} \kappa_k^\dagger \kappa_k = \mathbb{1}_d. \tag{2}$$

We can introduce the matrix $\mathbb{K} = [\kappa_1, ..., \kappa_m]^T \in \mathbb{C}^{md\times d}$, created by stacking the Kraus operators. This allows us to formulate Eq. (2) as

$$\mathbb{K}^\dagger\mathbb{K} = \mathbb{1}_d, \tag{3}$$

which is the defining equation of the Stiefel manifold

$$\mathrm{St}(md, d) = \{\mathbb{K} \in \mathbb{C}^{md\times d} : \mathbb{K}^\dagger\mathbb{K} = \mathbb{1}_d\}. \tag{4}$$

We can now use a method from optimization on smooth manifolds, namely Riemannian gradient descent (RGD), to optimize a smooth cost function $\mathcal{L}(\mathbb{K})$ [40]. A normal gradient descent method on the matrix $\mathbb{K}$ would in general lead out of the manifold defined by Eq. (4). Therefore, RGD makes use of a retraction $R_\mathbb{K}$, a mapping from the manifold's tangent bundle to the manifold, to bring $\mathbb{K}$ back to the manifold after a usual gradient descent step:

$$\mathbb{K}' = R_\mathbb{K}(-\epsilon \text{ grad } \mathcal{L}(\mathbb{K})). \tag{5}$$

Here, ϵ is the step size, also known as the learning rate. For a given manifold, multiple retractions might exist, which can be chosen from freely [34]. We use the Cayley transform together with the Sherman-Morrison-Woodbury formula leading to the update rule [48]

$$\mathbb{K}' = \mathbb{K} - \epsilon U \left(\mathbb{1} + \frac{\epsilon}{2} V^\dagger U\right)^{-1} V^\dagger \mathbb{K}, \tag{6}$$

where

$$U = [\tilde{G}, \mathbb{K}], \qquad V = [\mathbb{K}, -\tilde{G}],$$
$$\tilde{G} = \frac{G}{||G||}, \qquad G_{i,j} = \left(\frac{\partial \mathcal{L}}{\partial \mathbb{K}_{i,j}}\right)^*. \tag{7}$$

Here, $[\cdot,\cdot]$ represents the row vector of two matrices leading to a matrix of size $(md, 2d)$ and $||\cdot||$ the Frobenius norm. We use $\epsilon = 1$ throughout.

4 Regularization

In optimization procedures, regularization terms are included in the cost function to favor or penalize solutions with certain properties. These terms are often used to create simpler or unique solutions. The parametrization of the channel T by Kraus operators leads to an ambiguous solution. In fact, any channel has infinitely many different Kraus representations. We can use the idea of regularization to obtain a solution with a specific form.

For example, let us assume that we have included m Kraus operators in the Stiefel vector $\mathbb{K}$, but the optimal solution to the problem is a channel of rank $r < m$. In this case, our model is over-parametrized because the channel could also be represented by only r different Kraus operators. If we do not know the rank of the optimal solution, we cannot simply reduce the number of Kraus operators m in the Stiefel vector $\mathbb{K}$. However, we can try to penalize models with a large number of finite Kraus operators in $\mathbb{K}$ and favor those solutions where at least some Kraus operators are close to zero and therefore almost irrelevant for the channel.

The regularization amounts to an additional term $\mathcal{R}$ in the cost function. This term can depend on the representation $\mathbb{K}$ of the channel, in contrast to the cost term $\mathcal{L}_p$, which only depends on the channel T but is otherwise ignorant about its specific Kraus decomposition. The complete cost function then reads

$$\mathcal{L} = \mathcal{L}_p(T) + \gamma \mathcal{R}(\mathbb{K}) \tag{8}$$

where γ is a hyperparameter controlling the regularization strength.

4.1 Hilbert-Schmidt norm

The first regularization term we consider is given by the average Hilbert-Schmidt norm of all Kraus operators in the representation,

$$\mathcal{R}_{\mathrm{HS}} = \frac{1}{m} \sum_{k=1}^{m} \sqrt{\operatorname{tr} \kappa_k^\dagger \kappa_k} \tag{9}$$

This term favors representations with fewer nonzero Kraus operators. Indirectly, this also reduces the rank of the channel.

4.2 Logarithmic Choi state purity

As a second regularization term, we consider the logarithmic purity of the channel's Choi state. For a channel given by the Kraus operators κ_k, the Choi state reads

$$\chi = \sum_k (\kappa_k \otimes \mathbb{1}) |\Phi_+\rangle\langle\Phi_+| (\kappa_k^\dagger \otimes \mathbb{1}), \tag{10}$$

with the maximally entangled state

$$|\Phi_+\rangle = \frac{1}{\sqrt{d}} \sum_{j=0}^{d-1} |j\rangle \otimes |j\rangle \,. \tag{11}$$

The regularization term is then defined as the negative logarithmic purity of the Choi state χ,

$$\mathcal{R}_C = -\ln \operatorname{tr} \chi^2. \tag{12}$$

This term does not explicitly depend on the Kraus representation $\mathbb{K}$, since it only involves the unique Choi state of the channel. However, the purity term favors Choi states of low rank and therefore also Kraus representations with only a few independent Kraus operators.

4.3 L_1-norm

As a third term, we look at the L_1-norm of the Kraus vector.

$$\mathcal{R}_L = |\mathbb{K}|_1 = \max_j \sum_i |\mathbb{K}_{ij}| \tag{13}$$

This term is inspired by classical Lasso regularization [49] and has been used in Ref. [33] in a quantum process tomography context. However, there, the authors neither give a profound motivation for the term nor do they analyze its influence on the performance. We therefore include it here to fill this gap. The term can indeed improve the performance, it is however in general outperformed by the two previous choices.

5 Quantum process tomography

As a first example, let us apply the regularized Riemannian optimization to quantum process tomography (QPT), which is a special case of the general channel optimization setting outlined in Sec. 2. The aim of such a procedure is to model a quantum channel T which reproduces the experimentally obtained data for an unknown quantum channel $\mathcal{E}$. To do so, the input states ρ_α, used in the experiment, form a set of states that span the whole state space. The measurement $\{M_\beta\}$, performed after the application of the channel, is informationally complete [50]. This guarantees that there is a unique channel T whose statistics

(see Eq. (1)) reproduces the one of the experimental channel $\mathcal{E}$. It is assumed here that the input states ρ_α and the POVM elements M_β are perfectly known. Strictly speaking, this is never the case in a real experiment. However, this problem is independent of the question we are investigating in this paper.

Crucially, the probability distribution $p_m(\beta|\alpha)$ measured in the experiment is in general only an estimate of the true distribution due to the limited amount of measurement data. It has been shown that QPT methods nevertheless work in this regime, especially if the channel is not of full rank [22, 33]. However, gradient descent optimization techniques on sparse data are prone to overfitting. We will show that the regularization terms can help to reduce this behavior to a certain extent. In order to distinguish this effect from the regularization of the number of relevant Kraus operators in the Stiefel vector $\mathbb{K}$, we will first look into the case of perfect infinite measurement data (see Sec. 5.1), and only later turn toward the experimentally more relevant case of limited measurement data.

In order to optimize the channel T with the Riemannian gradient descent method, we need a suitable cost function. A standard cost function for the discrepancy between the measured probability distribution p_m and the modelled one p_T is given by the Kullback-Leibler divergence,

$$\mathcal{L}_p = \sum_\alpha p_0(\alpha) \sum_\beta p_m(\beta|\alpha) \ln \frac{p_m(\beta|\alpha)}{p_T(\beta|\alpha)}, \quad (14)$$

which we will use throughout this section. Here, $p_0(\alpha)$ is the prior distribution over the input states ρ_α, which we will assume to be uniform.

We want to investigate the influence of the regularization terms $\mathcal{R}$ introduced in Sec. 4. These terms are motivated by the idea to minimize the number of relevant Kraus operators in the channel representation $\mathbb{K}$. It can therefore be expected that the regularization works in particular for non-full-rank channels. A realistic quantum channel always has full rank. However, in many cases of practical relevance, only a few Kraus operators with significant magnitude are needed to well approximate the channel. This will in particular be true for channels that emerge from quantum circuits with limited depth and connectivity as they appear, e.g., in gate-based quantum computers. Therefore, in the following investigations, we will mainly focus on channels with intermediate rank and analyze how the performance of the regularization depends on the rank.

The overall procedure is as follows. We randomly sample n-qubit channels $\mathcal{E}$ of fixed rank and calculate the probability distribution $p_m(\beta|\alpha)$ that would be measured in an experiment. See App. A for details on the sampling of the channels. As input states ρ_α, we choose all combinations of eigenstates of the Pauli operators, i.e., each input state is of the form

$$\rho_\alpha = \Pi_{i_1} \otimes \ldots \otimes \Pi_{i_n}, \quad (15)$$

where Π_{i_j} is one of the six eigenstates of the Pauli operators σ_x, σ_y, σ_z on the jth qubit. Up to a normalizing prefactor, these states also form an informationally complete POVM which defines our measurement operators M_β.

We then initialize T in a random channel and use the Riemannian gradient descent method to fit T to the true channel $\mathcal{E}$. The figure of merit will in general be the infidelity

$$1 - F(\chi_T, \chi_\mathcal{E}) = 1 - \mathrm{tr}\sqrt{\sqrt{\chi_T}\chi_\mathcal{E}\sqrt{\chi_T}} \quad (16)$$

between the Choi states of the estimated channel T and the target channel $\mathcal{E}$.

5.1 Infinite shots

We start with the ideal case of infinite measurement data. Here, the experimentally measured probability distribution $p_m(\beta|\alpha)$ matches exactly the true statistics of the unknown target channel $\mathcal{E}$. The more realistic case of limited measurement data follows below.

We consider the tomography of quantum processes for $n = 2$ qubits. Without prior knowledge of the target quantum channels $\mathcal{E}$, $m = 4^n = 16$ Kraus operators have to be included in the Stiefel vector $\mathbb{K}$ to be able to cover two-qubit channels

up to the maximum rank. In Fig. 1, we plot the impact of the various regularization terms $\mathcal{R}$ and strengths γ for randomly sampled target channels with different ranks r. We see that for a suitably chosen regularization strength $\gamma = 10^{-3}$, the term $\mathcal{R}_{\text{HS}}$ leads to significant advantages up to a rank $r = 9$. The term $\mathcal{R}_C$ is advantageous even up to rank $r = 14$ for $\gamma = 10^{-4}$. The regularization term $\mathcal{R}_L$, based on the L_1-norm of the Stiefel vector $\mathbb{K}$, is of no help in the infinite shot regime and the unregularized case performs better on average for all channel ranks.

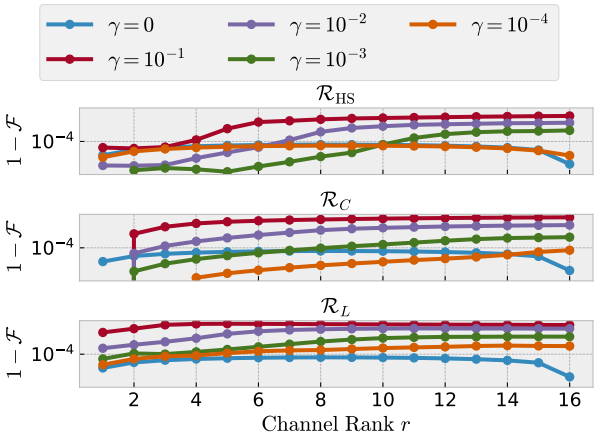

Figure 1: We plot the mean infidelity as a function of the rank of the target channel $\mathcal{E}$, trained with the regularization terms $\mathcal{R}$ after 10^5 epochs of training. For each rank, we sample 300 channels. In the infinite shot case considered here, both $\mathcal{R}_{\text{HS}}$ and $\mathcal{R}_C$ provide an improvement over the unregularized case for many target channel ranks and choices of γ. Thus, these terms can enhance the convergence properties of the optimization. By contrast, the term $\mathcal{R}_L$ leads to a disadvantage for any finite regularization strength γ.

The advantage of the regularization disappears when the true rank of the channel is known. In that case, one can set the number of Kraus operators m to the channel rank r and the convergence of the model to the target channel is much faster. This is illustrated in Fig. 2, where we plot the average training histories of channels with true rank $r = 5$. The regularization term leads to a lower infidelity on optimization models with $m = 16$ Kraus operators in $\mathbb{K}$. However, these models are outperformed by models where one sets $m = r = 5$.

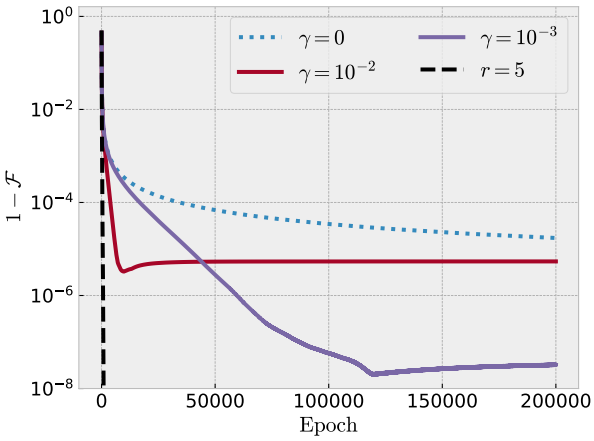

Figure 2: We plot the mean infidelity of 300 randomly sampled target channels with rank $r = 5$ in the infinite shot regime. The channel T is initialized as a random full-rank channel. Using the regularization term $\mathcal{R}_{\text{HS}}$, we compare multiple values of γ. Both non-zero values of γ lead to faster convergence compared to the unregularized case $\gamma = 0$. However, if the rank of the target channel $\mathcal{E}$ is known and the model T is constraint to it, the convergence is even better (black dashed line).

5.2 Finite shots

We now turn to the case of finite measurements, where each input state ρ_α is prepared s times and the evolved state is measured using the POVM $\{M_\beta\}$. The measurement process is simulated by drawing s times from a multinomial distribution with probability $p_{\mathcal{E}}(\beta|\alpha) = \text{tr}[M_\beta \mathcal{E}[\rho_\alpha]]$ for each input state ρ_α.

We start the finite shot case by examining the effect of the Hilbert-Schmidt regularization term $\mathcal{R}_{\text{HS}}$ on channels of true rank $r = 4$. Such channels might arise from a four-qubit unitary process after tracing out two of the four qubits. In the infinite shot case, we had seen that the optimization converges much faster and to a tighter value, if the true rank of the target channel is known and the number of Kraus operators in $\mathbb{K}$ are chosen accordingly (cf. black dashed line in Fig. 2). As for the convergence speed, this also applies to the case of finite shots. However, as shown in Fig. 3, with a suitable regularization strength γ, the optimization of a full rank model ($m = 16$ Kraus operators in $\mathbb{K}$) can reach the same mean infidelity $1 - \mathcal{F}$. Crucially, the unregularized op-

timization converges to a much higher value. The figure also shows that stronger regularization can lead to faster convergence, but this comes at the cost of poorer average infidelity.

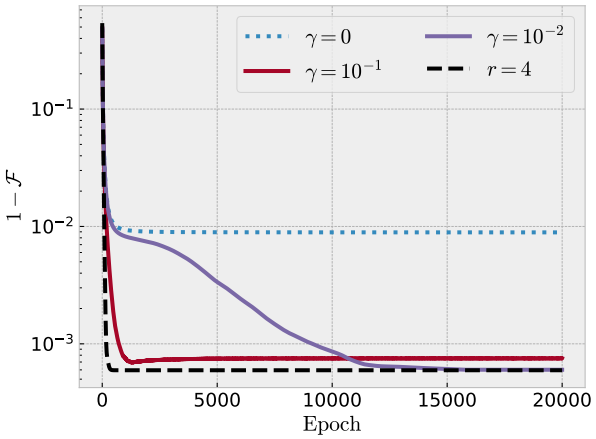

Figure 3: We plot the mean infidelity of 300 target channels with rank $r = 4$ for various values of γ using $\mathcal{R}_{\mathrm{HS}}$. The larger value $\gamma = 10^{-1}$ leads to faster convergence but does not reach the minimum infidelity. The optimal value $\gamma = 10^{-2}$ requires more epochs to converge but reaches the minimum that would also be obtained by a model with $m = r = 4$ Kraus operators (black dashed line). Here, $s = 10^4$ measurements are sampled for each input state ρ_α.

The optimal γ with the best mean infidelity in Fig. 3 has been found by a simple grid search over five values of $\gamma \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$. We perform the same search for the other two terms $\mathcal{R}_C$ and $\mathcal{R}_L$, and plot the infidelities for the respective best γ (with the lowest mean infidelity after 10^5 epochs) as a function of the training epoch in Fig. 4. Again, for comparison, we also plot the mean infidelity of a model with $m = r = 4$ Kraus operators in $\mathbb{K}$. Interestingly, in contrast to the infinite shot case, now also the term $\mathcal{R}_L$, based on the L_1-norm, provides an advantage over the unregularized case. However, this term as well as the Choi term $\mathcal{R}_C$ cannot compete with the Hilbert-Schmidt term $\mathcal{R}_{\mathrm{HS}}$ in the long run. On the other hand, for a small number of epochs (up to 8000 for the given example) the term $\mathcal{R}_C$ leads to the best results as it causes the fastest drop of infidelity in the beginning of the optimization.

The example shows that suitably chosen regularization terms in the cost function can be advantageous for Riemannian gradient descent approaches to quantum process tomography. However, the specific choice of a term may depend on whether the aim is to achieve rapid convergence of the algorithm or the highest possible fidelity with the target channel.

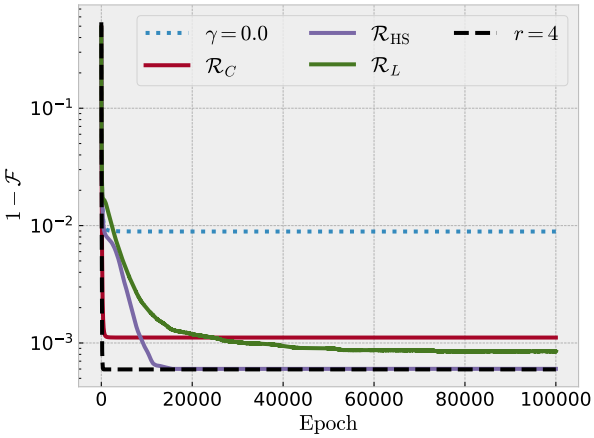

Figure 4: We plot the mean infidelity of models trained with different regularization terms $\mathcal{R}$. For each curve, 200 target channels are sampled. We perform a grid search over five values of γ, plotting the value with minimum mean infidelity for each regularization term after 10^5 epochs. The performance of models trained with the correct target channel rank $r = 4$ is plotted in black for comparison. In the finite shot case with number of shots $s = 10^4$, all regularization terms reach a significantly lower infidelity for some choice of γ than the unregularized case $\gamma = 0$. However, only $\mathcal{R}_{\mathrm{HS}}$ reaches the mean infidelity of models where the target channel rank is known in advance.

5.3 Optimization of the hyperparameter γ

Up until now, we have compared the impact of the regularization terms by examining the mean infidelity of the estimated channels T and target channels $\mathcal{E}$. In particular, the optimal γ for a specific term was determined from a minimization of the mean infidelity.

In an experimental setting, the infidelity is, however, not available as the target channel $\mathcal{E}$ is unknown. We thus need a different method to optimize the hyperparameter γ. Following Ref. [22], an often-used technique in machine learning to optimize hyperparameters splits the data set into

a training and a test set. The machine learning model is trained on the training data and its performance is tested on the unseen data in the test set. Similarly, we can split our measurement results into a training and test set to compare the performance of models with different values of γ. We use an 80/20 train/test split throughout all models trained in this paper. To that end, we independently draw the training and test set from a multinomial distribution $p(\beta|\alpha)$ for each input state ρ_α, splitting the number of shots s per input state.

In order to examine the performance of the regularization terms in an experimental setting where the infidelity is inaccessible, we perform a grid search over various values of γ as follows. We choose a list of 10 values (see App. C) for the regularization hyperparameter γ. For each γ, the models are trained on the training set. Then we choose the optimal γ^* by choosing the value with the lowest cost $\mathcal{L}$ on the test set. As we assume no knowledge of the target channels, we use full rank models to approximate the targets, setting $m = d^2$.

To benchmark this method, we then look at the difference $\Delta\mathcal{F} = \mathcal{F}_{\gamma^*} - \mathcal{F}_0$ between the fidelity of models with regularization strength γ^* and models with no regularization term. We plot the results for varying target channel rank in Fig. 5. For target channel ranks smaller than twelve, the grid search optimization of the regularization strength γ yields better results on average than training channels without such a term. This is in agreement with the fact that the regularization mainly works in scenarios where it can be assumed that the target channel is not of full rank. For the two-qubit case $d = 4$, the method performs particularly well for channels of ranks around $r = 5$. The largest advantages are obtained with the Hilbert-Schmidt term $\mathcal{R}_{\text{HS}}$. However, for ranks beyond $r = 12$ the Choi term $\mathcal{R}_C$ performs better. We stress that this possibility to optimize the hyperparameter γ based on experimentally accessible data makes the regularization a useful improvement of the Riemannian gradient descent method for quantum process tomography.

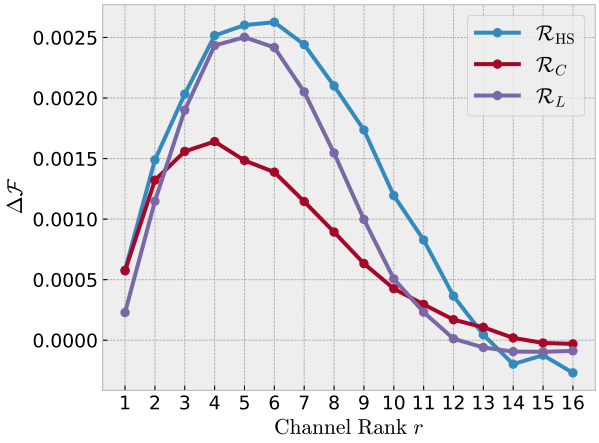

Figure 5: We plot the mean difference $\Delta\mathcal{F} = \mathcal{F}_{\gamma^*} - \mathcal{F}_0$ between the fidelity with optimized regularization strength γ^* and the fidelity of unregularized models. We consider this the experimentally relevant case where the fidelity is unknown and thus cannot be used to find the optimal γ. Instead, γ^* is optimized using a grid search over various γ, choosing the value with the lowest test set cost $\mathcal{L}$. All regularization terms provide an improvement over the unregularized case for some target channel ranks. Each model is trained for 10^5 epochs with $s = 10^5$.

The examples in this section show how regularization terms can improve the performance of Riemannian approaches to quantum process tomography. The method works best for rank-deficient channels. Note that a generic channel is of full rank. However, in many practically relevant settings such as quantum computers, the process is close to a low-rank channel. In this case, the regularization helps to suppress insignificant Kraus operators without fixing the rank of the optimized channel beforehand.

6 Quantum classification

Let us turn toward another example that fits into the framework of quantum channel optimization. A typical problem in quantum machine learning is the classification of classical data by means of a quantum mapping. Such a task is similar to the quantum process tomography setting we examined above. Instead of learning a channel T that reproduces a probability distribution $p_m(\beta|\alpha)$ given by measurement results of a to-

mography experiment, the aim here is to optimize a quantum transformation such that it learns a function $y = f(x)$ from a finite data set. This set $\{(x_i, y_i)\}$ includes inputs x_i sampled from some distribution $p(x)$ as well as their corresponding classes labelled by y_i. In order to demonstrate the impact of regularization terms added to the cost function of such a classification problem, we consider two different toy problems. Firstly, we use the Iris data set [51, 52] which comprises measurements of three species of the iris plant. Additionally, we consider the Wine data set [53], resulting from chemical analysis of different Italian wines. Both data sets have three target classes.

Each classical data vector $x \in \mathbb{R}^N$ is encoded into a quantum state $|\psi_x\rangle$. The choice of a suitable encoding is often crucial for the performance of a quantum machine learning task, both qualitatively and quantitatively [54, 55]. However, as we are not so much interested in the overall performance of the classification here but merely in the impact of the regularization, we do not optimize the encoding step in this paper. Instead, we choose a simple dense angle encoding which can consistently encode the data of both data sets. The classical data x is represented by an $\lceil N/2 \rceil$-qubit pure quantum state of the form [56]

$$|\psi_x\rangle = \bigotimes_{i=1}^{\lceil N/2 \rceil} \cos\frac{\pi}{2}x_{2i-1}\,|0\rangle + e^{i2\pi x_{2i}} \sin\frac{\pi}{2}x_{2i-1}\,|1\rangle\,, \tag{17}$$

where N is the total number of classical features. The encoded data is mapped by the channel T and subsequently measured. Contrary to the process tomography case, the POVM $\{M_\beta\}$ is given by projectors $\{|\beta\rangle\langle\beta|\}$ onto the computational basis. Thus, the measured data amounts to a probability distribution $p(\beta|x)$. The desired label y of the class is then given by the outcome β with the largest probability, i.e.,

$$\begin{aligned} y = f_T(x) &= \arg\max_\beta p(\beta|x) \\ &= \arg\max_\beta \langle\beta|\, T[|\psi_x\rangle\langle\psi_x|]\, |\beta\rangle\,. \end{aligned} \tag{18}$$

We can then use the Riemannian gradient descent method in order to optimize T such that its statistics yield a good approximation of the desired function, i.e., $f_T(x) \approx f(x)$.

The channel T determines the conditional probability $p(\beta|x_i)$ for each input sample. A typical cost function for our problem is then given by the cross entropy

$$\mathcal{L} = -\sum_i \ln p(\beta = y_i|x_i) + \gamma\mathcal{R}, \tag{19}$$

where $\mathcal{R}$ is again one of the three regularization terms defined in Sec. 4, and y_i is the correct classification of input x_i as given by the training data set. γ is a hyperparameter controlling the strength of the regularization.

We split the data into a training and a test set, optimize the channel T with the Riemannian gradient descent method on the training data, and apply the result to the test data in order to see how the classifier performs. The Iris data set consists of 150 data points with $N = 4$ features which are encoded in a two-qubit input state. The data belongs to three different classes, i.e., after applying the channel T to the input states, only three of the four projectors $|\beta\rangle\langle\beta|$ are considered to determine the classification outcome in Eq. (18). The objects in the Wine data set have 13 features. In a classical pre-processing, we reduce this number to $N = 8$ and encode them in a three-qubit input state (see App. B). The total number of 178 data points in this set belong to three different classes.

Unlike in process tomography, the input states $|\psi_x\rangle$ generally do not span the whole state space, nor is the measurement informationally complete. Thus, in general, the optimal channel T will not be unique. In particular, optimal T can differ in rank. We expect the regularization terms to steer the optimization toward a channel which solves the problem with as few Kraus operators as possible. Importantly, this regularization must not degrade the accuracy of the classification itself.

In order to see that this is indeed the case, we first plot the accuracy of the classification as a function of the regularization strength in Fig. 6. For both example data sets, we see that for sufficiently small regularization strength γ, the accu-

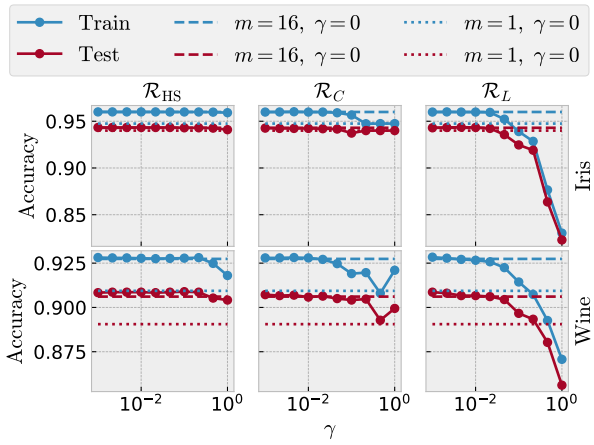

Figure 6: We plot the average accuracy of the classification on training and test data as a function of the regularization strength γ. For small γ, the accuracy is independent of the regularization, thus the additional term $\mathcal{R}$ in the cost function does not compromise the accuracy of the classification. For comparison, we plot the unregularized case (dashed line). The dotted lines correspond to a unitary model $m = 1$, which performs significantly worse than the general model with $m = 16$ Kraus operators. All plotted values are averages over 100 random splits of training and test data after 1500 (Iris data set) or 750 (Wine data set) epochs of training. Each optimization is initialized in a randomly sampled unitary channel.

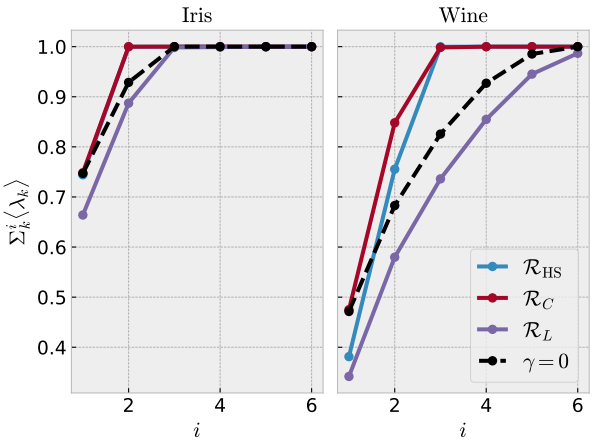

Figure 7: We plot the sum over the i largest eigenvalues of the channel's Choi state χ, averaged over 100 random training/test splits and initializations. The regularization terms $\mathcal{R}_C$ and $\mathcal{R}_{\mathrm{HS}}$ reduce the rank of the optimized channel T, particularly visible for the Wine data set. The unregularized case (black dashed line) converges to channels of rank six. For the regularized models, already the first three eigenvalues sum to unity, i.e., the channel can be represented by only three Kraus operators. For the Iris data set, the terms $\mathcal{R}_{\mathrm{HS}}$ and $\mathcal{R}_C$ reduce the rank from three to two. The term $\mathcal{R}_L$ shows the opposite behavior and is therefore not helpful. For the regularization with $\mathcal{R}_{\mathrm{HS}}$ we use $\gamma = 0.22$, for $\mathcal{R}_C$ and $\mathcal{R}_L$ we have $\gamma = 0.02$. These values lie in the plateau regions of Fig. 6. Thus, they are chosen so as not to compromise the accuracy of the classification.

racy both on the training and on the test set is independent of the regularization. For comparison, we plot the accuracy of the unregularized case $\gamma = 0$ (dashed lines). We also plot the accuracy that can be reached by a unitary model ($m = 1$ Kraus operator, dotted lines), which is significantly worse than the general case of a model with $m = 16$ Kraus operators. This shows that for a fixed dimension of the quantum system, a non-unitary transformation can in general be a better classifier than a unitary circuit. Clearly, according to Stinespring's dilation theorem, the same result could be obtained unitarily if sufficiently many ancilla qubits were added [57]. However, as we will see shortly, the regularized quantum channel model allows us to determine the minimal rank needed to accurately classify the data encoded in a quantum system of given dimension. This information can yield insight into the question of what complexity in a quantum circuit is actually needed to classify certain data sets.

For our toy problems, we find that the channel does not need to be full-rank. Instead, the Iris data can be classified by a channel of rank $r = 2$, while rank $r = 3$ is necessary for the Wine data set. This becomes visible in Fig. 7, where we plot the average sum of the i first eigenvalues of the channels Choi state χ. An optimization without regularization converges on average to a channel of rank $r = 3$ for the Iris data set and to a channel of rank $r = 6$ for the Wine toy problem, i.e., only the three (six) largest eigenvalues have a significant magnitude. However, with the regularization terms $\mathcal{R}_C$ and $\mathcal{R}_{\mathrm{HS}}$, only two nonzero eigenvalues remain in the Iris case and three in the Wine case. The regularization strength γ was chosen to be in the saturated regions of Fig. 6. Thus, this regularization decreases the rank of the channel without compromising its classification accuracy. The third regularization term $\mathcal{R}_L$

is of no help in this scenario, as it tends to increase the rank of the channel T. This is consistent with the findings for the infinite shot case of quantum process tomography in Sec. 5.1.

The examples discussed here highlight the significant role a regularization term can play for the numerical optimization of classifying quantum channels. Remarkably, the classification accuracy is invariant for a wide range of the regularization strength γ, as seen in Fig. 6. This can be attributed to the fact that the optimal channel for the classification problem is not unique, unlike in a quantum process tomography task. The regularization terms used here lift this ambiguity and favor a low-rank solution. The formalism is flexible and could in the same way include terms for other desired features of the channel, for example a particularly large overlap with a predefined channel that is easy to implement experimentally. Additionally, the regularization could include terms that bias the solution according to some classical knowledge we might have about the input data.

7 Conclusion

In this paper, we analyze the influence of different regularization terms on the performance of Riemannian optimization of quantum channels. We find that the use of such terms can be advantageous, especially in situations where the optimal solution of the problem is a channel that is not of full rank.

In quantum process tomography, the method leads to improvements for the accuracy of the optimization of channels of unknown rank. The regularization term in the cost function favors solutions with fewer non-zero Kraus operators in the Stiefel vector $\mathbb{K}$ and supports the convergence of the algorithm in particular for low-rank channels. Of the three terms considered, the Hilbert-Schmidt term $\mathcal{R}_{\mathrm{HS}}$ performs best in most cases. This becomes particularly visible for experimentally relevant scenarios of finite measurement data.

The method sensitively depends on the chosen regularization strength γ. We discuss how this hyperparameter can be optimized when the target channel is not known in advance, as is the case for quantum process tomography. By splitting the measured data into training and test sets, suitable values of γ can be determined, that outperform the unregularized case $\gamma = 0$ for most ranks r of the target channel. It has to be emphasized, though, that when the target channel rank is known, the regularization terms fail to provide an advantage. In this case, one can choose the number of Kraus operators in the Stiefel vector $\mathbb{K}$ to equal the target channel rank, eliminating the need for regularization.

As a second field of application of regularized Riemannian gradient descent, we investigate a simple quantum classification setting. Instead of classifying data by parametrized unitary circuits, we use a model of full-rank channels. Clearly, the implementation and in particular the training of a quantum channel is in general difficult compared to the parametrized unitary case. Thus, the method should not be seen as a tool to achieve practical quantum advantages. Instead, it can provide insight in the structure of a quantum circuit needed for the classification of specific data. In particular, the rank-decreasing form of the regularization terms can help to understand which minimum rank is needed to solve a certain classification problem for input data encoded in a quantum system of fixed dimension.

For both quantum process tomography and quantum machine learning, the computational costs for the method scale exponentially in the dimension of the involved quantum system, rendering the approach infeasible for many experimentally interesting scenarios. Extending the proof of concept presented in this paper to larger quantum systems is an important subject of ongoing research. Recently, channel representations in the form of matrix product states have been proposed to overcome the exponential scaling for process tomography [22]. The use of such compression techniques in the Riemannian gradient descent method will be a powerful tool for vari-

ous applications of channel optimization such as those presented here.

Acknowledgment

We are grateful to Oscar Dahlsten for valuable support and encouragement throughout this project. We thank Richard Hartmann for helpful advice on the numerical implementation. The computations were performed on a Cluster at the Center for Information Services and High Performance Computing (ZIH) at TU Dresden.

References

[1] Daniel A. Lidar and Todd A. Brun, editors. "Quantum Error Correction". Cambridge University Press. Cambridge (2013).

[2] Earl T. Campbell, Barbara M. Terhal, and Christophe Vuillot. "Roads towards fault-tolerant universal quantum computation". Nature **549**, 172–179 (2017).

[3] John Preskill. "Quantum Computing in the NISQ era and beyond". Quantum **2**, 79 (2018).

[4] Avimita Chatterjee, Koustubh Phalak, and Swaroop Ghosh. "Quantum Error Correction For Dummies" (2023). arxiv:2304.08678.

[5] V. V. Sivak, A. Eickbusch, B. Royer, S. Singh, I. Tsioutsios, S. Ganjam, A. Miano, B. L. Brock, A. Z. Ding, L. Frunzio, S. M. Girvin, R. J. Schoelkopf, and M. H. Devoret. "Real-time quantum error correction beyond break-even". Nature **616**, 50–55 (2023).

[6] Sascha Heußen, David F. Locher, and Markus Müller. "Measurement-Free Fault-Tolerant Quantum Error Correction in Near-Term Devices". PRX Quantum **5**, 010333 (2024).

[7] John M. Martinis. "Qubit metrology for building a fault-tolerant quantum computer". npj Quantum Information **1**, 1–3 (2015).

[8] Alexander Erhard, Joel J. Wallman, Lukas Postler, Michael Meth, Roman Stricker, Esteban A. Martinez, Philipp Schindler, Thomas Monz, Joseph Emerson, and Rainer Blatt. "Characterizing large-scale quantum computers via cycle benchmarking". Nature Communications **10**, 5347 (2019).

[9] Robin Harper, Steven T. Flammia, and Joel J. Wallman. "Efficient learning of quantum noise". Nature Physics **16**, 1184–1188 (2020).

[10] Isaac L. Chuang and M. A. Nielsen. "Prescription for experimental determination of the dynamics of a quantum black box". Journal of Modern Optics **44**, 2455–2467 (1997).

[11] J. F. Poyatos, J. I. Cirac, and P. Zoller. "Complete Characterization of a Quantum Process: The Two-Bit Quantum Gate". Physical Review Letters **78**, 390–393 (1997).

[12] G. M. D'Ariano and P. Lo Presti. "Quantum Tomography for Measuring Experimentally the Matrix Elements of an Arbitrary Quantum Operation". Physical Review Letters **86**, 4195–4198 (2001).

[13] Jaromír Fiurášek and Zdeněk Hradil. "Maximum-likelihood estimation of quantum processes". Physical Review A **63**, 020101 (2001).

[14] Massimiliano F. Sacchi. "Maximum-likelihood reconstruction of completely positive maps". Physical Review A **63**, 054104 (2001).

[15] Saleh Rahimi-Keshari, Artur Scherer, Ady Mann, A. T. Rezakhani, A. I. Lvovsky, and Barry C. Sanders. "Quantum process tomography with coherent states". New Journal of Physics **13**, 013006 (2011).

[16] Aamir Anis and A. I. Lvovsky. "Maximum-likelihood coherent-state quantum process tomography". New Journal of Physics **14**, 105021 (2012).

[17] G.A.L. White, F.A. Pollock, L.C.L. Hollenberg, K. Modi, and C.D. Hill. "Non-Markovian Quantum Process Tomography". PRX Quantum **3**, 020344 (2022).

[18] Leonardo Banchi, Jason Pereira, Seth Lloyd, and Stefano Pirandola. "Convex optimization of programmable quantum computers". npj Quantum Information **6**, 1–10 (2020).

[19] George C. Knee, Eliot Bolduc, Jonathan Leach, and Erik M. Gauger. "Quantum pro-

cess tomography via completely positive and trace-preserving projection". Physical Review A **98**, 062336 (2018).

[20] Trystan Surawy-Stepney, Jonas Kahn, Richard Kueng, and Madalin Guta. "Projected Least-Squares Quantum Process Tomography". Quantum **6**, 844 (2022).

[21] Adriano Macarone Palmieri, Egor Kovlakov, Federico Bianchi, Dmitry Yudin, Stanislav Straupe, Jacob D. Biamonte, and Sergei Kulik. "Experimental neural network enhanced quantum tomography". npj Quantum Information **6**, 1–5 (2020).

[22] Giacomo Torlai, Christopher J. Wood, Atithi Acharya, Giuseppe Carleo, Juan Carrasquilla, and Leandro Aolita. "Quantum process tomography with unsupervised learning and tensor networks". Nature Communications **14**, 2858 (2023).

[23] R. C. Bialczak, M. Ansmann, M. Hofheinz, E. Lucero, M. Neeley, A. D. O'Connell, D. Sank, H. Wang, J. Wenner, M. Steffen, A. N. Cleland, and J. M. Martinis. "Quantum process tomography of a universal entangling gate implemented with Josephson phase qubits". Nature Physics **6**, 409–413 (2010).

[24] Jerry M. Chow, Jay M. Gambetta, A. D. Córcoles, Seth T. Merkel, John A. Smolin, Chad Rigetti, S. Poletto, George A. Keefe, Mary B. Rothwell, J. R. Rozen, Mark B. Ketchen, and M. Steffen. "Universal Quantum Gate Set Approaching Fault-Tolerant Thresholds with Superconducting Qubits". Physical Review Letters **109**, 060501 (2012).

[25] Andrey V. Rodionov, Andrzej Veitia, R. Barends, J. Kelly, Daniel Sank, J. Wenner, John M. Martinis, Robert L. Kosut, and Alexander N. Korotkov. "Compressed sensing quantum process tomography for superconducting quantum gates". Physical Review B **90**, 144504 (2014).

[26] S. Krinner, S. Lazar, A. Remm, C.K. Andersen, N. Lacroix, G.J. Norris, C. Hellings, M. Gabureac, C. Eichler, and A. Wallraff. "Benchmarking Coherent Errors in Controlled-Phase Gates due to Spectator Qubits". Physical Review Applied **14**, 024042 (2020).

[27] L. C. G. Govia, G. J. Ribeill, D. Ristè, M. Ware, and H. Krovi. "Bootstrapping quantum process tomography via a perturbative ansatz". Nature Communications **11**, 1084 (2020).

[28] G. a. L. White, C. D. Hill, F. A. Pollock, L. C. L. Hollenberg, and K. Modi. "Demonstration of non-Markovian process characterisation and control on a quantum processor". Nature Communications **11**, 6301 (2020).

[29] J. L. O'Brien, G. J. Pryde, A. Gilchrist, D. F. V. James, N. K. Langford, T. C. Ralph, and A. G. White. "Quantum Process Tomography of a Controlled-NOT Gate". Physical Review Letters **93**, 080502 (2004).

[30] A. Shabani, R. L. Kosut, M. Mohseni, H. Rabitz, M. A. Broome, M. P. Almeida, A. Fedrizzi, and A. G. White. "Efficient Measurement of Quantum Dynamics via Compressive Sensing". Physical Review Letters **106**, 100401 (2011).

[31] M. Riebe, K. Kim, P. Schindler, T. Monz, P. O. Schmidt, T. K. Körber, W. Hänsel, H. Häffner, C. F. Roos, and R. Blatt. "Process Tomography of Ion Trap Quantum Gates". Physical Review Letters **97**, 220407 (2006).

[32] Yaakov S. Weinstein, Timothy F. Havel, Joseph Emerson, Nicolas Boulant, Marcos Saraceno, Seth Lloyd, and David G. Cory. "Quantum process tomography of the quantum Fourier transform". The Journal of Chemical Physics **121**, 6117–6133 (2004).

[33] Shahnawaz Ahmed, Fernando Quijandría, and Anton Frisk Kockum. "Gradient-descent quantum process tomography by learning kraus operators". Phys. Rev. Lett. **130**, 150402 (2023).

[34] Nicolas Boumal. "An introduction to optimization on smooth manifolds". Cambridge University Press. (2023).

[35] Jun Li, Li Fuxin, and Sinisa Todorovic. "Efficient riemannian optimization on the stiefel manifold via the cayley transform" (2020). arXiv:2002.01113.

[36] A. Shabani, R. L. Kosut, M. Mohseni, H. Rabitz, M. A. Broome, M. P. Almeida, A. Fedrizzi, and A. G. White. "Efficient measurement of quantum dynamics via compressive sensing". Phys. Rev. Lett. **106**, 100401 (2011).

[37] Steven T Flammia, David Gross, Yi-Kai Liu, and Jens Eisert. "Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators". New Journal of Physics **14**, 095022 (2012).

[38] C. A. Riofrío, D. Gross, S. T. Flammia, T. Monz, D. Nigg, R. Blatt, and J. Eisert. "Experimental quantum compressed sensing for a seven-qubit system". Nature Communications **8**, 15305 (2017).

[39] Y. S. Teo. "Objective compressive quantum process tomography". Physical Review A**101** (2020).

[40] Ilia A. Luchnikov, Alexander Ryzhov, Sergey N. Filippov, and Henni Ouerdane. "QGOpt: Riemannian optimization for quantum technologies". SciPost Phys. **10**, 079 (2021).

[41] Roeland Wiersema and Nathan Killoran. "Optimizing quantum circuits with riemannian gradient flow". Phys. Rev. A **107**, 062421 (2023).

[42] Edward Farhi and Hartmut Neven. "Classification with Quantum Neural Networks on Near Term Processors" (2018) arXiv:1802.06002.

[43] 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**, 2631 (2021).

[44] Jonas Kübler, Simon Buchholz, and Bernhard Schölkopf. "The Inductive Bias of Quantum Kernels". In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan, editors, Advances in Neural Information Processing Systems. Volume 34, pages 12661–12673. Curran Associates, Inc. (2021). arXiv:2106.03747.

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

[46] Maria Schuld and Nathan Killoran. "Is quantum advantage the right goal for quantum machine learning?". PRX Quantum **3**, 030101 (2022).

[47] Joseph Bowles, Shahnawaz Ahmed, and Maria Schuld. "Better than classical? the subtle art of benchmarking quantum machine learning models" (2024). arXiv:2403.07059.

[48] Hemant D Tagare. "Notes on Optimization on Stiefel Manifolds". Online resource (2011).

[49] Robert Tibshirani. "Regression Shrinkage and Selection Via the Lasso". Journal of the Royal Statistical Society: Series B (Methodological) **58**, 267–288 (1996).

[50] Ingemar Bengtsson and Karol Zyczkowski. "Geometry of Quantum States: An Introduction to Quantum Entanglement". Cambridge University Press. Cambridge (2006).

[51] Edgar Anderson. "The species problem in iris". Annals of the Missouri Botanical Garden **23**, 457–509 (1936).

[52] Ronald A. Fisher. "The use of multiple measurements in taxonomic problems". Annals of Eugenics **7**, 179–188 (1936).

[53] Stefan Aeberhard, Danny Coomans, and Olivier de Vel. "Comparative analysis of statistical pattern recognition methods in high dimensional settings". Pattern Recognition **27**, 1065–1077 (1994).

[54] Maria Schuld and Nathan Killoran. "Quantum machine learning in feature hilbert spaces". Phys. Rev. Lett. **122**, 040504 (2019).

[55] Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. "Effect of data encoding on the expressive power of variational quantum-machine-learning models". Phys. Rev. A **103**, 032430 (2021).

[56] Ryan LaRose and Brian Coyle. "Robust data encodings for quantum classifiers". Physical Review A **102**, 032420 (2020).

[57] W. Forrest Stinespring. "Positive functions on c*-algebras". Proceedings of the American Mathematical Society **6**, 211–216 (1955).

[58] Francesco Mezzadri. "How to generate random matrices from the classical compact groups" (2007). arXiv:math-ph/0609050.

[59] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. "Scikit-learn: Machine Learning in Python". The Journal of Machine Learning Research **12**, 2825–2830 (2011).

[60] Ian T. Jolliffe and Jorge Cadima. "Principal component analysis: a review and recent developments". Philosophical transactions. Series A, Mathematical, physical, and engineering sciences **374**, 20150202 (2016).

A Channel initialization

To initialize the optimization algorithm and to sample random quantum channels, we follow the method outlined in Ref. [58], which is based on the QR decomposition. This allows us to generate full-rank quantum channels on a given Stiefel manifold. In Sec. 6, we initialize the parameters $\mathbb{K}_0$ of the optimization algorithm as a unitary channel, regardless of the number of Kraus operators in the optimization manifold $\mathrm{St}(n_k \cdot d, d)$. To that end, we draw n_k real numbers $x_i \sim \mathcal{U}(0,1)$ as well as a unitary $u \in U(d)$ drawn from the Haar measure [58]. The initial parameters are then given by

$$\mathbb{K}_0 = [\sqrt{x_1}u, ..., \sqrt{x_{n_k}}u]^T / \sqrt{\Sigma_{i=1}^{n_k} x_i}. \tag{20}$$

B QML Data preprocessing

Both data sets are accessed through `scikit-learn` [59].

We encode classical data vectors $x \in \mathbb{R}^N$ using $\left\lceil \frac{N}{2} \right\rceil$ qubits and employ dense angle encoding as explained in the main text. Before the encoding, the classical data is transformed as

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \tag{21}$$

where $x_{\max}, x_{\min}$ are the element-wise maximum or minimum respectively. They are calculated on the training set only.

The Iris data set consists of four-dimensional classical data. No further preprocessing is needed, and the data can be encoded using two qubits and dense angle encoding.

The Wine data set however has 13 features. Here, we reduce the dimensionality of the classification problem by performing a principal component analysis on the training data [60]. We choose only the six most important principal components. This preprocessed data set can now be encoded using three qubits and dense angle encoding.

C Grid search values

In Sec. 5.3, the grid search is performed over the values [0.0, 0.0001, 0.000215, 0.000464, 0.001, 0.002154, 0.004642, 0.01, 0.021544, 0.046416, 0.1], where the best value of γ minimizes the test set Kullback-Leibler divergence.