Noise and trainability in quantum machine learning

Valentin Heyraud

Université Paris Cité

MPQ

Université Paris Cité

École doctorale 564 : Physique en Île-de-France

Laboratoire Matériaux et Phénomènes Quantiques (UMR 7162)

Thèse de doctorat en physique

Noise and trainability in quantum machine learning

présentée par

Valentin HEYRAUD

sous la direction de

Cristiano Ciuti

Jury

PR	Alessandra Di Pierro	Università di Verona	Rapporteuse
PR	Cristiano Ciuti	Université Paris Cité	Directeur de thèse
DR	Mazyar Mirrahimi	INRIA	Examinateur
PR	Zoë Holmes	EPFL	Rapporteuse

Présentée et soutenue publiquement à Paris le 7 Novembre 2023

This work was supported by Région Île-de-France in the framework of DIM SIRTEQ (Domaine d'Intérêt Majeur Science et Ingénierie en Région Île-de-France pour les Technologies Quantiques).

Summary

Noise and trainability in Quantum Machine Learning

This thesis is devoted to the exploration of the interface between quantum computing and machine learning, with an emphasis on the effects of noise and decoherence. In the first part, we investigate the use of open quantum systems to tackle classical pattern recognition tasks. In particular, we study the impact of noise on quantum kernel machines in a reservoir computing setup. The models we consider are based on the use of a large and uncontrolled quantum system, the reservoir, that is excited with an input signal to be processed. Measurements are then performed on the system, and a linear combination of the outcomes is optimized to achieve the desired processing task. Within the theoretical framework associated with kernel methods, we analyze the effect of dissipation on the expressive power of these models. We show that the noise affecting the reservoir can act as an implicit regularization that helps to prevent over-fitting. These findings are supported by a numerical study of a set of noisy kernel machines based on driven-dissipative chains of spins exhibiting decoherence and whose Markovian evolution is described by a Lindblad master equation. The second part of the thesis focuses on variational quantum algorithms. There, we present an efficient classical simulation scheme to estimate the trainability of a parameterized quantum circuit. We first study the quantum channels associated with the averages of random Z-rotations of one and two qubits. Upon some assumptions, we show that these average rotation channels can be decomposed into convex sums of Clifford channels. This result, which can be interpreted as an artificial decoherence induced by the random choice of the rotation angles, allows us to derive our efficient estimation scheme based on the celebrated Gottesman-Knill theorem. Among other figures of merits, this method enables to efficiently estimate the average amplitude of the cost-function gradient through classical simulations. This method is scalable and can be used to certify trainability for variational quantum circuits and explore design strategies that can overcome the barren plateau problem.

Keywords: open quantum systems, quantum information, machine learning, kernel methods, reservoir computing, decoherence, quantum computing, variational quantum algorithms, Clifford circuits

Résumé

Bruit et entraînabilité en apprentissage automatique quantique

Cette thèse est consacrée à l'exploration de l'interface entre l'informatique quantique et l'apprentissage automatique, en mettant l'accent sur les effets du bruit et de la décohérence. Dans la première partie, nous étudions l'utilisation de systèmes quantiques ouverts pour la réalisation de tâches classiques de reconnaissance des formes. En particulier, nous nous intéressons à l'impact du bruit sur les machines à noyau quantique dans une configuration de calcul à réservoir. Les modèles que nous considérons sont basés sur l'utilisation d'un système quantique de grande taille et non contrôlé, un réservoir, qui est excité par un signal d'entrée à traiter. Des mesures sont ensuite effectuées sur le système et une combinaison linéaire des résultats est optimisée afin de réaliser la tâche souhaitée. Dans le cadre théorique associé aux méthodes à noyau, nous analysons l'effet de la dissipation sur l'expressivité de ces modèles. Nous montrons que le bruit affectant le réservoir peut agir comme une régularisation implicite qui aide à prévenir l'ajustement excessif. Ces résultats sont accompagnés d'une étude numérique d'un ensemble de machines à noyaux bruitées basées sur des chaînes de spins quantiques sujets à de la décohérence, et dont l'évolution markovienne est décrite par une équation maîtresse de type Lindblad. Dans la deuxième partie de la thèse, nous nous concentrons sur les algorithmes quantiques variationnels. Nous y présentons un schéma de simulation classique efficace pour estimer des quantités moyennées à la sortie d'un circuit quantique paramétré dont les paramètres de rotation sont choisis aléatoirement. Nous étudions d'abord les canaux quantiques associés aux moyennes de rotations aléatoires d'un et de deux qubits autour d'un axe Z. Sous certaines hypothèses, nous montrons que ces canaux peuvent être décomposés en sommes convexes de canaux associés à des transformations unitaires de Clifford. Ce résultat peut être interprété comme un effet de décohérence artificiel induit par le choix aléatoire des angles de rotation. Il nous permet construire un schéma de simulation efficace basé sur le célèbre théorème de Gottesman-Knill. Cette méthode permet notamment d'estimer efficacement l'amplitude moyenne du gradient de la fonction de coût par des simulations classiques avec une complexité polynomiale en le nombre de qubits. Elle peut donc être utilisée pour explorer des stratégies de conception de circuits variationnels quantiques permettant d'éviter les difficultés d'entrainement due à une disparition du gradient.

Mots-clés : systèmes quantiques ouverts, information quantique, intelligence artificielle, apprentissage automatique, méthodes à noyau, informatique de réservoir, décohérence, calcul quantique, algorithmes quantiques variationnels, circuits Clifford

Publications

The scientific publications produced during this doctoral thesis are listed below and are indexed with Greek letters. The present manuscript is based mainly upon the results reported in [α] and [γ].

[α] V. Heyraud, Z. Li, Z. Denis, A. Le Boité and C. Ciuti,
"Noisy quantum kernel machines",
Physical Review A **106**, 052421 (2022).

[β] Z. Li, V. Heyraud, K. Donatella, Z. Denis and C. Ciuti,
"Machine learning via relativity-inspired quantum dynamics",
Physical Review A **106**, 032413 (2022).

[γ] V. Heyraud, Z. Li, K. Donatella, A. Le Boité and C. Ciuti,
"Efficient estimation of trainability for variational quantum circuits",
arXiv:2302.04649, 032413 (2023).

Contents

General introduction

Since its advent at the dawn of the 19th century, quantum mechanics has had a strong impact on our societies. The first quantum revolution brought the technologies in which the current age of information is rooted, namely transistors, lasers and atomic clocks [1]. The invention of the transistor by Bardeen, Shockley and Brattain [2, 3] in 1947 paved the way to the miniaturization of computers and to the exponential rise of the computational power that began in the late 1960's. Lasers developed at the end of the 1950's, following the works of Kastler on optical pumping [4], have found numerous applications and are nowadays crucial for transferring ever larger amounts of data at sufficient rates [1].

Besides its technological outcomes, quantum mechanics has dramatically changed our understanding of Nature. The development of the quantum theory required to abandon some rather natural assumptions inherent to a classical description of the universe. This led Einstein, Podolsky and Rosen to propose in their famous 1935 article that quantum mechanics might be incomplete [5]. Thirty years later, Bell proved the celebrated inequalities that would allow to adress this hypothesis experimentally [6, 7]. The experimental verification of the violation of Bell's inequalities by Clauser and Shimony in 1978 [8, 9] and Aspect in 1982 [10, 11] showed the bewildering implications of quantum mechanics to be features of Nature rather than artifacts of an incomplete theory.

In the same period, the first proposals to use quantum systems to process and transmit information appeared, giving birth to the field of quantum information [12]. In 1980, Paul Benioff formalized the notion of quantum computer by introducing the first model of quantum Turing machine [13, 14]. Soon after, Feynman and Manin proposed to use such quantum computers to simulate quantum systems [15, 16], a task requiring an exponential amount of ressources on a classical device. In the years following these proposals, the idea of a quantum computer blossomed, and the field of quantum computing started to define itself as a discipline, at the interface of quantum mechanics and computer science. In particular, the first theoretical quantum speedups over classical algorithms were proven with the works of Deutsch, Vazirani, Bernstein, Simon, Lloyd, Grover and Shor [17–22], culminating in the celebrated Shor's quantum algorithm for the factorization of prime numbers. All these results proved that the features specific to the quantum world could be used as a resource for technological purpose, thereby igniting the so-called second quantum revolution [23].

The existence of quantum algorithms able to solve problems that are practically out of reach of classical methods has generated a great deal of interest in quantum computing. Over the past thirty years, a lot of efforts have been made to build actual quantum computers able to demonstrate such a quantum advantage [24, 25]. However, constructing a quantum computer is a daunting task. On the one hand, quantum algorithms rely on

the use of delicate quantum coherent states [26]. As a result of the interactions of the devices with their environment, these coherences are progressively destroyed [27]. To avoid this decoherence mechanism, it is thus necessary to efficiently isolate the quantum systems. On the other hand, executing a quantum algorithm requires to manipulate the device state and to perform measurements on it, which can only be done through precisely controlled interactions. To meet this technological challenge, several hardware platforms have been proposed, including quantum dots [28, 29], quantum optical systems [30, 31], nuclear magnetic resonance platforms [32, 33], and the promising trapped ions [34–36] and superconducting circuits [37, 38]. In recent years, experimental demonstrations of a quantum advantage for specific tasks have been realized on some of these devices [39–41].

The search for a quantum advantage is a difficult task, and over the past decades significant advances have been made. It is clear that for a quantum algorithm to provide an advantage over classical methods, the algorithm outcomes must be hard to simulate with a classical computer. Hence, in the quest for a quantum advantage, an important effort has been devoted to finding ways to efficiently simulate quantum algorithms and systems using only classical resources. Efficient classical simulation schemes have been found for many cases [42–53], and some of these results have led to the questioning of the recent quantum advantage experiments [54, 55]. As a consequence, the range of cases where a quantum advantage remains possible has narrowed considerably.

As of today, we remain in the NISQ era described by Preskill five years ago [56], and the available quantum devices are noisy and limited in size. The detrimental effects of noise have been envisioned early in the history of quantum computing [57], and methods of quantum error corrections have been largely developed over the last decades [58–61]. Unfortunately, as for classical error correction, these methods rely on a redundant encoding of the information, which is difficult to implement given the quality and the size of the current devices. Despite their imperfect characteristics, current and near-term quantum devices may be able to provide a useful quantum advantage. Understanding their potential and limits and finding ways to exploit their power is a challenge, and much work has been devoted to this task over the last decade [62].

In parallel of the second quantum revolution, another field has enjoyed spectacular growth: machine learning. This field has faced multiple periods of stagnation, the so-called AI winters, that lasted until the end of the 1990s [63]. Machine learning has seen a revival at the beginning of the 21th century, pushed by the increase of computational power and the novel use of GPUs for machine learning calculations. The popularity of machine learning methods really exploded in the 2010s with impressive successes of Deep Learning techniques [64, 65]. Since then, the field has continued to develop, culminating in recent advances in reinforcement learning [66, 67], computer vision [68, 69], generative models [70–75] and large language models [76, 77]. Machine learning techniques have already deeply changed our societies, and they have largely diffused in other fields, including physics [50, 78–80].

Given this popularity and the current development of quantum technologies, the interaction between machine learning and quantum computing has naturally emerged as an exciting area of research. This new domain of research encompasses two dual approaches [81]. From one angle, machine learning methods could be of great use in developing a fault-tolerant quantum computer. For instance, reinforcement learning al-

gorithms have already shown potential in the context of quantum control and quantum error correction [82–93], and they have allowed striking recent experimental results in error correction [94]. Machine learning methods have also inspired the variational quantum algorithms [95], which are a class of versatile and shallow models that hold promise for harnessing the power of near-term devices.

The second approach, often called Quantum Machine Learning [96], focuses on a potential quantum advantage for machine learning related tasks. In this perspective, quantum algorithms have been envisioned to speed-up linear algebra sub-routines of classical machine learning algorithms [97]. Quantum generalizations of many existing machine learning methods have also been investigated, ranging from quantum neural networks to quantum kernel machines [95, 96, 98], with the hope that such quantum models could benefit from the large dimensions of the states spaces. Most of the proposed methods include a classical optimization step, a reason for which they are sometimes referred to as hybrid quantum-classical algorithms. A lot of progresses have been made in the understanding of the power and the limitations of these new algorithms [98–103]. However, despite the large amount of research dedicated to these topics, the potential advantage of these methods over their classical counterparts remains uncertain.

This thesis is devoted to the exploration of the interface between machine learning and quantum computing, with an emphasis on the role of noise and decoherence. It gathers results obtained during my PhD research on several topics and approaches described in the following.

Noisy quantum kernel machines

Kernel methods, also called Kernel machines, constitute a class of simple and versatile classical machine learning algorithms that are widely used in pattern analysis [104–106]. A prominent example of such methods are the Support Vector Machines, whose first quantum generalizations have been considered at the beginning of the 2000s [107]. The core idea of kernel methods is to extend linear classification techniques to non-linear setups, by making use of a set of non-linear transformations, or feature maps. The quantum analog of kernel methods, the so-called quantum kernel machines, have been widely studied [102, 108–117], and it has been shown that these methods encompass many supervised quantum machine learning algorithms [102, 114]. In the first part of this manuscript, we study the impact of the physical noise affecting current devices on the performances of quantum kernel machines. The main effect of noise is to reduce the expressive power of the kernel machines. This effect can be seen as an implicit regularization mechanism [1], for which we provide analytical evidences. We illustrate this phenomenon in a context of reservoir computing [118]. In this setup, the feature maps associated to the kernel machines are obtained by performing measurements on a quantum system, the reservoir, that is excited with a driving encoding an input signal to process. The previous effect is illustrated through numerical simulations of this scheme, where we take a set of one-dimensional driven-dissipative chains of coupled qubits as reservoirs.

[1] As such, this effect helps to avoid the well-known issue of over-fitting.

Trainability in Variational Quantum Algorithms

The second part of this thesis focuses on variational quantum algorithms [95]. These hybrid quantum-classical algorithms use the output of a parameterized quantum circuit as a variational ansatz. Depending on the task at hand, a cost function is measured on the circuit output, and the parameters of the ansatz are trained as to minimize the cost function using classical optimization techniques. This paradigm is well suited for near-term devices because it adapts easily to hardware constraints and allows the use of shallow quantum circuits, which limits the impact of the decoherence. Unfortunately, these algorithms can be difficult to train for large-scale problems due to the notorious barren plateaus phenomenon [119], in which the gradient of the cost function vanishes exponentially with the size of the circuit. This phenomenon has been extensively studied in the literature and has been shown to be caused by a wide variety of factors [120–123]. In this manuscript, we present an efficient classical method that allows to estimate the trainability, along with many figures of merit, for a specific class of variational quantum algorithms. This scheme is based on the celebrated Gottesman-Knill theorem [42, 45], a fundamental result which states that quantum algorithms associated with the so-called Clifford circuits can be simulated efficiently on a classical computer. The method is illustrated on a prototypical example of appearance of barren plateaus, and some limitations are identified. We also discuss the new prospects opened up by these results.

Structure of the manuscript

Chapter 1 provides an overview of the theory of open quantum systems and introduces the theoretical tools that will be used in the rest of this manuscript. We first revisit the statistical interpretation of quantum mechanics, and provide some reminders on the framework of quantum channels and density operators. The decoherence mechanism is then discussed through two simple models for a two-level system. We derive the master-equations describing the system's dynamics in these models, and we briefly discuss some of the conditions under which this description is valid.

Chapter 2 is devoted to the study of the impact of noise and decoherence on the performances of quantum kernel machines based on our work [α]. This chapter begins with the introduction of a general scheme of noisy quantum kernel machines relying on a reservoir computing approach. The role of noise and decoherence on the performances of these kernel machines is investigated analytically, and we show that the effect of noise can be seen as an intrinsic regularization of the corresponding models. Finally, these findings are supported by a numerical study of a class of noisy quantum kernel machines using driven-dissipative one-dimensional chains of spins.

Chapter 3 presents the results in our work [γ]. There we introduce a method to efficiently estimate the initial trainability of a class of variational quantum circuits through classical simulations. The general framework of variational quantum and trainability issues associated with these methods are introduced in the beginning. Then we presents technical results regarding quantum channels associated with random single-qubit rotations, on which our method is based. Building on these results we derive our method and

we prove its efficiency. The limits of the method are then discussed and the results are supported by a simple numerical study.

Finally, the general conclusion summarize the main results and discuss the perspectives and the ongoing research associated with the presented results.

1 Theory of open quantum systems

Perfectly isolated quantum systems are idealizations and in practice quantum systems are always interacting with their environment. This interactions can result from an imperfect isolation of the system or, more trivially, from the need to manipulate and measure it. In many situations, these couplings have a non-negligible effect on the system's dynamics. Understanding these effects is crucial as noise severely limits the practical applications of quantum devices, be it sensors or computer.

To model the dynamics of an open quantum system, it is both intractable and irrelevant to carry a complete microscopic description of the environment, which is often vast and uncontrolled. Fortunately, under fairly general assumptions one can provide a statistical descriptions of the system that allows the environment degrees of freedom to be removed. The theory of open quantum systems provides a general framework for describing this effective dynamics.

In this chapter we briefly review this framework with the underlying objective of introducing the theoretical tools that will be used through this thesis. In Sec. I we revisit the statistical interpretation of quantum mechanics and remind some useful results of classical probabilities. Sec. II then gives an overview of the framework of density operators and quantum channels, which allow to describe generic quantum systems. At last, in Sec. III we introduce the Lindblad master equation through the study of two simple models of decoherence.

I Classical probabilities and observables

In classical physics, the state of a system is described by a point in the phase space, whose coordinates represent physical quantities. It is sometimes useful to turn this description into a probabilistic one, for example if we need to take into account undesired interactions with the environment or uncertainties on the initial configuration of the system. In that case, the previous physical quantities are considered as random variables and the state of the system is represented by a joint probability of these variables.

In this view, quantum mechanics is fundamentally a statistical theory, as it provides such a probabilistic description for the outcomes of measurements made on a system. However, depending on both the measurements considered and the system state one might not be able to provide a classical joint probability law for the outcomes, a feature that distinguishes the quantum theory from a classical one.

In this section, we provide an overview of this statistical interpretation of quantum mechanics and introduces some of the mathematical notations and results that will be used through the text.

I.1 Mathematical preliminaries and notations

Let X be a real-valued random variable. We denote $\mathbb{P}_X$ its associated probability law. The *expectation* (or expected value) of a function h of X is denoted

$$\mathbb{E}\left[h(X)\right] = \int_{\mathbb{R}} h(x)\, \mathrm{d}\mathbb{P}_X(x)\ . \tag{1.1}$$

and the *characteristic function* of X is written

$$\varphi_X(t) = \mathbb{E}\left[e^{\mathrm{i}tX}\right],\ \forall t \in \mathbb{R}. \tag{1.2}$$

The characteristic function uniquely determines the law of X [124]. If X admits a density f_X with respect to the Lebesgue measure $\mathrm{d}x$, i.e. if $\mathrm{d}\mathbb{P}_X(x) = f_X(x)\mathrm{d}x$, then the characteristic function is simply the Fourier transform of f_X.

These definitions generalize to real random vectors taking values in $\mathbb{R}^n$ with $n \in \mathbb{N}$. For quantities depending on a random vector, the variables against which the expectation is taken are indicated in subscript. For example, for a function $h(X, Y)$ of the random vector (X, Y) we denote

$$\mathbb{E}_Y\left[h(X, Y)\right] = \int_{\mathbb{R}} h(X, y)\, \mathrm{d}\mathbb{P}_Y(y)\ , \tag{1.3}$$

where $\mathbb{P}_Y$ is the marginal probability of Y. The characteristic function of a random vector $\boldsymbol{X}$ with value in $\mathbb{R}^n$ becomes

$$\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = \mathbb{E}\left[e^{\mathrm{i}\boldsymbol{t}^T\boldsymbol{X}}\right],\ \forall \boldsymbol{t} \in \mathbb{R}^n, \tag{1.4}$$

where $\boldsymbol{t}^T$ is the transpose of $\boldsymbol{t}$.

I.2 Observables and statistical interpretation of quantum mechanics

The physical state of an isolated quantum system S is represented by a unit vector $|\phi\rangle$ belonging to a Hilbert space $\mathcal{H}_S$. In the canonical view of quantum mechanics, $|\phi\rangle$ encodes all the information about S and it is said to be a pure state [125]. A physical quantity A is represented by a self-adjoint operator $\hat{A}$ on $\mathcal{H}_S$, and both A and $\hat{A}$ are referred as *observables*. According to the von Neumann spectral theorem [126], any observable $\hat{A}$ can be uniquely written as

$$\hat{A} = \int_{\mathbb{R}} \lambda \mathrm{d}\hat{\mu}_A(\lambda), \tag{1.5}$$

where $\hat{\mu}_A$ is a projection-valued probability measure on $\mathbb{R}$. This allows to map any state $|\phi\rangle$ to a classical probability distribution $\mathbb{P}_A$ on $\mathbb{R}$ defined by

$$\mathbb{P}_A\left(E\right) = \langle\phi|\, \hat{\mu}_A\left(E\right) |\phi\rangle\ , \quad \forall E = [a, +\infty[\ ,\ a \in \mathbb{R}\,. \tag{1.6}$$

In the standard interpretation of quantum mechanics, this distribution is seen as the law of a real-valued random variable A representing the outcome of a measurement of the considered quantity. In particular, it holds that for any sufficiently well-behaved function f:

$$\begin{aligned} \mathbb{E}\left[f(A)\right] &= \int_{\mathbb{R}} f(\lambda) \mathrm{d}\mathbb{P}_A(\lambda) \\ &= \langle\phi| f(\hat{A}) |\phi\rangle \ . \end{aligned} \tag{1.7}$$

Moreover, the characteristic function of A can be written

$$\varphi_A(t) = \langle\phi| e^{\mathrm{i}t\hat{A}} |\phi\rangle \ . \tag{1.8}$$

Considering a set of n observables of the system gathered in a vector $\hat{\boldsymbol{A}} = (\hat{A}_1, \ldots, \hat{A}_n)$, it is natural to ask whether one can find a random vector $\boldsymbol{A} = (A_1, \ldots, A_n)$ with a joint probability distribution $\mathbb{P}_{\boldsymbol{A}}$ that would generalize the previous equations in some way. Extending Eq. (1.6) or Eq. (1.7) straightforwardly is fine if the $\hat{A}_i$ do commute. However, as emphasized by von Neumann in his foundational work [127], this extension is problematic for non-commuting observables, as in that case the order of the products of operators appearing in the generalized equations has an impact on the results, which is at odds with the corresponding classical expressions.

A potential workaround to this issue is to consider a generalization of Eq. (1.8) to $\hat{\boldsymbol{A}}$ using Eq. (1.4). In fact, defining the joint probability by its characteristic function would lift the ambiguity on the order of the operators products. This option is addressed by Nelson's theorem [128], according to which there exists a classical joint probability law which characteristic function generalizes Eq. (1.8) for every quantum state if and only if the observables commute. In particular, Nelson's theorem does not imply that for a set of non-commuting observables one can never find a classical joint probability satisfying Eq. (1.8) [1], but that there always exists a quantum state for which this is impossible. This impossibility is often seen a distinguishing feature of quantum mechanics [129], and it has remarkable connections with the existence of non-classical correlations in bipartite systems [130–132]. Specifically, it has been shown that for every set of non-commuting observables, there exists a bipartite system and a quantum state such that some Clauser-Horne-Shimony-Holt inequalities [8] involving the observables are violated [131]. On the other hand, a Bell-type experiment can result in a violation of Bell's inequalities only if the involved observables do not admit a joint probability [130].

Although questioning the existence of a coherent classical probabilistic description for the outcomes of measurements is an important aspect of the statistical interpretation of quantum mechanics, it does not tell much about the possibility of measuring different quantities in a real life experiment. The usual approach to this issue is to invoke Heisenberg's uncertainty theorem [125, 133], which imposes a shared constraint on the variances of non-commuting observables. Invoking to the wave-function collapse postulate, one sees that representing simultaneous measurements on a system by non-commuting observables is inconsistent, as it would result in a post-measurement state violating Heisenberg's in-

[1] For a quantum particle on the real line in a Gaussian state, the particle's Wigner function yields a valid joint probability distribution for the particle position and momentum.

equality [2]. With that regard, the theory only allows to jointly measure observables that commute, in which case they are said to be *compatible*. The previous argument heavily rely on the projection postulate. However, the general description of measurements uses positive operator valued measures rather than the projection operator valued measures considered above [137, 138]. In this framework, it is possible to define fuzzy measurements associated to non-commuting observables that can be performed jointly, in agreement with Heisenberg's principle [139].

Nevertheless, quantities associated to non-commuting observables can still be measured through independent experiments. To describe this setup, we consider that each observable acts on a separate copy of the system with the same state $|\phi\rangle$, thereby enforcing them to commute. As expected, this yields a joint law that is equal to the product of the probabilities given by Eq. (1.6) for each observable in the set, so that the observables are uncorrelated.

From the above, one can see that the randomness in the outcomes of measurements realised on quantum systems cannot be fully represented by means of classical probabilities. In the following, we stress this particularity by adopting the usual convention for quantum expectations and denote

$$\langle \hat{A} \rangle = \langle \phi | \, \hat{A} \, | \phi \rangle \, . \tag{1.9}$$

II Density operators and quantum channels

Like in classical physics, it is sometimes useful to turn a deterministic description of a system into a probabilistic one. For a quantum system, this statistical description takes the form of a classical mixture of pure states. These statistical ensembles and their evolution can be represented in the theoretical framework of density operators and quantum channels, which we briefly review in this section.

II.1 Density operators

A system S whose exact state is uncertain can be represented by a statistical mixture of pure states $\{|\phi_i\rangle \, , \; i \in I\}$ where to each state $|\phi_i\rangle$ is assigned a probability p_i. Such a system is said to be in a *mixed state* and it is described by a *density operator*

$$\hat{\rho} = \sum_{i \in I} p_i \, |\phi_i\rangle\langle\phi_i| \, . \tag{1.10}$$

As before, we denote $\langle \hat{A} \rangle$ the expectation of the measurement of an observable $\hat{A}$ on S, and we have

$$\langle \hat{A} \rangle = \sum_{i \in I} p_i \, \langle \phi_i | \, \hat{A} \, | \phi_i \rangle = \mathrm{Tr} \left[\hat{A} \hat{\rho} \right] , \tag{1.11}$$

[2] If Heisenberg's theorem is often seen as the signature of an unavoidable back-action due to the measurements, its exact interpretation and consequences on general joint measurements is a delicate topic [134–136]

so that $\hat{\rho}$ encodes all the statistical information on the system. From a probabilistic point of view, the previous statistical mixture is a random variables $|\varphi\rangle$ taking values in the set of pure states of $\mathcal{H}_S$, and the density operator is given by:

$$\hat{\rho} = \mathbb{E}_{\varphi}\left[|\varphi\rangle\langle\varphi|\right]. \tag{1.12}$$

For a pure state $|\phi\rangle$, taking $\hat{\rho} = |\phi\rangle\langle\phi|$ in Eq. (1.11) gives back Eq. (1.9). Therefore, density operators provide the most general mathematical representation of a quantum system. To define a *bona fide* density operator, a bounded operator $\hat{\rho}$ on $\mathcal{H}_S$ must present the following properties [137]:

- Self-adjointness: $\hat{\rho} = \hat{\rho}^{\dagger}$
- Positivity: $\hat{\rho} \geq 0$, that is $\forall\, |\phi\rangle \in \mathcal{H}_S,\ \langle\phi|\,\hat{\rho}\,|\phi\rangle \geq 0$
- Unit trace: $\mathrm{Tr}\left[\hat{\rho}\right] = 1$

We write $\mathcal{B}(\mathcal{H}_S)$ the vector space of bounded operators and $\mathcal{S}(\mathcal{H}_S)$ the set of density operators. The space $\mathcal{B}(\mathcal{H}_S)$ can be equipped with the *Hilbert-Schmidt inner product* that makes it a Hilbert space[3] [140]. This inner product and the associated norm[4] are given by:

$$\left\langle \hat{A}, \hat{B} \right\rangle_{\mathrm{HS}} = \mathrm{Tr}\left[\hat{A}^{\dagger}\hat{B}\right], \qquad \|\hat{A}\|_2 = \sqrt{\mathrm{Tr}\left[\hat{A}^{\dagger}\hat{A}\right]}. \tag{1.13}$$

Mixed states entropies

The space of density operators $\mathcal{S}(\mathcal{H}_S)$ is a convex subspace of $\mathcal{B}(\mathcal{H}_S)$. Pure and mixed states in $\mathcal{S}(\mathcal{H}_S)$ can be distinguished through their *purity* $\mathrm{Tr}\left[\hat{\rho}^2\right]$ since for pure states we have that $\mathrm{Tr}\left[\hat{\rho}^2\right] = 1$, while for mixed states $\mathrm{Tr}\left[\hat{\rho}^2\right] < 1$. From a geometric point of view, pure states are on the boundary of $\mathcal{S}(\mathcal{H}_S)$, while mixed states are in its interior. As a consequence, mixed states admit infinitely many convex decompositions of the form given in Eq. (1.10), whereas for pure states such decomposition is unique (up to a phase). The previous distinction can be seen from a probabilistic angle. Density operators can always be diagonalized into a countable sum of orthogonal pure states [137], taking the form:

$$\hat{\rho} = \sum_{i=0}^{\infty} \lambda_i\, |\phi_i\rangle\langle\phi_i|, \quad \langle\phi_i|\phi_j\rangle = \delta_{ij}, \tag{1.14}$$

where the eigenvalues $\{\lambda_i, i \in \mathbb{N}\}$ form a probability distribution. From this, one have that

$$-\log\left(\mathrm{Tr}\left[\hat{\rho}^2\right]\right) = -\log\left(\sum_{i=1}^{\infty} \lambda_i^2\right). \tag{1.15}$$

[3] In the general case, this is only true for the subspace of $\mathcal{B}(\mathcal{H})$ composed of the so-called Hilbert-Schmidt operators for which $\mathrm{Tr}\left[\hat{A}^{\dagger}\hat{A}\right]$ is well-defined and finite. In the following we will mostly consider finite dimensional spaces, for which every operator is both bounded and Hilbert-Schmidt.

[4] The Hilbert-Schmidt is the 2-norm of the broader familly of the Schatten p-norms, hence the notation [141].

The right-hand side of this equation is the second Rénye entropy of the previous probability distribution. The purity is thus measure of the degree of uncertainty in the statistical ensemble of perfectly distinguishable states associated to $\hat{\rho}$. In the same spirit, one defines the *von Neumann entropy* $S(\hat{\rho})$ as the Shannon entropy associated to the λ_i:

$$S(\hat{\rho}) = -\mathrm{Tr}\left[\hat{\rho}\log\left(\hat{\rho}\right)\right] = -\sum_{i=1}^{\infty}\lambda_i\log\left(\lambda_i\right) \tag{1.16}$$

Like the purity, the von Neumann entropy provides information on how much the state $\hat{\rho}$ is mixed.

Entanglement measures

Density operators allow to describe systems whose states are uncertain, a situation that is ubiquitous when dealing with components of composite systems.

Let us consider a bipartite system composed of two subsystems A and B, with associated Hilbert spaces $\mathcal{H}_A$ and $\mathcal{H}_B$. The density operator describing A is derived from the system state $\hat{\rho}\in\mathcal{S}\left(\mathcal{H}_A\otimes\mathcal{H}_B\right)$ by taking a partial trace over the degrees of freedom of B, that is $\hat{\rho}_A = \mathrm{Tr}_B\left[\hat{\rho}\right]$, and likewise for B [125]. The whole system is said to be in a *separable* state with respect to the partition (A,B) whenever $\hat{\rho}$ can be written as a convex sum of product states of the form $\hat{\rho}_A\otimes\hat{\rho}_B$. Otherwise, one says that the state is *entangled*. Entangled states are non-classical, in the sense that for these states the whole system contains more information than the partition's subsystems altogether. In particular, if the whole system is in a pure entangled state, the subsystems are necessary mixed. To see this, one can use the Schmidt decomposition theorem [142] which guarantee that any pure state of the global system $|\phi\rangle\in\mathcal{H}_A\otimes\mathcal{H}_B$ can be decomposed into:

$$|\phi\rangle = \sum_i \lambda_i \left|\varphi_i^A\right\rangle\otimes\left|\varphi_i^B\right\rangle\,,\quad \lambda_i\geq 0\,,\quad \sum_i\lambda_i^2 = 1\,, \tag{1.17}$$

where $\{\left|\varphi_i^A\right\rangle\}$ and $\{\left|\varphi_i^B\right\rangle\}$ are orthonormal basis of $\mathcal{H}_A$ and $\mathcal{H}_B$ respectively. As a consequence of the theorem, the states $\hat{\rho}_A$ and $\hat{\rho}_B$ share the same eigenvalues, and hence have the same von Neumann entropy. This entropy is called the *entanglement entropy* and it is a measure of the entanglement of the whole system. This measure of entanglement relies on the Schmidt decomposition, that only holds for pure states. For this reason, in the following we use a different entanglement witness, the *negativity* [143, 144], which is cheap to compute and adapted to mixed states. The negativity of a state $\hat{\rho}\in\mathcal{S}\left(\mathcal{H}_A\otimes\mathcal{H}_B\right)$ with respect to the partition (A,B) is given by:

$$\mathcal{N}\left(\hat{\rho}\right) = \frac{1}{2}\left(\|\hat{\rho}^{T_A}\|_1 - 1\right)\,, \tag{1.18}$$

where $\|A\|_1 = \mathrm{Tr}\left[\sqrt{A^\dagger A}\right]$[5], and $\hat{\rho}^{T_A}$ is the partial transpose of the density operator with respect to the degrees of freedom of the subsystem A.

[5] This is the trace norm, which is also the Schatten 1-norm.

II.2 Quantum channels

A quantum channel [141] is a map $\Phi : \mathcal{B}(\mathcal{H}_S) \mapsto \mathcal{B}(\mathcal{H}_S)$ [6] that is:

- Linear: $\Phi(\lambda\hat{\rho} + \hat{\sigma}) = \lambda\Phi(\hat{\rho}) + \Phi(\hat{\sigma})$,
- Trace preserving: $\mathrm{Tr}[\Phi(\hat{\rho})] = \mathrm{Tr}[\hat{\rho}]$,
- Completely positive: For any extra system E with associated Hilbert space $\mathcal{H}_E$ and $\mathcal{I}_E$ the identity on $\mathcal{B}(\mathcal{H}_E)$, the map $(\Phi \otimes \mathcal{I}_E) : \mathcal{B}(\mathcal{H}_S \otimes \mathcal{H}_E) \mapsto \mathcal{B}(\mathcal{H}_S \otimes \mathcal{H}_E)$ is positive.

Maps with these properties are said to be completely-positive trace-preserving (CPTP). These maps represent physically acceptable transformations of quantum systems, and the CPTP properties ensure that for any state $\hat{\rho} \in \mathcal{S}(\mathcal{H}_S)$, $\Phi(\hat{\rho})$ is a well-defined quantum states. In particular, the complete positivity requirement stems from the fact that any system can be seen as a component of a larger system, whose state must remains positive under a physical transformation.

Kraus and Choi representations

Quantum channels can be represented in different ways. Here we recall two theorems that will be used in the rest of this manuscript, namely the Kraus and the Choi representation theorems [141].

According to the Kraus theorem, every CPTP map $\Phi : \mathcal{B}(\mathcal{H}_S) \mapsto \mathcal{B}(\mathcal{H}_S)$ can be decomposed in the following Kraus form:

$$\Phi(\hat{\rho}) = \sum_{i\in I} \hat{K}_i \hat{\rho} \hat{K}_i^\dagger \,, \quad \sum_{i\in I} \hat{K}_i^\dagger \hat{K}_i = \hat{\mathbb{1}} \,, \tag{1.19}$$

where $\{\hat{K}_i, i \in I \subseteq \mathbb{N}\}$ is a countable [7] set of operators in $\mathcal{B}(\mathcal{H}_S)$, called the Kraus operators, and $\hat{\mathbb{1}}$ is the identity on $\mathcal{H}_S$. This decomposition is not unique, but one can show that two set of Kraus operators $\{\hat{K}_i\}, \{\hat{M}_j\}$ correspond to the same quantum map Φ if and only if there exists a unitary matrix (u_{ij}) such that $\hat{K}_i = \sum_j u_{ij}\hat{M}_j$ [137].

In the case where $\mathcal{H}_S$ is of finite dimension $D \in \mathbb{N}$, the Choi theorem provides an alternative representation. Given an orthonormal basis $\{|i\rangle, i \in [\![1, D]\!]\}$ of $\mathcal{H}_S$, this theorem states that any map $\Phi : \mathcal{B}(\mathcal{H}_S) \mapsto \mathcal{B}(\mathcal{H}_S)$ is uniquely represented by the following operator

$$\Lambda(\Phi) = \sum_{i,j=1}^{D} |i\rangle\langle j| \otimes \Phi(|i\rangle\langle j|) \in \mathcal{B}(\mathcal{H}_S \otimes \mathcal{H}_S) \,, \tag{1.20}$$

which is positive if and only if Φ is completely positive. The set $\{|i\rangle\langle j| \otimes |k\rangle\langle l|\}$ forms an orthonormal basis of $\mathcal{B}(\mathcal{H}_S \otimes \mathcal{H}_S)$ for the Hilbert-Schmidt inner product, so we can decompose $\Lambda(\Phi)$ as follow:

$$\Lambda(\Phi) = \sum_{i,j,k,l} \Lambda(\Phi)_{ijkl} |i\rangle\langle j| \otimes |k\rangle\langle l| \,, \tag{1.21}$$

[6] In full generality, a quantum channel can map operators acting on the Hilbert spaces of two different systems, but we will not consider that case in the following.

[7] If $\dim(\mathcal{H}_S) = D < \infty$, the channel can be represented with at most D^2 Kraus operators.

where the coefficients are given by:

$$\Lambda(\Phi)_{ijkl} = \left\langle |i\rangle\langle j| \otimes |k\rangle\langle l| \, , \Lambda(\Phi) \right\rangle_{\mathrm{HS}} = \langle k| \, \Phi(|i\rangle\langle j|) \, |l\rangle \ . \tag{1.22}$$

Mixed unitary channels

Among the simplest examples of quantum channels are the unitary channels of the form

$$\hat{\rho} \mapsto \hat{U}\hat{\rho}\hat{U}^\dagger \, , \tag{1.23}$$

for some unitary operator $\hat{U} \in \mathcal{B}(\mathcal{H}_S)$. Such channel typically describe the evolution of a closed system under a Hamiltonian $\hat{H}(t)$ over a given time interval $[0, T]$, in which case [137]

$$\hat{U} = \hat{U}(T, 0) = \mathcal{T}\left[\exp\left(-\mathrm{i}\int_0^T \hat{H}(t)\mathrm{d}t\right)\right] \, , \tag{1.24}$$

with $\mathcal{T}$ the time-ordering superoperator and $\hbar = 1$. In the following, we will be interested in a related class of quantum channels, namely the *mixed-unitary* channels. The latter represent statistical ensembles of unitary transformations. As such, they are to unitary channels what density operators are to pure states. Considering a statistical ensemble of unitary transformation $\{\hat{U}_i \, , \, i \in I\}$ with associated probabilities $\{p_i \, , \, i \in I\}$, the corresponding mixed-unitary channel is given by:

$$\Phi(\hat{\rho}) = \sum_{i \in I} p_i \hat{U}_i \hat{\rho} \hat{U}_i^\dagger \, . \tag{1.25}$$

As for density operators, it is equivalent to consider the mixed-unitary channel as the expectation value of a random variable $\hat{U}$ taking value in the set of unitary transformations, so that $\Phi(\hat{\rho}) = \mathbb{E}_{\hat{U}}\left[\hat{U}\hat{\rho}\hat{U}^\dagger\right]$. If $\mathcal{H}_S$ is of finite dimension $D \in \mathbb{N}$, one can show that any mixed-unitary channel [8] can be decomposed as a finite convex sum of at most $D^4 - 2D^2 + 2$ unitary channels [141].

Pauli channels

Mixed-unitary channels are widely used in the theory of quantum information, as they provides a simple and effective way to model the physical noise affecting quantum devices. Before introducing a generic model of such channels, we set some conventions for two level systems, or *qubits*, that will be used through the rest of this text. Let us consider a single qubit and choose a computational basis $\{|0\rangle \, , |1\rangle\}$ of its associated Hilbert space. The Pauli matrices in this basis are written:

$$\hat{X} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \, , \quad \hat{Y} = \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix} \, , \quad \hat{Z} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \, . \tag{1.26}$$

Depending on the context, we will sometimes use the old-fashioned notations $\hat{\sigma}_x, \hat{\sigma}_y$ and $\hat{\sigma}_z$. Fig. 1.1 recall the Bloch sphere representation of a qubit pure state. A generic qubit state $\hat{\rho}$ can always be decomposed in a similar way, as

$$\hat{\rho} = \frac{1}{2}\left(\hat{\mathbb{1}} + \boldsymbol{r}^T\hat{\boldsymbol{\sigma}}\right) \, , \tag{1.27}$$

[8] Including mixed unitary channels associated to unitary ensemble that are not countable.

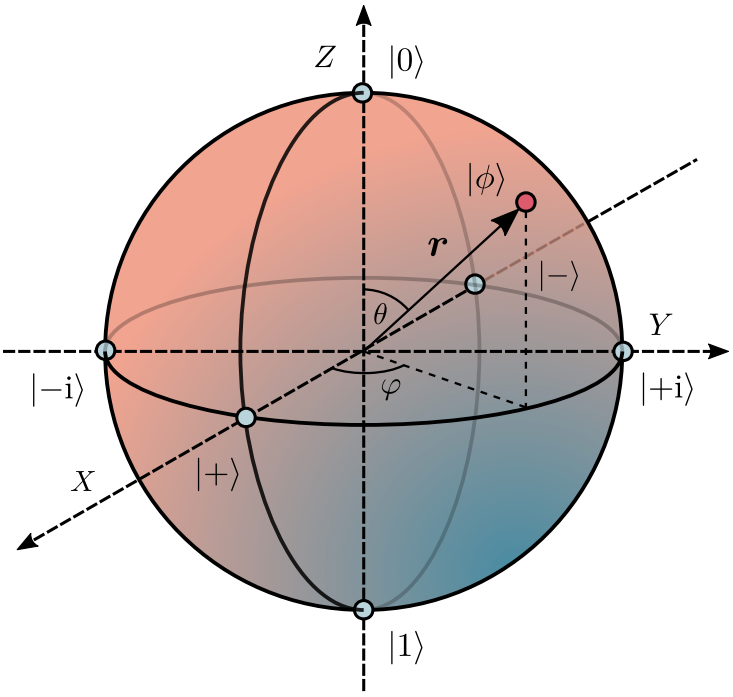


Figure 1.1: Bloch sphere representation of a qubit state $|\phi\rangle = \cos\left(\frac{\theta}{2}\right)|0\rangle + e^{i\varphi}\sin\left(\frac{\theta}{2}\right)|1\rangle$.

with $\hat{\boldsymbol{\sigma}}^T = (\hat{\sigma}_x, \hat{\sigma}_y, \hat{\sigma}_z)$ and $\boldsymbol{r} = \mathrm{Tr}\left[\hat{\rho}\hat{\boldsymbol{\sigma}}\right]$. That representation extends to states of multiple qubits. For a set of N qubits, we call *Pauli strings* (or *Pauli words*) the elements of $\{\mathbb{1}, \hat{X}, \hat{Y}, \hat{Z}\}^{\otimes N}$. The generalization of the previous equation then reads

$$\hat{\rho} = \frac{1}{2^N} \sum_{\hat{P}\in\{\mathbb{1},\hat{X},\hat{Y},\hat{Z}\}^{\otimes N}} \mathrm{Tr}\left[\hat{\rho}\hat{P}\right]\hat{P}\,. \tag{1.28}$$

Noise models for qubits often belong to the large class of the so-called *Pauli channels*. These are mixed unitary channels associated to ensembles of Pauli words, which can be written:

$$\hat{\rho} \mapsto \sum_{i=1}^{M} p_i \hat{P}_i \hat{\rho} \hat{P}_i\,, \quad \hat{P}_i \in \{\mathbb{1}, \hat{X}, \hat{Y}, \hat{Z}\}^{\otimes N}. \tag{1.29}$$

Let us remind the following three well-known examples of single qubits error channels in this class [142], for a probability of error $p \in [0, 1]$:

- *Bit flip channel*:
$$\hat{\rho} \mapsto (1-p)\hat{\rho} + p\hat{X}\hat{\rho}\hat{X}\,, \tag{1.30}$$
- *Phase flip channel*:
$$\hat{\rho} \mapsto (1-p)\hat{\rho} + p\hat{Z}\hat{\rho}\hat{Z}\,, \tag{1.31}$$
- *Depolarizing channel* [9]:
$$\hat{\rho} \mapsto (1-p)\hat{\rho} + \frac{p}{2}\hat{\mathbb{1}}\,. \tag{1.32}$$

[9] This is indeed a Pauli channel as we have $\left(\hat{\rho} + \hat{X}\hat{\rho}\hat{X} + \hat{Y}\hat{\rho}\hat{Y} + \hat{Z}\hat{\rho}\hat{Z}\right)/4 = \hat{\mathbb{1}}/2$.

Among these channels, the bit flip and the phase flip channels stands out by the fact that they act trivially on the eigenstates of respectively $\hat{X}$ and $\hat{Z}$. The existence of these preferential bases of states is characteristic of the physical phenomenon of *decoherence* on which we will focus in the next section.

III Qubit subject to decoherence

As explained in the introduction of this chapter, to represent the evolution of an open quantum system we often rely on a set of simplifying assumptions that enable to evacuate the environment degree's of freedom from the system's effective description. In cases where these conditions are fulfilled, the system's evolution is effectively described by a *master equation*. As for a classical system, this master equation can be *unravelled* into a set of stochastic equations of motion [10]. Such decomposition constitutes a *stochastic unravelling* of the master equation, and it provides an efficient means of simulating open quantum system [145–147].

In this section we provide an overview of these concepts through the study of a simple two-level system exhibiting decoherence. This choice is motivated by the fact that for many systems decoherence acts on time scales that are much shorter than the time scales associated to other types of noise, such as damping [137, 142]. In addition, decoherence will be helpful to interpret the results presented in the third chapter of this thesis.

III.1 Dynamics of a qubit under stochastic driving

We present a first simple model of decoherence for of a qubit, which is inspired by lecture notes of J. Preskill [148]. In this model, the qubit is driven by an external classical stochastic driving $(\xi(t))_{t\in\mathbb{R}_+}$. The system Hamiltonian takes the form

$$\hat{H}(t) = -\frac{1}{2}\omega_0\hat{\sigma}_z - \frac{1}{2}\xi(t)\hat{\sigma}_z\,. \tag{1.33}$$

This simple Hamiltonian is manifestly diagonal in the basis $\{|0\rangle, |1\rangle\}$ of the eigenstates of $\hat{\sigma}_z$. This allows to provide an intuitive description of the decoherence mechanism. Consider that the system is initially on the equator of the Bloch sphere, for instance we can take $|\phi(t=0)\rangle = |+\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$. After an evolution over a time duration T, the state will acquire a relative phase $\varphi(t) = \omega_0 t + \int_0^t \xi(s)\mathrm{d}s$ and the resulting pure state will have a random position on the equator of the Bloch sphere of Fig. 1.1. Averaging over multiple realizations, the radius of the mean state will thus shrink and the corresponding state will become mixed. Clearly, that effect only concerns states in a *coherent superposition* of $|0\rangle$ and $|1\rangle$, which are turned into *incoherent* mixtures of these states, hence the terminology.

[10] In a classical context, a simple example is given by a gaz of particles, which density evolves according to a diffusion equation. The density can then be decomposed using an average over a set of particles evolving according to a Langevin equation.

Evolution for a Gaussian noise

Let us now derive the evolution channel associated with this model. In this section we work in the interaction picture, which allows us to get rid of the bare Hamiltonian $\hat{H}_0 = -\frac{1}{2}\omega_0\hat{\sigma}_z$. The transformations associated with the corresponding change of reference frame are recalled in App. A. We denote $\hat{\rho}_\xi(t)$ the system's state at a time t associated with a given realization ξ of the driving. In the interaction picture, the system's evolution is given by the following *Liouville stochastic equation*:

$$\frac{\mathrm{d}\hat{\rho}_\xi}{\mathrm{d}t} = \mathrm{i}\frac{\xi(t)}{2}\left[\hat{\sigma}_z, \hat{\rho}_\xi(t)\right] . \tag{1.34}$$

Here we assume that at $t = 0$ the system is in an initial state $\hat{\rho}_0$ that is independent of ξ. Integrating Eq. (1.34), we obtain the evolved state at a time t through the associated unitary channel:

$$\hat{\rho}_\xi(t) = \left(\exp\left(\frac{\mathrm{i}}{2}\hat{\sigma}_z\int_0^t \xi(s)\mathrm{d}s\right)\right)\hat{\rho}_0\left(\exp\left(-\frac{\mathrm{i}}{2}\hat{\sigma}_z\int_0^t \xi(s)\mathrm{d}s\right)\right) . \tag{1.35}$$

where $\hat{\rho}_{\xi,0} = \hat{\rho}_\xi(t=0)$. In the following we denote $\Lambda^\xi_{(t,t')}$ the quantum channel mapping an initial system state at a time t to the corresponding evolved state at a later time t' for a given realization of ξ. Importantly, this channel describes the system evolution assuming that the initial state is independent of the driving. Using this notations and writing $\hat{\Pi}_i = |i\rangle\langle i|$, $i \in \{0,1\}$, we have $\hat{\sigma}_z = \Pi_0 - \Pi_1$ and the previous evolved state can be written as

$$\hat{\rho}_\xi(t) = \Lambda^\xi_{(t,0)}\left(\hat{\rho}_0\right) = \hat{\Pi}_0\hat{\rho}_0\hat{\Pi}_0 + \hat{\Pi}_1\hat{\rho}_0\hat{\Pi}_1 + \exp\left(\mathrm{i}\int_0^t \xi(s)\mathrm{d}s\right)\hat{\Pi}_0\hat{\rho}_0\hat{\Pi}_1 + h.c.^{11} . \tag{1.36}$$

Since we have assumed that $\hat{\rho}_0$ is independent of the driving, averaging the previous equations with respect to the driving yields a quantum channel $\Lambda_{(t,0)}$ that maps $\hat{\rho}_0$ to the average state

$$\hat{\rho}(t) = \mathbb{E}\left[\hat{\rho}_\xi(t)\right] = \mathbb{E}\left[\Lambda^\xi_{(t,0)}\left(\hat{\rho}_0\right)\right] = \Lambda_{(t,0)}\left(\hat{\rho}_0\right) . \tag{1.37}$$

To obtain an explicit expression of this state, a more precise characterization of the driving is required. For simplicity we assume that $\xi(t)$ is a centered Gaussian process. As such, it is uniquely determined by its covariance function $K : \mathbb{R}_+^2 \mapsto \mathbb{R}$, and for every finite sequence of times $S = (t_1, \ldots, t_n) \in \mathbb{R}_+^n$ the vector $\boldsymbol{\xi}_S = (\xi(t_1), \ldots, \xi(t_n))$ is a centered Gaussian vector with covariance matrix $(\mathbf{K}_S)_{ij} = K(t_i, t_j)$ [12]. In particular, we have

$$\mathbb{E}\left[\xi(t)\right] = 0 , \quad \mathbb{E}\left[\xi(t)\xi(s)\right] = K(t,s) , \quad \forall t, s \in \mathbb{R}_+ , \tag{1.38}$$

and in cases where $\mathbf{K}_S$ is non-singular, the vector $\boldsymbol{\xi}_S$ admits the usual Gaussian density

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n|\mathbf{K}_S|}}\exp\left(-\frac{1}{2}\boldsymbol{x}^T\mathbf{K}_S^{-1}\boldsymbol{x}\right) . \tag{1.39}$$

[11] In the rest of this manuscript, *h.c.* stands for the hermitian conjugate.

[12] The only constraint on the choice of K is that $\mathbf{K}_S$ must be a positive semi-definite matrix for every time sequence S.

For this simple noise, one can obtain an explicit expression for the expectation

$$\mathbb{E}\left[\exp\left(\mathrm{i}\int_0^t \xi(s)\mathrm{d}s\right)\right] = \sum_{k=0}^{+\infty} \frac{(\mathrm{i})^k}{k!}\mathbb{E}\left[\left(\int_0^t \xi(s)\mathrm{d}s\right)^k\right]. \tag{1.40}$$

Each term on the right-hand-side of the previous equation can be computed using Isserlis' theorem [149], which states that for a centered Gaussian vector $(\xi(s_1),\dots,\xi(s_k))$ we have

$$\mathbb{E}\left[\prod_{i=1}^k \xi(s_i)\right] = \begin{cases} \displaystyle\sum_{p\in\mathcal{P}_k^2}\prod_{(i,j)\in p}\mathbb{E}\left[\xi(s_i)\xi(s_j)\right] & \text{for } k \text{ even},\\ 0 & \text{for } k \text{ odd}, \end{cases} \tag{1.41}$$

where $\mathcal{P}_k^2$ denote the set of partitions of $[\![1,k]\!]$ into disjoint pairs. Hence we have for every k even:

$$\begin{aligned}\mathbb{E}\left[\left(\int_0^t \xi(s)\mathrm{d}s\right)^k\right] &= \int_{[0,t]^k}\mathbb{E}\left[\prod_{i=1}^k \xi(s_i)\right]\mathrm{d}\boldsymbol{s}\\ &= \sum_{p\in\mathcal{P}_k^2}\int_{[0,t]^k}\left(\prod_{(i,j)\in p}K(s_i,s_j)\right)\mathrm{d}\boldsymbol{s}\\ &= \sum_{p\in\mathcal{P}_k^2}\prod_{(i,j)\in p}\int_0^t\int_0^t K(s_i,s_j)\mathrm{d}s_i\mathrm{d}s_j\\ &= \sum_{p\in\mathcal{P}_k^2}\left(\int_0^t\int_0^t K(s_1,s_2)\mathrm{d}s_1\mathrm{d}s_2\right)^{k/2}\\ &= \frac{k!}{2^{k/2}(k/2)!}\left(\int_0^t\int_0^t K(s_1,s_2)\mathrm{d}s_1\mathrm{d}s_2\right)^{k/2}.\end{aligned} \tag{1.42}$$

The last expression follows from the fact that the set $\mathcal{P}_k^2$ contains $\frac{k!}{2^{k/2}(k/2)!}$ pairs. Injecting this expression into Eq. (1.40), we have

$$\mathbb{E}\left[\exp\left(\mathrm{i}\int_0^t \xi(s)\mathrm{d}s\right)\right] = \exp\left(-\frac{1}{2}\int_0^t\int_0^t K(s_1,s_2)\mathrm{d}s_1\mathrm{d}s_2\right). \tag{1.43}$$

We are now in position to give an explicit expression for the average system state. Using Eqs. (1.36), (1.37) and (1.43), we can write

$$\hat{\rho}(t) = \Pi_0\hat{\rho}_0\Pi_0 + \Pi_1\hat{\rho}_0\Pi_1 + e^{-\frac{1}{2}\int_0^t\int_0^t K(s_1,s_2)\mathrm{d}s_1\mathrm{d}s_2}\left(\Pi_0\hat{\rho}_0\Pi_1 + \Pi_1\hat{\rho}_0\Pi_0\right). \tag{1.44}$$

As expected, the average evolution results in a decay the non-diagnonal terms, i.e. the coherences, of the initial density operators by a factor

$$f(t) = e^{-\frac{1}{2}\int_0^t\int_0^t K(s_1,s_2)\mathrm{d}s_1\mathrm{d}s_2} = e^{-\Gamma(t)}, \tag{1.45}$$

where we have introduced the so-called the decoherence function [137]

$$\Gamma(t) = \frac{1}{2}\int_0^t\int_0^t K(s_1,s_2)\mathrm{d}s_1\mathrm{d}s_2. \tag{1.46}$$

Using the equalities

$$\begin{aligned}\frac{1}{2}\left(\hat{\sigma}_z\hat{\rho}\hat{\sigma}_z-\hat{\rho}\right)&=\Pi_0\hat{\rho}\Pi_1+\Pi_1\hat{\rho}\Pi_0\,,\\ \frac{1}{2}\left(\hat{\sigma}_z\hat{\rho}\hat{\sigma}_z+\hat{\rho}\right)&=\Pi_0\hat{\rho}\Pi_0+\Pi_1\hat{\rho}\Pi_1\,,\end{aligned} \tag{1.47}$$

we can rewrite Eq. (1.44) as follow:

$$\hat{\rho}(t)=\Lambda_{(t,0)}\left(\hat{\rho}_0\right)=\frac{1}{2}\left(1+f(t)\right)\hat{\rho}_0+\frac{1}{2}\left(1-f(t)\right)\hat{\sigma}_z\hat{\rho}_0\hat{\sigma}_z\,. \tag{1.48}$$

The channel $\Lambda_{(t,0)}$ that describes the average system evolution over the time interval $[0,t]$ is thus a phase-flip channel of the form of Eq. (1.31), with a flipping probability

$$p(t)=\frac{1}{2}\left(1-e^{-\Gamma(t)}\right)\,. \tag{1.49}$$

Markov assumption and Lindblad master equation

Let us now derive the master equation associated with the previous model. We simplify the problem further and assume that the noise is stationary. We can write

$$K(t,s)=K(t-s)=\int_{-\infty}^{+\infty}e^{-\mathrm{i}\omega(t-s)}\tilde{K}(\omega)\frac{\mathrm{d}\omega}{2\pi}\,, \tag{1.50}$$

where we have introduced the Fourier transform of the noise covariance, also known as the *spectral density* of the noise. As we consider a real classical noise, the covariance is an even function, as is the spectral density. With this assumption, we have

$$\begin{aligned}\int_0^t\int_0^t K(s_1,s_2)\mathrm{d}s_1\mathrm{d}s_2&=\frac{1}{2\pi}\int_0^t\int_0^t\int_{-\infty}^{+\infty}e^{-\mathrm{i}\omega(s_1-s_2)}\tilde{K}(\omega)\mathrm{d}\omega\mathrm{d}s_1\mathrm{d}s_2\\ &=\frac{t^2}{2\pi}\int_{-\infty}^{+\infty}\tilde{K}(\omega)\left(\mathrm{sinc}\left(\frac{\omega t}{2}\right)\right)^2\mathrm{d}\omega\,,\end{aligned} \tag{1.51}$$

where the sinc function is defined as $\mathrm{sinc}(x)=\frac{\sin(x)}{x}$. The squared sinc function appearing in the integral quickly vanishes outside of the interval $\left[0,\frac{2\pi}{t}\right]$. Thus for t sufficiently large we have

$$\frac{t^2}{2\pi}\int_{-\infty}^{+\infty}\tilde{K}(\omega)\left(\mathrm{sinc}\left(\frac{\omega t}{2}\right)\right)^2\mathrm{d}\omega\simeq\tilde{K}(\omega=0)t\int_{-\infty}^{+\infty}\mathrm{sinc}\,(x)^2\,\frac{\mathrm{d}x}{\pi}=2\gamma t\,, \tag{1.52}$$

where we denote $\gamma=\tilde{K}(0)/2$. In this regime, the decoherence function is thus simply given by

$$\Gamma(t)=\gamma t\,. \tag{1.53}$$

Roughly, we expect the characteristic noise correlation function $K(\tau)$ to vanish for times $|\tau|$ greater than some *correlation time* τ_c, or equivalently, that the spectral density is mostly supported by $[-\omega_c,\omega_c]$ with $\omega_c=1/\tau_c$. In this case, the previous approximation holds if we consider times that are large in front of the noise correlation time $t\gg\tau_c$, in which case we have

$$\hat{\rho}(t)=\frac{1}{2}\left(1+e^{-\gamma t}\right)\hat{\rho}_0+\frac{1}{2}\left(1-e^{-\gamma t}\right)\hat{\sigma}_z\hat{\rho}_0\hat{\sigma}_z\,. \tag{1.54}$$

In the foregoing discussion, we can safely replace the interval $[0,t]$ by the interval $[t,t+\delta t]$ and replace $\hat{\rho}_0$ by $\hat{\rho}(t)$. Provided we have $\tau_c \ll \delta t$, the previous equation now reads

$$\hat{\rho}(t+\delta t) = \frac{1}{2}\left(1+e^{-\gamma\delta t}\right)\hat{\rho}(t) + \frac{1}{2}\left(1-e^{-\gamma\delta t}\right)\hat{\sigma}_z\hat{\rho}(t)\hat{\sigma}_z\,. \tag{1.55}$$

Assuming that $\gamma\delta t \ll 1$, we have

$$\begin{aligned}\hat{\rho}(t+\delta t) &\simeq \left(1-\frac{\gamma}{2}\delta t\right)\hat{\rho}_0 + \left(\frac{\gamma}{2}\delta t\right)\hat{\sigma}_z\hat{\rho}_0\hat{\sigma}_z \\ &= \hat{\rho}(t) + \frac{\gamma}{2}\delta t\left(\hat{\sigma}_z\hat{\rho}(t)\hat{\sigma}_z - \hat{\rho}(t)\right)\end{aligned} \tag{1.56}$$

up to the second order in $\gamma\delta t$. To obtain this expansion, we need to use the *coarse grained* time scale

$$\tau_c \ll \delta t \ll 1/\gamma\,. \tag{1.57}$$

In particular, this time scale only exists if the decay rate of the system $1/\gamma$ is sufficiently large in front of the noise coherence time τ_c. This is the condition underlying the *Markov assumption* [137], which states that for the considered time scale, the noise is roughly memory-less and the system does not evolve much under its influence. As a consequence, it is possible to approximate the noise by a delta-correlated Gaussian white noise, i.e. $K(t-s) \simeq \gamma\delta(t-s)$. Under these conditions, the evolution of the density operator on the coarse grained time scale is given in the interaction picture by the following master equation:

$$\frac{\mathrm{d}\hat{\rho}(t)}{\mathrm{d}t} \simeq \frac{1}{\delta t}\left(\hat{\rho}(t+\delta t) - \hat{\rho}(t)\right) = \frac{\gamma}{2}\mathcal{D}\left[\hat{\sigma}_z\right]\left(\hat{\rho}(t)\right)\,, \tag{1.58}$$

where we have introduced the *dissipator* associated with a given *jump operator* $\hat{A}$

$$\mathcal{D}\left[\hat{A}\right] = \hat{A}\hat{\rho}\hat{A}^\dagger - \frac{1}{2}\left(\hat{A}^\dagger\hat{A}\hat{\rho} + \hat{\rho}\hat{A}^\dagger\hat{A}\right)\,. \tag{1.59}$$

Eq. (1.58) is an example of a generic *Lindblad master equation* of the form

$$\frac{\mathrm{d}\hat{\rho}}{\mathrm{d}t} = \mathcal{L}(t)\left(\hat{\rho}\right) = -\mathrm{i}\left[\hat{H}(t),\hat{\rho}(t)\right] + \sum_k \gamma_k(t)\mathcal{D}\left[\hat{A}_k(t)\right]\left(\hat{\rho}\right)\,, \tag{1.60}$$

where $\mathcal{L}$ denotes the system's *Lindbladian*, $\hat{H}$ is the system Hamiltonian and the factors $\{\gamma_k\}$ the system *dissipation rates*. It is possible to show [13] that for a system evolving according to a time-dependent Lindblad equation of the previous form, if the dissipation rates are positives, the corresponding evolution channels $\Lambda_{(t,t')}$ satisfy the relations [150]

$$\Lambda_{(t_2,t_1)}\Lambda_{(t_1,t_0)} = \Lambda_{(t_2,t_0)}\,, \quad \forall t_0 \leq t_1 \leq t_2\,. \tag{1.61}$$

This condition is sometimes called the *divisibility condition*. In cases where the Lindblad master equation is time-independent, the previous relations are turned into the stronger conditions

$$\begin{cases}\Lambda_{(t_2,t_1)} = \Lambda_{(t_2-t_1)} & , \forall t_1, \leq t_2 \\ \Lambda_{(\tau_1)}\Lambda_{(\tau_2)} = \Lambda_{(\tau_1+\tau_2)} & , \forall \tau_1,\tau_2 \geq 0,\end{cases} \tag{1.62}$$

[13] At least in the finite dimensional case.

along with some adequate notion of continuity in time. A family $\{\Lambda_{(\tau)}, \tau \in \mathbb{R}_+\}$ satisfying these conditions constitutes a so-called *dynamical semi-group*. The time-independent version of the result above was first proven by Gorini, Kossakowski and Sudarshan [151] and Lindblad [152], and it has played a fundamental role for the development of the theory of open quantum systems [153]. For the current thesis we will only deal with systems evolving according to a Lindblad equation with time-independent jump operators and dissipation rates.

The Lindblad master equation is intimately related to the notion of Markovianity. However, the very notion of Markovian evolution has multiple non-equivalent definitions in the literature [154–160]. One approach to this question lies in the previous notion of divisibility, and some authors defines Markovian systems as systems which evolution is given by a Lindblad equation with *positive dissipation rates* [14] [155, 156, 158]. With some adequate choice of driving, the previous model can exhibit a non-Markovian behavior, in the sense that we have just outlined. We provide such an example in App. B.

III.2 Coupling to a bosonic reservoir

We now turn to a fully quantum version of the previous model, where the classical stochastic driving is replaced by a bath of bosonic fields at thermal equilibrium. This model was first used in [57, 161] and here we follow the derivations of refs. [137] and [148] Consider a system composed of a qubit and a reservoir of harmonic oscillators, which Hamiltonian in the Schrödinger picture reads

$$\hat{H}_S = \frac{\omega_0}{2}\hat{\sigma}_z + \sum_{k\in\mathbb{N}} \omega_k \hat{b}_k^\dagger \hat{b}_k + \sum_{k\in\mathbb{N}} \hat{\sigma}_z \left(g_k \hat{b}_k^\dagger + g_k^* \hat{b}_k\right) , \tag{1.63}$$

where $\hat{b}_k^\dagger, \hat{b}_k$ are respectively the creation and annihilation operators associated to the different bosonic modes, which satisfy the commutation relations:

$$\left[\hat{b}_k, \hat{b}_{k'}^\dagger\right] = \delta_{kk'} . \tag{1.64}$$

As before, we work in the interaction picture to eliminate the bare Hamiltonian

$$\hat{H}_0 = \frac{\omega_0}{2}\hat{\sigma}_z + \sum_{k\in\mathbb{N}} \omega_k \hat{b}_k^\dagger \hat{b}_k . \tag{1.65}$$

We take $t = 0$ as a the time reference. Using the commutation relations we have

$$e^{\mathrm{i}\hat{H}_0 t} \hat{b}_k e^{-\mathrm{i}\hat{H}_0 t} = \hat{b}_k e^{-\mathrm{i}\omega_k t} . \tag{1.66}$$

The Hamiltonian in the interaction picture is thus given by

$$\hat{H}(t) = \sum_{k\in\mathbb{N}} \hat{\sigma}_z \left(g_k e^{\mathrm{i}\omega_k t} \hat{b}_k^\dagger + g_k^* e^{-\mathrm{i}\omega_k t} \hat{b}_k\right) . \tag{1.67}$$

[14] To a given Liouvillian can correspond multiple sets of jump operators and dissipation rates. To lift this uncertainty, one requires that the jump operators form a basis of the set of the traceless observables. This uniquely fixes a corresponding set of rates, and asking for this rates to be positive gives a well-defined condition of Markovianity. The Lindblad equation, when expressed with these specific dissipation rates and jump operators, is said to be in the *canonical form* [158].

Denoting $\hat{\rho}(t)$ the density operator of the whole system in the interaction picture and $\hat{\rho}_0 = \hat{\rho}(t=0)$, we have

$$\hat{\rho}(t) = \hat{U}(t)\hat{\rho}_0\hat{U}^\dagger(t)\,, \tag{1.68}$$

where $\hat{U}(t)$ is the evolution operator whose form is given by Eq. (1.24). This operator satisfy the Schrödinger equation

$$\frac{\mathrm{d}\hat{U}}{\mathrm{d}t} = -\mathrm{i}\hat{H}(t)\hat{U}(t)\,. \tag{1.69}$$

Unlike the case of a classical noise, the interaction Hamiltonian does not commute with itself at different times. If this was the case, the evolution operator would have been simply given by $\exp\left\{-\mathrm{i}\int_0^t \hat{H}(s)\mathrm{d}s\right\}$. Let us look for a solution of Eq. (1.69) of the form

$$\hat{U}(t) = \hat{C}(t)e^{-\mathrm{i}\int_0^t \hat{H}(s)\mathrm{d}s}\,. \tag{1.70}$$

The operator $\hat{C}(t)$ then satisfies

$$\frac{\mathrm{d}\hat{C}}{\mathrm{d}t} = \frac{\mathrm{d}\hat{U}}{\mathrm{d}t}e^{\mathrm{i}\int_0^t \hat{H}(s)\mathrm{d}s} + \hat{U}(t)\frac{\mathrm{d}}{\mathrm{d}t}\left(e^{\mathrm{i}\int_0^t \hat{H}(s)\mathrm{d}s}\right)\,. \tag{1.71}$$

The time derivative of the exponential operator of the right-hand side of this equation is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(e^{\mathrm{i}\int_0^t \hat{H}(s)\mathrm{d}s}\right) = \sum_{n=0}^{+\infty}\frac{\mathrm{i}^n}{n!}\frac{\mathrm{d}}{\mathrm{d}t}\left(\left(\int_0^t \hat{H}(s)\mathrm{d}s\right)^n\right)\,. \tag{1.72}$$

We have

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\left(\int_0^t \hat{H}(s)\mathrm{d}s\right)^n\right) = \sum_{k=0}^{n-1}\left(\int_0^t \hat{H}(s)\mathrm{d}s\right)^k \hat{H}(t)\left(\int_0^t \hat{H}(s)\mathrm{d}s\right)^{n-1-k}\,. \tag{1.73}$$

The commutator of the interaction Hamiltonian at different times can be written

$$\begin{aligned}\left[\hat{H}(s),\hat{H}(t)\right] &= \sum_{k,k'\in\mathbb{N}} g_k g_{k'}^* e^{\mathrm{i}(\omega_k s-\omega_{k'}t)}\left[\hat{b}_k^\dagger,\hat{b}_{k'}\right] + h.c.\\ &= -2\mathrm{i}\sum_{k\in\mathbb{N}}|g_k|^2\sin\left(\omega_k(s-t)\right)\,,\end{aligned} \tag{1.74}$$

which is a simple complex valued function. We denote

$$G(t) = \int_0^t\left(\sum_{k\in\mathbb{N}}|g_k|^2\sin\left(\omega_k(s-t)\right)\right)\mathrm{d}s\,, \tag{1.75}$$

and we have

$$\left[\int_0^t \hat{H}(s)\mathrm{d}s,\hat{H}(t)\right] = -2\mathrm{i}G(t)\,. \tag{1.76}$$

From this we obtain

$$\left[\left(\int_0^t \hat{H}(s)\mathrm{d}s\right)^k,\hat{H}(t)\right] = -2k\mathrm{i}G(t)\left(\int_0^t \hat{H}(s)\mathrm{d}s\right)^{k-1}\,, \tag{1.77}$$

such that Eq. (1.73) can be written

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\left(\int_0^t \hat{H}(s)\mathrm{d}s\right)^n\right) = n\hat{H}(t)\left(\int_0^t \hat{H}(s)\mathrm{d}s\right)^{n-1} - \mathrm{i}n(n-1)G(t)\left(\int_0^t \hat{H}(s)\mathrm{d}s\right)^{n-2}. \quad (1.78)$$

Injecting this expression in Eq. (1.72), we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(e^{\mathrm{i}\int_0^t \hat{H}(s)\mathrm{d}s}\right) = \left(\mathrm{i}\hat{H}(t) - \mathrm{i}G(t)\right)e^{\mathrm{i}\int_0^t \hat{H}(s)\mathrm{d}s}. \quad (1.79)$$

Using Eqs. (1.71) and (1.69), we can write

$$\frac{\mathrm{d}\hat{C}}{\mathrm{d}t} = -\mathrm{i}G(t)\hat{C}(t). \quad (1.80)$$

As $G(t)$ is a real valued function this equation is easily integrated, and the time evolution operator is given by

$$\hat{U}(t) = e^{-\mathrm{i}\int_0^t G(s)\mathrm{d}s}e^{-\mathrm{i}\int_0^t \hat{H}(s)\mathrm{d}s}. \quad (1.81)$$

In the following, we will denote

$$\begin{aligned}\hat{r}_k(t) &= 2\left(g_k e^{\mathrm{i}\omega_k t}\hat{b}_k^\dagger + g_k^* e^{-\mathrm{i}\omega_k t}\hat{b}_k\right),\\ \hat{r}(t) &= \sum_{k\in\mathbb{N}}\hat{r}_k(t).\end{aligned} \quad (1.82)$$

The state of the whole system at time t can be written

$$\hat{\rho}(t) = \left(\exp\left(\frac{\mathrm{i}}{2}\hat{\sigma}_z \otimes \int_0^t \hat{r}(s)\mathrm{d}s\right)\right)\hat{\rho}_0\left(\exp\left(-\frac{\mathrm{i}}{2}\hat{\sigma}_z \otimes \int_0^t \hat{r}(s)\mathrm{d}s\right)\right). \quad (1.83)$$

Reservoir in a thermal state

Let us now assume that the system is initially in a product state $\hat{\rho}_0 = \hat{\rho}_{S,0}\otimes\hat{\rho}_{\mathrm{th}}$, where

$$\hat{\rho}_{\mathrm{th}} = \bigotimes_{k\in\mathbb{N}}\left(1 - e^{-\beta\omega_k}\right)e^{-\beta\omega_k\hat{b}_k^\dagger\hat{b}_k} \quad (1.84)$$

is a thermal state of the bosonic bath at temperature $T = 1/\beta$. We write $\hat{\rho}_S(t) = \mathrm{Tr}_B\left[\hat{\rho}(t)\right]$ the reduced density operator of the system. By taking the partial trace over the bath degrees of freedom in the previous equation, and using the same decomposition as in the classical case, we obtain

$$\hat{\rho}_S(t) = \hat{\Pi}_0\hat{\rho}_{S,0}\hat{\Pi}_0 + \hat{\Pi}_1\hat{\rho}_{S,0}\hat{\Pi}_1 + \mathrm{Tr}\left[\hat{\rho}_{\mathrm{th}}e^{\mathrm{i}\int_0^t \hat{r}(s)\mathrm{d}s}\right]\hat{\Pi}_0\hat{\rho}_{S,0}\hat{\Pi}_1 + h.c.. \quad (1.85)$$

This is analog to the previous case, with the expectation over the classical noise replaced by the quantum average with respect to the bath thermal state:

$$\mathbb{E}\left[e^{\mathrm{i}\int_0^t \xi(s)\mathrm{d}s}\right] \to \left\langle e^{\mathrm{i}\int_0^t \hat{r}(s)\mathrm{d}s}\right\rangle_{\mathrm{th}}. \quad (1.86)$$

As for the classical case, the quantum noise for a bath in a thermal state is Gaussian, and we can write

$$\begin{aligned}\left\langle \exp\left(\mathrm{i}\int_0^t \hat{r}(s)\mathrm{d}s\right)\right\rangle_{\mathrm{th}} &= \exp\left(-\frac{1}{2}\int_0^t\int_0^t \langle \hat{r}(s_1)\hat{r}(s_2)\rangle_{\mathrm{th}}\,\mathrm{d}s_1\mathrm{d}s_2\right)\\ &= \exp\left(-\frac{1}{2}\sum_{k\in\mathbb{N}}\int_0^t\int_0^t \langle \hat{r}_k(s_1)\hat{r}_k(s_2)\rangle_{\mathrm{th}}\,\mathrm{d}s_1\mathrm{d}s_2\right)\,.\end{aligned} \tag{1.87}$$

A detailed proof of this result is provided in App. C. If we denote [15]

$$\begin{aligned}K_k(t-s) &= \langle \hat{r}_k(t)\hat{r}_k(s)\rangle_{\mathrm{th}}\,,\\ K(t-s) &= \sum_{k\in\mathbb{N}} K_k(t,s)\,,\end{aligned} \tag{1.88}$$

then we have

$$K_k(t-s) = 4|g_k|^2\left(e^{\mathrm{i}\omega_k(t-s)}\left\langle \hat{b}_k^\dagger\hat{b}_k\right\rangle_{\mathrm{th}} + e^{-\mathrm{i}\omega_k(t-s)}\left\langle \hat{b}_k\hat{b}_k^\dagger\right\rangle_{\mathrm{th}}\right) \tag{1.89}$$

A simple calculation gives [16]

$$\begin{aligned}\left\langle \hat{b}_k^\dagger\hat{b}_k\right\rangle_{\mathrm{th}} &= \frac{1}{2}\coth\left(\frac{\beta\omega_k}{2}\right) - \frac{1}{2}\,,\\ \left\langle \hat{b}_k\hat{b}_k^\dagger\right\rangle_{\mathrm{th}} &= \frac{1}{2}\coth\left(\frac{\beta\omega_k}{2}\right) + \frac{1}{2}\,.\end{aligned} \tag{1.90}$$

We can then approximate the discrete set of modes by a continuum with an associated frequency density

$$\sum_{k\in\mathbb{N}}|g_k|^2 \to \int_0^{+\infty} J(\omega)\mathrm{d}\omega\,, \tag{1.91}$$

and we have [17]

$$\tilde{K}(\omega) = \int_{-\infty}^{+\infty} K(\tau)e^{\mathrm{i}\omega\tau}\mathrm{d}\tau = \begin{cases}\pi J(\omega)\left(\coth\left(\dfrac{\beta\omega}{2}\right)+1\right) & \text{if } \omega>0\,,\\ \pi J(\omega)\left(\coth\left(\dfrac{\beta\omega}{2}\right)-1\right) & \text{if } \omega<0\,.\end{cases} \tag{1.92}$$

Unlike for the classical noise, the spectral density of the quantum noise is asymetric. This asymetry results from the non-commutativity of the field operators, and it corresponds to the Kubo-Martin-Schwinger condition

$$\tilde{K}\left(-\omega\right) = e^{-\beta\omega}\tilde{K}\left(\omega\right)\,, \tag{1.93}$$

which the bath state must satisfy to be a stationary state [137].

[15] The quantum noise is stationary since the reservoir is in a thermal stationary state.

[16] $\coth(x) = \frac{e^x+e^{-x}}{e^x-e^{-x}}$.

[17] Here $J(\omega)$ is defined as the continuous approximation of $|g_k|^2$, such that $J(\omega) = J(-\omega)$.

Short-time regime	Vacuum regime	Thermal regime
$t \ll \Omega^{-1}, \tau_B$	$\Omega^{-1} \ll t \ll \tau_B$	$\tau_B \ll t$
$\Gamma(t) \simeq \frac{A}{2}\Omega^2 t^2$	$\Gamma(t) \simeq A \ln(\Omega t)$	$\Gamma(t) \simeq At/\tau_B$

Table 1.1: Different regimes for the decoherence function.

Thermal and vacuum regimes

Let us now estimate the decoherence function for this model. Using Eqs. (1.51) and (1.92) we obtain

$$\Gamma(t) = \int_0^{+\infty} J(\omega)\frac{t^2}{2}\mathrm{sinc}\left(\frac{\omega t}{2}\right)^2 \coth\left(\frac{\beta\omega}{2}\right)\mathrm{d}\omega\,. \tag{1.94}$$

In the following we take

$$J(\omega) = A\omega e^{-\omega/\Omega}\,, \tag{1.95}$$

where Ω is a high-frequency cutoff and the linear increase in ω corresponds to a quantum optical regime for which $g_k \propto \sqrt{\omega_k}$. The decoherence function can be decomposed in a vacuum part $\Gamma_{\mathrm{vac}}(t)$ corresponding to a zero temperature and a thermal part $\Gamma_{\mathrm{th}}(t)$. Denoting $\tau_B = \beta/\pi$ and assuming $\tau_B \gg \Omega^{-1}$, we have [18]

$$\Gamma_{\mathrm{vac}}(t) = \frac{A}{2}\ln\left(1+\Omega^2 t^2\right), \qquad \Gamma_{\mathrm{th}}(t) = A\ln\left(\frac{\sinh(t/\tau_B)}{t/\tau_B}\right). \tag{1.96}$$

Three time regimes can be identified from these expressions, which are summarized in Tab. 1.1. In the short-time regime, the fluctuations of the bath are negligible and the system is barely affected by decoherence, as $\Gamma(t) \propto (\Omega t)^2 \ll 1$. In the vacuum regime, the decoherence is mainly due to the vacuum fluctuations of the bosonic fields. The thermal regime corresponds to times larger than the bath auto-correlation time $\tau_B = 1/(\pi k_B T)$, where the decoherence is mostly thermal. This last regime is Markovian [19], as it yields a constant decoherence rate $\gamma = A/\tau_B$. As in the classical case, we can derive a Lindblad master equation in this regime for the coarse-grained time scale $\tau_B \ll \delta t \ll \gamma^{-1}$ which only exists if $A \ll 1$, which is a condition of weak-coupling between the system and the bath.

Decoherence

The models presented in this section are models of decoherence. As such, they are characterized by the fact that they single out a specific basis of states, which is determined by the coupling between the system and its environment. This basis of states remains unaffected by the noise, while superpositions of these states are progressively destroyed. The decoherence mechanism thus provides physical explanation to the transition from

[18] The proofs of these formulas can be found in App. C.

[19] In the sense that it yields a constant dissipation rate. The instantaneous dissipation rate is positive in the three time regimes, so in that sense all the regimes could be called Markovian.

quantum superpositions to classical mixtures, an effect that has been observed experimentally [162]. This mechanism is crucial in measurement theory to explain the existence of macroscopic measurement appartus measuring given appartus [20] [137, 163].

Decoherence has also an impact on the computational complexity associated with the simulation of a quantum system on a classical computer. To store and modify the density operator of a given quantum system requires an amount memory and computational resources that scales exponentially in the number of degrees of freedom of the system. Roughly, if the system considered is subject to large enough decoherence, many of the non-diagonal entries of the density operator vanishes and the system can be approximately described by a density operator with a small number of non-zero entries. Put in another way, the decoherence destroys entanglement and coherences, typically rendering the system more classical and easier to simulate [44, 54, 161].

[20] When a measurement is realized, a macroscopic appartus couples to the system to measure. A given coupling can corresponds to the measurements multiple incompatible observables, but the decoherence affecting the measurement appartus distinguish the effectively measured observable.

2 Noisy quantum kernel machines

In recent years, machine learning has blossomed in a wide variety of fields and delivered a large number of applications driven by the achievements of deep artificial neural networks [64, 65]. However, the growing demand for computational resources and energy for training such deep architectures on ever-increasing amounts of data makes its long-term sustainability uncertain [164]. In this context, devolving computationally demanding tasks to machine-learning devices with suitable physical systems acting as hardware is emerging as a relevant alternative. Unfortunately, while the neural-network sequential architecture is well suited for software implementations on standard computers, the great number of parameters to be tuned during training remains in practice an obstacle to physical implementations. A simpler alternative approach is provided by the category of "shallow models", such as reservoir-computing [118] or extreme learning machines [165], which have led to physical proposals [166, 167] and experimental realizations [168, 169]. In such machines, the input data are encoded in the dynamics of a physical system and the associated predictions are obtained by considering a linear combination of measured observables, weighted by a set of trainable parameters to be optimized by training [1]. Importantly, this is done while keeping the parameters of the physical system fixed, hence requiring hardly any degree of control over the system. Kernel machines, whose trial functions can be represented in terms of positive semi-definite and symmetric kernel functions [104], belong to this category. More generally, kernel theory has proved to be a very useful tool to understand a wide range of machine-learning algorithms. Recently, a close connection between kernel machines and deep neural networks in the infinite width limit has been established [170], further extending the relevance of these methods.

With the emergence of the field of quantum machine learning, quantum "shallow" machines have also been put forward, such as those based on quantum reservoir-computing, extreme learning and quantum kernels [102, 103, 111–117, 171–175]. Most often, the investigations of these models have been focusing on isolated quantum systems with unitary dynamics. At present, however, most quantum devices within practical reach are subject to a significant degree of dissipation and/or decoherence. An important problem is therefore to understand the impact of realistic noise on these settings. The literature on the subject is yet in its very infancy. For time-dependent tasks, one study on quantum reservoir computing suggested that dissipation increases the processing capacity and the non-

[1] In the litterature different names correspond to this type of approach. The general trend seems to be to call such a scheme a *reservoir computing* method whenever the task to tackle is time-dependent, and to call them *extreme learning* methods when it is not.

linearity of the embedding, at the price of a reduced memory capacity of the system [176]. An advantageous scaling of the performance of a quantum reservoir-computing scheme, as compared to its classical counterpart, was recently reported [177]. For quantum neural networks, noise have been linked the existence of trainability issues [120], at topic which is discussed in chapter 3. Regarding quantum kernel machines, a recent work have showed that noise induces a concentration of the kernel functions, leading to trainability issues of the same kind [117].

This chapter presents the results in our work [α], which investigate the use of open quantum systems as noisy quantum kernel machines. It is structured as follows. In Sec. I, we describe a general noisy quantum kernel machine scheme based on a reservoir computing approach. In Sec. II, we analyze this scheme within the kernel-theory framework. In particular we study the link between the kernel spectrum and important properties of machine-learning models, such as the expressive power and the generalization capacity. We introduce and study the effective kernel rank. Within a statistical-learning approach, we provide an upper-bound on the generalization error for noisy quantum kernels. In Sec. III, we describe a class of noisy quantum kernel machines based on driven-dissipative chains of spins. We report a comprehensive study of the dependence of the performance metrics on the system size and noise for this class of models in section III.2. Finally, conclusions and perspectives are drawn in Sec. IV.

I General scheme

In this section we present a scheme of noisy quantum kernel machine based on a reservoir computing approach, in a supervised learning setup. The objective of supervised learning is to approximate a causal relation between elements $\boldsymbol{x}$ of an input set $\mathcal{X}$ and some target quantities $y \in \mathcal{Y}$, based upon a set of known training examples $\mathcal{S} = \{(\boldsymbol{x}_i, y_i) \,|\, i \in [\![1, N_{\text{train}}]\!]\}$. The input features are considered as independent realizations of a random variable following a probability distribution $p(x)$ on $\mathcal{X}$. Upon assuming the inputs and target quantities are related according to an unknown ground-truth function $y_i = y(\boldsymbol{x}_i)$, we aim to approximate it using a trial function f parameterized by $\boldsymbol{w}$, to be optimized using the training set $\mathcal{S}$. The specific form of f depends on the considered model architecture. Here, we describe noisy quantum kernel machines exploiting the dynamics of open quantum systems to generate such a trial function. This scheme is summarized pictorially in Fig. 2.1.

I.1 Encoding on the quantum system

Let us consider a system S initially prepared in a state $\hat{\rho}_0$ at $t = 0$. For each element of the input space, represented by a vector $\boldsymbol{x} \in \mathcal{X}$, a procedure can be defined to encode it into the non-unitary dynamics of a generic open quantum system. As will be shown in Sec. III, this can be achieved, for instance, by encoding the input vector in a proper modulation of the driving fields acting on the system.

We consider an open quantum system S whose dynamic can be described by a time-independent Lindblad master equation of the form given in Eq. (1.60). This master

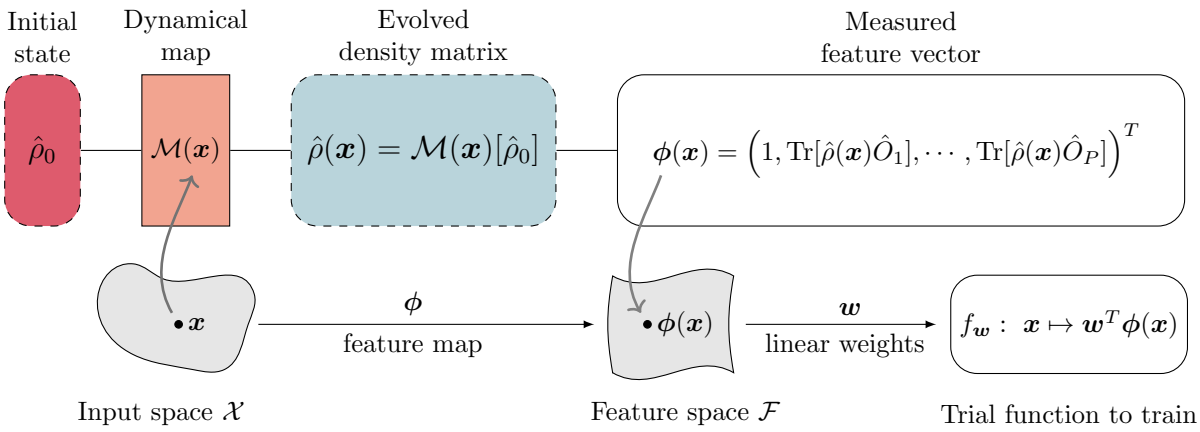


Figure 2.1: Scheme of a noisy quantum kernel machine. An element $\boldsymbol{x}$ of the input space $\mathcal{X}$ is encoded into a density matrix $\hat{\rho}(\boldsymbol{x})$ obtained by evolving in time a fixed initial state described by the density matrix $\hat{\rho}_0$ [see Fig. 2.2 for a specific example of the encoding process described by the evolution map $\mathcal{M}(\boldsymbol{x})$]. The measured features are represented by a vector of observables $\boldsymbol{\phi}(\boldsymbol{x})$ (with an added 1 corresponding to the unity operator as first element to create an offset term) that belongs to the feature space $\mathcal{F}$. The trial function is obtained by applying a linear transformation to the feature vector (depending nonlinearly on $\boldsymbol{x}$) with a vector of weights $\boldsymbol{w}$ that is optimized via the training procedure described in the main text.

equation describes the evolution from the initial density matrix into a final density matrix:

$$\hat{\rho}(\boldsymbol{x}, t) = \mathcal{M}(\boldsymbol{x}, t)[\hat{\rho}_0], \tag{2.1}$$

where $\mathcal{M}(\boldsymbol{x}, t)$ is the quantum channel representing the evolution of $\hat{\rho}_0$. It depends on $\boldsymbol{x}$ via the encoding procedure: If the input is encoded in driving fields, as we will consider later, the Hamiltonian, and consequently the density matrix at any time, bears a dependence on the input. In principle, one could also encode the input into a modulation of the loss rates, although we will not treat this case here [2]. In what follows, when considering a fixed final time t_f for the time-evolution, we denote $\mathcal{M}(\boldsymbol{x}) = \mathcal{M}(\boldsymbol{x}, t_f)$ and $\hat{\rho}(\boldsymbol{x}) = \hat{\rho}(\boldsymbol{x}, t_f)$ to simplify the notation.

I.2 Decoding through measurements

At time t_f, after the encoding procedure, we extract the processed information by performing a set of measurements of the system. Given the density matrix $\hat{\rho}(\boldsymbol{x})$ and a set of system observables $\mathcal{O} = \{\hat{O}_j \,|\, j = 1, \ldots, P\}$, information about the response of the open quantum system to the input $\boldsymbol{x}$ is contained in the following vector

$$\boldsymbol{\phi}(\boldsymbol{x}) \equiv \left(1, \langle \hat{O}_1 \rangle_{\boldsymbol{x}}, \ldots, \langle \hat{O}_P \rangle_{\boldsymbol{x}}\right)^T, \tag{2.2}$$

[2] Using the models presented in Chap. 1, such an encoding could be realized using a random driving which amplitude and correlation function would depend on the input $\boldsymbol{x}$.

where

$$\langle \hat{O}_j \rangle_{\boldsymbol{x}} = \mathrm{Tr}\left[\hat{O}_j \hat{\rho}(\boldsymbol{x})\right]. \tag{2.3}$$

The vector $\boldsymbol{\phi}(\boldsymbol{x})$ belongs to the feature space $\mathcal{F} \subseteq \mathbb{R}^P$ and depends on the input $\boldsymbol{x}$, generally in a nonlinear fashion. Note that the constant component 1, which ensures that the trial function can fit a biased target function, can be seen as the measurement of the identity observable, since the density matrix $\hat{\rho}(\boldsymbol{x})$ always has unit trace.

Finally, the trial function f of the noisy quantum kernel machine is given by the affine transformation:

$$f : \boldsymbol{x} \mapsto \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}), \tag{2.4}$$

where the vector $\boldsymbol{w} \equiv (b, w_1, \ldots, w_P)^T \in \mathbb{R}^{P+1}$ contains the parameters of the linear transformation and b represents the bias term. An alternative approach to the construction of the feature vector, based on time-multiplexing measurements, will be presented in Sec. III.

I.3 Training procedure

The previous trial function characterized will be optimized using the l2-regularized least-squares loss function over a training set $(\boldsymbol{x}_i, y_i) \in \mathcal{S}$ consisting of N_{train} inputs $\boldsymbol{x}_i \in \mathcal{X}$ and labels $y_i \in \mathcal{Y}$, namely:

$$\mathcal{L}(\boldsymbol{w} \mid \mathcal{S}) := \frac{1}{2N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} \left(y_i - \boldsymbol{w}^T \boldsymbol{\phi}(x_i)\right)^2 + \frac{\lambda}{2} \|\boldsymbol{w}\|_2^2 . \tag{2.5}$$

The second term in Eq. (2.5) is a regularization penalty that helps to prevent over-fitting. The strength of this penalty is controlled by the corresponding regularization parameter λ. Adding such a regularization bias is on average equivalent to adding a centered Gaussian noise of variance λ to the measurement features before the optimization [65].

Such a classifier is sometimes referred to as a least-square support-vector classifier [178] [3]. Although most classification problems are commonly treated with other loss functions [105], using the least-squares loss function allows us to perform the optimization analytically [4]. Indeed, upon introducing the $(P+1) \times N_{\text{train}}$ matrix $\boldsymbol{\Phi}$, whose columns are the quantum feature vectors $\boldsymbol{\phi}(\boldsymbol{x}_i)$ associated to the training input $\boldsymbol{x}_i$, and $\boldsymbol{y}$, the column vector of size N_{train} containing the corresponding labels, the optimal weights are given by [5]

$$\boldsymbol{w}^* = \left(\boldsymbol{\Phi}\boldsymbol{\Phi}^T + N_{\text{train}}\lambda \mathbb{1}\right)^{-1} \boldsymbol{\Phi}\boldsymbol{y} . \tag{2.6}$$

[3] More generally, models using a l2-regularized least square loss function are called ridge regression models in the literature [105].

[4] Usually, the choice of the loss function is guided by type of task to be performed, since a given loss can be associated with a probabilistic model of the relationship between the inputs and the targets. This link is briefly reminded in App. D

[5] Note that in certain scenario, one may prefer to exclude the first component of $\vec{w}$, which is a constant intercept term, in the regularization. Then it suffices to replace the $(1,1)$ entry of $\mathbb{1}$ by 0 in Eq. (2.6) to obtain the optimal weights. This is equivalent to using a centered kernel, as discussed in App. E.

II Quantum kernel and decoherence

The generic encoding-decoding scheme encompasses a large class of quantum machine-learning models. Here, we describe a decoding based on a linear combination of measurements, but other decoding methods were proposed in the literature. In particular, it was recently shown that quantum neural networks with alternated encoding and parameterized layers can be mapped to models with an encoding/decoding structure [179], where the decoding is achieved by optimizing a single parameterized measurement.

Models described by the previous scheme can be analyzed in the framework of kernel theory, which provides useful tools to understand properties such as expressivity, trainability and capacity to generalize to a test sample of unseen data. In this section we first concisely introduce the kernel framework. A overview of classical kernel methods is given in App. F. We then specialize our discussion to noisy quantum kernels, and show how we can link the role of dissipation and decoherence to the kernel's main figures of merit.

We aim at determining the largest class of functions that can be approximated by our trial function f. This class depends on the type of decoding used, that is on the specific set of measurements that are performed on the quantum system. When measuring a set $\mathcal{O}$ of observables, this function space reads

$$\mathcal{H}(\mathcal{O}) = \{f : \boldsymbol{x} \mapsto \mathrm{Tr}\left[\hat{\rho}(\boldsymbol{x})\hat{A}\right] \mid \hat{A} \in \mathrm{Span}(\mathcal{O})\}\,. \tag{2.7}$$

In this case, the feature vector $\boldsymbol{\phi}(\boldsymbol{x})$ gives rise to a positive semi-definite and symmetric function which we call the feature kernel:

$$k_{\mathcal{O}}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{\phi}(\boldsymbol{x}'). \tag{2.8}$$

This kernel function, together with the probability distribution p of inputs $\boldsymbol{x} \in \mathcal{X}$ [6], uniquely determines a specific set of real-valued functions, the so-called reproducing kernel Hilbert space (RKHS):

$$\mathrm{Span}\{f : \boldsymbol{x} \mapsto k_{\mathcal{O}}(\boldsymbol{x}, \boldsymbol{x}') \mid \boldsymbol{x}' \in \mathcal{X}\}. \tag{2.9}$$

The RKHS associated to $k_{\mathcal{O}}$ can be shown to be exactly the space of hypothesis functions $\mathcal{H}(\mathcal{O})$ (see App. F and the references therein.). Hence, the study of the kernel function allows one to investigate the structure of $\mathcal{H}(\mathcal{O})$. In particular, it follows that one can use the eigendecomposition of the kernel function as a basis of the class of functions that can be represented by our model. This useful property motivates the adoption of a kernel standpoint in what follows.

II.1 Quantum kernel

In order to discuss the expressive power of our model, we introduce the largest class of transformations $\mathcal{H}_{\mathrm{full}}$ that can be achieved for a given encoding strategy [101]:

$$\mathcal{H}_{\mathrm{full}} = \{f : \boldsymbol{x} \mapsto \mathrm{Tr}\left[\hat{\rho}(\boldsymbol{x})\hat{A}\right] \mid \hat{A} \in \mathcal{B}\left(\mathcal{H}_S\right),\, \hat{A} = \hat{A}^\dagger\}\,. \tag{2.10}$$

[6] The probability measure p on the input space $\mathcal{X}$ is important here as it determines the scalar product on the space of real-valued functions on $\mathcal{X}$ through $\langle f, g\rangle = \mathbb{E}_{\boldsymbol{x}\sim p}\left[f(\boldsymbol{x})g(\boldsymbol{x})\right]$. The reproducing property, crucial to link the RKHS and its kernel, relies on such a well-defined scalar product.

The class of transformation yielded by a set of measurements $\mathcal{O}$ is necessarily included in this maximal class $\mathcal{H}(\mathcal{O}) \subseteq \mathcal{H}_{\text{full}}$; the equality holds whenever $\mathcal{O}$ spans the set of all the system's observables. In the following, we will use the term "full tomography" to refer to this ideal implementation. It turns out that $\mathcal{H}_{\text{full}}$ is the RKHS of a particular kernel, the quantum kernel, that solely depends on the feature map $\hat{\rho}(\boldsymbol{x})$ [7] [102, 114]:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \text{Tr}\left[\hat{\rho}(\boldsymbol{x})\hat{\rho}(\boldsymbol{x}')\right] . \tag{2.11}$$

This kernel arises naturally from the Hilbertian structure of the space of quantum states. As it represents the maximal achievable class of transformation an encoding can give, the quantum kernel provides insight on the expressive power of our model. Note that this kernel can be identified with the previous feature kernel $k_{\mathcal{O}}$ provided that the measurements $\mathcal{O}$ form an orthonormal basis $\mathcal{B} = \{B_j\}_j$ of the space of observables, i.e. $k = k_{\mathcal{B}}$ with $\text{Tr}\left[\hat{B}_i\hat{B}_j\right] = \delta_{ij}$ and we impose $\hat{B}_0 \propto \hat{\mathbb{1}}$ by convention.

In what follows, it will be useful to work with a "centered" version of the quantum kernel. Centering the kernel is equivalent to working with hypothesis functions that have zero mean value on the input set. As we will show, this is convenient for interpreting some of the key quantities we will introduce in terms of probabilistic quantities. In App. E, we show that, at least for balanced data, the use of the L2 loss function allows us to work with a centered version of the quantum kernel without lack of generality. We define the centered version of a quantity $f(\boldsymbol{x})$ as

$$\delta f(\boldsymbol{x}) = f(\boldsymbol{x}) - \mathbb{E}_p\left[f\right] . \tag{2.12}$$

With this notation, the centered kernel is given by

$$k_c(\boldsymbol{x}, \boldsymbol{x}') = \text{Tr}\left[\delta\hat{\rho}(\boldsymbol{x})\delta\hat{\rho}(\boldsymbol{x}')\right] , \tag{2.13}$$

and the corresponding RKHS is

$$\mathcal{H}_{k,c} = \text{Span}\{f : \boldsymbol{x} \mapsto k_c(\boldsymbol{x}, \boldsymbol{x}') \mid \boldsymbol{x}' \in \mathcal{X}\} . \tag{2.14}$$

The constant feature we introduced in Eq. (2.2) becomes irrelevant when using centered quantities, so we drop it and define

$$\delta\boldsymbol{\phi}(\boldsymbol{x}) \equiv \left(\delta\langle\hat{O}_1\rangle_{\boldsymbol{x}}, \ldots, \delta\langle\hat{O}_P\rangle_{\boldsymbol{x}}\right)^T . \tag{2.15}$$

We can also correspondingly drop the weight term b, so that the weight vectors can be redefined as $\boldsymbol{w} = (w_1, \ldots, w_P)^T \in \mathbb{R}^P$. The space $\mathcal{H}_{k,c}$ can be rewritten as

$$\mathcal{H}_{k,c} = \{f : \boldsymbol{x} \mapsto \boldsymbol{w}^T\delta\boldsymbol{\phi}(\boldsymbol{x}), \quad \boldsymbol{w} \in \mathbb{R}^P\}, \tag{2.16}$$

where the centered quantum kernel reads

$$k_c(\boldsymbol{x}, \boldsymbol{x}') = \delta\boldsymbol{\phi}(\boldsymbol{x})^T\delta\boldsymbol{\phi}(\boldsymbol{x}') \tag{2.17}$$

with the choice of $\mathcal{O} = \mathcal{B}$. The quantum feature matrix $\boldsymbol{\Phi}$ is then replaced by a $P \times N_{\text{train}}$ matrix $\delta\boldsymbol{\Phi}$, whose columns are the centered feature vectors $\delta\boldsymbol{\phi}(\boldsymbol{x_i})$.

[7] For a closed system, this quantum kernel can be directly evaluated through measurement [112] and the trial function can be expressed in terms of the quantum kernel and optimized in an equivalent way. This corresponds to the dual picture approach, which is reminded in App. F.

II.2 Kernel eigen-decomposition

Under general assumptions, the centered quantum kernel admits a decomposition into an orthonormal family of eigenfunctions [180]:

$$k_c(\boldsymbol{x}, \boldsymbol{x}') = \sum_i \lambda_i \delta\psi_i(\boldsymbol{x}) \delta\psi_i(\boldsymbol{x}') \,, \quad \mathbb{E}_p \left[\delta\psi_i \delta\psi_j\right] = \delta_{ij} \,, \tag{2.18}$$

where $\{\lambda_i\}_i$ are positive eigenvalues sorted in a decreasing order, namely $\lambda_{i+1} \leq \lambda_i$, $\forall i$. When necessary, we can complete this orthonormal family into a basis with eigenfunctions associated to zero eigenvalues. In the case of the uncentered quantum kernel, the kernel eigenfunctions correspond to an orthonormal basis of system observables [103]. When the kernel is centered, the basis of kernel eigenfunctions corresponds to an orthonormal basis $\{\hat{E}_i\}_i$ of the space of zero-trace observables, which we call eigenobservables. Such operators satisfy the following properties:

$$\mathrm{Tr}\left[\hat{E}_i \hat{E}_j\right] = \delta_{ij} \,, \quad \mathrm{Tr}\left[\hat{E}_i\right] = 0 \,. \tag{2.19}$$

The eigenfunctions are given by

$$\delta\psi_i(\boldsymbol{x}) = \frac{1}{\sqrt{\lambda_i}} \mathrm{Tr}\left[\delta\hat{\rho}(\boldsymbol{x}) \hat{E}_i\right] = \frac{1}{\sqrt{\lambda_i}} \delta\langle \hat{E}_i \rangle_{\boldsymbol{x}} \,. \tag{2.20}$$

The corresponding eigenvalues are then given by the variances of the eigenobservable measurements over the input set, namely:

$$\lambda_i = \mathbb{E}_p \left[\delta\langle \hat{E}_i \rangle_{\boldsymbol{x}}^2\right] = \mathrm{Var}_p \left[\langle \hat{E}_i \rangle_{\boldsymbol{x}}\right] \,. \tag{2.21}$$

One can see this eigen-decomposition of the kernel as a principal-component analysis in the space of quantum features, as it yields an orthogonal basis of measurement functions ordered by their variances on the input set. We stress that these are variances of the observables expectation values over the quantum states representing the different inputs, and thus are very different from the quantum variance of the corresponding observable for a specific state.

The previous decomposition of the kernel is very useful for grasping the learning mechanism and the model expressivity. Upon working with centered features, the loss function introduced in Eq. (2.5) becomes

$$\mathcal{L}_c\left(\boldsymbol{w} \mid \mathcal{S}\right) = \frac{1}{2N_{\mathrm{train}}} \sum_{i=1}^{N_{\mathrm{train}}} \left(y_i - \boldsymbol{w}^T \delta\boldsymbol{\phi}(x_i)\right)^2 + \frac{\lambda}{2} \|\boldsymbol{w}\|_2^2 \,. \tag{2.22}$$

Following [105], we can decompose the trial function $f(\boldsymbol{x}) = \boldsymbol{w}^T \delta\boldsymbol{\phi}(\boldsymbol{x})$ in the basis of the kernel eigenfunctions, namely as $f(\boldsymbol{x}) = \sum_j \beta_j \delta\psi_j(\boldsymbol{x})$. Exploiting such decomposition, the loss function becomes

$$\mathcal{L}_c\left(\boldsymbol{\beta} \mid \mathcal{S}\right) = \frac{1}{2N_{\mathrm{train}}} \sum_{i=1}^{N_{\mathrm{train}}} \left[y_i - \sum_j \beta_j \delta\psi_j(\boldsymbol{x}_i)\right]^2 + \frac{\lambda}{2} \sum_j \frac{\beta_j^2}{\lambda_j} \,. \tag{2.23}$$

Note that in the regularization term the components of the trial function on the eigenbasis are weighted by the corresponding kernel eigenvalues. The lower the variance of an

eigenobservable, the more the corresponding eigenfunction is penalized. Hence the regularization parameter λ acts as a smooth cutoff on the basis of the kernel eigenfunctions, which are then used to approximate the target function.

The spectrum of the kernel characterizes the generalization capacity and the expressivity of our model. It also finds applications in understanding many other machine learning scenarios. For instance, in the context of classical neural networks it has links with learning curves [181, 182]. Moreover, the kernel (or the neural tangent kernel in the context of classical and quantum neural networks [170, 174, 183]) shares its spectrum with the Fisher information matrix, of particular relevance for quantum neural networks [100].

II.3 Role of decoherence on expressivity and generalization error

The exponential growth of the Hilbert space dimension with the number of qubits in a network and the complex dynamics of quantum systems have created hope for a quantum advantage in the field of quantum machine learning. However, it is known that having a very high-dimensional feature space does not necessarily guarantee high machine-learning performances [105, 184]. Indeed, recent investigations within the quantum kernel framework somehow mitigated the hope for a general quantum advantage [99, 103, 113]. Yet, a clear quantum advantage has been demonstrated for some specific tasks [174, 175], again by exploiting the quantum-kernel formalism. In order for a quantum-kernel-based model to perform well on a given task, the set of transformations achieved must be well "aligned" with the target function $y(\boldsymbol{x})$. This notion of alignment is mathematically encapsulated in the kernel-target-alignment measure [185] which reads, for the centered quantum kernel,

$$\begin{aligned} A(k_c, y) &= \frac{\mathbb{E}_p\left[y(\boldsymbol{x})k_c(\boldsymbol{x},\boldsymbol{x}')y(\boldsymbol{x}')\right]}{\mathbb{E}_p\left[k_c(\boldsymbol{x},\boldsymbol{x}')^2\right]^{1/2}\mathbb{E}_p\left[y(\boldsymbol{x})^2\right]} \\ &= \frac{\sum_i \lambda_i \mathbb{E}_p\left[\delta\psi_i(\boldsymbol{x})y(\boldsymbol{x})\right]^2}{(\sum_i \lambda_i^2)^{1/2}\mathbb{E}_p\left[y(\boldsymbol{x})^2\right]} \,. \end{aligned} \tag{2.24}$$

Although the kernel-target alignment measures how well a kernel and the associated embedding fits a specific function, in this article we introduce another figure of merit that does not depend on a specific task, namely the "effective kernel rank" $R_{\text{eff}}(k)$, which quantifies the effective number of independent transformations that a given kernel can yield. Such a quantity is defined as:

$$\sqrt{R_{\text{eff}}(k_c)} = \sum_j A(k_c, g_j)\,, \tag{2.25}$$

where $\{g_j\}_j$ is any orthonormal basis of functions on the input space. As shown in App. I, for the centered quantum kernel, the effective kernel rank can be also expressed in terms of variances of the quantum expectation values of the measured observables:

$$\sqrt{R_{\text{eff}}(k_c)} = \frac{\sum_{i=1}^P \text{Var}_p\left[\langle\hat{O}_i\rangle_{\boldsymbol{x}}\right]}{\left(\sum_{i,j=1}^P \text{Cov}_p\left[\langle\hat{O}_i\rangle_{\boldsymbol{x}}, \langle\hat{O}_j\rangle_{\boldsymbol{x}}\right]^2\right)^{\frac{1}{2}}} \,. \tag{2.26}$$

Note that the denominator acts as a normalization and can be seen as a measure of the redundancy of the embedding when expressed in terms of $\hat{O}_i$. In Sec. III.2, we will investigate in a rather general class of physical models how the kernel effective rank scales with the system size and with noise.

In App. I, we also provide the proof showing that the kernel effective rank can be expressed in terms of the kernel spectrum:

$$\sqrt{R_{\text{eff}}(k)} = \frac{\sum_i \lambda_i}{\sqrt{\sum_i \lambda_i^2}} \,. \tag{2.27}$$

This expression is reminiscent of the reciprocal of the inverse participation ratio. The kernel effective rank provides information about the size of its support. Moreover, we have the following inequality:

$$R_{\text{eff}}(k) \leq |\{\lambda_i \neq 0\}| \,. \tag{2.28}$$

This is saturated when all the non-zero eigenvalues are all equal. The numerator in the expression for the square-root of the effective kernel rank is the kernel trace, which can be rewritten as:

$$\sum_i \lambda_i = \mathbb{E}_p \left[\text{Tr} \left[\hat{\rho}(\boldsymbol{x})^2 \right] \right] - \text{Tr} \left[\mathbb{E}_p \left[\hat{\rho}(\boldsymbol{x}) \right]^2 \right] \,. \tag{2.29}$$

In this expression, we recognize the difference between the average purity of the embedded density matrices over the input space and the purity of the average embedding matrix. The first term is of great relevance to our study, as it crucially depends on the dissipation and decoherence affecting the noisy quantum system: indeed, a low purity is the consequence of the openness of the quantum system. The second term instead measures the diversity of the embedding map; its importance is discussed in [103].

We emphasize that the kernel trace also appears to be relevant when investigating the ability of the model to perform well on unseen data, hence on its generalization properties. To measure the performance of a model on a binary classification task we use the accuracy $\mathcal{A}$. Given a prediction function f, the accuracy is given by the fraction of samples for which f assigns the right label and it can be defined as the expectation of a 0-1 loss function:

$$\mathcal{A}(f) = \mathbb{E} \left[\mathbb{1}_{y(\boldsymbol{x}) f(\boldsymbol{x}) \geq 0} \right] \,. \tag{2.30}$$

Since during the training we only have access to the data set $\mathcal{S}$ and not to the true distribution p, expectations values can only be approximated using the empirical distribution $\hat{p}$ on $\mathcal{S}$. The corresponding empirical expectations are given by $\mathbb{E}_{\hat{p}} \left[f(\boldsymbol{x}) \right] = \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} f(\boldsymbol{x}_i)$. From this, we can define the empirical accuracy $\mathcal{A}$ on the training set $\mathcal{S}$ and the true accuracy $\mathcal{A}^*$. Correspondingly, we can introduce the risk $\mathcal{R}^*$, also called error or inaccuracy, as $\mathcal{R}^* = 1 - \mathcal{A}^*$ (its empirical counterpart is defined analogously). It is convenient to introduce slightly modified versions of the risk and inaccuracy that depend on a margin-parameter $\eta > 0$. We introduce the η-margin loss as:

$$\Phi_\eta(y) = \begin{cases} 1 & \text{if} \quad y \leq 0 \\ 1 - \frac{y}{\eta} & \text{if} \quad 0 \leq y \leq \eta \\ 0 & \text{if} \quad \eta \leq y \end{cases} \,. \tag{2.31}$$

Correspondingly, we can introduce the empirical η-margin risk as:

$$\mathcal{R}_\eta(f) = \mathbb{E}_{\hat{p}}\left[\Phi_\eta(y(\boldsymbol{x})f(\boldsymbol{x}))\right] . \tag{2.32}$$

The η-margin-risk and the risk satisfy the following inequality:

$$\mathcal{R}(f) \leq \mathcal{R}_\eta(f) \leq \mathbb{E}_{\hat{p}}\left[\mathbb{1}_{y(\boldsymbol{x})f(\boldsymbol{x})\leq\eta}\right] . \tag{2.33}$$

The ability of the model to generalize well on unseen data is then quantified by the generalization error:

$$\mathcal{E} = \mathcal{R}^* - \mathcal{R} . \tag{2.34}$$

For kernel methods with kernel k, the generalization error admits an upper-bound involving the $N_{\text{train}} \times N_{\text{train}}$ empirical kernel matrix $\mathbf{K}$ whose entries are defined as $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. This bound depends on the specific task under consideration and on the exact space of trial functions used (details on the bound used and its derivation can be found in [186] and in App. II).

To derive the upper-bound, we fix a class of trial functions of the form $f : \boldsymbol{x} \mapsto \boldsymbol{w}^T\delta\boldsymbol{\phi}(\boldsymbol{x})$ where $\delta\boldsymbol{\phi}(\boldsymbol{x})$ corresponds to measurements of an orthonormal basis of observable: $\delta\phi_i(\boldsymbol{x}) = \delta\langle\hat{B}_i\rangle_{\boldsymbol{x}}$. We further constrain this class by choosing a parameter $\Lambda \geq 0$, and require that the trial function's parameters $\boldsymbol{w}$ satisfy $\|\boldsymbol{w}\|^2\Lambda \leq 1$. By exploiting Eq. (2.29), we get that, for such functions, the following inequality holds with probability at least $1-\delta$ on the training set $\mathcal{S}$:

$$\mathcal{R}^*(f) - \mathcal{R}_\eta(f) \leq \frac{2}{\eta}\left(\frac{\mathbb{E}_{\hat{p}}\left[\mathrm{Tr}\left[\hat{\rho}^2\right]\right] - \mathrm{Tr}\left[\mathbb{E}_{\hat{p}}\left[\hat{\rho}\right]^2\right]}{N_{\text{train}}\Lambda}\right)^{\frac{1}{2}} + 3\sqrt{\frac{\log(\frac{2}{\delta})}{2N_{\text{train}}}} . \tag{2.35}$$

Other generalization bounds can be established, in particular the authors of [187] found another bound using a quantum information theory standpoint, and their conclusions are in agreement with our results. Let us make a few important comments on the meaning of this inequality. The inequality has a probabilistic character controlled by $\delta > 0$. If we set this parameter to 0^+, the bound is always satisfied although it becomes trivial. The same goes with the margin parameter η: as $\eta \to 0^+$ the margin-error $\mathcal{R}_\eta(f)$ tends to the training error $\mathcal{R}(f)$, but again the right-hand side of the inequality diverges. The parameter Λ is another sort of regularization parameter, as the parameter λ: if $\Lambda \to 0^+$, the norm of the weight vector $\|\boldsymbol{w}\|$ can be arbitrarily large and over-fitting is not limited. Correspondingly, the right-hand side diverges and the bound becomes trivial. The most important crucial physical quantity involved in the upper bound is the kernel trace given in Eq. (2.29). Such a quantity accounts for the model expressivity. This duality between expressivity and generalization is crucial in machine learning [105]. What is relevant to our study is that this expressivity measure involves the mean purity of the embedded states and hence is affected by dissipation and decoherence acting on the noisy quantum kernel machine. The appearance of the regularization parameter Λ in this upper bound is also relevant as it allows us to establish a link with experimental constraints, such as imperfect measurements. In fact, as we will see in Section III.2, adding a Gaussian error of standard deviation σ to the observable measurements is equivalent to working with an infinitely precise measurement apparatus while replacing the regularization parameter λ with $\lambda + \sigma^2$ [65].

III Noisy quantum kernel machines with driven-dissipative spin chains

As an illustrative example, we here numerically simulate noisy quantum kernel machines based on 1D chains of spins subject to both driving and decoherence. The simulation of such an open quantum system for a large number of inputs, various choices of the number of sites and distinct disorder realizations is a computationally daunting task [8]. Indeed, this requires to exactly integrate a large set of corresponding Lindblad master equations of the form of Eq. (1.60). Hence, we have considered a simplified classification task involving only a subset of the MNIST dataset, namely classifying images of handwritten digits corresponding to the digits 3, 6 and 8, which share common shapes.

III.1 Encoding and decoding methods

Encoding through driving

A schematic description of the task and of the feature encoding through driving of the considered physical system is presented in Fig. 2.2. The original MNIST dataset consists of 28×28-pixel images. Encoding such high-dimensional features in the state of a quantum system is not an easy task. Therefore, we first linearly down-sample the raw images from 28×28 to 8×8 pixels, thereby reducing the dimension of the input features. The down-sampled images, viewed as vectors, are then multiplied by a random $8^2 \times 10$ matrix $\mathbf{W}$, whose entries are uniformly drawn over the interval $[-1, 1]$, yielding vectors $\boldsymbol{x'} = (x'_1, \ldots, x'_M)^T$ of $M = 10$ random-projection features. These are finally normalized by 3 times the standard deviation of the set $\{x'_i \mid i = 1, \ldots, M,\ x \in \mathcal{S}\}$. At the end of this procedure, every image in the dataset is represented by a vector $\boldsymbol{x}$ of size $M = 10$, which will be used as inputs in the following. These are computed only once and reused throughout this article, except in section III.2.

This encoding is designed so as to fix the amount of information fed to the system, independently from its number of sites. It allows us to perform a fair comparison of models associated to quantum systems of increasing sizes. In particular, this ensures that any observed increase of the performance with the system size is solely due to an intrinsic enhancement of the model expressive power. In Section III.2, we lift the above-defined "information bottleneck" and use a different encoding, where the number of encoded features M scales with the system size N. Therein, we show that this results in competitive performances, as compared to classical reservoir-computing settings involving hundreds to thousands of degrees of freedom [166, 168].

In what follows, we denote $\mathcal{X} \subseteq \mathbb{R}^M$ the input space consisting of the random-projection features representing the images to classify, and $\mathcal{Y} = \{3, 6, 8\}$ the set of corresponding labels. Our dataset consists of 17,000 images, which we split into a training set of $N_{\text{train}} = 15{,}000$ images and a testing set of $N_{\text{test}} = 2000$ images. As before, the training set is denoted as $\mathcal{S} = \{(\boldsymbol{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y} \,|\, i = 1, \ldots, N_{\text{train}}\}$.

[8] In total, the results presented in this work required approximately 800,000 scalar hours ($\sim$ 90 years) of computation and 5 terabytes of storage on the acknowledged French National High Performance Computing facility (GENCI). During the simulations, we used approximately up to $30,000$ cores simultaneously.

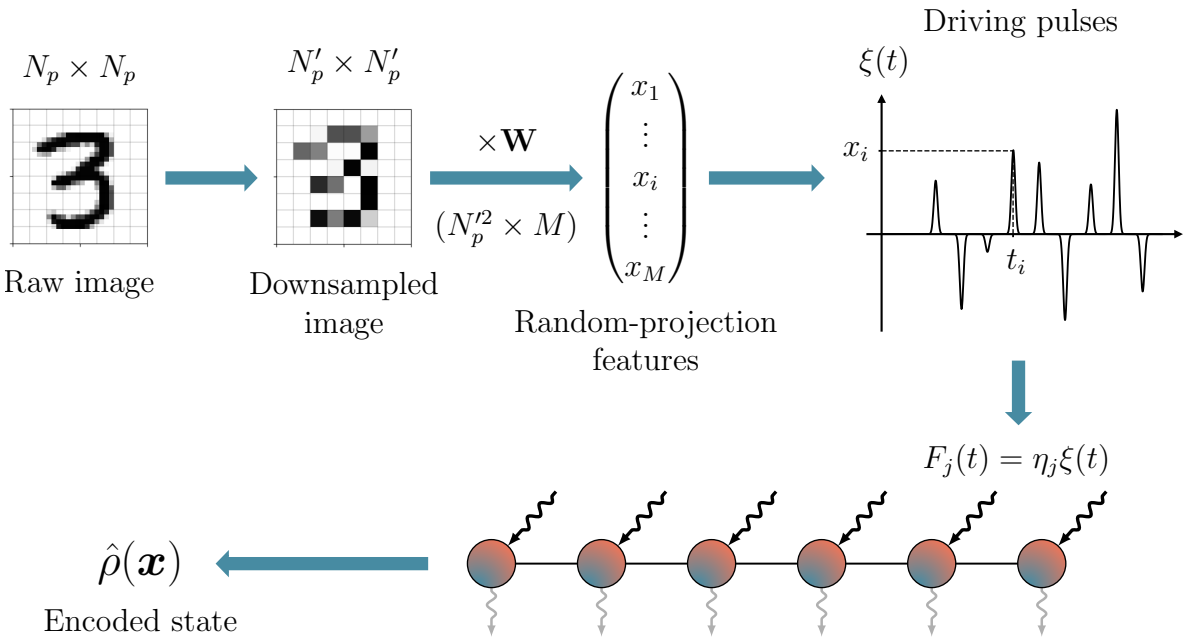


Figure 2.2: Schematic representation of the encoding procedure for the MNIST classification task. The input grayscale image, of original size $N_p \times N_p$, with $N_p = 28$, is first downsampled to a size of $N'_p \times N'_p$ (or $N'^2_p \times 1$ when viewed as a column vector), with $N'_p = 8$, and linearly transformed by a fixed $N'^2_p \times M$ random projection filter $\mathbf{W}$ to yield the vectors $\boldsymbol{x}'$ containing $M = 10$ random-projection features. Those features are normalized by 3 times the standard deviation over the set of all features for all images in the training set, and we denote $\boldsymbol{x}$ the normalized vectors representing the images. The vector $\boldsymbol{x}$ is then encoded into a sequence of driving pulses $\xi(t)$, where the amplitude of the ith pulse (at time t_i) is proportional to the input's ith component x_i. Finally, the pulses are used to drive a spin chain (initially prepared in the state $\hat{\rho}_0$), where the driving amplitude at site j is $F_j(t) = \eta_j \xi(t)$ with η_j a random site-dependent scale factor. We define the state of the spin chain immediately after the driving sequence to be the encoded state, represented by its density matrix $\hat{\rho}(\boldsymbol{x})$.

The system in which we encode the previous features is a driven-dissipative one-dimensional chain of N spins-1/2 described by the following Heisenberg XYZ Hamiltonian:

$$\hat{H}(t;\boldsymbol{x}) = \frac{\hbar}{2}\sum_{i=1}^{N}\left(F_i(t;\boldsymbol{x})\hat{\sigma}_x^i + \Delta_i\hat{\sigma}_z^i\right) - \frac{\hbar}{2}\sum_{\langle i,j\rangle}(J_{ij}^x\hat{\sigma}_x^i\hat{\sigma}_x^j + J_{ij}^y\hat{\sigma}_y^i\hat{\sigma}_y^j + J_{ij}^z\hat{\sigma}_z^i\hat{\sigma}_z^j), \tag{2.36}$$

with $F_i(t;\boldsymbol{x})$ an input-dependent driving field, Δ_i an on-site frequency detuning, and J_{ij}^k the symmetric coupling rate between nearest neighbors. Here, indices $\langle i,j\rangle$ run over all pairs of nearest neighbors. Parameters J_{ji}^k and Δ_i are uniformly drawn at random in the interval $[0, 2J]$. $1/J$ will be used as unit of time in the numerical plots. We prepare the system in an initial state with all spins down $\hat{\rho}_0 = \bigotimes_{i=1}^{N} |0\rangle\langle 0|$.

The encoding of the input $\boldsymbol{x}$ corresponding to a given image into the system state is performed by driving the system with a series of $M = 10$ sharp Gaussian pulses, whose

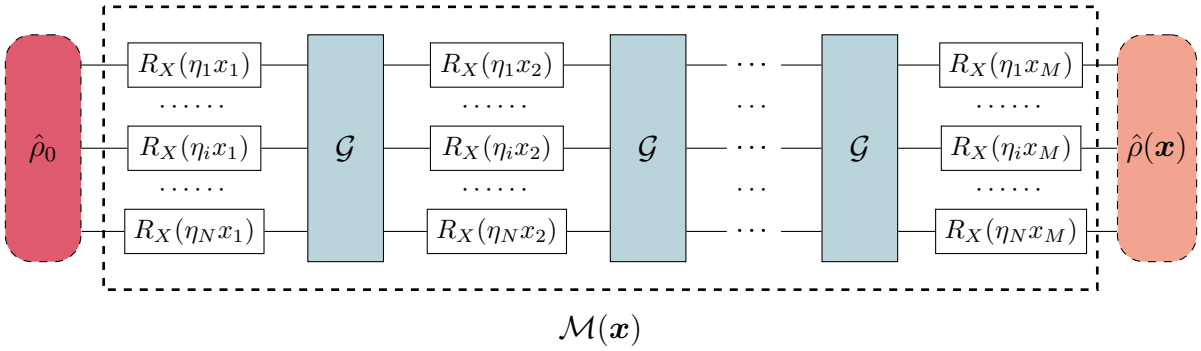


Figure 2.3: Equivalent circuit of the encoding procedure for the MNIST classification task. If the driving pulses are sharp enough, the encoding process of Fig. 2.2 can be equivalently seen as a quantum circuit, where the i-th driving pulse on site j is effectively a single-qubit X-rotation gate $R^X(\eta_j x_i)$, and the pulses at different times are separated by the gate $\mathcal{G}$ generated by the free dynamics of the spin system in the absence of the drive. Note that the entire process between $\hat{\rho}_0$ and $\hat{\rho}(x)$ serves as the dynamical map $\mathcal{M}(x)$ shown in Fig. 2.1.

amplitudes are proportional to the input vector elements, as illustrated in Fig. 2.2. We first define a generic driving $\xi(t;\boldsymbol{x})$ from the feature $\boldsymbol{x}$:

$$\begin{aligned} \xi(t;\boldsymbol{x}) &= \sum_{k=1}^{N} \frac{x_k}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t-t_k)^2}{2\sigma^2}\right), \\ t_k &= (k-1)\Delta t + 10\sigma, \quad \forall k = 1,\ldots,M, \end{aligned} \tag{2.37}$$

where the time interval between two successive pulses is $\Delta t = 1/(2J)$ and the width of each pulse is $\sigma = 1/(50J)$. Then the driving on site i is taken to be proportional to this generic driving:

$$F_i(t;\boldsymbol{x}) = \eta_i \xi(t;\boldsymbol{x}), \tag{2.38}$$

where the η_i are random factors uniformly distributed in the interval $[-\pi,\pi]$. Under these driving conditions, the coherent part of the system dynamics can be thought of as that of an equivalent quantum circuit alternating between a set of local X-rotation gates, of the form $R_i^X(\eta_i x_k)$, and a deep block generating entanglement among qubits [9], as illustrated in Fig. 2.3. The scaling factors η_i prevent the spins from rotating all together. This procedure, where a random-projection feature is fed to the system every Δt, is in close analogy with the repeated-encoding prescription in variational quantum circuits, which is known to improve the expressivity of a model [101].

Shortly after the last pulse of the driving ends, at time $\tau = 30\sigma + M\Delta t$, we get the final encoded state represented by the density matrix $\hat{\rho}(\boldsymbol{x})$. This encoding procedure acts as a non-linear map from the input space of images to the high-dimensional space of N-spin mixed quantum states.

[9] This can be explicitly expressed as a D-deep circuit via Trotterization as $\mathcal{G} = \left[\prod_{i=1}^{N} R_i^Z\left(-\frac{2\Delta_i \Delta t}{D}\right) \prod_{\langle i,j\rangle} \prod_{K\in\{X,Y,Z\}} R_{ij}^{KK}\left(\frac{2J_{ij}^K \Delta t}{D}\right)\right]^D + O(J_0^2 \Delta t^2/D^2)$

Concerning the non-unitary dynamics due to the openness of the quantum kernel machine, we will consider spin dephasing as the source of decoherence. Within the Lindblad master equation formalism [Eq. (1.60)], this process is described by the jump operators $\hat{A}^j = \hat{\sigma}_z^j$, and we consider a uniform dephasing rate for each site $\gamma_i = \gamma,\ \forall i \in \{1, \cdots, N\}$.

Note that while the considered illustrative task involves three classes, it can be reduced to a set of binary classification problems by changing the labels $y \in \{3, 6, 8\}$ into vector labels of the form $(y_1, y_2, y_3)^T$ with $y_j \in \{-1, 1\}^3$. For example an outcome $(-0.3, -0.2, 0.9)$ would correspond to the digit 8. This "One-vs-Rest" approach is equivalent to training three binary classifiers, one for each class, and takes the highest output among the three classifiers as a prediction. However, for the sake of simplicity, we will use binary classification notations in the following, and consider that the labels belong to $\{-1, 1\}$.

Regarding the measurements of the system observables, we will consider two measurement protocols:

(i) A *full tomography* of the output density matrix. In this case we consider that the measurements are made without delay after the end of the encoding, and the extracted features are exactly the components of the generalized Bloch vector $\boldsymbol{\phi}(\boldsymbol{x})$ by considering a complete set of observables.

(ii) A *time-multiplexing* measurement protocol, where the output is obtained by sequential measurements at different times of a set of local observables.

Full tomography

Any Hermitian operator of the considered spin system can be decomposed on the orthogonal (for the Hilbert-Schmidt inner product) basis of Pauli strings. For a system of N spins, we write this basis $\{\hat{O}_i \mid i = 0, \ldots, P\}$, with $P = 4^N - 1$. The corresponding observables are such that

$$\begin{aligned} &\hat{O}_i = \bigotimes_{k=1}^{N} \hat{\sigma}_{i_k}^k, \quad i_k \in \{0, 1, 2, 3\}; \\ &\mathrm{Tr}\left[\hat{O}_i^\dagger \hat{O}_j\right] = 2^N \delta_{ij}, \quad \forall i, j\,, \end{aligned} \tag{2.39}$$

with $\hat{O}_0 = \hat{\mathbb{1}}$, and thus any observable $\hat{A}$ is decomposed in this basis through the expansion:

$$\hat{A} = \frac{1}{2^N}\left(\mathrm{Tr}\left[\hat{A}\right]\hat{\mathbb{1}} + \sum_{i=1}^{P} \mathrm{Tr}\left[\hat{O}_i \hat{A}\right]\hat{O}_i\right). \tag{2.40}$$

The density matrix associated to the input $\boldsymbol{x}$ can also be decomposed into this basis:

$$\hat{\rho}(\boldsymbol{x}) = \frac{1}{2^N}\left(\hat{\mathbb{1}} + \sum_{i=1}^{P} \langle \hat{O}_i \rangle_{\boldsymbol{x}} \hat{O}_i\right), \tag{2.41}$$

and hence any density matrix is uniquely characterized by its associated generalized Bloch vector. For the full-tomography decoding we take these Bloch vectors as the quantum features, which is equivalent to rescaling the quantum kernel function [Eq. (2.11)] by a constant factor of 2^N.

The encoding method we use leads to embedded states that exhibit entanglement. Fig. 2.4a shows that the average entanglement negativity quickly increases during the encoding, and then eventually decays at a rate depending on γ. In parallel, as we see from Fig. 2.4b, there is a finite von Neumann entropy of the system due to mixed character of the state. In Section III.2, we will show how these processes affect the performances of noisy quantum kernel machines.

Time multiplexing measurements

A simplified and experimentally less demanding decoding is obtained by measuring all the single-site observables (i.e., the three local Pauli spin operators) at different times after the end of the encoding. In the following, we will denote $N_{\rm rep}$ the number of repetitions of these measurements. Hence, for a system of N spins, a total number $3N \times N_{\rm rep}$ of measurements have been performed after $N_{\rm rep}$ repetitions. We use measurements of the on-site observables for each spin, which correspond to the components of the Bloch vectors of the reduced density matrices on each site. We consider corresponding observables in the Heisenberg picture. The new feature vector $\tilde{\boldsymbol{\phi}}(\boldsymbol{x})$ in the time-multiplexing protocol have entries of the form $\langle B_i(t + k\delta t_m)\rangle_{\boldsymbol{x}}$ with $1 \leq i \leq 3N, 1 \leq k \leq N_{\rm rep}$, where δt_m is the time interval between two consecutive measurements. Similar methods were used in previous works to perform an approximate tomography of the system state [188, 189]. Note that the time-multiplexing procedure can only decrease the model expressive power when compared to the full tomography, as information leaks into the system's environment as the system evolves between successive measurement times (see App. III).

III.2 Numerical results

In this section, we discuss the numerical results on the noisy quantum kernel machines obtained by considering the model spin Hamiltonian, dephasing channels, input encoding via driving, decoding protocol through measurement and the classification task detailed in the previous section.

Performances, noise and system size

The main goal is to determine how the performance of the noisy quantum kernel machine scales with the amount of noise and the number of chain sites, i.e. network nodes. To provide a fair comparison, it is necessary to ensure that the same amount of information is fed into the system for all the system sizes. This is achieved by keeping fixed the number M of projections and resolution of the images. As it will be shown in Section III.2, the performance can be greatly enhanced when this information bottleneck is lifted and the amount of encoded information is varied.

The first point to address is the trainability and generalization properties. In Fig. 2.5, we show the dependence of the training and testing errors on the generalization parameter λ. The curves in (a) are obtained assuming a full tomography and ideal measurements. Panel (b) instead presents the same results, but with imperfect measurements (see caption for more details). In panel (a), the training error (dashed lines) drops to zero as $\lambda \to 0^+$; this is a manifestation of over-fitting and indicates that, thanks to the high dimensionality

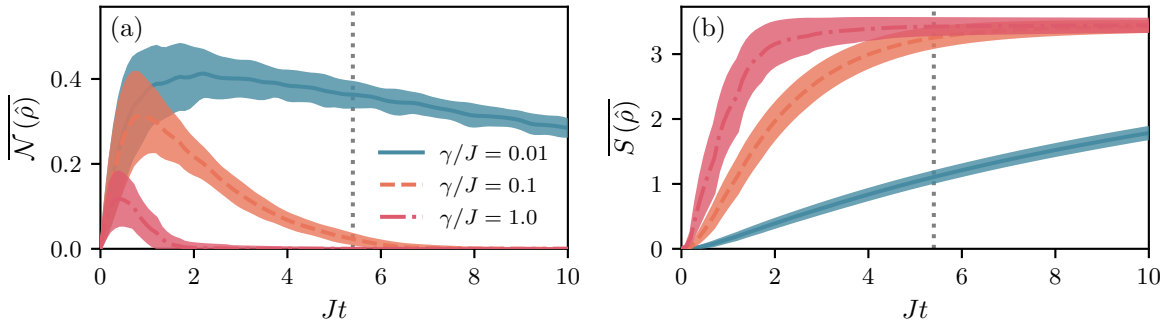


Figure 2.4: Time dynamics of the average entanglement negativity $\overline{\mathcal{N}(\hat{\rho})}$ (a) and the average von Neumann entropy $\overline{S(\hat{\rho})}$(b) in presence of pure dephasing with different values of the corresponding rates γ. At the initial time $t = 0$ the system is in the pure state $\hat{\rho}_0$ defined in the text. Note that the driving sequence finishes at the time $Jt \simeq 5$ indicated by the vertical dotted lines on the figures. The time is expressed in units of $1/J$ where J is the average value of the spin coupling. We define $\overline{\mathcal{N}(\hat{\rho})}$ as the average over all the sites of the negativities associated to the system partitions having the form $\{\{\text{site } i\}, \{\text{site } j \mid j \neq i\}\}$. This quantity is averaged over 20 inputs $\boldsymbol{x} \in \mathcal{X}$, 5 disordered configurations of spin couplings and for a chain of $N = 5$ spins. The filled areas correspond to a one standard deviation confidence interval.

of the quantum feature space, the system is able to completely fit the training data. Instead, the testing error (solid lines and markers) has a minimum value for some optimal value of λ, which depends on the dephasing rate γ (different markers denote different rates). For large enough values of λ the testing and training error curves eventually overlap. For increasing γ the minimum shifts to vanishing values of λ. A remarkable result is that the minimal testing error is very little affected by the dephasing rate. As shown in Fig. 2.5b, the situation changes in the presence of imperfect measurements. Indeed, the minimum of the testing error is obtained for a finite value of λ even for large values of γ. Importantly, the minimum error increases with increasing dephasing noise.

In panels (c) and (d) of Fig. 2.5, we report the dependence of the minimal testing error as a function of the number of spins N for increasing values of the dephasing rate. Again, panel (c) corresponds to ideal measurements, while curves in panel (d) are obtained under imperfect measurements. Panel (c) shows that the testing error diminishes as a function of the number of spins and increases with dephasing rate. Note that also for very small dephasing rate the minimal testing error appears to saturate at large system sizes. This is hardly surprising as the input images have been preprocessed and considerably down-sampled. This deliberate choice aims at making the task harder in order to gauge the expressivity of the machine without overloading the input information. As shown in panel (d) of Fig. 2.5, by considering imperfect measurements the role of dephasing is dramatically amplified.

As we have described in the analytical discussion in Section II, the quantum kernel spectrum allows us to assess the capacity of our model independently from the specific task one wants to achieve. Fig. 2.6a shows the dependence of the quantum kernel's effective

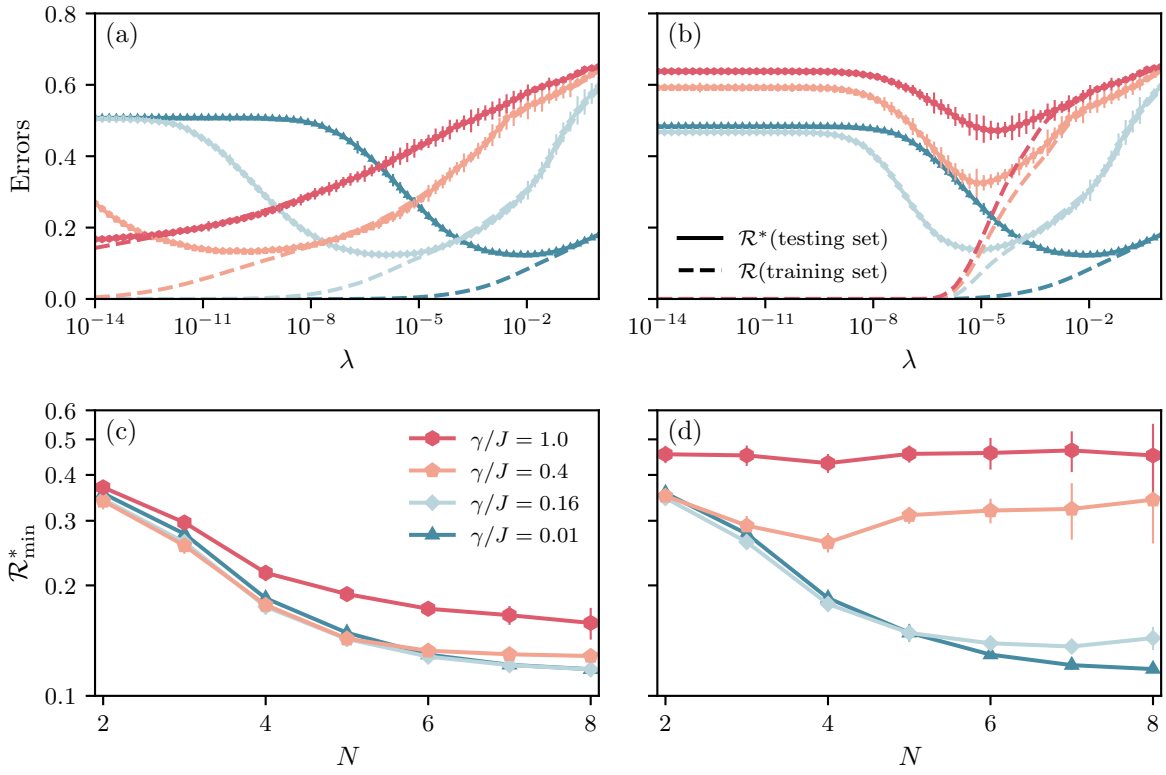


Figure 2.5: (a) Training error $\mathcal{R}$ (dashed lines) and testing error $\mathcal{R}^*$ (solid lines and markers) as a function of the regularization parameter λ for a chain with $N = 7$ spins in the presence of pure dephasing for different values of the corresponding rate γ (different markers in the legend) in units of the average spin coupling J. Measurements are assumed to be ideal . (b) Same as (a) with an extra random Gaussian noise of width $\sigma = 10^{-3}$ added to the observable expectation values to account for imperfect measurements. For each value of λ the corresponding errors are averaged over 15 disordered configurations, and the error bars are bootstrap estimates of the standard deviation for the estimated mean values. We use 10 bootstrap sets, each consisting of 15 samples randomly drawn with replacement from the original set of 15 disorder realizations. (c) Minimal testing error as a function of the number of spins N for different values of the dephasing rate γ. For each disorder configuration, the regularization parameter λ is chosen to minimize the testing error and the resulting minimum is averaged over the disorder. The error bars are derived using the same bootstrap procedure. Number of disorder configurations: 50 for $N = 2$ to $N = 5$ spins, 25 for $N = 6$, 15 for $N = 7$ and 5 for $N = 8$. (d) Same as (c) with an extra random Gaussian noise of width $\sigma = 10^{-3}$ added to the observable expectation values to account for imperfect measurements.

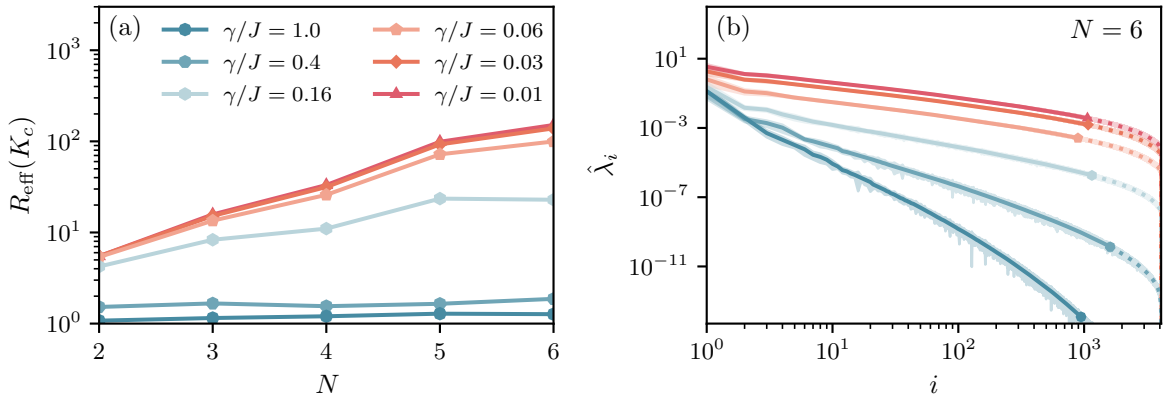


Figure 2.6: (a) Kernel effective rank for the full tomography decoding as a function of the number of spins N and for different values of the dephasing rate γ. We have used the empirical representation of the kernel matrix on the training set. The results are averaged over over the same numbers of disorder realizations as for Fig. 2.5. (b) Kernel empirical spectrum for the full tomography decoding, $N = 6$ spins and for different values of the dephasing rate γ. The markers correspond to the optimal generalization parameter λ. The curves have been obtained via the kernel empirical representation on the training set. Results are averaged over 25 disorder configurations. The filled area corresponds to twice the estimated standard error on the averaged value, using the same bootstrap method as for Fig. 2.5.

rank $R_{\text{eff}}(\mathbf{K}_c)$ on the system size and noise strength. For vanishing values of the dephasing rate γ, we see that this figure of merit first increases exponentially with the number of spins before saturating. For increasing γ, the effective quantum kernel rank decreases approaching one in the limit of very large γ.

The same behavior is observed in the empirical spectrum in Fig. 2.6b as the noise rate is varied. For increasing values of γ, we observe a faster decrease of the empirical kernel eigenvalues as a function of the eigenvalue number. For comparison, we have indicated with markers the largest eigenvalue below the optimal generalization parameter. This gives a rough estimate of the number of kernel eigenfunctions required to correctly approximate the target function. Note that in the context of imperfect measurements the generalization parameter is bounded from below, and hence some of the kernel eigenfunctions becomes out-of-reach. This shows a clear link between the kernel eigenvalues and the expressivity of the machine.

The results discussed relied on full tomography. As we have explained in Section III, it is possible to design a simplified and less expensive measurement protocol based on a time-multiplexing procedure where a set of local spin observables are measured at N_{rep} different times. The results obtained with such an approach are summarized in Fig. 2.7. As appears from Fig. 2.7a, by increasing the number of repetitions N_{rep} the error diminishes. For small enough values of dephasing γ, the error converges to the value in the ideal case of full-tomography. For increasing γ, however, the saturating value departs from the ideal one given by full tomography, showing that the time-multiplexing expressivity

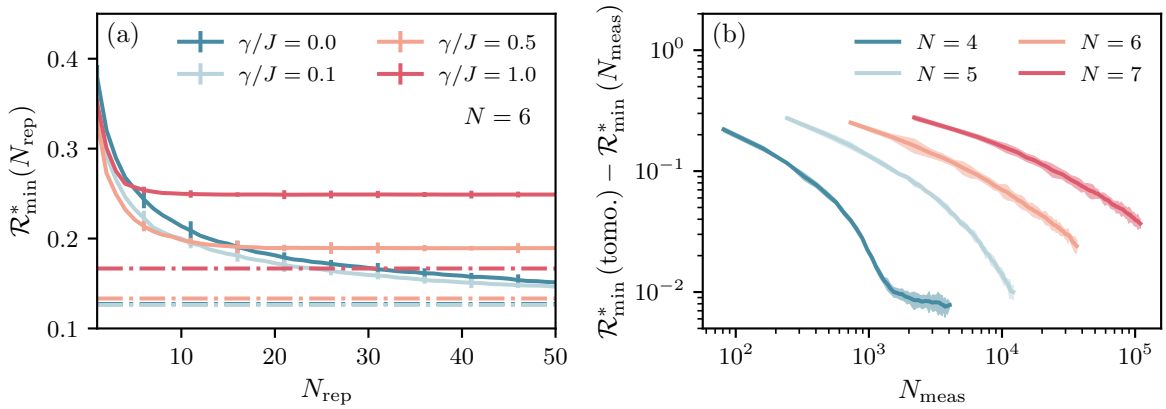


Figure 2.7: (a) Minimal testing error as a function of the number of repetitions $N_{\rm rep}$ of all the local spin measurements for $N = 6$ spins and different values of the dephasing noise γ. The regularization parameter value is chosen to minimize the testing error. The horizontal dashed lines represent the errors obtained using a full tomography decoding for the corresponding dephasing rates. The results are averaged over 50 disorder realizations. (b) Difference between the testing error of the time-multiplexing decoding and the one for full tomography as a function of the number $N_{\rm meas} = 3NN_{\rm rep}$ of local measurements performed for $N_{\rm rep} = 50$, dephasing rate $\gamma/J = 0.01$ and different values of the number N of spins. For each realization of the disorder, the regularization parameter value is chosen to minimize the testing error. The results are averaged over 25 disorder realizations.

deteriorates more than that of the full tomography for larger noise. This trend is further elucidated in Fig. 2.7b, where the difference between the time-multiplexing error and the full-tomography error is reported as a function of the total number of measured observables. By increasing the number of spins and hence the dimension of the Hilbert space for a given dephasing rate, the required number of repetitions increases.

Optimizing the encoding

In this section we investigate an alternative encoding scheme for which the amount of information fed to the system scales with the system size. The embedding studied in the previous sections involved a set of $M = N_{\rm pulse}$ random-projection features derived from the down-sampled images. Here we derive a number $M = N \times N_{\rm pulse}$ of such features and split them in N sequences of $N_{\rm pulse}$ features, which we use to drive the N sites. In particular, the driving sequences sent to different sites are unique, while for the previous encoding those sequences were proportional to each other. The new encoding procedure is presented in Fig. 2.8 in the form of its equivalent circuit. In Fig. 2.9 we report the evolution of the performances given by this new encoding as a function of the number of driving pulses. As the number of pulses rises the corresponding number of encoded features $M = N \times N_{\rm pulse}$ increases and so does the amount of encoded information. The corresponding maximal testing accuracy reaches an optimum of 94.5% for $N_{\rm pulse} = 3$. For large number of pulses $N_{\rm pulse}$ the performances drop. This effect is due to the fact that

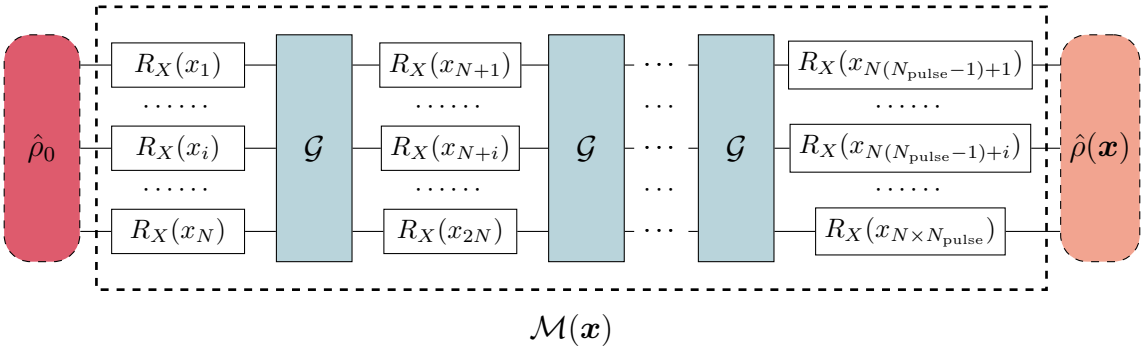


Figure 2.8: Equivalent circuit of the encoding with the information bottleneck removed. Instead of injecting a single random projection feature x_i per time step as represented in Fig. 2.3, where a total number of $M = N_{\text{pulse}}$ random projection features are fed into the kernel machine, here we inject N random projection features in each time step, with a total number of $M = N \times N_{\text{pulse}}$ random projection features injected by the end of the encoding sequence.

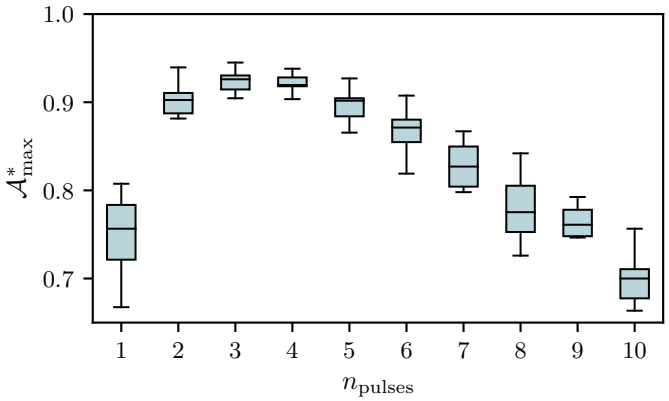


Figure 2.9: Optimal testing accuracy as a function of the number of driving pulses, for the encoding presented on Fig. 2.8, $N = 6$ spins and $\gamma/J = 0.01$. The regularization parameter is chosen as to maximize the testing accuracy. Note that for this encoding the number of features M yielded by the preprocessing is $M = N \times N_{\text{pulse}}$. The results are shown for 10 realizations of the disorder on both the preprocessing and the system parameters. The maximal accuracy obtained is 94.5% for $N_{\text{pulse}} = 3$. The boxes extend from the first to the third quartile of the distributions, the middle line indicates the median and the wiskers indicates the extreme values of the distributions.

for such parameters the transformations yielded by the encoding are poorly adapted to the task at hand, i.e. the kernel and the target function become less "aligned". Note that the performance is very sensitive to both the encoding method and the physical system parameters. While controlling the physical parameters might be hard, it appears that a careful design of the encoding procedure can significantly boost the performances. This makes the research of tailored encoding procedures a promising avenue of research.

IV Conclusion

In this chapter, we have presented a quantum machine-learning model based on the quantum-kernel paradigm. Within the formalism of kernel theory, we have characterized the expressivity and generalization capacity of this model. We have linked the relevant figure of merits to the spectrum of the associated centered quantum kernel. In particular, we presented an upper bound on the generalization error involving the average purity of quantum states representing the data to classify. This upper-bound shows that dissipation and decoherence act as a regularization for the quantum kernel machines. By considering an illustrating example of a driven-dissipative spin chain as the noisy quantum kernel machine, we have shown how the expressivity and generalization capacity are controlled by both the dephasing rate and by experimental uncertainties on the measurements. Moreover, we have shown how the performances of the noisy quantum kernel machines are modified when the full-tomography measurement protocol is replaced by a time-multiplexing procedure requiring only local observables, and how the openness of the system mitigates the efficiency of this protocol. We observed a qualitative improvement in the processing performance of our model when going from a scenario where the system is fed a constant amount of information to one where the inputs are encoded at a finite information rate that scales extensively with the system size. How to design tailored encoding strategies able to harness the full power of quantum kernel machines remains an open question. In particular, investigating encoding schemes that would allow to inject information at a rate scaling exponentially in the system size seems promising. The concepts presented here and the unavoidable role of the decoherence in any realistic physical system are relevant for a wide range of quantum machine-learning models, ranging from quantum extreme-learning machines to quantum neural networks.

3 Efficient estimation of trainability for VQC

Inspired by the success of machine-learning methods, variational quantum algorithms [78, 95, 98] have emerged as a promising way to harness the power of quantum computing in various domains ranging from quantum chemistry [190–192] to combinatorial optimization problems [193–195]. These algorithms use the output of parameterized quantum circuits as variational ansätze, whose parameters are classically optimized through gradient-based methods. In the emerging field of quantum machine learning, a large amount of work has been devoted to finding quantum analogs to neural-network models [196, 197], and variational quantum circuits have appeared as natural candidates for such a generalization [198]. Thus, these models are sometimes called quantum neural networks in this context [199].

Variational quantum circuits can suffer from trainability issues caused by the existence of barren plateaus [119], a limitation that has been extensively studied in the recent literature [120–123, 200–221]. It is characterized by an exponential vanishing of the cost function's gradient with the system size that makes training variational quantum circuits impossible for a large number of qubits. Although many strategies have been proposed to avoid barren-plateaus [201–215], tackling this fundamental issue remains an important theoretical challenge.

In this chapter, we propose an alternative approach to the problem by providing an efficient method to estimate the average gradients and their variance for a wide class of variational quantum algorithms. The associated results are gathered in our work [γ]. This chapter is structured as follows. Sec. I introduces the theoretical framework of variational quantum algorithms, and we provide a brief review of the existing results related to the barren plateaus. In Sec. II we give an overview of the efficient simulation scheme presented in this chapter and discuss its conditions of applicability. In addition, we show some numerical experiments to illustrate our method on examples of random circuits and faithfully reproduce the exponential suppression of the variance first found in Ref. [119] with polynomial resources. The technical results are then presented in Secs. III and IV. We study the general quantum channel associated to a random single-qubit rotation in Sec. III. There, we prove that under some simple conditions the first and second moments can be expressed as mixed-unitary channels [141] composed of Clifford gates [142]. Then, in Sec. IV, we demonstrate that upon some additional general assumptions for the random angles distribution, this decomposition allows to exactly map randomly initialized circuits composed of Clifford gates and parameterized rotations to an ensemble of Clifford circuits. Moreover, we prove that the obtained ensemble can be efficiently sampled to compute

quantities of interest, such as the variance of the gradient or the average of the cost function over the initial random parameters. Making use of the celebrated Gottesman-Knill theorem [42, 45], we analytically prove the efficiency of our method that can be implemented on a classical computer with a complexity scaling polynomially in both the number of variational parameters and the system size.

I Theoretical framework

In this section we introduce the framework associated with variational quantum algorithms along with a prompt discussion of the barren plateaus phenomenon. We also define the notions of first- and second-order averaged quantities that will be used in this chapter.

I.1 Variational problem

In variational quantum algorithms, a parameterized unitary transformation $\hat{U}(\boldsymbol{\theta})$ acting on n qubits is used as a variational ansatz to achieve a task expressed as the minimization of a cost function

$$C(\boldsymbol{\theta}) = \mathrm{Tr}\left[\hat{U}(\boldsymbol{\theta})\hat{\rho}\hat{U}^{\dagger}(\boldsymbol{\theta})\hat{O}\right] \tag{3.1}$$

for some observable $\hat{O}$ and some initial n-qubit state $\hat{\rho}$. This formulation is general and encompasses typical tasks, such as the preparation of a target state $|\psi\rangle$ (setting $\hat{O} = -|\psi\rangle\langle\psi|$) or ground state search for some Hamiltonian $\hat{\mathcal{H}}$ (setting $\hat{O} = \hat{\mathcal{H}}$).

The considered parameterized unitaries are typically composed of a succession of parameterized gates and fixed layers. Here we consider a generic ansatz of the form

$$\hat{U}(\boldsymbol{\theta}) = \prod_{i=1}^{M} \hat{U}_i(\theta_i)\hat{W}_i\,, \qquad \hat{U}_i(\theta_i) = e^{-\mathrm{i}\frac{\theta_i}{2}\hat{P}_i}\,, \tag{3.2}$$

where each unitary $\hat{U}_i(\theta_i)$ is a single qubit rotation associated to a given Pauli operator $\hat{P}_i \in \{\hat{X}, \hat{Y}, \hat{Z}\}$, while the $\hat{W}_k$ are fixed layers composed of a sequence of unparameterized gates that can act on multiple qubits. Upon absorbing Clifford gates in the fixed layers, we can transform all the parameterized gates into Z rotations. In fact, let us denote $\hat{H}$ the Hadamard gate and $\hat{S}$ the phase gate, whose matrices in the computational basis $\{|0\rangle, |1\rangle\}$ reads

$$\hat{H} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\,, \quad \hat{S} = \begin{pmatrix} 1 & 0 \\ 0 & \mathrm{i} \end{pmatrix}\,. \tag{3.3}$$

Both of these gates belong to the Clifford group, and we have that

$$\begin{aligned} \hat{X} &= \hat{H}\hat{Z}\hat{H}^{\dagger}\,, \\ \hat{Y} &= (\hat{S}\hat{H})\hat{Z}(\hat{S}\hat{H})^{\dagger}\,, \end{aligned} \tag{3.4}$$

so that we can write

$$\begin{aligned} e^{-\mathrm{i}\frac{\theta_i}{2}\hat{X}} &= \hat{H}e^{-\mathrm{i}\frac{\theta_i}{2}\hat{Z}}\hat{H}^{\dagger}\,, \\ e^{-\mathrm{i}\frac{\theta_i}{2}\hat{Y}} &= (\hat{S}\hat{H})e^{-\mathrm{i}\frac{\theta_i}{2}\hat{Z}}(\hat{S}\hat{H})^{\dagger}\,. \end{aligned} \tag{3.5}$$

Hence, if $\hat{U}_i(\theta_i)$ is a X-rotations, we can replace $\hat{W}_i$ and $\hat{W}_{i+1}$ respectively by $\hat{H}\hat{W}_i$ and $\hat{W}_{i+1}\hat{H}$ to get another ansatz with the same form as the original one and with only Y and Z rotations. Proceeding likewise for Y-rotations, we obtain an equivalent ansatz with only Z rotations [1]. As result, there is no loss of generality to consider ansatz based only on Z-rotations, and in the following we will focus on this class of circuits.

The unitary transformation $\hat{U}(\boldsymbol{\theta})$ depends on M continuous parameters gathered in the vector $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_{M-1})$. These rotation parameters can be optimized using classical gradient-descent techniques. The gradient of the cost function with respect to the k-th parameter can be conveniently estimated using the parameter-shift rule [171, 222]:

$$\partial_k C(\boldsymbol{\theta}) = \frac{1}{2}\left(C(\boldsymbol{\theta} + \frac{\pi}{2}\boldsymbol{e}_k) - C(\boldsymbol{\theta} - \frac{\pi}{2}\boldsymbol{e}_k)\right), \tag{3.6}$$

where $\boldsymbol{e}_k$ is the canonical vector along the component k. It is worth noticing that the $\pm\pi/2$ shifts in the parameter θ_k can be factored out and seen as an extra Clifford gate added to the fixed layer $\hat{W}_k$. In fact, remarking that

$$\hat{S} = e^{\mathrm{i}\frac{\pi}{4}} e^{-\mathrm{i}\frac{\pi}{2}\hat{Z}}, \tag{3.7}$$

and assuming $\hat{P}_k = \hat{Z}$, we have:

$$\hat{U}_k(\theta_k + \pi/2)\hat{W}_k = e^{-\mathrm{i}\frac{\theta_k}{2}\hat{Z}} e^{-\mathrm{i}\frac{\pi}{2}\hat{Z}}\hat{W}_k = e^{-\mathrm{i}\frac{\pi}{4}}\hat{U}_k(\theta_k)\hat{S}\hat{W}_k. \tag{3.8}$$

We define

$$\hat{W}_{k,+} = e^{-\mathrm{i}\frac{\pi}{4}}\hat{S}\hat{W}_k, \qquad \hat{W}_{k,-} = e^{+\mathrm{i}\frac{\pi}{4}}\hat{S}^\dagger\hat{W}_k, \tag{3.9}$$

such that we can write

$$\hat{U}_k(\theta_k \pm \pi/2)\hat{W}_k = \hat{U}_k(\theta_k)\hat{W}_{k,\pm}. \tag{3.10}$$

We denote

$$\hat{U}_\pm(\boldsymbol{\theta}) = \hat{U}\left(\boldsymbol{\theta} \pm \frac{\pi}{2}\boldsymbol{e}_k\right) \tag{3.11}$$

the shifted unitaries appearing in the parameter-shift-rule. From what preceeds we have

$$\hat{U}_\pm(\boldsymbol{\theta}) = \prod_{i=1}^{M} \hat{U}_i(\theta_i)\hat{V}_{i,\pm}, \quad \text{with} \quad \hat{V}_{i,\pm} = \begin{cases} \hat{W}_{k,\pm} & \text{if } i = k \\ \hat{W}_i & \text{otherwise} \end{cases}. \tag{3.12}$$

I.2 Unitary ensembles and t-fold channels

To start the optimization process the rotation angles are randomly initialized according to some probability distribution $p(\boldsymbol{\theta})$. The initialized circuit can then be represented by a unitary ensemble $\mathbb{U} = \{\hat{U}, \mathbb{P}(\hat{U})\}$, where $\mathbb{P}$ is a probability measure on $\mathbb{U}$. One is often interested in computing averages of quantities that are polynomial of a given order t in the entries of $\hat{U}$. Such quantities can be completely determined by the knowledge of the t-fold channel [223]

$$\Phi_{\mathbb{U}}^{(t)}(\hat{\rho}) = \int_{\mathbb{U}} \hat{U}^{\otimes t}\hat{\rho}\hat{U}^{\dagger\otimes t}\mathrm{d}\mathbb{P}(\hat{U}), \tag{3.13}$$

[1] Note that in the case of the last layer one of the extra gates must be absorbed in the cost function observable to get the same ansatz structure.

where $\hat{\rho}$ is an initial state of t copies of the original n-qubit system.

In this chapter, we will be primarily interested in the unitary ensemble arising from the random initialization of the ansatz parameters, which we write

$$\mathbb{U} = \left\{ \hat{U}(\boldsymbol{\theta}),\ \mathbb{R}^M \ni \boldsymbol{\theta} \sim p(\boldsymbol{\theta}) \right\}. \tag{3.14}$$

We denote $\Phi_{\boldsymbol{\theta}}^{(t)}$ the t-fold channel associated with this ensemble. In the following, we will give examples of quantities involving the 1- and 2-fold channels, and we introduce the related *first- and second-order quantities*.

More generally, one can characterize the expressivity of a given ansatz by comparing its t-fold channels to ones obtained for a Haar (uniform) distribution over the whole unitary group [123, 183, 224]. Unitary ensemble whose t-fold channel matches the t-fold channels for the Haar measure, the so-called t-designs, have played a crucial role in the original discovery of the barren plateaus phenomenon [119]. Moreover in multiple cases random quantum circuits are approximate t-designs [225–227].

I.3 Barren Plateaus

As mentioned earlier, variational quantum algorithms can suffer from trainability issues caused by an exponential vanishing of the cost function's gradient with the system size. This issue, known as the Barren Plateaus phenomenon, can originate from various and fundamentally different phenomena. Its emergence was first shown in Ref. [119] for 2-designs (random unitary transformation matching the Haar distribution up to the second moment). Recent works linked barren plateaus to the expressibility of the ansatz [123] as well as noise [120] and entanglement. In particular, the authors of Ref. [122] showed that for architectures that can be split into a hidden and a visible subsystem, such as quantum Boltzmann machines or feed-forward quantum neural networks, an excess of entanglement between the two subsystems would result in a highly mixed state for the visible subsystem. This can lead to a flat landscape for the cost function. The effect of the structure of the cost function on the appearance of barren plateaus was also investigated in other works [121, 200], and it was shown that global cost functions are more prone to exhibit barren plateaus. Note that shallow models such as quantum kernel machines [101, 108, 179] and reservoir computing models [167, 169, 228, 229], while often easier to train than variational quantum algorithms, might also suffer from trainability issues of a similar nature [117].

For a unitary ensemble $\mathbb{U}$ that describes parameterized ansätze $\hat{U}(\boldsymbol{\theta})$ with random continuous parameters $\boldsymbol{\theta}$ and a possibly random architecture, a cost function $C(\hat{U}(\boldsymbol{\theta}))$ is said to exhibit a barren plateau if the probability of obtaining a gradient that deviates from zero by some $\epsilon > 0$ vanishes exponentially with the system size n. More precisely, $\mathbb{P}_{\mathbb{U}}(|\partial_k C| > \epsilon) \leq \mathcal{O}(\exp(-\alpha n))$ for some $\alpha > 0$.

In many cases, the average value of the gradient vanishes exactly, for instance when the rotation parameters are initialized uniformly in $[-\pi, \pi]$. However this does not imply a vanishing of the gradient amplitude on average, and thus does not tell much about the trainability of the model. In this unbiased case, the variance is a relevant quantity. Due to the Chebyshev inequality, one has $\mathbb{P}_{\mathbb{U}}(|\partial_k C| > \epsilon) \leq \mathrm{Var}\left[\partial_k C\right] / \epsilon^2$, so that a vanishing

variance implies the existence of a barren plateau. On the other hand a non-vanishing variance guarantees large fluctuations and thus a good trainability, independently of the gradient bias. Higher-order moments may also help diagnose the trainability of variational quantum algorithms [230].

Numerous investigations have proposed strategies to address the barren-plateau issue. In the context of entanglement-induced barren plateaus, most strategies rely on limiting the amount of entanglement [201–206]. Other methods make use of tailored distributions of the initial circuits parameters and carefully designed circuits architectures [207–213]. Yet, only a handful of configurations offer trainability guarantees and robustness against barren plateaus [214, 215]. It is worth noticing that some authors argue that barren plateaus do not always hinder an efficient training of the ansatz [231]. Also, other challenges related to the optimization of variational quantum algorithms remains [232, 233].

I.4 First- and second-order quantities

First-order quantities

Let us denote

$$\mathcal{U}_i(\theta_i)(\hat{\rho}) = \hat{U}_i(\theta_i)\hat{\rho}\hat{U}_i^\dagger(\theta_i)\,, \quad \mathcal{W}_i(\hat{\rho}) = \hat{W}_i\hat{\rho}\hat{W}_i^\dagger\,, \tag{3.15}$$

the unitary channels associated to the different layers of the circuit. With these notations, the whole circuit unitary transformation reads

$$\begin{aligned}\mathcal{U}(\boldsymbol{\theta})(\hat{\rho}) &= \mathcal{U}_M(\theta_M)\circ\cdots\circ\mathcal{W}_1(\hat{\rho}) \\ &= \bigcirc_{i=1}^{M}\left(\mathcal{U}_i(\theta_i)\circ\mathcal{W}_i\right)(\hat{\rho}).\end{aligned} \tag{3.16}$$

The cost function can be written

$$C(\boldsymbol{\theta}) = \mathrm{Tr}\left[\mathcal{U}(\boldsymbol{\theta})(\hat{\rho})\hat{O}\right] \tag{3.17}$$

and its expectation with respect to $\boldsymbol{\theta}$ is given by

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}\left[C(\boldsymbol{\theta})\right] &= \mathbb{E}_{\boldsymbol{\theta}}\left[\mathrm{Tr}\left[\mathcal{U}(\boldsymbol{\theta})(\hat{\rho})\hat{O}\right]\right] \\ &= \mathrm{Tr}\left[\mathbb{E}_{\boldsymbol{\theta}}\left[\mathcal{U}(\boldsymbol{\theta})(\hat{\rho})\right]\hat{O}\right] \\ &= \mathrm{Tr}\left[\mathbb{E}_{\boldsymbol{\theta}}\left[\hat{U}(\boldsymbol{\theta})\hat{\rho}\hat{U}^\dagger(\boldsymbol{\theta})\right]\hat{O}\right] \\ &= \mathrm{Tr}\left[\int_{\mathbb{R}^M}\hat{U}(\boldsymbol{\theta})\hat{\rho}\hat{U}^\dagger(\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\hat{O}\right] \\ &= \mathrm{Tr}\left[\Phi_{\boldsymbol{\theta}}^{(1)}(\hat{\rho})\hat{O}\right].\end{aligned} \tag{3.18}$$

Here, we used both the linearity of the expectation and the definition of the 1-fold channel from Eq. (3.13). The cost function expectation can thus be obtained from the knowledge of the complete 1-fold channel $\Phi_{\boldsymbol{\theta}}^{(1)}$. Assuming that the angles $\{\theta_i\}$ are independent from each other, the expectation against $\boldsymbol{\theta}$ can be factored in expectations against the θ_i's, which allows to write:

$$\Phi_{\boldsymbol{\theta}}^{(1)}(\hat{\rho}) = \mathbb{E}_{\boldsymbol{\theta}}\left[\mathcal{U}(\boldsymbol{\theta})(\hat{\rho})\right] = \bigcirc_{i=1}^{M}\left(\mathbb{E}_{\theta_i}\left[\mathcal{U}_i(\theta_i)\right]\circ\mathcal{W}_i\right)(\hat{\rho}). \tag{3.19}$$

As explained in the foregoing discussion, we can consider without loss of generality all the rotations to be Z-rotations. Then the channels $\mathbb{E}_{\theta_i}\left[\mathcal{U}_i(\theta_i)\right]$ are exactly 1-fold channels associated to a Z-rotation acting on a single qubit. In the following, we will refer to quantities that can be obtained from the knowledge of the 1-fold channels associated to each rotations of the ansatz as *first-order quantities*. Hence the average cost function $\mathbb{E}_{\boldsymbol{\theta}}\left[C(\boldsymbol{\theta})\right]$ is a first-order quantity.

Another example of an interesting first-order quantity is the average of the gradient. Using Eq. (3.6) and the linearity of the expectation, we have:

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\partial_k C(\boldsymbol{\theta})\right] = \frac{1}{2}\mathbb{E}_{\boldsymbol{\theta}}\left[C\left(\boldsymbol{\theta}+\frac{\pi}{2}\boldsymbol{e}_k\right)\right] - \frac{1}{2}\mathbb{E}_{\boldsymbol{\theta}}\left[C\left(\boldsymbol{\theta}-\frac{\pi}{2}\boldsymbol{e}_k\right)\right]. \tag{3.20}$$

Denoting

$$\mathcal{V}_{i,\pm}(\hat{\rho}) = \hat{V}_{i,\pm}\hat{\rho}\hat{V}_{i,\pm}^{\dagger} \tag{3.21}$$

the channels associated with the unitary transformations of Eq. (3.12), we can write

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}\left[\hat{U}_{\pm}(\boldsymbol{\theta})\hat{\rho}\hat{U}_{\pm}^{\dagger}(\boldsymbol{\theta})\right] &= \mathbb{E}_{\boldsymbol{\theta}}\left[\bigcirc_{i=1}^{M}\left(\mathcal{U}_i(\theta_i)\circ\mathcal{V}_{i,\pm}\right)(\hat{\rho})\right] \\ &= \bigcirc_{i=1}^{M}\left(\mathbb{E}_{\theta_i}\left[\mathcal{U}_i(\theta_i)\right]\circ\mathcal{V}_{i,\pm}\right)(\hat{\rho}).\end{aligned} \tag{3.22}$$

The average gradient is therefore a first-order quantity, namely depending on 1-fold channels only.

Second-order quantities

We now turn our attention to the mean value of the squared cost function. This is given by

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}\left[C(\boldsymbol{\theta})^2\right] &= \mathbb{E}_{\boldsymbol{\theta}}\left[\mathrm{Tr}\left[\mathcal{U}(\boldsymbol{\theta})(\hat{\rho})\hat{O}\right]^2\right] \\ &= \mathbb{E}_{\boldsymbol{\theta}}\left[\mathrm{Tr}\left[\left(\mathcal{U}(\boldsymbol{\theta})(\hat{\rho})\hat{O}\right)^{\otimes 2}\right]\right] \\ &= \mathbb{E}_{\boldsymbol{\theta}}\left[\mathrm{Tr}\left[\mathcal{U}^{(2)}(\boldsymbol{\theta})(\hat{\rho}^{\otimes 2})\hat{O}^{\otimes 2}\right]\right].\end{aligned} \tag{3.23}$$

For every state $\hat{\rho}$ of a system of $2n$ qubits (i.e., a doubled version of the original system where the copy is not connected by gates to the original circuit), we define

$$\mathcal{U}^{(2)}(\boldsymbol{\theta})(\hat{\rho}) = \hat{U}^{\otimes 2}(\boldsymbol{\theta})\hat{\rho}\hat{U}^{\dagger\otimes 2}(\boldsymbol{\theta}). \tag{3.24}$$

Likewise we can define the doubled version of the circuit layers as

$$\begin{aligned}\mathcal{U}_i^{(2)}(\theta_i)(\hat{\rho}) &= \hat{U}_i^{\otimes 2}(\theta_i)\hat{\rho}\hat{U}_i^{\otimes 2}(\theta_i), \\ \mathcal{W}_i^{(2)}(\hat{\rho}) &= \hat{W}_i^{\otimes 2}\hat{\rho}\hat{W}_i^{\otimes 2},\end{aligned} \tag{3.25}$$

giving

$$\mathcal{U}^{(2)}(\boldsymbol{\theta})(\hat{\rho}) = \bigcirc_{i=1}^{M}\left(\mathcal{U}_i^{(2)}(\theta_i)\circ\mathcal{W}_i^{(2)}\right)(\hat{\rho}). \tag{3.26}$$

Thus for independent rotations we have

$$\begin{aligned}\Phi_{\boldsymbol{\theta}}^{(2)}(\hat{\rho}) &= \mathbb{E}_{\boldsymbol{\theta}}\left[\mathcal{U}^{(2)}(\boldsymbol{\theta})(\hat{\rho})\right] \\ &= \bigcirc_{i=1}^{M}(\mathbb{E}_{\theta_i}\left[\mathcal{U}_i^{(2)}(\theta_i)\right] \circ \mathcal{W}_i^{(2)})(\hat{\rho})\,.\end{aligned} \tag{3.27}$$

As for first-order quantities, we refer to quantities that can be obtained from the knowledge of the average 2-fold channels of the rotations layers $\mathbb{E}_{\theta_i}\left[\mathcal{U}_i^{(2)}(\theta_i)\right]$ as *second-order quantities*.

The average of the squared cost function is thus a second-order quantity, and as for the first order case, we can show that the squared gradient is also a second-order quantity. In fact, by making use of the parameter-shift rule, we see that to obtain the average of the squared gradient we have to compute the following four terms

$$\mathbb{E}_{\boldsymbol{\theta}}\left[C(\boldsymbol{\theta}+a_1\boldsymbol{e}_k)C(\boldsymbol{\theta}+a_2\boldsymbol{e}_k)\right]\,, \qquad a_1,a_2 \in \left\{\frac{\pi}{2},-\frac{\pi}{2}\right\}\,. \tag{3.28}$$

As before, it suffices to replace the $\mathcal{W}_i^{(2)}$ in Eq. (3.27) with

$$\mathcal{V}_{i,a_1,a_2}^{(2)}(\hat{\rho}) = \left(\hat{V}_{i,a_1}\otimes\hat{V}_{i,a_2}\right)\hat{\rho}(\hat{V}_{i,a_1}^{\dagger}\otimes\hat{V}_{i,a_2}^{\dagger}). \tag{3.29}$$

As a result, the gradient variance can be computed as

$$\mathrm{Var}_{\boldsymbol{\theta}}\left[\partial_k C(\boldsymbol{\theta})\right] = \mathbb{E}_{\boldsymbol{\theta}}\left[\partial_k C(\boldsymbol{\theta})^2\right] - \mathbb{E}_{\boldsymbol{\theta}}\left[\partial_k C(\boldsymbol{\theta})\right]^2\,, \tag{3.30}$$

which is the sum of a first and a second-order quantity.

II Overview of the results and discussion

The main finding of this theoretical work is that under some rather general assumptions on the distribution of the rotation parameter θ, it is possible to map the 1-fold and 2-fold channels of a random rotation $\hat{R}_Z(\theta)$ to a finite unitary ensemble of Clifford gates. Moreover, we prove that such mapping allows us to estimate quantities of interest such as the gradient variance using only Clifford circuits. Finally, we illustrate our rigorous proofs through numerical experiments. The detailed mathematical proofs are presented in Secs. III and IV.

II.1 Exact mapping and efficient sampling

As mentioned earlier, we will focus on the class of variational quantum circuits composed of fixed Clifford gates alternated with single qubit parameterized rotations along the X, Y or Z directions, such as the one depicted in Fig. 3.1. As explained in Sec. I.1, we will restrict our study to rotations along Z, as we can obtain the cases of rotations along Y and X by adding extra Clifford gates to the different fixed layers of the considered ansatz. Let us consider a rotation along the Z-axis with a distribution that is symmetric about

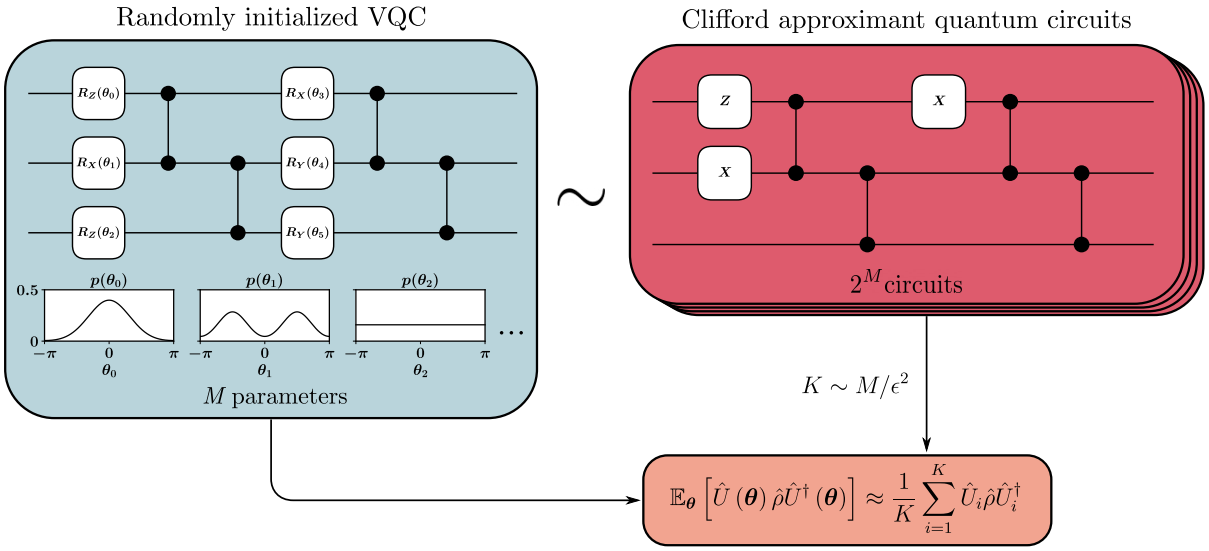


Figure 3.1: A schematic representation of the mapping from a parameterized quantum circuit with random parameters to Clifford approximant circuits for first-order quantities (quantities that requires only the knowledge of the rotation 1-fold channels to be computed, see App. I.4). For a circuit with M parameters, a sample size of the order of M/ϵ^2 is enough to get an approximation of the average on the initial circuit with a precision ϵ on the observables mean values (see App. IV).

the $\theta = 0$ angle [2]. We will show in Sec. III that the 1-fold channel corresponding to a first order average can be written as a convex sum of the unitary channels associated to the identity and the Pauli Z gates, as schematically represented on the upper part of Fig. 3.2. Note that this result has been derived and used in [234] in the case of a uniform probability distribution for θ in order to analyze a variational ansatz through the lens of ZX-calculus.

Estimation of first order quantities

To compute the 1-fold channel for the randomly initialized ansatz of the form given in Eq. (3.2) with independent rotation parameters, one can simply compose the 1-fold channels associated to each rotation, interwined with the unitary channels associated to the fixed gates W_k. We find that the 1-fold channel of the ansatz is a convex sum of 2^M Clifford unitary channels, where M is the number of rotations. One can view this convex sum as an average over a finite ensemble of Clifford approximant circuits. Examples of such circuits are provided in App. H for a simple architecture similar to the one in Fig. 3.1. Although the number of Clifford approximant circuits in this ensemble is exponential in the number of parameters, we will show in Sec. IV.2 that a number of samples polynomial in M/ϵ^2 is sufficient to approximate the average of an observable expectation value (or more generally of any first-order quantity) to any desired precision ϵ. This result relies

[2] This encompasses distributions that are symmetric about the angle $k\pi/2$ for $k \in \mathbb{Z}$. In this case the bias can be factored out in the form of an extra fixed Clifford gate.

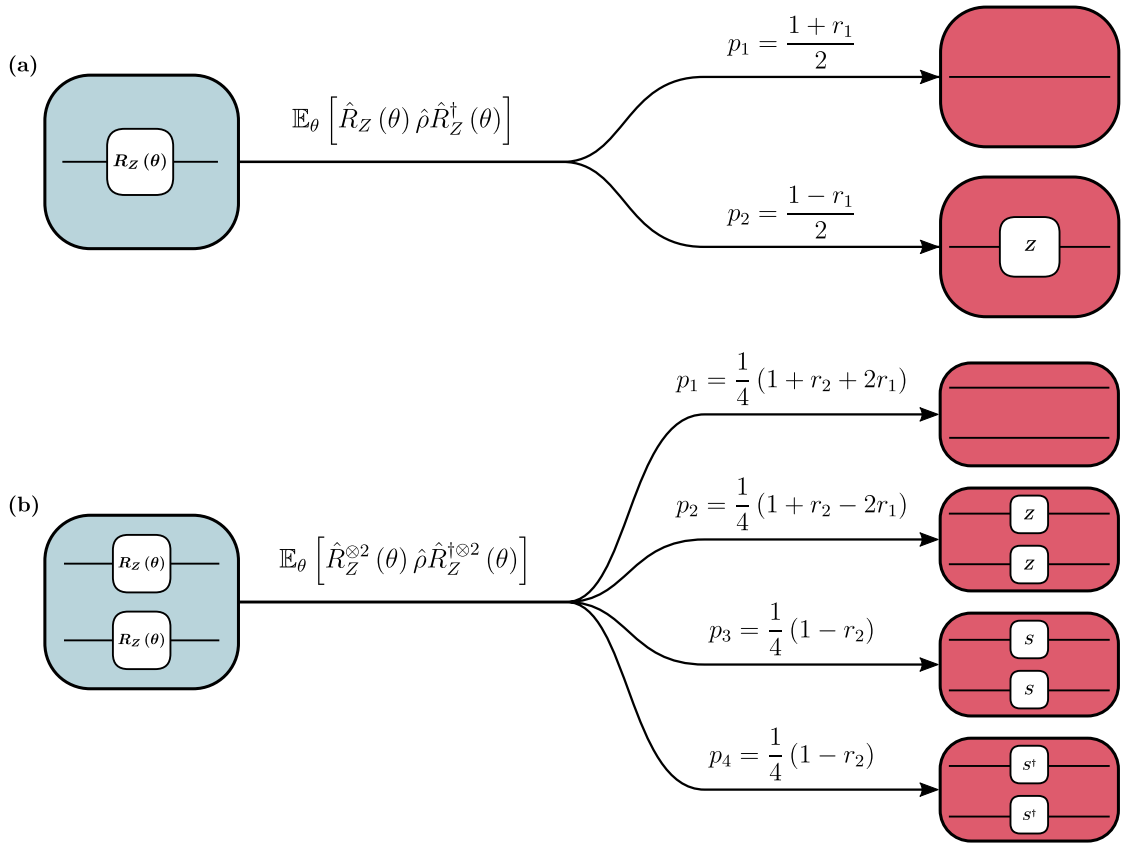


Figure 3.2: Schematic representation of the mapping rules from Z-rotations with a random parameter to unitary ensembles composed of Clifford gates. Panels (a) and (b) respectively are for first and second-order averages. The mapping here is for probability distributions that are even with respect to θ: we denote $r_1 = \mathbb{E}_\theta[\cos(\theta)]$ and $r_2 = \mathbb{E}_\theta[\cos(2\theta)]$. The coefficients p_i are the probabilities dictating how the corresponding Clifford circuits are sampled.

on a classical concentration argument, and is schematically represented in Fig. 3.1 for a simple circuit at the first order. From this, one can estimate the expectation value of the gradient, as it suffices to replace $\hat{U}(\boldsymbol{\theta})$ by $\hat{U}_\pm(\boldsymbol{\theta})$ (as defined in Sec. I.1) in the 1-fold channel definition to obtain the expectation of $C(\boldsymbol{\theta} \pm \pi/2)$. This gives the expectation of the gradient thanks to the parameter-shift rule.

Estimation of second order quantities

As we will prove in Sec. III, the 2-fold channel associated to a random Z-rotation is also a linear combination of Clifford channels, provided that the probability distribution is an even function of θ. This result is depicted in Fig. 3.2(b). When the inequalities $\mathbb{E}_\theta[f_+(\theta)] \geq 0$ and $\mathbb{E}_\theta[f_+(\theta)] \geq 0$ with $f_\pm(\theta) = \cos\theta(\cos\theta \pm 1)$ are satisfied, the previous linear combination is in fact a convex sum. Equivalently, the 2-fold channel of a Z-rotation with an even angular probability distribution is a Clifford mixed-unitary channel if

$$\mathbb{E}_\theta\left[\cos^2\theta\right] \geq |\mathbb{E}_\theta[\cos\theta]|. \tag{3.31}$$

The zeros of f_+, f_- are the angles $\{k\pi/2, k \in \mathbb{Z}\}$ for which $\hat{R}_Z(\theta)$ matches a Clifford gate (see Sec. III). Indeed, if the distribution of θ is a convex sum of Dirac distributions at these angles, the average over θ becomes a discrete average over the corresponding Clifford unitaries. Hence, the associated 2-fold channel is indeed a convex sum of Clifford channels. One can also verify that the previous conditions are satisfied for distributions that are both even with respect to the angle θ and π-periodic. For example, the uniform distribution is included. In the case of a centered Gaussian distribution, the previous conditions are satisfied if and only if the corresponding width is large enough.

Provided the distributions of the rotation angles satisfy the conditions discussed above, the scheme can be extended to the second order, allowing to approximate second-order quantities such as the average of the squared cost function $\mathbb{E}_{\boldsymbol{\theta}}\left[C(\boldsymbol{\theta})^2\right]$ using a set of Clifford approximant circuits. By the parameter-shift rule, the expectation of the squared gradient can be calculated from the knowledge of four quantities of the form $\mathbb{E}_{\boldsymbol{\theta}}\left[C(\boldsymbol{\theta} \pm (\pi/2)\boldsymbol{e}_k)C(\boldsymbol{\theta} \pm (\pi/2)\boldsymbol{e}_k)\right]$. The latter can be estimated with Clifford approximants by replacing the $\hat{U}^{\otimes 2}$ term in the definition of the 2-fold channel by $\hat{U}_\pm \otimes \hat{U}_\pm$. Hence the scheme covers the estimation of the gradient variance. Note that at the second order, the approximant circuits are obtained by replacing the rotation 2-fold channels by one of the four 2-qubit Clifford gates depicted on Fig. 3.2, yielding an ensemble of 4^M possible Clifford circuits. As for first-order quantities, a number of samples scaling in M is enough to guarantee convergence. These rigorous results are summarized in the following theorem, whose detailed proof is shown in Secs. III and IV.

Theorem II.1. *For a variational ansatz composed of fixed Clifford gates and of M parameterized rotations along the X, Y or Z direction, if the random variational parameters $(\theta_1, \ldots, \theta_M)$ are independent and symmetric with respect to one of the Clifford angles, i.e. $\in \{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$, then for any error $\epsilon > 0$ and a probability $1 - \delta$ to meet such accuracy, any first-order quantity can be computed using*

$$K \geq O\left(\frac{M}{\epsilon^2} \log\left(\frac{2}{\delta}\right)\right)$$

Clifford approximant circuits. Moreover, if the distribution of θ_i satisfies the inequality

$$\mathbb{E}_{\theta_i}\left[\cos^2(\theta_i - \mathbb{E}_{\theta_i}\left[\theta_i\right])\right] \geq \left|\mathbb{E}_{\theta_i}\left[\cos(\theta_i - \mathbb{E}_{\theta_i}\left[\theta_i\right])\right]\right|, \quad \forall i \in \{1, \ldots, M\}$$

then the same holds for any second-order quantity.

Finally, making use of the Gottesman-Knill theorem, which states that for a Clifford unitary $\hat{U}$ and an observable $\hat{O}$ acting non-trivially on N_O qubits, the expectation value $\mathrm{Tr}\left[|0\rangle\langle 0|^{\otimes n}\, \hat{U}^\dagger \hat{O} \hat{U}\right]$ can be classically computed with a polynomial complexity in both n and N_O. Our method inherits this complexity, and in particular we can classically estimate the gradient expectation and variance with a polynomial complexity in n, N_O and M.

Extensions of the scheme

In Sec. III, we present different expressions for the general case of t-fold channels associated with random Z-rotations. In particular, we show that in the case where the distribution of

the random angle does not satisfy the convex condition of Eq. (3.31), the 2-fold channels can still be expressed as a linear combination of Clifford channels. In that case, it is still possible to use Clifford approximant circuits to estimate second-order quantities, but this comes at the price of an exponential complexity in the number of variational parameters M, as shown in Sec. IV. This result is based on a standard sampling method of quasi-probability distributions [235, 236]. In that context, the method allows to trade an exponential complexity in the system size for an exponential complexity in the number of variational parameters M. This added complexity results from the existence of negative terms in the Clifford-channels decomposition of the circuits [235].

We also provide two decompositions of the N-fold channel associated to random Z-rotations in Sec. III. The first decomposition enables to write the N-fold channel as a linear combination of Clifford unitary channels. From this decomposition, we derive a condition for the N-fold channel to be a convex sum of Clifford unitaries by imposing the coefficients of the combination to be positive. However we show that the obtained decomposition is not unique, so that the derived condition is sufficient but not necessary. Moreover, we have no guarantee that the derived condition can indeed be fulfilled, and finding a sufficient and necessary condition on the distribution of a random angle that guarantees that the corresponding N-fold unitary is a Clifford mixed unitary remains an open problem. We give a second decomposition of the N-fold channel into a discrete sum of Z-rotations that generalizes some of the formulas found for the 2-fold channel.

II.2 Numerical simulations

To illustrate the applications of our exact mapping and the ensuing estimation method, we have performed numerical experiments on concrete examples. Let us consider a simple variational quantum circuit composed of layers of single-qubit rotations along either the X, Y or Z axes, alternated with fixed layers of Control-Z gates. Such an ansatz is shown for three qubits in Fig. 3.1. We further assume that the rotation angles are independent and identically distributed according to the uniform law over $[0, 2\pi]$. Moreover, we assume that the cost function is of the form in Eq. (3.1) with $\hat{O} = |0\rangle\langle 0|^{\otimes n}$.

We consider these architectures with random directions of the rotation gates. Up to a different fixed first layer, such random circuits have been showed to exhibit barren plateaus in Ref. [237]. Note that in this particular case the averaging was done on both the rotations angles and the rotations directions. Here we reproduce this result using Clifford approximants. To do so we sample both the exact circuit architecture by randomly selecting the rotation directions uniformly from $\{X, Y, Z\}$, and then we either sample the rotation angles directly or we sample a Clifford approximant circuit. For a uniform distribution we have $\forall k \in \mathbb{Z}$, $\mathbb{E}_\theta [\cos k\theta] = 0$ so the sampling of the replacement Clifford gates is uniform (as represented on Fig. 3.2 for $r_1 = r_2 = 0$). Moreover, by the parameter-shift rule (Eq. 3.6) it is clear that for uniformly distributed rotations the average gradient is analytically zero, thus it suffices to estimate the average of the squared gradient as $\mathrm{Var}_{\boldsymbol{\theta}} [\partial_k C(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}} [\partial_k C(\boldsymbol{\theta})^2]$.

In Fig. 3.3 the estimations of the average squared gradient using either direct evaluations or by sampling Clifford approximants are shown. Note that the average is taken over both the random rotation angles and the variable architecture (i.e. the random

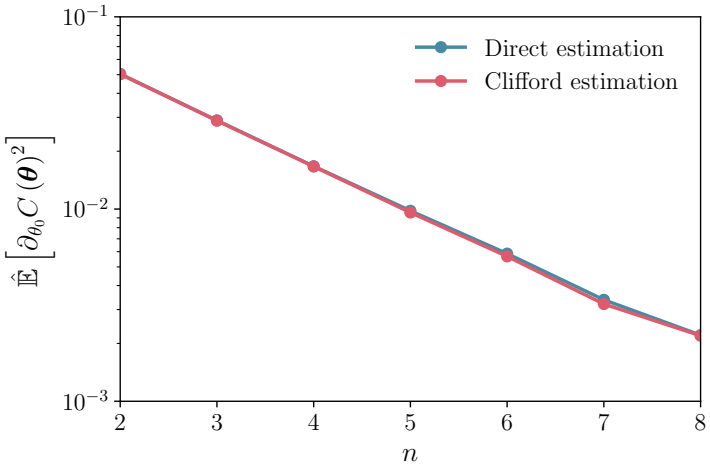


Figure 3.3: Estimated average of the squared gradient of the cost function with respect to the first variational parameter versus the number n of qubits. We emphasize that derivatives with respect to the other angles θ_k give similar results (not shown). The results are for random circuits composed of a single layer of gates, with one rotation per qubit. Such rotations are randomly chosen among R_X, R_Y, R_Z. The rotation layer is followed by a layer of alternated CZ gates (note that this is the same type of architecture as that represented on Fig. 3.1). The random rotation angles are independent and follow the uniform probability distribution on the interval $[0, 2\pi]$. In order to get the estimation, we have randomly sampled 500 different circuit architectures. For each gate architecture, we have computed the average of the squared gradient assuming a uniform distribution of the rotation angles, using both a direct estimation and our method based on the mapping to Clifford approximant circuits. In particular, we have sampled 500 vectors of angle parameters for the direct estimation and 500 Clifford circuits for our method. Note that for the uniform distribution the average gradient vanishes, thus estimating the squared gradient is equivalent to estimate the gradient variance.

direction of the rotation gates). The estimation obtained from Clifford approximants accurately matches the direct estimation and the average squared gradient vanishes exponentially with the number of qubits, as expected. In addition, the evolution of the bias of the Clifford estimation with the number of approximant circuits K is shown in Fig. 3.4. The bias decreases polynomially with K. As appears in Fig. 3.5, the same trend holds for the variance of the Clifford estimators. These results are in agreement with the analytical scaling derived in Sec. IV.

It is worth noticing that probing an exponential vanishing of the cost function gradient requires to simulate an exponential number of Clifford approximant circuits. This is a direct consequence of the scaling provided in Thm. II.1. Although this might seems like an important constain, we argue that on an hardware platform the gradient estimation would be limited to some experimental precision ϵ_{exp}. As a result, any variational quantum circuits exhibiting an average gradient amplitude below ϵ_{exp} would be hard to train. More generally, estimating quantities to this experimental accuracy is sufficient for all practical purpose.

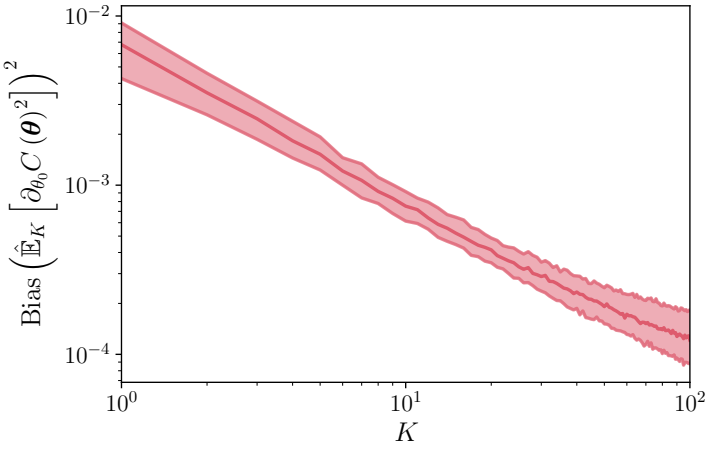


Figure 3.4: Squared statistical bias of the estimator considered in Fig. 3.3 for random circuits with $n = 5$ qubits versus the number K of Clifford approximant circuits. The results have been obtained with 500 randomly drawn circuit architectures. For each sample size K, we consider a bootstrap batch of 100 estimators (each estimator is obtained by sampling K circuits from a set of 2000 Clifford approximant circuits for each choice of the rotation directions). Then for each K, the statistical bias is derived from the bootstrap batch. The estimator true expected value is provided by the direct estimation of the average squared gradient with 4000 samples. The shaded area corresponds to the interval between the 20 and 80 percentiles of the estimated biases for the 500 random architectures.

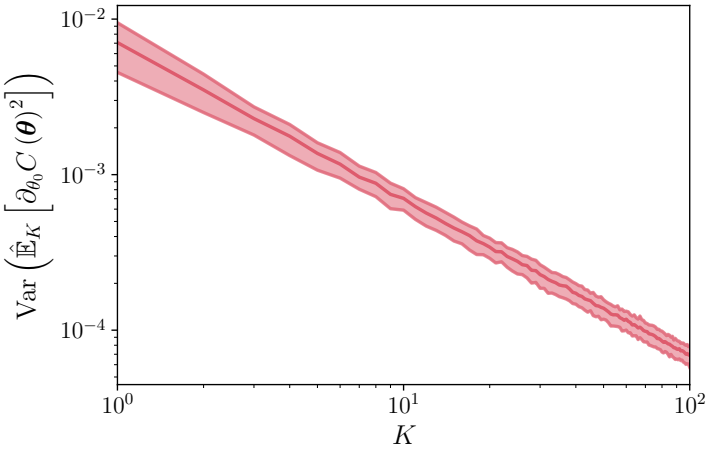


Figure 3.5: Variance of the estimator of the expected squared gradient with respect to the first parameter θ_0 versus the number K of Clifford approximant circuits. Same type of random circuits as in Fig. 3.3 with $n = 5$ qubits. We have used the same bootstrap procedure as in Fig. 3.4. The shaded area corresponds to the interval between the 20 and 80 percentiles of the estimated biases for 500 random architectures.

III t-fold channels of a random Z-rotation

In this section we study the t-fold channels associated with random Z-rotations. The rotations around the X and Y axis can then be obtained by combination with Hadamard and phase gates.

III.1 1-fold channel

As in Sec. III, we denote $\hat{\Pi}_0 := |0\rangle\langle 0|$ and $\hat{\Pi}_1 := |1\rangle\langle 1|$. We have

$$\hat{R}_Z(\theta) = e^{-\mathrm{i}\frac{\theta}{2}}\hat{\Pi}_0 + e^{\mathrm{i}\frac{\theta}{2}}\hat{\Pi}_1 \,, \tag{3.32}$$

such that

$$\hat{R}_Z(\theta)\hat{\rho}\hat{R}_Z^\dagger(\theta) = \hat{\Pi}_0\hat{\rho}\hat{\Pi}_0 + \hat{\Pi}_1\hat{\rho}\hat{\Pi}_1 + e^{\mathrm{i}\theta}\hat{\Pi}_1\hat{\rho}\hat{\Pi}_0 + e^{-\mathrm{i}\theta}\hat{\Pi}_0\hat{\rho}\hat{\Pi}_1 \,. \tag{3.33}$$

Averaging over the random angle, we obtain

$$\mathbb{E}_\theta\left[\hat{R}_Z(\theta)\hat{\rho}\hat{R}_Z^\dagger(\theta)\right] = \hat{\Pi}_0\hat{\rho}\hat{\Pi}_0 + \hat{\Pi}_1\hat{\rho}\hat{\Pi}_1 + \mathbb{E}_\theta\left[e^{\mathrm{i}\theta}\right]\hat{\Pi}_1\hat{\rho}\hat{\Pi}_0 + \mathbb{E}_\theta\left[e^{-\mathrm{i}\theta}\right]\hat{\Pi}_0\hat{\rho}\hat{\Pi}_1 \,. \tag{3.34}$$

We recognize the characteristic function of the distribution of θ, namely

$$\phi(t) := \mathbb{E}_\theta\left[e^{\mathrm{i}t\theta}\right] .$$

Even distribution

Assuming this probability distribution is even in θ, we have $\phi(t) \in \mathbb{R}$, $\forall t$ and we can define $r_1 = \phi(1) = \phi(1)^* = \phi(-1)$. As we have $\mathbb{1} = \hat{\Pi}_0 + \hat{\Pi}_1$ and $\hat{Z} = \hat{\Pi}_0 - \hat{\Pi}_1$, we get

$$\begin{aligned} \hat{\rho} &= \left(\hat{\Pi}_0\hat{\rho}\hat{\Pi}_0 + \hat{\Pi}_1\hat{\rho}\hat{\Pi}_1\right) + \left(\hat{\Pi}_1\hat{\rho}\hat{\Pi}_0 + \hat{\Pi}_0\hat{\rho}\hat{\Pi}_1\right) , \\ \hat{Z}\hat{\rho}\hat{Z} &= \left(\hat{\Pi}_0\hat{\rho}\hat{\Pi}_0 + \hat{\Pi}_1\hat{\rho}\hat{\Pi}_1\right) - \left(\hat{\Pi}_1\hat{\rho}\hat{\Pi}_0 + \hat{\Pi}_0\hat{\rho}\hat{\Pi}_1\right) , \end{aligned} \tag{3.35}$$

and hence

$$\frac{1+r_1}{2}\hat{\rho} + \frac{1-r_1}{2}\hat{Z}\hat{\rho}\hat{Z} = \mathbb{E}_\theta\left[R_Z(\theta)\hat{\rho}R_Z^\dagger(\theta)\right] . \tag{3.36}$$

As expected, the 1-fold channel for the considered random Z-rotation is a dephasing channel. This result is the time-independent analog of Eq. 1.48 obtained for the decoherence model studied in Sec. III. In the present case, the decoherence is artificially induced by the random choice of the rotation angle θ.

The channel of Eq. (3.36) is a convex sum of Clifford channels under the condition that $r_1 \in [-1, 1]$, which is always satisfied. For distributions that are symmetric with respect to a Clifford angles $\in \{k\pi/2, k \in \{0, 1, 2, 3\}\}$, we can factor out the corresponding rotation, which is (up to a phase) a Clifford gate. This way we can fall back to the case of an unbiased even distribution, i.e. symmetric with respect to the zero angle. Note that in the particular case of the uniform distribution over $[0, 2\pi]$, we have $r_1 = 0$.

Generic distribution

Here we provide a generalization of the previous expressions to the case of a *generic* distribution of the rotation angle. Let us denote $\mathbb{E}_\theta\left[\mathrm{e}^{\mathrm{i}\theta}\right] := r_1 + \mathrm{i}s_1$, where we have introduced $s_1 = \mathbb{E}_\theta\left[\sin\theta\right]$. $\mathbb{E}_\theta\left[\mathrm{e}^{2\mathrm{i}\theta}\right] := r_2 + \mathrm{i}s_2$. Developing Eq. (3.34), we get:

$$\begin{aligned}\mathbb{E}_\theta\left[\hat{R}_Z(\theta)\hat{\rho}\hat{R}_Z^\dagger(\theta)\right] =& \ \hat{\Pi}_0\hat{\rho}\hat{\Pi}_0 + \hat{\Pi}_1\hat{\rho}\hat{\Pi}_1 + (r_1+\mathrm{i}s_1)\hat{\Pi}_1\hat{\rho}\hat{\Pi}_0 + (r_1-\mathrm{i}s_1)\hat{\Pi}_0\hat{\rho}\hat{\Pi}_1 \\ =& \frac{1+r_1}{2}\mathcal{E}[\mathbb{1}](\hat{\rho}) + \frac{1-r_1}{2}\mathcal{E}[\hat{Z}](\hat{\rho}) + \frac{s_1}{2}\mathcal{E}[\hat{S}](\hat{\rho}) - \frac{s_1}{2}\mathcal{E}[\hat{S}^\dagger](\hat{\rho})\,,\end{aligned} \tag{3.37}$$

where $\hat{S} = \hat{\Pi}_0 + \mathrm{i}\hat{\Pi}_1$ is the phase gate, and one can use this definition together with Eq. (3.35) to verify the equation above.

Here, the parameter s_1 can be understood as a measure of asymmetry in the probability distribution of θ. In the symmetric case, we have $s_1 = 0$ and the sum reduces to the convex one given by Eq. (3.36). If $s_1 \neq 0$, one can still decompose the 1-fold channel as a non-convex combination of Clifford gates.

III.2 2-fold channel

We now turn our attention to the 2-fold channel. In this section we will make use of the Choi representation of quantum channels, which is reminded in Sec. II. This allows to represent channels acting on two-qubits states by 16×16 matrices. For a quantum channel $\mathcal{E}$, the Choi operator is defined by:

$$\Lambda(\mathcal{E}) = \sum_{i,j,k,l=0}^{1} |ij\rangle\langle kl| \otimes \mathcal{E}\left(|ij\rangle\langle kl|\right). \tag{3.38}$$

Its corresponding matrix entries are:

$$\begin{aligned}\Lambda(\mathcal{E})_{(ijkl),(mnpq)} =& \mathrm{Tr}\left[\Lambda(\mathcal{E})^\dagger\left(|ij\rangle\langle kl| \otimes |mn\rangle\langle pq|\right)\right] \\ =& \mathrm{Tr}\left[\mathcal{E}\left(|ij\rangle\langle kl|\right)^\dagger |mn\rangle\langle pq|\right]\end{aligned} \tag{3.39}$$

In the following, we will write

$$\mathcal{E}[\hat{U}](\hat{\rho}) := \hat{U}\hat{\rho}\hat{U}^\dagger \tag{3.40}$$

for the quantum channel associated to a unitary transformation $\hat{U}$. We assume that $\hat{U}$ is diagonal in the computational basis, so that we can write

$$\hat{U} = \sum_{i,j=0}^{1} \lambda_{ij}\hat{\Pi}_{ij}\,, \tag{3.41}$$

where we define the projectors $\hat{\Pi}_{ij} := \hat{\Pi}_i \otimes \hat{\Pi}_j$. For $\hat{U}$ unitary, we have

$$\hat{U}\hat{U}^\dagger = \hat{\mathbb{1}} = \sum_{i,j} \lambda_{ij}\lambda_{ij}^*\hat{\Pi}_{ij}\,, \tag{3.42}$$

and hence $\lambda_{ij} = e^{\mathrm{i}\theta_{ij}}$, $\forall i, j$. Therefore we have

$$\begin{aligned}\mathcal{E}[\hat{U}](|ij\rangle\langle kl|) &= \sum_{m,n,p,q} \lambda_{mn}\lambda^*_{pq}\hat{\Pi}_{mn}\,|ij\rangle\langle kl|\,\hat{\Pi}_{pq} \\ &= \lambda_{ij}\lambda^*_{kl}\,|ij\rangle\langle kl| \\ &= e^{\mathrm{i}(\theta_{ij}-\theta_{kl})}\,|ij\rangle\langle kl|\,.\end{aligned} \tag{3.43}$$

Thus the Choi matrix of $\mathcal{E}[\hat{U}]$ is diagonal whenever $\hat{U}$ is of the form given in Eq. (3.41). We can represent it by a 4×4 matrix M, whose entries are defined by

$$M_{(ij),(kl)} := \Lambda_{(ijkl),(ijkl)}. \tag{3.44}$$

Note that the matrix M is Hermitian and that its diagonal entries are always equal to one, due to Eq. (3.43). In the following we will represent each channel $\mathcal{E}$ whose Choi matrix is diagonal by its associated matrix $M(\mathcal{E})$ in the basis $(00), (01), (10), (11)$. If $\mathcal{E} = \mathcal{E}\left[\hat{U}\right]$, we simply write $M(\hat{U})$.

As done earlier, we will focus on rotations around the Z axis. We have

$$\Phi^{(2)}_Z(\hat{\rho}) := \mathbb{E}_\theta\left[(\hat{R}_Z(\theta)\otimes\hat{R}_Z(\theta))\hat{\rho}(\hat{R}^\dagger_Z(\theta)\otimes\hat{R}^\dagger_Z(\theta))\right] \tag{3.45}$$

and

$$\hat{R}_Z(\theta)\otimes\hat{R}_Z(\theta) = \left(e^{-\mathrm{i}\theta}\hat{\Pi}_0\otimes\hat{\Pi}_0 + e^{\mathrm{i}\theta}\hat{\Pi}_1\otimes\hat{\Pi}_1\right) + \left(\hat{\Pi}_0\otimes\hat{\Pi}_1 + \hat{\Pi}_1\otimes\hat{\Pi}_0\right). \tag{3.46}$$

Defining

$$\begin{aligned}\Gamma_\theta &= \left(e^{-\mathrm{i}\theta}\hat{\Pi}_{00} + e^{\mathrm{i}\theta}\hat{\Pi}_{11}\right), \\ \Xi &= \hat{\Pi}_{01} + \hat{\Pi}_{10},\end{aligned} \tag{3.47}$$

we can write

$$\Phi^{(2)}_Z(\hat{\rho}) = \mathbb{E}_\theta\left[\Xi\hat{\rho}\Xi^\dagger\right] + \mathbb{E}_\theta\left[\Gamma_\theta\hat{\rho}\Gamma^\dagger_\theta\right] + \mathbb{E}_\theta\left[\Gamma_\theta\hat{\rho}\Xi^\dagger\right] + \mathbb{E}_\theta\left[\Xi\hat{\rho}\Gamma^\dagger_\theta\right]. \tag{3.48}$$

Uniform distribution

For θ uniformly distributed in $[0, 2\pi]$, we have $\mathbb{E}_\theta\left[e^{\pm\mathrm{i}\theta}\right] = \mathbb{E}_\theta\left[e^{\pm 2\mathrm{i}\theta}\right] = 0$, and thus:

$$\begin{aligned}\mathbb{E}_\theta\left[\Gamma_\theta\hat{\rho}\Xi^\dagger\right] &= 0\,, \\ \mathbb{E}_\theta\left[\Gamma_\theta\hat{\rho}\Gamma^\dagger_\theta\right] &= \hat{\Pi}_{00}\hat{\rho}\hat{\Pi}_{00} + \hat{\Pi}_{11}\hat{\rho}\hat{\Pi}_{11}\,, \\ \mathbb{E}_\theta\left[\Xi\hat{\rho}\Xi^\dagger\right] &= \hat{\Pi}_{01}\hat{\rho}\hat{\Pi}_{01} + \hat{\Pi}_{10}\hat{\rho}\hat{\Pi}_{10} + \hat{\Pi}_{01}\hat{\rho}\hat{\Pi}_{10} + \hat{\Pi}_{10}\hat{\rho}\hat{\Pi}_{01}\,.\end{aligned} \tag{3.49}$$

Finally, we get

$$\Phi^{(2)}_Z(\hat{\rho}) = \hat{\Pi}_{00}\hat{\rho}\hat{\Pi}_{00} + \hat{\Pi}_{11}\hat{\rho}\hat{\Pi}_{11} + \hat{\Pi}_{01}\hat{\rho}\hat{\Pi}_{01} + \hat{\Pi}_{10}\hat{\rho}\hat{\Pi}_{10} + \hat{\Pi}_{01}\hat{\rho}\hat{\Pi}_{10} + \hat{\Pi}_{10}\hat{\rho}\hat{\Pi}_{01}\,. \tag{3.50}$$

We can represent $\Phi_Z^{(2)}$ by its associated matrix

$$M(\Phi_Z^{(2)}) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{3.51}$$

One can verify that the following channels also have a diagonal Choi matrix, and we can use the same representation of their diagonals, giving

$$\begin{aligned} M(\hat{S}\otimes\hat{S}) &= \begin{pmatrix} 1 & \mathrm{i} & \mathrm{i} & -1 \\ -\mathrm{i} & 1 & 1 & \mathrm{i} \\ -\mathrm{i} & 1 & 1 & \mathrm{i} \\ -1 & -\mathrm{i} & -\mathrm{i} & 1 \end{pmatrix}, \quad M(\hat{Z}\otimes\hat{Z}) = \begin{pmatrix} 1 & -1 & -1 & 1 \\ -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}, \\ M(\hat{S}^\dagger\otimes\hat{S}^\dagger) &= \begin{pmatrix} 1 & -\mathrm{i} & -\mathrm{i} & -1 \\ \mathrm{i} & 1 & 1 & -\mathrm{i} \\ \mathrm{i} & 1 & 1 & -\mathrm{i} \\ -1 & \mathrm{i} & \mathrm{i} & 1 \end{pmatrix}, \quad M(\mathbb{1}) = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}. \end{aligned} \tag{3.52}$$

Gathering all together, we have

$$M(\Phi_Z^{(2)}) = \frac{1}{4}\left(M(\mathbb{1}) + M(\hat{Z}\otimes\hat{Z}) + M(\hat{S}\otimes\hat{S}) + M(\hat{S}^\dagger\otimes\hat{S}^\dagger)\right). \tag{3.53}$$

The final result presented in Sec. II then follows by linearity and uniqueness of the Choi matrix.

Even distribution

Let us consider an even probability distribution of θ (i.e. a distribution for which θ has the same law as $-\theta$). For such distributions we again have that $\phi_\theta(t) = \phi_\theta(-t) \in [-1,1] \subset \mathbb{R}$ for all $t \in \mathbb{R}$ and thus

$$\phi_\theta(t) = \frac{1}{2}\left(\phi_\theta(t) + \phi_\theta(-t)\right) = \mathbb{E}_\theta\left[\cos(t\theta)\right]. \tag{3.54}$$

Defining $r_1 = \phi_\theta(1)$ and $r_2 = \phi_\theta(2)$, we can write

$$\begin{aligned} \mathbb{E}_\theta\left[\Gamma_\theta\hat{\rho}\Xi^\dagger\right] &= r_1\left(\hat{\Pi}_{00}+\hat{\Pi}_{11}\right)\hat{\rho}\left(\hat{\Pi}_{01}+\hat{\Pi}_{10}\right), \\ \mathbb{E}_\theta\left[\Gamma_\theta\hat{\rho}\Gamma_\theta^\dagger\right] &= \hat{\Pi}_{00}\hat{\rho}\hat{\Pi}_{00} + \hat{\Pi}_{11}\hat{\rho}\hat{\Pi}_{11} + r_2\left(\hat{\Pi}_{00}\hat{\rho}\hat{\Pi}_{11} + \hat{\Pi}_{11}\hat{\rho}\hat{\Pi}_{00}\right), \\ \mathbb{E}_\theta\left[\Xi\hat{\rho}\Xi^\dagger\right] &= \hat{\Pi}_{01}\hat{\rho}\hat{\Pi}_{01} + \hat{\Pi}_{10}\hat{\rho}\hat{\Pi}_{10} + \hat{\Pi}_{01}\hat{\rho}\hat{\Pi}_{10} + \hat{\Pi}_{10}\hat{\rho}\hat{\Pi}_{01}. \end{aligned} \tag{3.55}$$

Hence we obtain:

$$M(\Phi_Z^{(2)}) = \begin{pmatrix} 1 & r_1 & r_1 & r_2 \\ r_1 & 1 & 1 & r_1 \\ r_1 & 1 & 1 & r_1 \\ r_2 & r_1 & r_1 & 1 \end{pmatrix}. \tag{3.56}$$

We can express $M(\Phi_Z^{(2)})$ as a linear combination of the matrices of Eq. (3.52), giving

$$M(\Phi_Z^{(2)}) = aM(\mathbb{1}) + bM(\hat{Z}\otimes\hat{Z}) + \frac{c}{2}\left(M(\hat{S}\otimes\hat{S}) + M(\hat{S}^\dagger\otimes\hat{S}^\dagger)\right). \tag{3.57}$$

The coefficients a, b, c can be found by solving the linear system

$$\begin{cases} a+b+c=1 \\ a-b=r_1 \\ a+b-c=r_2 \end{cases}, \tag{3.58}$$

and one finds

$$\begin{aligned} M(\Phi_Z^{(2)}) = &\frac{1}{4}\left(1+r_2+2r_1\right)M(\mathbb{1}) \\ &+\frac{1}{4}\left(1+r_2-2r_1\right)M(\hat{Z}\otimes\hat{Z}) \\ &+\frac{1}{4}\left(1-r_2\right)M(\hat{S}\otimes\hat{S}) \\ &+\frac{1}{4}\left(1-r_2\right)M(\hat{S}^\dagger\otimes\hat{S}^\dagger). \end{aligned} \tag{3.59}$$

Therefore, the associated channel is

$$\begin{aligned} \Phi_Z^{(2)}(\hat{\rho}) = &\frac{1}{4}\left(1+r_2+2r_1\right)\hat{\rho} \\ &+\frac{1}{4}\left(1+r_2-2r_1\right)\left(\hat{Z}\otimes\hat{Z}\right)\hat{\rho}\left(\hat{Z}\otimes\hat{Z}\right) \\ &+\frac{1}{4}\left(1-r_2\right)\left(\hat{S}\otimes\hat{S}\right)\hat{\rho}\left(\hat{S}^\dagger\otimes\hat{S}^\dagger\right) \\ &+\frac{1}{4}\left(1-r_2\right)\left(\hat{S}^\dagger\otimes\hat{S}^\dagger\right)\hat{\rho}\left(\hat{S}\otimes\hat{S}\right). \end{aligned} \tag{3.60}$$

Remark. Defining $CZ = \hat{\Pi}_0\otimes\mathbb{1} + \hat{\Pi}_1\otimes\hat{Z}$ the control-Z gate and $CZ_X = (\hat{X}\otimes\hat{X})CZ(\hat{X}\otimes\hat{X})$, we have

$$M(CZ) = \begin{pmatrix} 1 & 1 & 1 & -1 \\ 1 & 1 & 1 & -1 \\ 1 & 1 & 1 & -1 \\ -1 & -1 & -1 & 1 \end{pmatrix}, \quad M(CZ_X) = \begin{pmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \end{pmatrix}, \tag{3.61}$$

and thus

$$\mathcal{E}[\hat{S}\otimes\hat{S}] + \mathcal{E}[\hat{S}^\dagger\otimes\hat{S}^\dagger] = \mathcal{E}[CZ] + \mathcal{E}[CZ_X]. \tag{3.62}$$

Therefore the decomposition of $\Phi_Z^{(2)}$ into a convex sum of Clifford channels of Eq. (3.60) is not unique.

The decomposition obtained in Eq. (3.60) is a convex sum if one assume that

$$(1+r_2-2r_1) \geq 0 \quad \text{and} \quad (1+r_2+2r_1) \geq 0\,. \tag{3.63}$$

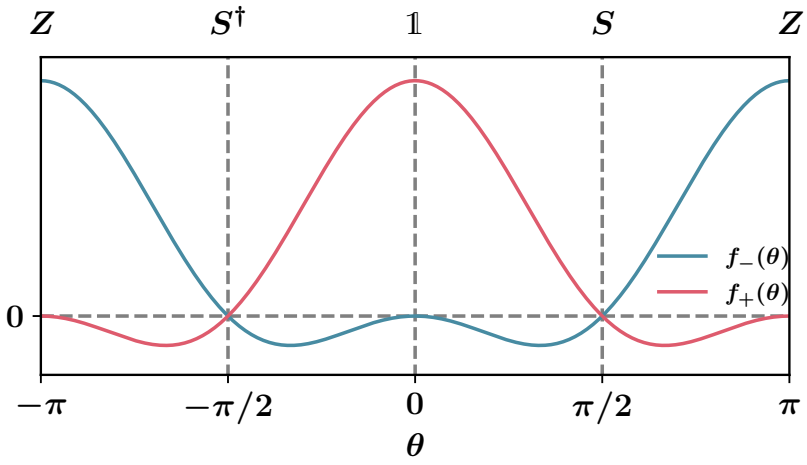


Figure 3.6: Plot of $f_\pm(\theta) = \cos\theta(\cos\theta \pm 1)$ versus θ. The condition in Eq. (3.65) is fulfilled if and only if $\mathbb{E}_\theta\left[f_+(\theta)\right] \geq 0$ and $\mathbb{E}_\theta\left[f_-(\theta)\right] \geq 0$. $\hat{S}$ is the phase gate.

This condition holds if and only if

$$\mathbb{E}_\theta\left[\frac{1}{2}(1+\cos 2\theta) \pm \cos\theta\right] \geq 0, \tag{3.64}$$

namely if and only if

$$\mathbb{E}_\theta\left[\cos^2\theta\right] \geq \left|\mathbb{E}_\theta\left[\cos\theta\right]\right|. \tag{3.65}$$

It is fulfilled for the distributions that are π-periodic, as in that case we have $\mathbb{E}_\theta\left[\cos\theta\right] = 0$. Another example of distribution that satisfy this constraint is a Gaussian distribution with a large enough variance. In fact for a centered Gaussian distribution of variance σ^2, we have $r_1 = e^{-\sigma^2/2}$ and $r_2 = e^{-2\sigma^2}$, so that the condition becomes

$$1 + e^{-2\sigma^2} - 2e^{-\sigma^2/2} \geq 0. \tag{3.66}$$

One can show that this condition is equivalent to $\sigma^2 \geq \sigma_0^2$ for some specific $\sigma_0 \in \mathbb{R}$, yielding a requirement on the width of the gaussian.

Let us remark that, again, the effect of the averaging over random angles is to artificially introduce decoherence. In fact, the two-fold channel corresponding to the matrix in Eq. (3.56) is a time-independent version of the evolution channel of an open two-qubits system of the same type as the ones presented in Sec.. III. The general evolution equations for such model can be found in [137].

Generic distribution

We now generalize the previous discussion to the case of a generic distribution of θ. We write $s_2 = \mathbb{E}_\theta\left[\sin 2\theta\right]$. The 2-fold channel for a single-qubit Z-rotation is given by Eq. (3.48):

$$\Phi_Z^{(2)}(\hat{\rho}) = \mathbb{E}_\theta\left[\Xi\hat{\rho}\Xi^\dagger\right] + \mathbb{E}_\theta\left[\Gamma_\theta\hat{\rho}\Gamma_\theta^\dagger\right] + \mathbb{E}_\theta\left[\Gamma_\theta\hat{\rho}\Xi^\dagger\right] + \mathbb{E}_\theta\left[\Xi\hat{\rho}\Gamma_\theta^\dagger\right]. \tag{3.67}$$

For a generic probability distribution of θ, we have

$$\begin{aligned}
\mathbb{E}_\theta\left[\Xi\hat{\rho}\Xi^\dagger\right] &= \hat{\Pi}_{01}\hat{\rho}\hat{\Pi}_{01} + \hat{\Pi}_{10}\hat{\rho}\hat{\Pi}_{10} + \hat{\Pi}_{01}\hat{\rho}\hat{\Pi}_{10} + \hat{\Pi}_{10}\hat{\rho}\hat{\Pi}_{01}\,, \\
\mathbb{E}_\theta\left[\Gamma_\theta\hat{\rho}\Xi^\dagger\right] &= r_1\left(\hat{\Pi}_{00}+\hat{\Pi}_{11}\right)\hat{\rho}\left(\hat{\Pi}_{01}+\hat{\Pi}_{10}\right) + \mathrm{i}s_1\left(\hat{\Pi}_{11}-\hat{\Pi}_{00}\right)\hat{\rho}\left(\hat{\Pi}_{01}+\hat{\Pi}_{10}\right)\,, \\
\mathbb{E}_\theta\left[\Gamma_\theta\hat{\rho}\Gamma_\theta^\dagger\right] &= \hat{\Pi}_{00}\hat{\rho}\hat{\Pi}_{00} + \hat{\Pi}_{11}\hat{\rho}\hat{\Pi}_{11} + r_2\left(\hat{\Pi}_{00}\hat{\rho}\hat{\Pi}_{11} + \hat{\Pi}_{11}\hat{\rho}\hat{\Pi}_{00}\right) \\
&\qquad + \mathrm{i}s_2\left(\hat{\Pi}_{11}\hat{\rho}\hat{\Pi}_{00} - \hat{\Pi}_{00}\hat{\rho}\hat{\Pi}_{11}\right)\,.
\end{aligned} \tag{3.68}$$

As one can verify, the Choi representation of the above terms are all diagonal, so that their sum can be represented via the M matrix as before:

$$M(\Phi_Z^{(2)}) = \begin{pmatrix} 1 & r_1+\mathrm{i}s_1 & r_1+\mathrm{i}s_1 & r_2+\mathrm{i}s_2 \\ r_1-\mathrm{i}s_1 & 1 & 1 & r_1+\mathrm{i}s_1 \\ r_1-\mathrm{i}s_1 & 1 & 1 & r_1+\mathrm{i}s_1 \\ r_2-\mathrm{i}s_2 & r_1-\mathrm{i}s_1 & r_1-\mathrm{i}s_1 & 1 \end{pmatrix}\,. \tag{3.69}$$

This can again be decomposed as a weighted sum of the channels $\mathcal{E}[\mathbb{1}]$, $\mathcal{E}[\hat{Z}\otimes\hat{Z}]$, $\mathcal{E}[\hat{S}\otimes\hat{S}]$ and $\mathcal{E}[\hat{S}^\dagger\otimes\hat{S}^\dagger]$ given in Eq. (3.52) and of the following Clifford channels:

$$\begin{aligned}
M\left(\mathbb{1}\otimes\hat{S}\right) &= \begin{pmatrix} 1 & \mathrm{i} & 1 & \mathrm{i} \\ -\mathrm{i} & 1 & -\mathrm{i} & 1 \\ 1 & \mathrm{i} & 1 & \mathrm{i} \\ -\mathrm{i} & 1 & -\mathrm{i} & 1 \end{pmatrix}\,, \quad M\left(\mathbb{1}\otimes\hat{S}^\dagger\right) = \begin{pmatrix} 1 & -\mathrm{i} & 1 & -\mathrm{i} \\ \mathrm{i} & 1 & \mathrm{i} & 1 \\ 1 & -\mathrm{i} & 1 & -\mathrm{i} \\ \mathrm{i} & 1 & \mathrm{i} & 1 \end{pmatrix}\,, \\
M\left(\hat{S}\otimes\mathbb{1}\right) &= \begin{pmatrix} 1 & 1 & \mathrm{i} & \mathrm{i} \\ 1 & 1 & \mathrm{i} & \mathrm{i} \\ -\mathrm{i} & -\mathrm{i} & 1 & 1 \\ -\mathrm{i} & -\mathrm{i} & 1 & 1 \end{pmatrix}\,, \quad M\left(\hat{S}^\dagger\otimes\mathbb{1}\right) = \begin{pmatrix} 1 & 1 & -\mathrm{i} & -\mathrm{i} \\ 1 & 1 & -\mathrm{i} & -\mathrm{i} \\ \mathrm{i} & \mathrm{i} & 1 & 1 \\ \mathrm{i} & \mathrm{i} & 1 & 1 \end{pmatrix}\,, \\
M\left(\hat{Z}\otimes\hat{S}\right) &= \begin{pmatrix} 1 & \mathrm{i} & -1 & -\mathrm{i} \\ -\mathrm{i} & 1 & \mathrm{i} & -1 \\ -1 & -\mathrm{i} & 1 & \mathrm{i} \\ \mathrm{i} & -1 & -\mathrm{i} & 1 \end{pmatrix}\,, \quad M\left(\hat{Z}\otimes\hat{S}^\dagger\right) = \begin{pmatrix} 1 & -\mathrm{i} & -1 & \mathrm{i} \\ \mathrm{i} & 1 & -\mathrm{i} & -1 \\ -1 & \mathrm{i} & 1 & -\mathrm{i} \\ -\mathrm{i} & -1 & \mathrm{i} & 1 \end{pmatrix}\,, \\
M\left(\hat{S}\otimes\hat{Z}\right) &= \begin{pmatrix} 1 & -1 & \mathrm{i} & -\mathrm{i} \\ -1 & 1 & -\mathrm{i} & \mathrm{i} \\ -\mathrm{i} & \mathrm{i} & 1 & -1 \\ \mathrm{i} & -\mathrm{i} & -1 & 1 \end{pmatrix}\,, \quad M\left(\hat{S}^\dagger\otimes\hat{Z}\right) = \begin{pmatrix} 1 & -1 & -\mathrm{i} & \mathrm{i} \\ -1 & 1 & \mathrm{i} & -\mathrm{i} \\ \mathrm{i} & -\mathrm{i} & 1 & -1 \\ -\mathrm{i} & \mathrm{i} & -1 & 1 \end{pmatrix}\,.
\end{aligned} \tag{3.70}$$

Note that the channels listed above are all diagonal in the Choi representation and hence the M matrices capture all their nonzero entries. Following the same reasoning as above, we solve a linear system to obtain the following decomposition:

$$\begin{aligned}
\Phi_Z^{(2)}(\hat{\rho}) = &\frac{s_2}{8}\left(\mathcal{E}\left[\hat{S}\otimes\mathbb{1}\right](\hat{\rho}) + \mathcal{E}\left[\mathbb{1}\otimes\hat{S}\right](\hat{\rho}) + \mathcal{E}\left[\hat{Z}\otimes\hat{S}^\dagger\right](\hat{\rho}) + \mathcal{E}\left[\hat{S}^\dagger\otimes\hat{Z}\right](\hat{\rho})\right) \\
&- \frac{s_2}{8}\left(\mathcal{E}\left[\hat{S}^\dagger\otimes\mathbb{1}\right](\hat{\rho}) + \mathcal{E}\left[\mathbb{1}\otimes\hat{S}^\dagger\right](\hat{\rho}) + \mathcal{E}\left[\hat{Z}\otimes\hat{S}\right](\hat{\rho}) + \mathcal{E}\left[\hat{S}\otimes\hat{Z}\right](\hat{\rho})\right) \\
&+ \frac{1-r_2+2s_1}{4}\mathcal{E}\left[\hat{S}\otimes\hat{S}\right](\hat{\rho}) + \frac{1-r_2-2s_1}{4}\mathcal{E}\left[\hat{S}^\dagger\otimes\hat{S}^\dagger\right](\hat{\rho}) \\
&+ \frac{1+r_2+2r_1}{4}\mathcal{E}\left[\mathbb{1}\right](\hat{\rho}) + \frac{1+r_2-2r_1}{4}\mathcal{E}\left[\hat{Z}\otimes\hat{Z}\right](\hat{\rho})\,.
\end{aligned} \tag{3.71}$$

Remark. Denoting

$$\widehat{CNOT} = \hat{\Pi}_0 \otimes \mathbb{1} + \hat{\Pi}_1 \otimes \hat{X}$$

the CNOT gate and

$$\widehat{CNOT}_X := (\hat{X} \otimes \hat{X})\widehat{CNOT}(\hat{X} \otimes \hat{X})$$

its conjugation by the $\hat{X} \otimes \hat{X}$ gate, we have

$$\begin{aligned} M\left(\widehat{CNOT}(\hat{S} \otimes \hat{S})\widehat{CNOT}\right) &= \begin{pmatrix} 1 & \mathrm{i} & -1 & \mathrm{i} \\ -\mathrm{i} & 1 & \mathrm{i} & 1 \\ -1 & -\mathrm{i} & 1 & -\mathrm{i} \\ -\mathrm{i} & 1 & \mathrm{i} & 1 \end{pmatrix}, \\ M\left(\widehat{CNOT}_X(\hat{S} \otimes \hat{S})\widehat{CNOT}_X\right) &= \begin{pmatrix} 1 & -\mathrm{i} & 1 & \mathrm{i} \\ \mathrm{i} & 1 & \mathrm{i} & -1 \\ 1 & -\mathrm{i} & 1 & \mathrm{i} \\ -\mathrm{i} & -1 & -\mathrm{i} & 1 \end{pmatrix}. \end{aligned} \tag{3.72}$$

Again, by solving a linear system one finds another decomposition of the two-fold channel that involves the above channels, namely:

$$\begin{aligned} \Phi_Z^{(2)}(\hat{\rho}) = &\frac{s_2}{4}\left(\mathcal{E}\left[\widehat{CNOT}(\hat{S} \otimes \hat{S})\widehat{CNOT}\right](\hat{\rho}) + \mathcal{E}\left[\widehat{CNOT}_X(\hat{S} \otimes \hat{S})\widehat{CNOT}_X\right](\hat{\rho})\right) \\ &- \frac{s_2}{4}\left(\mathcal{E}\left[\hat{Z} \otimes \hat{S}\right](\hat{\rho}) + \mathcal{E}\left[\mathbb{1} \otimes \hat{S}^\dagger\right](\hat{\rho})\right) \\ &+ \frac{1 - r_2 + 2s_1}{4}\mathcal{E}\left[\hat{S} \otimes \hat{S}\right](\hat{\rho}) + \frac{1 - r_2 - 2s_1}{4}\mathcal{E}\left[\hat{S}^\dagger \otimes \hat{S}^\dagger\right](\hat{\rho}) \\ &+ \frac{1 + r_2 + 2r_1}{4}\mathcal{E}\left[\mathbb{1}\right](\hat{\rho}) + \frac{1 + r_2 - 2r_1}{4}\mathcal{E}\left[\hat{Z} \otimes \hat{Z}\right](\hat{\rho}). \end{aligned} \tag{3.73}$$

III.3 N-fold channel

In this section we give two decompositions of the N-fold channel associated with a generic random Z-rotation. First, we provide a decomposition into a real sum of Clifford unitary channels. We also exhibit a *sufficient* condition on the distribution of the random angle θ for the decomposition to be a convex one. Then we provide a decomposition into a discrete convex sum of common Z-rotations of the N qubits on which the channel acts.

Decomposition into Clifford channels

In Eq. (3.37) we obtained a decomposition of the 1-fold channel of a Z-rotation in terms of Clifford unitary channels for a generic distribution of the random angle, namely

$$\mathbb{E}_\theta\left[\hat{R}_Z(\theta)\hat{\rho}\hat{R}_Z^\dagger(\theta)\right] = \frac{1 + r_1}{2}\mathcal{E}[\mathbb{1}](\hat{\rho}) + \frac{1 - r_1}{2}\mathcal{E}[\hat{Z}](\hat{\rho}) + \frac{s_1}{2}\mathcal{E}[\hat{S}](\hat{\rho}) - \frac{s_1}{2}\mathcal{E}[\hat{S}^\dagger](\hat{\rho}). \tag{3.74}$$

More generally, we have that

$$\begin{aligned} \hat{R}_Z(\theta)\hat{\rho}\hat{R}_Z^\dagger(\theta) = &\frac{1 + \cos\theta}{2}\mathcal{E}[\mathbb{1}](\hat{\rho}) + \frac{1 - \cos\theta}{2}\mathcal{E}[\hat{Z}](\hat{\rho}) \\ &+ \frac{\sin\theta}{2}\mathcal{E}[\hat{S}](\hat{\rho}) - \frac{\sin\theta}{2}\mathcal{E}[\hat{S}^\dagger](\hat{\rho}), \end{aligned} \tag{3.75}$$

for any $\theta \in \mathbb{R}$. This can be seen as a consequence of Eq. (3.74) for a Dirac probability measure centered at θ. On can directly generalize this equation to obtain an expression of the N-fold channel as a real sum of Clifford unitary channels, as

$$\hat{R}_Z^{\otimes N}(\theta)\hat{\rho}\hat{R}_Z^{\otimes N\dagger}(\theta) = \sum_{I=(i_1,\ldots,i_n)} \lambda_I(\theta)\mathcal{E}\left[\otimes_{j=1}^N \hat{U}_{i_j}\right](\hat{\rho}) \tag{3.76}$$

where the sum goes over all the multi-indices $I = (i_1, \ldots, i_N) \in \{0,1,2,3\}$, and $\hat{U}_0 = \mathbb{1}$, $\hat{U}_1 = \hat{Z}, \hat{U}_2 = \hat{S}^\dagger$ and $\hat{U}_3 = \hat{S}$. The coefficient $\lambda_I(\theta)$ for a multi-index I representing a product of numbers m_i of the $\hat{U}_i$ gates is given by

$$\lambda_I(\theta) = \frac{1}{2^N}(1+\cos\theta)^{m_0}(1-\cos\theta)^{m_1}\sin^{m_2}(-\theta)\sin^{m_3}(\theta)\ , \tag{3.77}$$

with $m_0 + m_1 + m_2 + m_3 = N$. As a result, the N-fold channel is given by a real combination of 4^N unitary Clifford channels that are composed of products of the gates $\mathbb{1}, \hat{Z}, \hat{S}$ and $\hat{S}^\dagger$. This gives us a trivial sufficient condition for the N-fold channel to be a convex sum of Clifford unitary channels, namely it suffices that the expectation values of all the coefficient $\mathbb{E}_\theta[\lambda_I(\theta)]$ be positive.

Although this condition is sufficient, it is *not necessary*. In particular, in the case of the 2-fold channel, the expectation of coefficients associated to the multi-indices $(2,3)$ and $(3,2)$ is given by $\mathbb{E}_\theta[-\sin^2\theta]$, which is always negative. However, we proved that a convex decomposition exists for the uniform distribution. This is due to the fact that the decomposition of Eq. (3.76) is not unique. In fact, the family of channels

$$\mathcal{P} := \{\mathcal{E}[\hat{U}\otimes\hat{V}]\ :\ \hat{U},\hat{V} \in \{\mathbb{1}, \hat{Z}, \hat{S}, \hat{S}^\dagger\}\} \tag{3.78}$$

is not linearly independent. Consider two single qubit unitaries $\hat{U}$ and $\hat{V}$ that are diagonal in the computational basis. As we are free to chose the global phase of these unitaries, we can always write them as $\hat{U} = e^{i\theta_U/2}\hat{\Pi}_0 + e^{-i\theta_U/2}\hat{\Pi}_1$ and $\hat{V} = e^{i\theta_V/2}\hat{\Pi}_0 + e^{-i\theta_V/2}\hat{\Pi}_1$. We saw in App. III that the product unitary $\hat{U}\otimes\hat{V}$ can be represented by the diagonal of the associated Choi matrix, written as a 4-by-4 matrix M:

$$M(\hat{U}\otimes\hat{V}) = \begin{pmatrix} 1 & e^{-i\theta_V} & e^{-i\theta_U} & e^{-i(\theta_U+\theta_V)} \\ e^{i\theta_V} & 1 & e^{-i(\theta_U-\theta_V)} & e^{-i\theta_U} \\ e^{i\theta_U} & e^{i(\theta_U-\theta_V)} & 1 & e^{-i\theta_V} \\ e^{i(\theta_U+\theta_V)} & e^{i\theta_U} & e^{i\theta_V} & 1 \end{pmatrix}. \tag{3.79}$$

This shows that for a tensor product of single-qubit unitaries, the matrices M in the basis $((00),(01),(10),(11))$ are symmetric with respect to the anti-diagonal transposition. Therefore, the channels in $\mathcal{P}$ belong to a real vector space of dimension 9 (1 dimension for the diagonal, 2×3 dimensions for the complex exponentials of the first row, and 2 dimensions for the third term of the second row). As there are 16 channels in $\mathcal{P}$, the family is not linearly independent. The condition that all the $\mathbb{E}_\theta[\lambda_I(\theta)]$ be positive is clearly too restrictive. One way to extend it to find back the condition we previously derived is to use the fact that

$$\mathcal{E}[\mathbb{1}] + \mathcal{E}[\hat{Z}] = \mathcal{E}[\hat{S}] + \mathcal{E}[\hat{S}^\dagger] \tag{3.80}$$

to absorb the $\mathbb{E}_\theta[-\sin^2\theta]$ factors into the coefficients associated to other channels.

Remark. To obtain the previous relation, we used the following channels:

$$M\left(\mathbb{1}\otimes\hat{Z}\right)=\begin{pmatrix}1&-1&1&-1\\-1&1&-1&1\\1&-1&1&-1\\-1&1&-1&1\end{pmatrix},\quad M\left(\hat{S}\otimes\hat{S}^\dagger\right)=\begin{pmatrix}1&-\mathrm{i}&\mathrm{i}&1\\\mathrm{i}&1&-1&\mathrm{i}\\-\mathrm{i}&-1&1&-\mathrm{i}\\1&-\mathrm{i}&\mathrm{i}&1\end{pmatrix},$$
$$M\left(\hat{Z}\otimes\mathbb{1}\right)=\begin{pmatrix}1&1&-1&-1\\1&1&-1&-1\\-1&-1&1&1\\-1&-1&1&1\end{pmatrix},\quad M\left(\hat{S}^\dagger\otimes\hat{S}\right)=\begin{pmatrix}1&\mathrm{i}&-\mathrm{i}&1\\-\mathrm{i}&1&-1&-\mathrm{i}\\\mathrm{i}&-1&1&\mathrm{i}\\1&\mathrm{i}&-\mathrm{i}&1\end{pmatrix}. \tag{3.81}$$

We showed that the N-fold channel associated to Z-rotations can always be decomposed as a real linear combination of diagonal Clifford unitary channels. However, it remains an open problem to find necessary and/or sufficient conditions under which the N-fold channel can be decomposed a convex combination of Clifford unitaries, i.e. conditions under which the N-fold channel is as a Clifford mixed-unitary channel [3]. The knowledge of such conditions could allow to extend the scheme proposed in this work to ansätze with correlated rotation parameters.

Decomposition into Z-rotations

It is possible to derive another general formula for the N-fold channel that extends the previous results. The single Z-rotation unitary is given by Eq. (3.32), namely

$$\hat{R}_Z(\theta)=e^{-\mathrm{i}\frac{\theta}{2}}\hat{\Pi}_0+e^{\mathrm{i}\frac{\theta}{2}}\hat{\Pi}_1\,. \tag{3.82}$$

The N-th tensor power of this unitary can be written

$$\begin{aligned}\hat{R}_Z(\theta)^{\otimes N}&=\bigotimes_{k=1}^{N}\left(e^{-\mathrm{i}\frac{\theta}{2}}\hat{\Pi}_0^k+e^{\mathrm{i}\frac{\theta}{2}}\hat{\Pi}_1^k\right)\\&=\sum_{\boldsymbol{\alpha}\in\{0,1\}^N}e^{\mathrm{i}\frac{\theta}{2}(2|\boldsymbol{\alpha}|-1)}\hat{\Pi}_{\boldsymbol{\alpha}}\,,\end{aligned} \tag{3.83}$$

where $|\boldsymbol{\alpha}|$ is the Hamming weight of $\boldsymbol{\alpha}$, namely its number of non-zero entries, $\hat{\Pi}_i^k$ is the projector $\hat{\Pi}_i$ on the k-th qubit, and

$$\hat{\Pi}_{\boldsymbol{\alpha}}=\bigotimes_{k=1}^{N}\hat{\Pi}_{\alpha_k}^k\,. \tag{3.84}$$

From this we obtain

$$\begin{aligned}\mathcal{E}\left[\hat{R}_Z(\theta)^{\otimes N}\right](\hat{\rho})&=\hat{R}_Z(\theta)^{\otimes N}\hat{\rho}\hat{R}_Z^\dagger(\theta)^{\otimes N}\\&=\sum_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\{0,1\}^N}e^{\mathrm{i}\theta(|\boldsymbol{\alpha}|-|\boldsymbol{\beta}|)}\hat{\Pi}_{\boldsymbol{\alpha}}\hat{\rho}\hat{\Pi}_{\boldsymbol{\beta}}\,.\end{aligned} \tag{3.85}$$

[3] In general, estimating the degree of "stabilizerness" of a quantum channel is difficult. This issue has been investigated in [238, 239].

Taking the expectation of this equation, we get

$$\Phi_Z^{(N)}(\hat{\rho}) = \sum_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\{0,1\}^N} \phi_\theta\left(|\boldsymbol{\alpha}| - |\boldsymbol{\beta}|\right) \hat{\Pi}_{\boldsymbol{\alpha}} \hat{\rho} \hat{\Pi}_{\boldsymbol{\beta}}\,, \tag{3.86}$$

where ϕ_θ is the characteristic function of θ. Since $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \{0,1\}^N$, the difference between the Hamming weights of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is an integer in $[\![-N, N]\!]$. As a result the N-fold channel can be expressed from the sole knowledge of the restriction of ϕ_θ to the set $[\![-N, N]\!]$. Let us denote $\tilde{\phi}_\theta$ this restriction. Taking the discrete Fourier transform of this restriction, we have

$$\tilde{\phi}_\theta(k) = \frac{1}{2N} \sum_{n=-N}^{N-1} c_n e^{\mathrm{i}\frac{kn\pi}{N}}\,, \quad \forall k \in [\![-N, N]\!]\,, \tag{3.87}$$

with

$$c_n = \sum_{k=-N}^{N-1} \tilde{\phi}_\theta(k) e^{-\mathrm{i}\frac{kn\pi}{N}}\,. \tag{3.88}$$

Recall that this decomposition is a direct consequence of the identity

$$\frac{1}{2N} \sum_{k=-N}^{N-1} e^{\mathrm{i}\frac{k(n-m)\pi}{N}} = \delta_{n,m}\,. \tag{3.89}$$

Injecting the previous decomposition in Eq. (3.86), we get

$$\begin{aligned} \Phi_Z^{(N)}(\hat{\rho}) &= \frac{1}{2N} \sum_{n=-N}^{N-1} c_n \left(\sum_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\{0,1\}^N} e^{\mathrm{i}\frac{(|\boldsymbol{\alpha}|-|\boldsymbol{\beta}|)n\pi}{N}} \hat{\Pi}_{\boldsymbol{\alpha}} \hat{\rho} \hat{\Pi}_{\boldsymbol{\beta}} \right) \\ &= \frac{1}{2N} \sum_{n=-N}^{N-1} c_n \mathcal{E}\left[\hat{R}_Z\left(\frac{n\pi}{N}\right)^{\otimes N} \right](\hat{\rho})\,, \end{aligned} \tag{3.90}$$

where we have used Eq. (3.85). As a result, the N-fold channel of a random Z-rotation can be decomposed into a linear combination of the N-qubits Z-rotation channels

$$\mathcal{E}\left[\hat{R}_Z\left(\frac{n\pi}{N}\right)^{\otimes N} \right]\,, \quad n \in [\![-N, N-1]\!]\,. \tag{3.91}$$

For $N = 1$ and $N = 2$, the previous equation reduces to a decomposition of the 1- and 2-fold channel into Clifford gates. In fact, up to a phase, the Z-rotations with angles $\frac{n\pi}{2}$, $n \in \{-1, 0, 1, 2\}$ are respectively the gates $\hat{S}^\dagger, \mathbb{1}, \hat{S}$ and $\hat{Z}$. Moreover, assuming the distribution of θ is even, Eq. (3.90) allows to recover the decomposition provided in Eqs. (3.36) and (3.60).

For the $N \geq 3$, the rotations involved in Eq. (3.90) are not Clifford gates, and it is unclear whether this decomposition can be used to provide an efficient sampling scheme.

IV Sampling of Clifford circuits and efficiency

In this section we prove that to obtain an estimation of any first or second-order quantity for a given ansatz up to a precision ϵ and probability $\delta \in [0, 1]$ to meet this precision,

it suffices to sample a number of Clifford approximant circuits $K \sim \log(2\delta)M/\epsilon^2$. By invoking the Gottesman-Knill theorem, we then obtain an estimation of any first- or second-order quantity with a complexity polynomial in both the size of the system and the number of variational parameters of the considered ansatz.

IV.1 Details on the mapping

Here we give details on the mapping of the randomly initialized parameterized circuit to Clifford approximants.

Remark. We use the notations adapted to first-order quantities. The generalization to the second order and the shifted versions is straightforward as it suffices to replace each channel by its doubled and/or shifted version, as done in Sec. I.4.

Assuming that the θ_i are independent from each other, averaging $\mathcal{U}(\boldsymbol{\theta})$ over $\boldsymbol{\theta}$ amount to replace each rotation channel $\mathcal{U}_i(\theta_i)$ by a convex sum of m Clifford unitary channels $\mathcal{U}_{ij}$ with associated weight p_{ij}. Thus $\mathbb{E}_{\boldsymbol{\theta}}\left[\mathcal{U}(\boldsymbol{\theta})(\hat{\rho})\right]$ is replaced by a discrete average over m^M Clifford unitary channels (with $m = 2$ for the 1-fold channel and $m = 4$ for the 2-fold channel):

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\mathcal{U}(\boldsymbol{\theta})(\hat{\rho})\right] = \bigcirc_{i=1}^{M}\left(\sum_{j=1}^{m} p_{ij}\mathcal{U}_{ij}\circ\mathcal{W}_i\right)(\hat{\rho}). \tag{3.92}$$

As we want to sample from that sum, we can define for each i a discrete random variable X_i taking values in $\{1,\dots,m\}$ such that $\mathbb{P}(X_i = j) = p_{ij}$. This represents a choice of a given unitary in the previous convex sum. Gathering these for all k we get a random vector $\boldsymbol{X} = (X_1,\dots,X_M) \in \{1,\dots,m\}^M$ that completely defines a unique unitary $\mathcal{U}(\boldsymbol{X})$ through:

$$\mathcal{U}(j_1,\dots,j_M) = \bigcirc_{i=1}^{M}\mathcal{U}_{ij_i}\circ\mathcal{W}_i. \tag{3.93}$$

Thus we have:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}\left[\mathcal{U}(\boldsymbol{\theta})(\hat{\rho})\right] &= \mathbb{E}_{\boldsymbol{X}}\left[\mathcal{U}(\boldsymbol{X})(\hat{\rho})\right] \\ &= \bigcirc_{i=1}^{M}\left(\sum_{j=1}^{m} p_{ij}\mathcal{U}_{ij}\circ\mathcal{W}_i\right)(\hat{\rho}).\end{aligned} \tag{3.94}$$

The main idea is now to approximate the k-fold channels by an empirical average over K samples of the previous Clifford circuits, namely:

$$\hat{\Phi}(\hat{\rho}) := \frac{1}{K}\sum_{i=1}^{K}\mathcal{U}(\boldsymbol{X}_i)(\hat{\rho}). \tag{3.95}$$

IV.2 Sampling efficiency

Our result relies on classical arguments for the sampling of bounded functions depending on a set of random variables using the McDiarmid's concentration inequality [186, 240], which we remind below.

Definition IV.1 (Bounded difference property)**.** A function $f : \mathcal{X}^M \to \mathbb{R}$ satisfies the bounded difference property if and only if there exists coefficients $(c_i)_{i\in[\![1,M]\!]} \in \mathbb{R}_+^M$ such that $\forall i \in [\![1, M]\!]$, $\forall\, (x_k)_{k\in[\![1,M]\!]}$:

$$\sup_{x_i^{'}\in\mathcal{X}} \left|f(x_1, .., x_i, .., x_M) - f(x_1, .., x_i^{'}, .., x_M)\right| < c_i\,.$$

Theorem IV.1 (McDiarmid's inequality)**.** *Let $f : \mathcal{X}^M \to \mathbb{R}$ satisfy the bounded difference property with bounds $\{c_1, \ldots, c_M\}$, and a random vector $\boldsymbol{X} = (X_1, \ldots, X_M)$ taking values in $\mathcal{X}^M$, then $\forall \epsilon > 0$*

$$\mathbb{P}\left(|f(\boldsymbol{X}) - \mathbb{E}_{\boldsymbol{X}}\left[f(\boldsymbol{X})\right]| \geq \epsilon\right) \leq 2\text{exp}\left(-\frac{2\epsilon^2}{\sum_{i=1}^M c_i^2}\right).$$

We will show that the quantities we want to estimate satisfy the bounded difference property and apply the McDiarmid's inequality to prove that our previous sampling is efficient. In the following we define

$$f(\boldsymbol{x}) = \text{Tr}\left[\mathcal{U}(\boldsymbol{x})(\hat{\rho})\hat{O}\right], \tag{3.96}$$

where $\hat{O}$ is the cost function observable defined in the main text and, as in the previous section, $\mathcal{U}(\boldsymbol{x})$ the unitary channel associated to a given Clifford approximant circuit that is completely specified by a discrete vector $\boldsymbol{x} = (x_1, \ldots, x_i, \ldots, x_M) \in \{1, \ldots, m\}^M$. By the Cauchy-Schwarz inequality for the Hilbert-Schmidt inner product, f is upper-bounded:

$$|f(\boldsymbol{x})| \leq \|\hat{\rho}\|_2 \|\hat{O}\|_2 \tag{3.97}$$

with $\|A\|_2 = \sqrt{\text{Tr}\left[A^\dagger A\right]}$ the Hilbert-Schmidt norm. Defining a second vector for which only the i-th component is changed $\boldsymbol{x'} = (x_1, \ldots, x_i^{'}, \ldots, x_M)$, we get by using the triangle inequality

$$\begin{aligned}\left|f(\boldsymbol{x}) - f(\boldsymbol{x}^{'})\right| &\leq |f(\boldsymbol{x})| + |f(\boldsymbol{x'})| \\ &\leq 2\|\hat{O}\|_2 \|\hat{\rho}\|_2.\end{aligned} \tag{3.98}$$

Hence f satisfies the bounded difference property with $c_i = c = 2\|\hat{O}\|_2\|\hat{\rho}\|_2$, and we can apply McDiarmid's inequality, which gives almost the desired result. To go further, we define

$$\begin{aligned}f_K(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K) &= \sum_{j=1}^K f(x_{j1}, \ldots, x_{jM}) \\ &= \sum_{j=1}^K \text{Tr}\left[\mathcal{U}(\boldsymbol{x}_j)(\hat{\rho})\hat{O}\right] \\ &= K\text{Tr}\left[\hat{\Phi}(\hat{\rho})\hat{O}\right].\end{aligned} \tag{3.99}$$

Clearly, f_K satisfies the bounded difference property with the same bound c [4]. Thus

[4] To see this, we take all x_{ij} equal except for x_{kl}, and it follows that the difference $f_K(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K) - f_K(\boldsymbol{x}_1^{'}, \ldots, \boldsymbol{x}_K^{'})$ is simply $f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k^{'})$.

McDiarmid's inequality applies to f_K, which is a function of KM parameters:

$$\begin{aligned}\mathbb{P}(|f_K(\boldsymbol{X}) - \mathbb{E}_{\boldsymbol{X}}[f_K(\boldsymbol{X})]| \geq K\epsilon) &= \mathbb{P}\left(\left|\frac{1}{K}f_K(\boldsymbol{X}) - \mathbb{E}_{\boldsymbol{X}}\left[\frac{1}{K}f_K(\boldsymbol{X})\right]\right| \geq \epsilon\right) \\ &= \mathbb{P}\left(\left|\mathrm{Tr}\left[\hat{\Phi}(\hat{\rho})\hat{O}\right] - \mathbb{E}_{\boldsymbol{\theta}}\left[\mathrm{Tr}\left[\mathcal{U}(\boldsymbol{\theta})(\hat{\rho})\hat{O}\right]\right]\right| \geq \epsilon\right) \\ &\leq 2\exp\left(-\frac{2K^2\epsilon^2}{KMc^2}\right) \\ &= 2\exp\left(-\frac{K\epsilon^2}{2M\|\hat{\rho}\|_2^2\|\hat{O}\|_2^2}\right).\end{aligned} \tag{3.100}$$

Therefore, choosing a precision $\epsilon > 0$ and a probability $1-\delta \in [0,1]$ to meet this precision, we get

$$\mathbb{P}\left(\left|\mathrm{Tr}\left[\hat{\Phi}(\hat{\rho})\hat{O}\right] - \mathbb{E}_{\boldsymbol{\theta}}\left[\mathrm{Tr}\left[\mathcal{U}(\boldsymbol{\theta})(\hat{\rho})\hat{O}\right]\right]\right| \leq \epsilon\right) \geq 1-\delta \tag{3.101}$$

whenever the number of sampled Clifford circuits K is

$$K \geq \frac{2}{\epsilon^2}\log\left(\frac{2}{\delta}\right) M\|\hat{\rho}\|_2^2\|\hat{O}\|_2^2 = O(M)\ . \tag{3.102}$$

Note that in Eq. (3.100), replacing the observable $\hat{O}$ by its normalized counterpart $\hat{O}/\|\hat{O}\|$ with an associated precision $\tilde{\epsilon}$ gives the same scaling for K, as in that case $\tilde{\epsilon} = \epsilon/\|\hat{O}\|$. Hence we can always work with a normalized observable. However, if one is interested in the scaling with the system size n, we have to consider a sequence of observables $\hat{O}_n$, whose norms can present a particular scaling in n, so the presence of the norm of $\hat{O}$ in Eq. (3.102) allows to keep track of this effect. In many situations of interest, the observables considered scale polynomialy in the system size, and so does K. Finally, one can use the Gottesman-Knill theorem which states that for a Clifford unitary $\hat{U}$ and an observable $\hat{O}$ acting non-trivially on N_O qubits, the expectation value $\mathrm{Tr}\left[|0\rangle\langle 0|^{\otimes n}\hat{U}^\dagger\hat{O}\hat{U}\right]$ can be classically computed with a complexity polynomial in both N_O and the number of qubits n [210]. Our scheme inherits this scaling and we can estimate the gradient variance $\mathrm{Var}_{\boldsymbol{\theta}}[\partial_k C(\boldsymbol{\theta})]$ for each k with a classical computer in a complexity in $O(n^p N_O^q M)$ with M the number of parameters in the variational quantum circuit.

IV.3 Generalizations to the nonconvex cases

Relations to previous works

In this section we extend the previous scheme to more general distributions. The method we present have been proposed by the authors of [235] and [236], and it falls in the more general class quasi-probability methods [241, 242].

We first discuss the scaling of the sampling complexity with the convexity condition relaxed, i.e. where we no longer require the decomposition of the 2-fold channel [Eq. (3.60)] to be a convex sum and only assume that the distribution of θ is even. Then, we study the case of an arbitrary distribution of the rotation angles, which is not necessarily symmetrically distributed. Finally, we show that our previous scheme still applies at the price of an exponential factor in the number of variational parameters M in the number

of Clifford approximant circuits to be sampled. Compared to a brute-force simulation, this method can be used to trade an exponential complexity in the system size for an exponential complexity in the number of variational parameters.

Even distributions

Here we consider distributions of rotation angle θ that are even, but do not satisfy the convexity condition of Eq. (3.65). In this case, our decomposition of the 1-fold channel remains convex while the 2-fold channel becomes a nonconvex sum, hence the coefficients for the Clifford channels can no longer be interpreted as probabilities. We first show how one can still estimate such nonconvex sums via probabilistic sampling [235, 236]. Denoting

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\mathcal{U}(\boldsymbol{\theta})(\hat{\rho})\right] = \bigcirc_{k=1}^{M}\left(\sum_{j=1}^{m} q_{kj}\mathcal{U}_{kj}\circ\mathcal{W}_k\right)(\hat{\rho})\,, \tag{3.103}$$

we hereby assume

$$q_{kj}\in\mathbb{R}\,,\quad \sum_{j=1}^{M} q_{kj} = 1\,,\quad \forall k\,. \tag{3.104}$$

Defining

$$\gamma_k := \Sigma_{j=1}^{M}|q_{kj}|\,,\quad \tilde{p}_{kj} := |q_{kj}|/\gamma_k\,, \tag{3.105}$$

Eq. (3.103) can be rewritten in terms of convex sums:

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\mathcal{U}(\boldsymbol{\theta})\right] = \bigcirc_{k=1}^{M}\sum_{j=1}^{m}\tilde{p}_{kj}\left[\gamma_k\,\mathrm{sgn}(q_{kj})\right]\mathcal{U}_{kj}\circ\mathcal{W}_k\,. \tag{3.106}$$

Similar to Sec. IV.1, we now define the random vector $\tilde{\boldsymbol{X}} = (\tilde{X}_1,\ldots,\tilde{X}_M)\in\{1,\ldots,m\}^M$, with probabilities $\mathbb{P}(\tilde{X}_k = j) = \tilde{p}_{kj}$, and the rescaled random unitary channel $\tilde{\mathcal{U}}(\tilde{\boldsymbol{X}})$ through

$$\tilde{\mathcal{U}}(j_1,\ldots,j_M) = \bigcirc_{k=1}^{M}\left[\gamma_k\,\mathrm{sgn}(q_{kj_k})\right]\mathcal{U}_{kj_k}\circ\mathcal{W}_k\,. \tag{3.107}$$

Therefore, we recover the form of an expectation value similar to Eq. (3.94):

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\mathcal{U}(\boldsymbol{\theta})(\hat{\rho})\right] = \mathbb{E}_{\tilde{\boldsymbol{X}}}\left[\tilde{\mathcal{U}}(\tilde{\boldsymbol{X}})(\hat{\rho})\right]\,. \tag{3.108}$$

This allows us to apply the same arguments as in Sec. IV.2 by considering the function

$$\tilde{f}(\boldsymbol{x}) = \mathrm{Tr}\left[\tilde{\mathcal{U}}(\boldsymbol{x})(\hat{\rho})\hat{O}\right] \tag{3.109}$$

instead of $f(\boldsymbol{x})$ defined in Eq. (3.96). The function bound (3.97) should be rescaled accordingly:

$$\left|\tilde{f}(\boldsymbol{x})\right| \leq \gamma\|\hat{\rho}\|_2\|\hat{O}\|_2\,, \tag{3.110}$$

where the scaling factor is defined as

$$\gamma := \prod_{k=1}^{M}\gamma_k\,. \tag{3.111}$$

The number of sampled Clifford circuits previously derived in Eq. (3.102) should therefore be scaled with the same factor:

$$K \geq \frac{2}{\epsilon^2}\log\left(\frac{2}{\delta}\right)\gamma M\|\hat{\rho}\|_2^2\|\hat{O}\|_2^2\,. \tag{3.112}$$

Note that the factor $\gamma_k \geq 1$ can be regarded as a measure of "nonconvexity" in the decomposition of the k-th channel. In the case of a convex sum, where $q_{kj} > 0,\ \forall k, j$, the scaling factor is simply $\gamma = 1^M = 1$ and we recover the previous results.

We now show that γ_k is upper-bounded. Following our discussion in Sec. III, it suffices to consider the 2-fold channel for a single-qubit Z-rotation, where the decomposition can be possibly nonconvex. Without loss of generality, let us rewrite Eq. (3.60) as

$$\Phi_Z^{(2)}(\hat{\rho}) = q_{k1}\hat{\rho} + q_{k2}(Z\otimes Z)\,\hat{\rho}\,(Z\otimes Z) + q_{k3}CZ\hat{\rho}CZ + q_{k4}CZ_X\hat{\rho}CZ_X \tag{3.113}$$

for some k, where

$$\begin{aligned} q_{k1} &= \mathbb{E}_\theta\left[\frac{1}{4}(1+\cos 2\theta + 2\cos\theta)\right], \\ q_{k2} &= \mathbb{E}_\theta\left[\frac{1}{4}(1+\cos 2\theta - 2\cos\theta)\right], \\ q_{k3} &= q_{k4} = \mathbb{E}_\theta\left[\frac{1}{4}(1-\cos 2\theta)\right]. \end{aligned} \tag{3.114}$$

Defining the non-negative function

$$\varphi(\theta) := \left|\frac{1}{4}(1+\cos 2\theta + 2\cos\theta)\right| + \left|\frac{1}{4}(1+\cos 2\theta - 2\cos\theta)\right| + 2\times\left|\frac{1}{4}(1-\cos 2\theta)\right|, \tag{3.115}$$

We then get

$$\begin{aligned} \gamma_k &= |q_{k1}| + |q_{k2}| + |q_{k3}| + |q_{k4}| \\ &\leq \mathbb{E}_\theta\left[\varphi(\theta)\right] \\ &\leq \mathbb{E}_\theta\left[\sup_{\theta'}\varphi(\theta')\right] \\ &= \sup_{\theta'}\varphi(\theta') \\ &= \frac{5}{4}\,. \end{aligned} \tag{3.116}$$

Here the function $\varphi(\theta)$ reaches its maximum for $\theta = \pm\frac{\pi}{3}, \pm\frac{2\pi}{3}$. Therefore, the factor γ_k reaches its upper bound $\frac{5}{4}$ if the distribution of θ is a sum of Dirac-delta distributions peaked at $\theta = \pm\frac{\pi}{3}$ and/or $\theta = \pm\frac{2\pi}{3}$, in which case we obtain the worst-case scaling of the number of sampled Clifford circuits (3.112):

$$K \geq \frac{2}{\epsilon^2}\log\left(\frac{2}{\delta}\right)\gamma M\|\hat{\rho}\|_2^2\|\hat{O}\|_2^2 = O(\gamma M)\,, \quad \gamma \leq \left(\frac{5}{4}\right)^M. \tag{3.117}$$

Combining the above result with the Gottesman-Knill theorem, for a cost-function observable $\hat{O}$ acting non-trivially on N_O qubits, our scheme implies a complexity of at most $O\left(n^p N_O^q (\frac{5}{4})^M M\right)$ for the estimation of gradient variance $\mathrm{Var}_{\boldsymbol{\theta}}\left[\partial_k C(\boldsymbol{\theta})\right]$ for each k on a

classical computer in the general scenario, where n is the number of qubits, M is the number of parameters in the variational quantum circuit and p, q are some constants inherited from the Gottesman-Knill theorem.

From the above, it is clear that the negativity of the considered quasi-distribution strongly impact the complexity of the method, which has been highlighted by the authors of [235]. More generally, the role of the negativity in general quasi-probability methods have been investigated in [243–245].

Let us now, we extend our scheme to the most generic case, by considering an arbitrary probability distribution for the rotation angles $\boldsymbol{\theta}$, and derive the corresponding sampling complexity. As before, one needs only to consider the one- and two-fold channels for a single-qubit Z-rotation gate.

1-fold channel

Consider the decomposition of the 1-fold channel for a generic distribution that is given by Eq. (3.37), namely:

$$\mathbb{E}_\theta \left[\hat{R}_Z(\theta)\hat{\rho}\hat{R}_Z^\dagger(\theta)\right] = \frac{1+r_1}{2}\mathcal{E}[\mathbb{1}](\hat{\rho}) + \frac{1-r_1}{2}\mathcal{E}[\hat{Z}](\hat{\rho}) + \frac{s_1}{2}\mathcal{E}[\hat{S}^\dagger](\hat{\rho}) - \frac{s_1}{2}\mathcal{E}[\hat{S}](\hat{\rho})\,, \quad (3.118)$$

Following the same procedure as above, this decomposition allows to express the 1-fold channel of the ansatz as a (possibly nonconvex) linear combination of Clifford channels, which can be estimated via sampling. As before the number of required samples should be scaled, according to the nonconvexity of the sum, by a factor $\gamma = \Pi_{k=1}^{M}\gamma_k$ [5]. We now derive an upper bound for $\gamma_k^{(1)}$, the scaling factor associated to a single (the k-th) 1-fold Z-rotation channel. We proceed by applying the same argument as in Eqs. (3.113)-(3.116):

$$\begin{aligned}
\gamma_k^{(1)} &= \left|\frac{1+r_1}{2}\right| + \left|\frac{1-r_1}{2}\right| + \left|\frac{s_1}{2}\right| + \left|-\frac{s_1}{2}\right| \\
&= \left|\mathbb{E}_\theta\left[\frac{1+\cos\theta}{2}\right]\right| + \left|\mathbb{E}_\theta\left[\frac{1-\cos\theta}{2}\right]\right| + \left|\mathbb{E}_\theta\left[\sin\theta\right]\right| \\
&\leq \mathbb{E}_\theta\left[\left|\frac{1+\cos\theta}{2}\right| + \left|\frac{1-\cos\theta}{2}\right| + \left|\sin\theta\right|\right] \\
&\leq \sup_\theta\left\{\left|\frac{1+\cos\theta}{2}\right| + \left|\frac{1-\cos\theta}{2}\right| + \left|\sin\theta\right|\right\} \\
&= 2\,.
\end{aligned} \quad (3.119)$$

This implies that the number of samples $K^{(1)}$ required for the estimation of the generic 1-fold channel [see Eq. (3.117)] scales as

$$K^{(1)} \sim O(\gamma^{(1)}M)\,, \quad \gamma^{(1)} = \prod_{k=1}^{M}\gamma_k^{(1)} \leq 2^M\,. \quad (3.120)$$

Note that the bound derived above depends on the specific choice of the Clifford channels in the decomposition. As the Clifford group does not form a linearly independent set, it should be possible to find a different decomposition that yields a different upper bound and further optimize the complexity.

[5] See definition in Eqs. (3.103)-(3.105) and Eq. (3.111)

2-fold channel

Similar to our treatment with the 1-fold channel, let us derive an upper bound for $\gamma_k^{(2)}$, the scaling factor for the number of samples required for the estimation of the generic 2-fold k-th Z-rotation channel:

$$\begin{aligned}
\gamma_k^{(2)} &= 4\left|\frac{s_2}{8}\right| + 4\left|-\frac{s_2}{8}\right| \\
&\quad + \left|\frac{1+r_2+2r_1}{4}\right| + \left|\frac{1+r_2-2r_1}{4}\right| \\
&\quad + \left|\frac{1-r_2+2s_1}{4}\right| + \left|\frac{1-r_2-2s_1}{4}\right| \\
&= \left|\mathbb{E}_\theta\left[\sin 2\theta\right]\right| \\
&\quad + \left|\mathbb{E}_\theta\left[\frac{1+\cos 2\theta+2\cos\theta}{4}\right]\right| \\
&\quad + \left|\mathbb{E}_\theta\left[\frac{1+\cos 2\theta-2\cos\theta}{4}\right]\right| \\
&\quad + \left|\mathbb{E}_\theta\left[\frac{1-\cos 2\theta+2\sin\theta}{4}\right]\right| \\
&\quad + \left|\mathbb{E}_\theta\left[\frac{1-\cos 2\theta-2\sin\theta}{4}\right]\right|. \\
&\leq \sup_\theta \left\{\left|\sin 2\theta\right|\right. \\
&\quad + \left|\frac{1+\cos 2\theta+2\cos\theta}{4}\right| \\
&\quad + \left|\frac{1+\cos 2\theta-2\cos\theta}{4}\right| \\
&\quad + \left|\frac{1-\cos 2\theta+2\sin\theta}{4}\right| \\
&\quad \left. + \left|\frac{1-\cos 2\theta-2\sin\theta}{4}\right|\right\} \\
&= 1+\sqrt{2}\,.
\end{aligned} \tag{3.121}$$

This implies that the number of samples $K^{(2)}$ required for the estimation of the generic 2-fold channel scales as

$$K^{(2)} \sim O(\gamma^{(2)} M)\,, \quad \gamma^{(2)} = \prod_{k=1}^{M} \gamma_k^{(2)} \leq (1+\sqrt{2})^M\,, \tag{3.122}$$

which is dominant over the complexity of the estimation of the 1-fold channel [Eq. (3.120)] since $1+\sqrt{2} > 2$.

Again, combining the above result with the Gottesman-Knill theorem, for a cost-function observable $\hat{O}$ acting non-trivially on N_O qubits, our scheme implies a complexity of no more than $O\left(n^p N_O^q (1+\sqrt{2})^M M\right)$ for the estimation of gradient variance

$\mathrm{Var}_{\boldsymbol{\theta}}\left[\partial_k C(\boldsymbol{\theta})\right]$ for each k on a classical computer in the most generic case, where n is the number of qubits, M is the number of parameters in the variational ansatz and p, q are some constants inherited from the Gottesman-Knill theorem.

V Conclusions and perspectives

In this chapter, we presented a classically efficient method to estimate first and second-order expectation values for a large class of randomly initialized variational quantum circuits. This includes estimating the average gradient of the cost function and its variance, which can be used to estimate the trainability. Our method applies to the large class of circuits whose architecture is composed of fixed Clifford gates and single-qubit parameterized rotations, provided that the rotation angles are independent and that their distribution are symmetric with respect to an angle $\theta_0 \in \{k\pi/2,\ k \in \mathbb{Z}\}$ and satisfy $\mathbb{E}_\theta\left[\cos^2(\theta-\theta_0)\right] \geq \left|\mathbb{E}_\theta\left[\cos(\theta-\theta_0)\right]\right|$. The method relies on an exact mapping of randomly initialized variational quantum circuits to ensembles of Clifford circuits and on the Gottesman-Knill theorem. We provide rigorous convergence guarantees, and in particular we show that the complexity of the method scales polynomially in both the system size and the number of parameters of the considered ansatz. We investigated the generalization of the proposed scheme to the case of N-fold channels, and showed that the N-fold average of random Z-rotations can be expressed either as a real combination of diagonal Clifford unitaries or as a discrete sum of N-qubits Z-rotations. However, such a decomposition is not unique, and finding a sufficient and necessary condition for the considered N-fold channel to be a Clifford mixed-unitary channel remains an open problem. Solving this problem is of great interest as it could allow to generalize the scheme presented in this work to ansätze with correlated variational parameters.

We believe that such a tool will prove very useful in future applications, as it could be employed to conduct classical optimization of architectures and initialization of large scale variational quantum circuits. As the absence of barren plateaus can be guaranteed by a large enough variance of the gradient, regardless of the exact origin of the potential barren plateaus, this method could be used to certify trainability for system with a very large number of qubits.

General conclusion

In this manuscript, we have explored the rich interface between machine learning and quantum computing, with an emphasis on the effects of noise and decoherence. This work was structured around two main axes of research:

- the study of noisy quantum kernel machines exploiting open quantum systems,
- an efficient classical estimation of the trainability of variational quantum circuits.

In Chap. 1, we provided a brief overview of the theory of open quantum systems. Along the path, we promptly discussed the statistical interpretation of quantum mechanics and we introduced the theoretical tools that were used in the different chapters. We presented the decoherence phenomenon through the analysis of two simple models. The first model involved a single qubit subject to a random classical driving, while the second model replaced the driving by an interaction with a bosonic bath in a thermal state. We derived in details the reduced dynamic of the qubit in both cases, and we discussed the conditions under which this dynamics can be described by a Lindblad master equation. We also discussed the connections between both models as well as the links between the Lindblad equation and the notion of Markov evolution.

Following this introductory chapter, Chap. 2 presented the results of our work [α]. In this work, we introduced a scheme of noisy quantum kernel machine. In this setup, we considered models with an encoding-decoding structure, where the encoding procedure is described by a mapping of the input data to a dynamical map characterizing the evolution of the quantum device. The decoding part then consists in an optimization of a linear combination of observables of the evolved system. We analyzed the expressivity and the generalization capacity of the proposed scheme in the framework of kernel theory. In particular, we introduced a figure of merit based on the eigen-spectrum of the model kernel, the kernel effective rank, that allowed us to quantify the model expressivity. From this, we showed how the noise affecting the device contributes to the reduction of the model expressive power. Using an adequate classical bound on the generalization error, we showed that the noise can play the role of an intrinsic regularization, which helps avoiding over-fitting by diminishing the model expressiveness. This theoretical investigation was supported by a numerical study of models involving driven-dissipative chains of spins subject to decoherence, with an encoding consisting in the application of a series of driving pulses with properly modulated amplitudes. We considered two decoding methods: a first method based on a full tomography of the evolved state, and a second one using series of independent measurements at different times. The obtained numerical results are in agreement with our theoretical heuristics. Surprisingly, these results shows

similar performances for both decoding schemes. In this regard, it would be interesting to investigate further decoding procedures based on repeated measurements in light of the recent breakthroughs associated with the use of random measurements and the classical shadows procedure [246–249]. More generally, the investigation of tailored embedding and decoding strategies able to harness the power of current quantum devices is a promising research avenue [98].

Then, we presented the results of our work [γ] in Chap. 3. There, we provided an efficient classical simulation scheme to estimate the initial trainability of a class of variational quantum algorithms. To derive our method, we studied the 1- and 2-fold quantum channels associated with averages of random single-qubit Z-rotations. Under some conditions on the random angle distribution, we showed that these channels can be decomposed as convex sums of Clifford channels. The existence of such a decomposition can be interpreted as a consequence of an artificial decoherence induced by the random selection of the rotation angle. For a variational circuit composed of fixed Clifford gates and independent single-qubit rotations, this result allows to map the ensemble of unitary associated with random choices of the rotation parameters to an ensemble of Clifford circuits. Relying on a standard concentration inequality, we then showed that this Clifford ensemble can be efficiently sampled to estimate averages of first and second-order quantities measured on the output of the considered variational. Making use of the Gottesman-Knill theorem, we proved the efficiency of this method that can be implemented on a classical computer with a complexity scaling polynomially in both the number of variational parameters and the system size. In particular, our methods provides a scalable way to estimate the average of the gradient of the cost function and its variance, which can be used to evaluate the initial trainability of a given ansatz. The method we proposed can be applied to a wide class of independent random rotations. Using existing quasi-probability sampling methods, we extended our scheme to the case of generic independent rotations, at the price of an additional exponential complexity in the number of rotations. We also investigated the generalization of the results obtained for the 1- and 2-fold Z-rotation channels to the general case of a N-fold channel. However, although we obtained interesting decomposition, finding such a decomposition or proving its nonexistence remains an open problem. Solving this problem is of great interest as it could allow to generalize the presented scheme to ansätze with correlated variational parameters.

The results obtained in Chap. 3 allow to obtain information on the trainability of a given ansatz in a scalable way. This could be very useful to conduct classical optimization of architectures of large scale variational quantum circuits. In particular, reinforcement learning methods have recently proved very useful for multiple task related to quantum computing [82–94], and in particular for the optimization of quantum circuits [90, 91]. However, the training procedure of these algorithms often requires large quantities of data, which could render the presented methods impractical. Our scheme could provide a solution to this scalability issue for the optimization of the architectures of variational ansätze with reinforcement learning. This proposal is the object of current investigations.

A Interaction picture

This appendix briefly recall the transformations associated with the Schrödinger and the interaction pictures. Consider a quantum system which Hamiltonian in the Schrödinger picture is given by

$$\hat{H}_S(t) = \hat{H}_0 + \hat{H}_S^{\text{int}}(t)\,, \tag{A.1}$$

where $\hat{H}_0(t)$ and $\hat{H}_S^{\text{int}}$ are the bare and the interaction Hamiltonian respectively. Denoting $\hat{\rho}_S(t)$ the density operator of the system in Schrödinger's picture, the Liouville equation reads

$$\frac{\mathrm{d}\hat{\rho}_S}{\mathrm{d}t} = -\mathrm{i}\left[\hat{H}_S(t), \hat{\rho}_S(t)\right]\,. \tag{A.2}$$

We choose $t = 0$ as a reference time at which the descriptions in the Schrödinger and in the interaction picture coincide. The unitary operator associated with the evolution for the bare Hamiltonian is given by

$$\hat{U}_0(t) = e^{-\mathrm{i}\hat{H}_0 t}\,. \tag{A.3}$$

Writing $\hat{\rho}_I(t)$ for the system density operator in the interaction picture, we have

$$\hat{\rho}_I(t) = \hat{U}_0^\dagger(t)\hat{\rho}_S(t)\hat{U}_0(t)\,. \tag{A.4}$$

The time evolution in the interaction picture is thus given by

$$\begin{aligned}
\frac{\mathrm{d}\hat{\rho}_I}{\mathrm{d}t} &= \mathrm{i}\left[\hat{H}_0, \hat{\rho}_I(t)\right] + \hat{U}_0^\dagger(t)\frac{\mathrm{d}\hat{\rho}_S}{\mathrm{d}t}\hat{U}_0(t) \\
&= \mathrm{i}\left[\hat{H}_0, \hat{\rho}_I(t)\right] - \mathrm{i}\hat{U}_0^\dagger(t)\left[\hat{H}_0(t) + \hat{H}_S^{\text{int}}, \hat{\rho}_S(t)\right]\hat{U}_0(t) \\
&= \mathrm{i}\left[\hat{H}_0, \hat{\rho}_I(t)\right] - \mathrm{i}\left[\hat{U}_0^\dagger(t)\left(\hat{H}_0(t) + \hat{H}_S^{\text{int}}\right)\hat{U}_0(t), \hat{\rho}_I(t)\right] \\
&= -\mathrm{i}\left[\hat{H}_I(t), \hat{\rho}_I\right]\,,
\end{aligned} \tag{A.5}$$

where

$$\hat{H}_I(t) = \hat{U}_0^\dagger(t)\hat{H}_S^{\text{int}}(t)\hat{U}_0(t) \tag{A.6}$$

is the Hamiltonian in the interaction picture.

B Example of non-Markovian evolution

In this appendix we give an example of a non-Markovian system based on the simple model developed in Sec. III of the main text. The method used in this appendix rely on the formalism of time-convolutionless master equations [137]. Let us start from the general evolution channel of Eq. (1.48), which we recall below:

$$\hat{\rho}(t) = \Lambda_{(t,0)}\left(\hat{\rho}_0\right) = \frac{1}{2}\left(1 + f(t)\right)\hat{\rho}_0 + \frac{1}{2}\left(1 - f(t)\right)\hat{\sigma}_z\hat{\rho}_0\hat{\sigma}_z\,, \tag{B.1}$$

where $f(t) = e^{-\Gamma(t)} = e^{-\frac{1}{2}\int_0^t\int_0^t K(s_1,s_2)\mathrm{d}s_1\mathrm{d}s_2}$. Deriving this equation with respect to the time t gives

$$\frac{\mathrm{d}\hat{\rho}}{\mathrm{d}t} = \frac{\gamma(t)}{2}f(t)\left(\hat{\sigma}_z\hat{\rho}_0\hat{\sigma}_z - \hat{\rho}\right)\,, \tag{B.2}$$

with

$$\gamma(t) = \partial_t\Gamma(t) = \int_0^t K(t,s)\mathrm{d}s\,. \tag{B.3}$$

As a linear map, the channel $\Lambda_{(t,0)}$ is invertible, although its inverse is not necessarily trace preserving nor completely positive. Notice that we have $0 < f(t) < 1$, and a calculation gives

$$\hat{\rho}_0 = \Lambda^{-1}_{(t,0)}\left(\hat{\rho}(t)\right) = \frac{1}{2}\left(1 + \frac{1}{f(t)}\right)\hat{\rho}(t) + \frac{1}{2}\left(1 - \frac{1}{f(t)}\right)\hat{\sigma}_z\hat{\rho}(t)\hat{\sigma}_z\,. \tag{B.4}$$

Injecting this equation in Eq. (B.2), we obtain

$$\frac{\mathrm{d}\hat{\rho}}{\mathrm{d}t} = \frac{\gamma(t)}{2}\left(\hat{\sigma}_z\hat{\rho}(t)\hat{\sigma}_z - \hat{\rho}(t)\right) \tag{B.5}$$

This equation is a time-dependent Lindblad equation with an associated instantaneous dissipation rate $\gamma(t) = \int_0^t K(t,s)\mathrm{d}s$. The jump operator $\hat{\sigma}_z$ belongs to the Pauli basis of the single qubit traceless observables, so that the previous equation is in the so-called standard form. Hence, if $\gamma(t) < 0$ for some time t, the system will exhibits a non-Markovian behavior (in the sense briefly discussed in the main text).

Let us remark that we obtained the Lindblad equation above without using any approximation. In particular, it seems that assuming that $t \gg \tau_c$, one can proceed as in the main text and approximate

$$\int_0^t\int_0^t K(s_1,s_2)\mathrm{d}s_1\mathrm{d}s_2 \simeq \tilde{K}(\omega = 0)t = 2\gamma t\,, \tag{B.6}$$

and recover the Lindblad equation of the main text from the sole assumption that $t \gg \tau_c$, thereby dropping the constrain $\delta t \ll 1/\gamma$. However, the two equations do not describe the same thing. In the main text, we obtained the master equation starting from the evolved state at a time $t' = t + \delta t$

$$\hat{\rho}(t') = \Lambda_{(t',t)}(\hat{\rho}(t))\,. \tag{B.7}$$

This state corresponds to the assumption that the evolution started at time t from an initial state $\hat{\rho}(t)$ independent of the driving. In term of trajectories, the corresponding state is computed as

$$\hat{\rho}(t') = \mathbb{E}\left[\Lambda^{\xi}_{(t',t)}\right](\hat{\rho}(t))\ . \tag{B.8}$$

On the contrary, the master equation we just derived describes the evolution of the system when assuming that the initial state $\hat{\rho}_0$ at $t = 0$ is independent of the driving. In this case, the state at a time t' is given by

$$\hat{\rho}(t') = \begin{cases} \mathbb{E}\left[\Lambda^{\xi}_{(t',t)}\left(\hat{\rho}_\xi(t)\right)\right]\ , \\ \mathbb{E}\left[\Lambda^{\xi}_{(t',0)}\right](\hat{\rho}_0)\ . \end{cases} \tag{B.9}$$

For the two pictures to agree, we must have

$$\mathbb{E}\left[\Lambda^{\xi}_{(t',t)}\left(\hat{\rho}_\xi(t)\right)\right] = \mathbb{E}\left[\Lambda^{\xi}_{(t',t)}\left(\mathbb{E}\left[\hat{\rho}_\xi(t)\right]\right)\right]\ , \tag{B.10}$$

which is approximately ensured (up to the second order in $t' - t = \delta t$) if we work with the coarse grained time scale of the main text. This can be seen as a classical equivalent of the *Born-Markov approximation*, where the Born approximation is the assumption that in the case where the classical noise is replaced by a quantum reservoir, the state of the global system is well approximated by the tensor product of the reduced states of the system and the reservoir.

To obtain a non-Markovian behavior in the previous model, we have to chose a covariance function of the noise that yields a negative rate $\gamma(t)$. A covariance function that enables to do so is the following linear kernel

$$K(t,s) = (t-1)(s-1)\,. \tag{B.11}$$

This kernel provides a well defined covariance, as we have $\forall S = (s_1, \ldots, s_n) \in \mathbb{R}^n_+$, $\forall \boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$:

$$\boldsymbol{x}^T \mathbf{K}_S \boldsymbol{x} = \sum_{i,j=1}^{n} x_i x_j (s_i - 1) j (s_j - 1) = \left(\sum_{i=1}^{n} x_i (s_i - 1)\right)^2 \geq 0\,. \tag{B.12}$$

The associated dissipation rate is given by

$$\gamma(t) = (t-1)\left(\frac{t^2}{2} - 1\right) < 0\,, \quad \forall t \in (1, \sqrt{2})\,. \tag{B.13}$$

Physically, the considered noise presents an anti-correlation for the times greater and lower than 1, i.e.

$$\mathbb{E}\left[\xi(t)\xi(s)\right] < 0\,, \quad 0 \leq s < 1\,,\, t > 1\,. \tag{B.14}$$

The effect of this anti-correlation is to partially cancel the random relative phase acquired by the qubit, thereby restoring a part of the coherence lost under the influence of the random driving. Other examples of non-Markovian behaviors for the present model with non-Gaussian noises have been studied in [250].

C Quantum thermal noise

Characteristic function for the quantum thermal noise

We provide a proof of the form of the characteristic function for a quantum thermal noise. This material is inspired of [251]. Let us recall some notations introduced in Sec. III.2:

$$\begin{aligned}\hat{r}_k(t) &= 2\left(g_k e^{\mathrm{i}\omega_k t}\hat{b}_k^\dagger + g_k^* e^{-\mathrm{i}\omega_k t}\hat{b}_k\right)\\ \hat{r}(t) &= \sum_{k\in\mathbb{N}} \hat{r}_k(t)\end{aligned} \tag{C.1}$$

We also write

$$\int_0^t \hat{r}_k(s)\mathrm{d}s = \frac{1}{\mathrm{i}}\left(\alpha_k(t)\hat{b}_k^\dagger - \alpha_k^*(t)\hat{b}_k\right) . \tag{C.2}$$

where $\alpha_k(t) = 2g_k\left(e^{\mathrm{i}\omega_k t}-1\right)/\omega_k$. The reservoir is assumed to be in a thermal state at temperature $T = 1/\beta$:

$$\hat{\rho}_{\mathrm{th}} = \bigotimes_{k\in\mathbb{N}} \left(1-e^{-\beta\omega_k}\right) e^{-\beta\omega_k \hat{b}_k^\dagger \hat{b}_k} . \tag{C.3}$$

The noise characteristic function can be written

$$\mathrm{Tr}\left[\hat{\rho}_{\mathrm{th}} e^{\mathrm{i}\int_0^t \hat{r}(s)\mathrm{d}s}\right] = \Pi_{k\in\mathbb{N}}\left(\left(1-e^{-\beta\omega_k}\right)\mathrm{Tr}\left[e^{-\beta\omega_k\hat{b}_k^\dagger\hat{b}_k} e^{\mathrm{i}\int_0^t \hat{r}_k(s)\mathrm{d}s}\right]\right) . \tag{C.4}$$

Thus we simply need to determine the single-mode characteristic function, which reads (dropping the index k):

$$\chi_{\mathrm{th}}(\alpha) = \left(1-e^{-\beta\omega}\right)\mathrm{Tr}\left[e^{-\beta\omega\hat{b}_k^\dagger\hat{b}_k} e^{\alpha\hat{b}^\dagger - \alpha^*\hat{b}}\right] . \tag{C.5}$$

Using the Baker-Campbell-Hausdorff formula, we have

$$e^{\alpha\hat{b}^\dagger-\alpha^*\hat{b}} = e^{\alpha\hat{b}^\dagger} e^{-\alpha^*\hat{b}} e^{-\frac{|\alpha|^2}{2}} . \tag{C.6}$$

Inserting a resolution of the identity and expanding the exponentials, the trace appearing in Eq. (C.5) can be rewritten

$$\begin{aligned}\mathrm{Tr}\left[e^{-\beta\omega\hat{b}_k^\dagger\hat{b}_k} e^{\alpha\hat{b}^\dagger} e^{-\alpha^*\hat{b}}\right] &= \sum_{n,m=0}^{+\infty} e^{-\beta\omega}\left\langle n\right| e^{\alpha\hat{b}^\dagger}\left|m\right\rangle\!\left\langle m\right| e^{-\alpha^*\hat{b}}\left|n\right\rangle\\ &= \sum_{n,m,k,l=0}^{+\infty} (-1)^l \frac{\alpha^{*l}}{l!}\frac{\alpha^k}{k!} e^{-\beta\omega}\left\langle n\right|\left(\hat{b}^\dagger\right)^k\left|m\right\rangle\!\left\langle m\right|\left(\hat{b}\right)^l\left|n\right\rangle .\end{aligned} \tag{C.7}$$

Recall that the powers of creation and annihilation operators acts on the Fock states as follow:

$$\begin{cases} \left(\hat{b}\right)^l |n\rangle = \sqrt{\dfrac{n!}{(n-l)!}}\,|n-l\rangle & , \\ \left(\hat{b}^\dagger\right)^k |m\rangle = \sqrt{\dfrac{(m+k)!}{m!}}\,|m+k\rangle & . \end{cases} \tag{C.8}$$

Hence we have

$$\begin{aligned} \langle n|\left(\hat{b}^\dagger\right)^k |m\rangle\langle m|\left(\hat{b}\right)^l |n\rangle &= \sqrt{\frac{n!(m+k)!}{(n-l)!m!}}\,\langle n|m+k\rangle\,\langle m|n-l\rangle \\ &= \frac{n!}{(n-k)!}\delta_{m,n-k}\delta_{k,l} \end{aligned} \tag{C.9}$$

and the previous trace reads

$$\mathrm{Tr}\left[e^{-\beta\omega\hat{b}_k^\dagger\hat{b}_k}e^{\alpha\hat{b}^\dagger}e^{-\alpha^*\hat{b}}\right] = \sum_{n=0}^{+\infty} e^{-\beta\omega n}\sum_{k=0}^{n}(-1)^k\frac{|\alpha|^{2k}}{k!}\binom{n}{k} \tag{C.10}$$

The sum appearing on the right handside of this equation is the Laguerre polynomial of order n:

$$L_n\left(|\alpha|^2\right) = \sum_{k=0}^{n}(-1)^k\frac{|\alpha|^{2k}}{k!}\binom{n}{k}. \tag{C.11}$$

We have

$$\begin{aligned} \mathrm{Tr}\left[e^{-\beta\omega\hat{b}_k^\dagger\hat{b}_k}e^{\alpha\hat{b}^\dagger}e^{-\alpha^*\hat{b}}\right] &= \sum_{n,k=0}^{+\infty} e^{-\beta\omega n}(-1)^k\frac{|\alpha|^{2k}}{k!}\binom{n}{k}\mathbb{1}_{(0,+\infty]}(n-k) \\ &= \sum_{m,k=0}^{+\infty} e^{-\beta\omega(m+k)}(-1)^k\frac{|\alpha|^{2k}}{k!}\binom{m+k}{k} \\ &= \sum_{k=0}^{+\infty} e^{-\beta\omega k}(-1)^k\frac{|\alpha|^{2k}}{k!}\sum_{m=0}^{+\infty}\left(e^{-\beta\omega}\right)^m\binom{k+m}{m} \\ &= \sum_{k=0}^{+\infty} e^{-\beta\omega k}(-1)^k\frac{|\alpha|^{2k}}{k!}\frac{1}{\left(1-e^{-\beta\omega}\right)^{k+1}}, \end{aligned} \tag{C.12}$$

where $\mathbb{1}_{(0,+\infty]}$ is the indicator function of $(0,+\infty]$, $m = n-k$ and we have used the identity

$$\sum_{m=0}^{+\infty} x^m\binom{k+m}{m} = \frac{1}{(1-x)^{k+1}}. \tag{C.13}$$

The single-mode characteristic function can therefore be written

$$\begin{aligned} \chi_{\mathrm{th}}(\alpha) &= \exp\left(-\frac{|\alpha|^2}{2}\left(1+2\frac{e^{-\beta\omega}}{(1-e^{-\beta\omega})}\right)\right) \\ &= \exp\left(-\frac{|\alpha|^2}{2}\left(1+2\langle\hat{n}\rangle_{\mathrm{th}}\right)\right) \\ &= \exp\left(-\frac{|\alpha|^2}{2}\left\langle\{\hat{b}^\dagger,\hat{b}\}\right\rangle_{\mathrm{th}}\right) \end{aligned} \tag{C.14}$$

and the multi-mode characteristic function follows:

$$\mathrm{Tr}\left[\hat{\rho}_{\mathrm{th}} e^{\mathrm{i}\int_0^t \hat{r}(s)\mathrm{d}s}\right] = e^{-\frac{1}{2}\sum_{k\in\mathbb{N}}|\alpha_k|^2\left\langle\{\hat{b}_k^\dagger,\hat{b}_k\}\right\rangle_{\mathrm{th}}}. \tag{C.15}$$

Then, we have

$$\begin{aligned}
\int_0^t\int_0^t \left\langle\hat{r}(s_1)\hat{r}(s_2)\right\rangle_{\mathrm{th}}\mathrm{d}s_1\mathrm{d}s_2 &= \sum_{k,k'\in\mathbb{N}}\int_0^t\int_0^t \left\langle\hat{r}_k(s_1)\hat{r}_{k'}(s_2)\right\rangle_{\mathrm{th}}\mathrm{d}s_1\mathrm{d}s_2 \\
&= \sum_{k\in\mathbb{N}}\left\langle\int_0^t\int_0^t \hat{r}_k(s_1)\hat{r}_k(s_2)\mathrm{d}s_1\mathrm{d}s_2\right\rangle_{\mathrm{th}} \\
&= \sum_{k\in\mathbb{N}}\left\langle\left(\int_0^t \hat{r}_k(s)\mathrm{d}s\right)^2\right\rangle_{\mathrm{th}} \\
&= \sum_{k\in\mathbb{N}}|\alpha_k|^2\left\langle\{\hat{b}_k^\dagger,\hat{b}_k\}\right\rangle_{\mathrm{th}},
\end{aligned} \tag{C.16}$$

where we have used the equalities $\left\langle\hat{r}_k(s)\right\rangle_{\mathrm{th}} = 0$ and $[\hat{r}_k(s),\hat{r}_k(s')] = 0\ \forall s,s'$. Finally, we obtain

$$\mathrm{Tr}\left[\hat{\rho}_{\mathrm{th}} e^{\mathrm{i}\int_0^t \hat{r}(s)\mathrm{d}s}\right] = e^{-\frac{1}{2}\int_0^t\int_0^t\left\langle\hat{r}(s_1)\hat{r}(s_2)\right\rangle_{\mathrm{th}}\mathrm{d}s_1\mathrm{d}s_2}. \tag{C.17}$$

Vacuum and thermal decoherence functions

Here we prove the forms of the vacuum and thermal parts of the decoherence function for the density [1]

$$J(\omega) = \omega e^{-\omega/\Omega}. \tag{C.18}$$

The vaccum and thermal decoherence functions are defined as follow:

$$\begin{aligned}
\Gamma_{\mathrm{vac}}(t) &= \int_0^{+\infty} J(\omega)\frac{t^2}{2}\mathrm{sinc}\left(\frac{\omega t}{2}\right)^2\mathrm{d}\omega, \\
\Gamma_{\mathrm{th}}(t) &= \int_0^{+\infty} J(\omega)\frac{t^2}{2}\mathrm{sinc}\left(\frac{\omega t}{2}\right)^2\left(\coth\left(\frac{\beta\omega}{2}\right)-1\right)\mathrm{d}\omega.
\end{aligned} \tag{C.19}$$

For the vacuum part, we have

$$\begin{aligned}
\Gamma_{\mathrm{vac}}(t) &= \int_0^{+\infty}\frac{(1-\cos(\omega t))}{\omega}e^{-\omega/\Omega}\mathrm{d}\omega \\
&= \int_0^{+\infty}\left(\sum_{k=1}^{+\infty}(-1)^k\frac{t^{2k}}{2k!}\omega^{2k-1}e^{-\omega/\Omega}\right)\mathrm{d}\omega \\
&= \sum_{k=1}^{+\infty}(-1)^k\frac{t^{2k}}{2k!}\int_0^{+\infty}\left(\omega^{2k-1}e^{-\omega/\Omega}\right)\mathrm{d}\omega \\
&= \sum_{k=1}^{+\infty}(-1)^k\frac{(\Omega t)^{2k}}{2k!}\Gamma(2k) \\
&= \frac{1}{2}\sum_{k=1}^{+\infty}\frac{(-\Omega^2t^2)^k}{k} \\
&= \frac{1}{2}\ln\left(1+\Omega^2t^2\right),
\end{aligned} \tag{C.20}$$

[1] We take $A = 1$ in this appendix.

where we have used the power series

$$\cos(x) = \sum_{k=0}^{+\infty} (-1)^k \frac{x^{2k}}{2k!} \qquad \ln(x) = \sum_{k=1}^{+\infty} \frac{x^k}{k}\,, \tag{C.21}$$

and the Gamma function evaluated at $k \in \mathbb{N}$:

$$\Gamma(k) = \int_0^{+\infty} t^{k-1} e^{-t} \mathrm{d}t = (k-1)!\,. \tag{C.22}$$

For the thermal part, we follow [137]. Setting $x = \beta\omega$, we have

$$\begin{aligned} \Gamma_{\mathrm{th}}(t) &= \int_0^{+\infty} \frac{(1-\cos(\omega t))}{\omega} \left(\coth\left(\frac{\beta\omega}{2}\right) - 1\right) e^{-\omega/\Omega} \mathrm{d}\omega \\ &= \frac{1}{\beta} \int_0^{+\infty} \left(\int_0^t \sin\left(\frac{sx}{\beta}\right) \mathrm{d}s\right) \left(\coth\left(\frac{x}{2}\right) - 1\right) e^{-x/(\beta\Omega)} \mathrm{d}x\,. \end{aligned} \tag{C.23}$$

As in the main text, we assume $\beta\Omega \gg 1$, so that for $x \gg \beta\Omega$ we have $\left(\coth\left(\frac{x}{2}\right) - 1\right) \simeq 0$, and the term $e^{-x/(\beta\Omega)}$ can be neglected. In this regime we obtain

$$\begin{aligned} \Gamma_{\mathrm{th}}(t) &\simeq \frac{1}{\beta} \int_0^{+\infty} \left(\int_0^t \sin\left(\frac{sx}{\beta}\right) \mathrm{d}s\right) \left(\coth\left(\frac{x}{2}\right) - 1\right) \mathrm{d}x \\ &= \frac{1}{\beta} \int_0^t \left(\int_0^{+\infty} \sin\left(\frac{sx}{\beta}\right) \left(\coth\left(\frac{x}{2}\right) - 1\right) \mathrm{d}x\right) \mathrm{d}s \\ &= \frac{2}{\beta} \int_0^t \left(\int_0^{+\infty} \frac{\sin(sx/\beta)}{(e^x - 1)} \mathrm{d}x\right) \mathrm{d}s \end{aligned} \tag{C.24}$$

Using the series expansions [2]

$$\sin(x) = \sum_{n=0}^{+\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!}\,, \quad \frac{1}{2}\left(\coth(x) - \frac{1}{x}\right) = \sum_{n=1}^{+\infty} \left(\frac{x}{x^2 + \pi^2 n^2}\right)\,, \tag{C.25}$$

we calculate

$$\begin{aligned} \int_0^{+\infty} \frac{\sin(sx/\beta)}{(e^x - 1)} \mathrm{d}x &= \sum_{n,k=0}^{+\infty} (-1)^n \frac{(s/\beta)^{2n+1}}{(2n+1)!} \int_0^{+\infty} x^{2n+1} e^{-x(k+1)} \mathrm{d}x \\ &= \sum_{n,k=0}^{+\infty} \frac{(-1)^n}{k+1} \left(\frac{s}{\beta(k+1)}\right)^{2n+1} \frac{\int_0^{+\infty} x^{2n+1} e^{-x} \mathrm{d}x}{(2n+1)!} \\ &= \sum_{n,k=0}^{+\infty} \frac{(-1)^n}{k+1} \left(\frac{s}{\beta(k+1)}\right)^{2n+1} \\ &= \sum_{k=0}^{+\infty} \frac{s}{\beta(k+1)^2} \frac{1}{1 + s^2/(\beta(k+1))^2} \\ &= \sum_{k=1}^{+\infty} \frac{s/\beta}{s^2/\beta^2 + k^2} \\ &= \frac{\pi}{2} \left(\coth\left(\frac{\pi s}{\beta}\right) - \frac{\beta}{\pi s}\right)\,. \end{aligned} \tag{C.26}$$

[2] The serie expansion for $\coth(x)$ can be obtained from the Fourier series expansion of e^{ax} over $(-\pi, \pi)$.

From this we get

$$\begin{aligned}\Gamma_{\mathrm{th}}(t) &= \frac{\pi}{\beta}\int_0^t \left(\coth\left(\frac{\pi s}{\beta}\right) - \frac{\beta}{\pi s}\right) \mathrm{d}s \\ &= \int_0^{t/\tau_B} \left(\coth(x) - \frac{1}{x}\right) \mathrm{d}x\,,\end{aligned} \tag{C.27}$$

where we have introduced $\tau_B = \pi/\beta$. This last integral is easily calculated and we obtain

$$\Gamma_{\mathrm{th}}(t) = \ln\left(\frac{\sinh\left(t/\tau_B\right)}{t/\tau_B}\right)\,. \tag{C.28}$$

D Probabilistic model and loss function

This appendix recalls the probabilistic modeling associated with a choice of trial function and loss function. The loss function $l(y, \boldsymbol{x}, \boldsymbol{w})$ represents the discrepancy between the trial function f and the target y for the input $\boldsymbol{x}$, where f depends on the parameters $\boldsymbol{w}$. In general, the choice of the loss and the trail functions is guided by probabilistic considerations. In fact, the loss function can be interpreted as a negative log-likelihood

$$l(y, \boldsymbol{x}, \boldsymbol{w}) = -\log p(y|\boldsymbol{x}, \boldsymbol{w}) \,, \tag{D.1}$$

with $p(y|\boldsymbol{x}, \boldsymbol{w})$ is the conditional probability density of $\boldsymbol{y}$ knowing the input $\boldsymbol{x}$ and for a given set of weights $\boldsymbol{w}$. With this identification, choosing a loss function amount to chose a model for the conditional probability distribution of the targets knowing the inputs [105, 252]. In a supervised setting, the optimal weights are then chosen as to minimize the average loss function on the training set:

$$\mathcal{L}\left(\boldsymbol{w}|\mathcal{S}\right) = \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} l(y_i, \boldsymbol{x}_i, \boldsymbol{w}) \,. \tag{D.2}$$

In this view, the optimization procedure of the trial function is a maximum likelihood estimation for the conditional probability density $p(y|\boldsymbol{x})$ with respect to the training set $\mathcal{S} = \{(\boldsymbol{x}_i, y_i) \,|\, i = 1, \ldots, N_{\text{train}}\}$.

For example, for a logistic regression model associated with a binary classification task with target space $\mathcal{Y} = \{0, 1\}$, the loss function is given by the cross-entropy

$$l(y, \boldsymbol{x}, \boldsymbol{w}) = -y \log\left(f(\boldsymbol{x})\right) - (1 - y) \log\left(1 - f(\boldsymbol{x})\right) \,, \tag{D.3}$$

with the associated trial function

$$f(\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w}^T \boldsymbol{x}}} \,. \tag{D.4}$$

Using Eq. (D.1), we have

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = f(\boldsymbol{x})^y (1 - f(\boldsymbol{x}))^{1-y} \,, \tag{D.5}$$

from which we can identify

$$\begin{cases} p(y = 1|\boldsymbol{x}, \boldsymbol{w}) = f(\boldsymbol{x}) & , \\ p(y = 0|\boldsymbol{x}, \boldsymbol{w}) = 1 - f(\boldsymbol{x}) & . \end{cases} \tag{D.6}$$

This usually motivates the choice of the cross-entropy loss function for classification problems, but other types of loss can be chosen.

From the above, using a least square loss function amount to modeling the conditional probability of the targets knowing the inputs as a Gaussian distribution of the form

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(y - \boldsymbol{w}^T\boldsymbol{x}\right)\right) . \tag{D.7}$$

In that case, adding a l2-regularization to the least square loss can be seen as a choice of a Gaussian prior distribution for the weights [252]. In fact, using Bayes rules we have

$$p(\boldsymbol{w}|y, \boldsymbol{x}) = \frac{p(y|\boldsymbol{x}, \boldsymbol{w})p(\boldsymbol{x}, \boldsymbol{w})}{p(y|\boldsymbol{x})} . \tag{D.8}$$

Assuming $\boldsymbol{w}$ to be independent of $\boldsymbol{x}$, we get

$$\log(p(\boldsymbol{w}|y, \boldsymbol{x})) = \log(p(y|\boldsymbol{x}, \boldsymbol{w})) + \log(p(\boldsymbol{w})) + C(\boldsymbol{x}, y) , \tag{D.9}$$

where the terms independent of $\boldsymbol{w}$ are absorbed in $C(\boldsymbol{x}, y)$. Identifying the left-hand side of this equation with the opposite of the l2-regularized least square loss (2.5), and injecting Eq. (D.7) in the right-hand side, we obtain:

$$\log(p(\boldsymbol{w})) \propto -\frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w} , \tag{D.10}$$

which is the desired result.

E Intercept and kernel centering

The initial optimization problem is to find a weight vector $\boldsymbol{w} = (b, w_1, \ldots, w_P)^T$ that minimizes the regularized loss function of Eq. (2.5). We slightly change our notation and drop the first constant term of the embedding map $\boldsymbol{\phi}(\boldsymbol{x})$ to explicitly seperate the bias b from the weight $\boldsymbol{w}$, so that the loss can be rewritten:

$$\mathcal{L}\left(\boldsymbol{w}, b \mid \mathcal{S}\right) = \frac{1}{2N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} \left(y_i - \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_i) - b\right)^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2 \,. \tag{E.1}$$

The optimal intercept b is found by imposing $\frac{\partial \mathcal{L}}{\partial b} = 0$. The solution reads:

$$b^* = \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} y_i - \boldsymbol{w}^T \left(\frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} \boldsymbol{\phi}(\boldsymbol{x}_i) \right) . \tag{E.2}$$

We see that the optimal intercept consists of two terms: one that has the effect of centering the labels, while the other centers the features. Assuming the dataset we use are balanced, we have $\sum_{i=1}^{N_{\text{train}}} y_i \simeq 0$. Plugging back the optimal intercept into the previous regularized loss function, we get a new effective loss:

$$\mathcal{L}^*(\boldsymbol{w} \mid \mathcal{S}) = \frac{1}{2N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} \left(y_i - \boldsymbol{w}^T \left(\boldsymbol{\phi}(x_i) - \mathbb{E}_{\hat{p}}\left[\boldsymbol{\phi}\right]\right)\right)^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2 \,. \tag{E.3}$$

If the data are not balanced one can simply replace the labels y_i by their centered counterpart $y_i' = y_i - \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} y_i$ such that $\sum_{i=1}^{N_{\text{train}}} y_i' = 0$. Note that this might lead to issues when using the accuracy metric with unbalanced labels. This issue can be fixed, e.g. by changing the metric used for a balanced one. Thus, working with the quantum kernel without regularizing the intercept term is equivalent to working with the centered kernel and centered labels.

F Kernel methods in a nutshell

This appendix provides some elementary results on kernel methods. The discussion in this appendix is inspired by the references [105, 180, 186, 253]. Kernel methods allow to extend algorithms using linear models to a non-linear setup, by transforming (sometimes implicitly) the original input data into new features in a non-linear way. Here we illustrate some basic results associated with these methods on a simple example of a linear ridge regression of the same type as the one used in the main text.

Linear Ridge Regression

Recall that we work in a supervised learning setup, where one tries to represent a relation between an input variable $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^{M+1}$ and a target variable $y \in \mathcal{Y} \subset \mathbb{R}$ based upon a training set $\mathcal{S} = \{(\boldsymbol{x}_i, y_i) \,|\, i = 1, \dots, N_{\text{train}}\}$. Assuming that the relationship between inputs and targets is affine, we can use a linear model whose trial function takes the form [1]:

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} \,. \tag{F.1}$$

We consider a simple linear ridge regression model [105], for which the loss function is a l2-regularized least-square loss that reads

$$\mathcal{L}\left(\boldsymbol{w} \mid \mathcal{S}\right) := \frac{1}{2N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} \left(y_i - \boldsymbol{w}^T \boldsymbol{x}_i\right)^2 + \frac{\lambda}{2} \boldsymbol{w}^T \boldsymbol{w} \,. \tag{F.2}$$

Define the $(M+1) \times N_{\text{train}}$ input features matrix $\mathbf{X}$ whose columns are the input features $\boldsymbol{x}_i$. The optimal weights $\boldsymbol{w}^*$ [2] for this model can be derived analytically, and we have

$$\boldsymbol{w}^* = \left(\mathbf{X}\mathbf{X}^T + N_{\text{train}} \lambda \mathbb{1}\right)^{-1} \mathbf{X}\boldsymbol{y} \,. \tag{F.3}$$

Primal and dual pictures

The optimal function f^* corresponding the weights $\boldsymbol{w}^*$ can be written as a linear combination of the form

$$f^*\left(\boldsymbol{x}\right) = \boldsymbol{\alpha}^{*T} \mathbf{X}^T \boldsymbol{x} = \sum_{i=1}^{N_{\text{train}}} \alpha_i^* \boldsymbol{x}_i^T \boldsymbol{x} \,, \tag{F.4}$$

[1] For the sake of simplicity, we proceed as in Sec. I.3 and assume that the inputs $\boldsymbol{x}$ have a constant first coordinate equal to one, so that the intercept can be absorbed in the weights vector.

[2] In this appendix the $*$ symbol denotes optimal quantities and should not be confused with a complex conjugate.

	Primal picture	**Dual picture**
Trial function	$f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$	$f(\boldsymbol{x}) = \boldsymbol{\alpha}^T\mathbf{X}^T\boldsymbol{x}$
Model parameters	$\boldsymbol{w} = \mathbf{X}\boldsymbol{\alpha}$	$\boldsymbol{\alpha}$
Regularization term	$\frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w}$	$\frac{\lambda}{2}\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha}$
Dimension	$M+1$	N_{train}

Table F.1: Primal and dual picture for the Linear Ridge Regression model.

where we have [3]

$$\boldsymbol{\alpha}^* = \left(\mathbf{X}^T\mathbf{X} + N_{\text{train}}\lambda\mathbb{1}\right)^{-1}\boldsymbol{y}\,, \tag{F.5}$$

such that

$$\boldsymbol{w}^* = \mathbf{X}\boldsymbol{\alpha}^*\,. \tag{F.6}$$

As a consequence, the optimal trial function can be obtained by solving another optimization problem expressed in terms of the new weights $\boldsymbol{\alpha}$, namely by minimizing the loss function

$$\tilde{\mathcal{L}}\left(\boldsymbol{\alpha} \mid \mathcal{S}\right) := \frac{1}{2N_{\text{train}}}\sum_{i=1}^{N_{\text{train}}}\left(y_i - \boldsymbol{\alpha}^T\mathbf{X}^T\boldsymbol{x}_i\right)^2 + \frac{\lambda}{2}\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha}\,, \tag{F.7}$$

where we have introduced the *kernel matrix* $\mathbf{K} = \mathbf{X}^T\mathbf{X}$. Both optimization models are equivalents, and they define the so-called *primal and dual pictures*. These are summarized in Tab. F.1. The optimization problems in these two pictures have different dimensions. Depending on the problem at hand, one might prefer to work in a picture or the other, usually choosing the picture with the lowest dimension. Crucially, the optimization problem in the dual picture only requires the knowledge of the kernel matrix $\mathbf{K}$, while the primal picture optimization rely on the knowledge of input features.

Feature space embedding

A straightforward way to generalize the ridge regression model to a non-linear setup is to transform the original input variables $\boldsymbol{x} \in \mathcal{X}$ into new variables $\boldsymbol{\phi}(\boldsymbol{x})$ belonging to some feature space $\mathcal{F}$. This transformation is achieved through a non-linear *feature space embedding* $\boldsymbol{\phi} : \mathcal{X} \mapsto \mathcal{F}$. This is the approach followed in the main text. Assuming that the feature space $\mathcal{F}$ is a finite dimensional inner-product space, the generalization then simply consists in replacing the inner products on $\mathbb{R}^{M+1}$ by the inner product on $\mathcal{F}$. In particular, denoting $\left\langle\,,\,\right\rangle_{\mathcal{F}}$ the inner product on $\mathcal{F}$, the trial function reads

$$f(\boldsymbol{x}) = \left\langle\tilde{\boldsymbol{w}}, \boldsymbol{\phi}(\boldsymbol{x})\right\rangle_{\mathcal{F}}, \quad \tilde{\boldsymbol{w}} \in \mathcal{F}\,, \tag{F.8}$$

[3] This is a direct consequence of the trivial identity $\mathbf{X}(\mathbf{X}^T\mathbf{X} + \mathbb{1}) = (\mathbf{X}\mathbf{X}^T + \mathbb{1})\mathbf{X}$.

and the kernel matrix becomes

$$\mathbf{K}_{ij} = \Big\langle \boldsymbol{\phi}(\boldsymbol{x}_i), \boldsymbol{\phi}(\boldsymbol{x}_j) \Big\rangle_{\mathcal{F}} . \tag{F.9}$$

If the dimension of the feature space is too large, the optimization of this model in the primal picture can become impractical. In this case, the dual picture offers a valuable alternative.

Kernel functions, RKHS and the Mercer theorem

Another way to generalize the linear ridge regression algorithm to a non-linear setup is to work in the dual picture. In this approach, instead of using a transformation of the input variables, we replace the kernel matrix. To do so, one uses a *kernel function* $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ that yields a positive symmetric matrix

$$\mathbf{K}_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{F.10}$$

for every set of inputs $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N) \in \mathcal{X}^N$, $N \in \mathbb{N}$ [4]. The corresponding trial function in the dual picture is given by

$$f(\boldsymbol{x}) = \sum_{i=1}^{N_{\text{train}}} \alpha_i k(\boldsymbol{x}, \boldsymbol{x}_i) . \tag{F.11}$$

This function belongs to the real vector space

$$\tilde{\mathcal{H}} = \text{Span}\left(\{ f : \boldsymbol{x} \mapsto k(\boldsymbol{x}, \boldsymbol{x'}) \,|\, \boldsymbol{x'} \in \mathcal{X} \} \right) , \tag{F.12}$$

which be equipped with the following inner product:

$$\begin{aligned} \Big\langle h, g \Big\rangle_{\tilde{\mathcal{H}}} &= \sum_{i,j} \alpha_i \beta_j k(\boldsymbol{x}_i, \boldsymbol{x_j}) , \\ \text{for} \quad h : \boldsymbol{x} \mapsto \sum_i \alpha_i k(\boldsymbol{x}, \boldsymbol{x}_i) , &\quad g : \boldsymbol{x} \mapsto \sum_j \beta_j k(\boldsymbol{x}, \boldsymbol{x}_j) . \end{aligned} \tag{F.13}$$

The completion of $\tilde{\mathcal{H}}$ yields the *Reproducing Kernel Hilbert Space* (RKHS) [5] associated with k [253], which we write $\mathcal{H}$. Introducing the RKHS allows to see the choice of the kernel k in the dual picture as a choice of a feature space embedding in the primal picture. In fact, identifying $\mathcal{F} = \mathcal{H}$ in the previous paragraph, we can define the feature space embedding

$$\boldsymbol{\Phi} : \boldsymbol{x} \mapsto k(\,\cdot\,, \boldsymbol{x}) \in \mathcal{H} . \tag{F.14}$$

The kernel function can then be rewritten

$$k(\boldsymbol{x}, \boldsymbol{x'}) = \Big\langle \boldsymbol{\Phi}(\boldsymbol{x}), \boldsymbol{\Phi}(\boldsymbol{x'}) \Big\rangle_{\mathcal{H}} , \tag{F.15}$$

and the trial function is the primal picture reads:

$$f(\boldsymbol{x}) = \Big\langle \boldsymbol{w}, \boldsymbol{\Phi}(\boldsymbol{x}) \Big\rangle_{\mathcal{H}} , \quad \boldsymbol{w} \in \mathcal{H} . \tag{F.16}$$

[4] This is the same requirement as for the covariance function introduced in Sec. III of Chap. 1.

[5] By the *Moore-Aronszajn* theorem, the RKHS associated to a kernel k is unique [180].

Feature space	$\mathcal{H}$	$\ell^2(\mathbb{R})$
Embedding map	$\boldsymbol{\phi}(\boldsymbol{x}) = k(\cdot, \boldsymbol{x})$	$\boldsymbol{\phi}(\boldsymbol{x}) = \left(\sqrt{\lambda_i}\varphi_i(\boldsymbol{x})\right)_{i\in[\![1,N_\mathcal{H}]\!]}$
$k(\boldsymbol{x}, \boldsymbol{x'})$	$\left\langle \boldsymbol{\Phi}(\boldsymbol{x}), \boldsymbol{\Phi}(\boldsymbol{x'}) \right\rangle_\mathcal{H}$	$\sum_{i=1}^{N_\mathcal{H}} \lambda_i \varphi_i(\boldsymbol{x})\varphi_i(\boldsymbol{x'})$

Table F.2: Feature space embeddings associated with a kernel function k.

Therefore the dual picture generalization allows, through the introduction of an adequate kernel function, to effectively optimize a linear form on a feature space of potentially infinite dimension. The RKHS $\mathcal{H}$ being an abstract functional space, the previous realization of the feature space embedding associated with a choice of k is not practical. The *Mercer theorem* provides a more concrete embedding [180, 253]. Let $N_\mathcal{H}$ denotes the dimension of $\mathcal{H}$. The Mercer theorem states that, under some quite general assumptions, the kernel k admits an eigen-decomposition

$$k(\boldsymbol{x}, \boldsymbol{x'}) = \sum_{i=1}^{N_\mathcal{H}} \lambda_i \varphi_i(\boldsymbol{x})\varphi_i(\boldsymbol{x'}) \,, \tag{F.17}$$

where $(\varphi_i)_{i\in[\![1,N_\mathcal{H}]\!]}$ is an orthonormal basis of $L^2_p(\mathcal{X})$ [6] and $\lambda_{i+1} \geq \lambda_i \geq 0$. Thus, provided that the Mercer theorem is valid, the kernel k yields an embedding in the space of square-integrable real sequences $\mathcal{F} = \ell^2(\mathbb{R})$:

$$\boldsymbol{\Phi} : \boldsymbol{x} \mapsto \left(\sqrt{\lambda_i}\varphi_i(\boldsymbol{x})\right)_{i\in[\![1,N_\mathcal{H}]\!]} \,. \tag{F.18}$$

The two embedding maps presented here are summarized in Tab. F.2.

The representer theorem and the kernel trick

The ongoing discussion can be generalized to a large class of models beyond the ridge regression. Indeed, the *representer theorem* [186] guarantees that for a generic loss function $\mathcal{L} : \mathbb{R}^N \mapsto \mathbb{R}_+ \cup \{+\infty\}$, a non-decreasing function $G : \mathbb{R} \mapsto \mathbb{R}$ and a kernel k on $\mathcal{X}$ with an associated RKHS $\mathcal{H}$, the optimization problem

$$\underset{f\in\mathcal{H}}{\operatorname{argmin}} \left(\left(\mathcal{L}\left(f\left(\boldsymbol{x}_1\right), \ldots, f\left(\boldsymbol{x}_N\right)\right)\right) + G\left(\|f\|_\mathcal{H}\right)\right) , \tag{F.19}$$

admits a solution of the form

$$f^*(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i^* k(\boldsymbol{x}, \boldsymbol{x}_i) \,. \tag{F.20}$$

[6] The Hilbert space of read square-integrable functions on $\mathcal{X}$ for the probability density $p(\boldsymbol{x})$, which inner-product is given by

$$\langle h, g \rangle_{L^2_p(\mathcal{X})} = \int_\mathcal{X} h(\boldsymbol{x}) g(\boldsymbol{x}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \,.$$

Hence, provided that the trial function lies in a RKHS $\mathcal{H}$ and that the regularization term is a suitable function of $\|f\|_{\mathcal{H}}$, one can always perform the optimization in the dual picture. Moreover, for such optimization problem the kernel can be freely chosen, a feature known as the *kernel trick*.

G Expressivity and generalization for noisy quantum kernels

I Expressivity and kernel effective rank

To measure the ability of a kernel k to learn a function $y(\boldsymbol{x})$, we have introduced in Eq. (2.24) the *kernel target alignment* $A(k, y)$. We then defined the *kernel effective rank* R_{eff} by considering a set of orthonormal basis functions $\{g_i\}$, that gives the following equalities:

$$\begin{aligned}\sqrt{R_{\text{eff}}(k)} &= \sum_j A(k, g_j) \\ &= \frac{1}{(\sum_i \lambda_i^2)^{1/2}} \sum_j \sum_i \lambda_i \mathbb{E}_p \left[\psi_i(\boldsymbol{x}) g_j(\boldsymbol{x})\right]^2 \\ &= \frac{1}{(\sum_i \lambda_i^2)^{1/2}} \sum_i \lambda_i \mathbb{E}_p \left[\psi_i(\boldsymbol{x})^2\right] \\ &= \frac{\sum_i \lambda_i}{(\sum_i \lambda_i^2)^{1/2}} .\end{aligned} \tag{G.1}$$

Note that the final expression concerns only the spectrum of the kernel and is independent of the choice of the basis functions $\{g_i\}$. From the Cauchy-Schwarz inequality, we have

$$R_{\text{eff}}(k) \leq |\{\lambda_i \neq 0\}| , \tag{G.2}$$

where the equality is attained if and only if all non-zero eigenvalues of the kernel are equal. Therefore, it provides information about the flatness of the spectrum of the kernel.

Given a training sample of size N_{train}, the kernel spectrum can be empirically computed using the $N_{\text{train}} \times N_{\text{train}}$ kernel matrix $\mathbf{K}$ associated to the kernel k, whose entries are $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. The eigenvalues λ_i of the kernel k can then be approximated by those of the matrix $\mathbf{K}/N_{\text{train}}$ [254]. For the centered quantum kernel k_c with the associated kernel matrix $\mathbf{K}_c$, we can compute the effective rank empirically as:

$$\sqrt{R_{\text{eff}}(\mathbf{K}_c)} = \frac{\text{Tr}\left[\mathbf{K}_c\right]}{\sqrt{\text{Tr}\left[\mathbf{K}_c^2\right]}} = \frac{\sum_i \hat{\lambda}_i}{\sqrt{\sum_i \hat{\lambda}_i^2}} , \tag{G.3}$$

where the $\hat{\lambda}_i$ are the empirical eigenvalues [1].

[1] The hat symbol is used for estimators such as the empirical eigenvalues, as customary in statistical theory. The hat must not confused with the one used for the quantum operators

The numerator can be expressed using the empirical kernel eigenobservables. In order to keep light notations we use the same notations as in the main text, i.e. the empirical kernel eigenobservables are denoted $\hat{E}_i$. Whether this notation refers to the exact or the empirical observable should be clear from the context. We have for the numerator:

$$\sum_i \hat{\lambda}_i = \mathbb{E}_{\hat{p}}\left[\sum_i \delta\langle \hat{E}_i\rangle_{\boldsymbol{x}}^2\right] = \mathbb{E}_{\hat{p}}\left[\sum_i \mathrm{Tr}\left[\delta\hat{\rho}(\boldsymbol{x})\hat{E}_i\right]^2\right] . \tag{G.4}$$

Since $\mathrm{Tr}\left[\delta\hat{\rho}(\boldsymbol{x})\right] = 0$, $\delta\hat{\rho}(\boldsymbol{x})$ can be decomposed onto the eigenobservable basis $\{\hat{E}_i\}$ through the expression:

$$\delta\hat{\rho}(\boldsymbol{x}) = \sum_i \mathrm{Tr}\left[\delta\hat{\rho}(\boldsymbol{x})\hat{E}_i\right]\hat{E}_i\,. \tag{G.5}$$

Consequently, the squared Hilbert-Schmidt norm reads:

$$\mathrm{Tr}\left[\delta\hat{\rho}(\boldsymbol{x})^2\right] = \sum_i \mathrm{Tr}\left[\delta\hat{\rho}(\boldsymbol{x})\hat{E}_i\right]^2 . \tag{G.6}$$

Eq. (G.4) therefore becomes:

$$\begin{aligned}\sum_i \hat{\lambda}_i &= \mathbb{E}_{\hat{p}}\left[\mathrm{Tr}\left[\delta\hat{\rho}(\boldsymbol{x})^2\right]\right] \\ &= \mathrm{Tr}\left[\mathbb{E}_{\hat{p}}\left[\left(\hat{\rho}(\boldsymbol{x}) - \mathbb{E}_{\hat{p}}\left[\hat{\rho}(\boldsymbol{x})\right]\right)^2\right]\right] \\ &= \mathbb{E}_{\hat{p}}\left[\mathrm{Tr}\left[\hat{\rho}(\boldsymbol{x})^2\right]\right] - \mathrm{Tr}\left[\mathbb{E}_{\hat{p}}\left[\hat{\rho}(\boldsymbol{x})\right]^2\right] ,\end{aligned} \tag{G.7}$$

giving Eq. (2.29) in the main text (as the same relation holds between the true eigenvalues λ_i and the distribution p). This quantity can also be written in terms of the measured observables (note that $\mathcal{O} = \mathcal{B}$ for a quantum kernel):

$$\begin{aligned}\sum_i \hat{\lambda}_i =& \frac{\mathrm{Tr}\left[\mathbf{K}_c\right]}{N_{\text{train}}} \\ =& \frac{1}{N_{\text{train}}}\sum_i k_c(\boldsymbol{x}_i, \boldsymbol{x}_i) \\ =& \frac{1}{N_{\text{train}}}\sum_i\sum_k \delta\phi_k(\boldsymbol{x}_i)\delta\phi_k(\boldsymbol{x}_i) \\ =& \sum_k\left(\frac{1}{N_{\text{train}}}\sum_i \delta\phi_k(\boldsymbol{x}_i)\delta\phi_k(\boldsymbol{x}_i)\right) \\ =& \sum_k \mathbb{E}_{\hat{p}}\left[\delta\phi_k(\boldsymbol{x})^2\right] \\ =& \sum_k \mathbb{E}_{\hat{p}}\left[\delta\langle\hat{O}_k\rangle_{\boldsymbol{x}}^2\right] \\ =& \sum_k \mathrm{Var}_{\hat{p}}\left[\langle\hat{O}_k\rangle_{\boldsymbol{x}}\right] .\end{aligned} \tag{G.8}$$

Similarly, in the denominator of Eq. (G.3), we get:

$$\begin{aligned}\sum_i \hat{\lambda}_i^2 &= \frac{\mathrm{Tr}\left[\mathbf{K}_c^2\right]}{N_{\mathrm{train}}^2} \\ &= \frac{1}{N_{\mathrm{train}}^2}\sum_{i,j} k_c\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)^2 \\ &= \frac{1}{N_{\mathrm{train}}^2}\sum_{i,j}\left(\sum_k \delta\phi_k(\boldsymbol{x}_i)\delta\phi_k(\boldsymbol{x}_j)\right)^2 \\ &= \sum_{k,l}\left(\frac{1}{N_{\mathrm{train}}}\sum_i \delta\phi_k(\boldsymbol{x}_i)\delta\phi_l(\boldsymbol{x}_i)\right)^2 \\ &= \sum_{k,l}\mathbb{E}_{\hat{p}}\left[\delta\langle\hat{O}_k\rangle_{\boldsymbol{x}}\delta\langle\hat{O}_l\rangle_{\boldsymbol{x}}\right]^2 \\ &= \sum_{k,l}\mathrm{Cov}_{\hat{p}}\langle\hat{O}_k\rangle_{\boldsymbol{x}}, \langle\hat{O}_l\rangle_{\boldsymbol{x}}{}^2 \,.\end{aligned} \tag{G.9}$$

Finally, we get the general expression:

$$\sqrt{R_{\mathrm{eff}}(\mathbf{K}_c)} = \frac{\sum_{i=1}^P \mathrm{Var}_{\hat{p}}\langle\hat{O}_i\rangle_{\boldsymbol{x}}}{\left(\sum_{i,j=1}^P \mathrm{Cov}_{\hat{p}}\langle\hat{O}_i\rangle_{\boldsymbol{x}}, \langle\hat{O}_j\rangle_{\boldsymbol{x}}{}^2\right)^{\frac{1}{2}}} \,. \tag{G.10}$$

Note that this relation also holds for the true (non-empirical) effective rank $R_{\mathrm{eff}}(k_c)$ provided that the variances and the covariances are taken with respect to the true probability distribution p instead of the empirical one $\hat{p}$.

II Generalization and Rademacher complexity

Here we give the detailed derivation of Eq. (2.35) using methods of statistical learning theory applied to the specific case of a noisy centered quantum kernel [186].

As in the main text, we work in a standard statistical learning setup, where the inputs $\boldsymbol{x} \in \mathcal{X}$ are considered as a random variable following a probability distribution $p(\boldsymbol{x})$. We define the target function $y : \mathcal{X} \mapsto \mathcal{Y}$ that assigns to each input its right label. We will consider the case of a binary classification, for which $\mathcal{Y} = \{-1, 1\}$. In practice the true distribution p of the inputs is unknown and during the training we only have access to a finite training dataset $\mathcal{S} = \{(\boldsymbol{x}_i, y_i) \mid i = 1, \ldots, N_{\mathrm{train}}\}$. The elements of the dataset are considered as realizations of a set of independent and identically distributed random variables following p. The empirical distribution associated to this training set is given by:

$$\hat{p}(\boldsymbol{x}) = \frac{1}{N_{\mathrm{train}}}\sum_{i=1}^{N_{\mathrm{train}}} \delta(\boldsymbol{x} - \boldsymbol{x}_i) \,. \tag{G.11}$$

We rely on this empirical distribution to evaluate expectations of any function $f(\boldsymbol{x})$, namely:

$$\mathbb{E}_p\left[f(\boldsymbol{x})\right] = \int_{\mathcal{X}} f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \,. \tag{G.12}$$

The expectation value is approximated by its empirical counterpart:

$$\mathbb{E}_{\hat{p}}\left[f(\boldsymbol{x})\right] = \int_{\mathcal{X}} f(\boldsymbol{x})\hat{p}(\boldsymbol{x})d\boldsymbol{x} = \frac{1}{N_{\text{train}}}\sum_{i=1}^{N_{\text{train}}} f(\boldsymbol{x}_i) \tag{G.13}$$

A common question in statistical learning is to know how a model trained on a given set of data will perform on any other set of unseen data. For a binary classification task with balanced data one can use the accuracy as measure of the model performance. Given a trial function $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x})$ that has been optimized using the training set $\mathcal{S}$, we define the corresponding prediction function as $\tilde{f}(\boldsymbol{x}) = \text{sign}[f(\boldsymbol{x})]$. An input $\boldsymbol{x}$ is correctly classified if $\tilde{f}(\boldsymbol{x}) = y(\boldsymbol{x})$. The true accuracy $\mathcal{A}^*(f)$ is defined as the probability that any input in $\mathcal{X}$ is correctly classified by $\tilde{f}$:

$$\begin{aligned}\mathcal{A}^*(f) &= \mathbb{E}_p\left[\mathbb{1}_{y(\boldsymbol{x})=\tilde{f}(\boldsymbol{x})}\right] = \mathbb{E}_p\left[\mathbb{1}_{y(\boldsymbol{x})f(\boldsymbol{x})\geq 0}\right] \\ &= 1 - \mathbb{E}_p\left[\mathbb{1}_{y(\boldsymbol{x})f(\boldsymbol{x})\leq 0}\right] = 1 - \mathcal{R}^*(f)\,,\end{aligned} \tag{G.14}$$

where we define the risk (also called error or inaccuracy) as $\mathcal{R}^*(f) = 1 - \mathcal{A}^*(f)$. The corresponding empirical quantities $\mathcal{A}(f)$ and $\mathcal{R}(f)$ are defined in an analogous way using the empirical distribution $\hat{p}$ instead of p. The ability to perform well on new data is measured by the generalization error:

$$\mathcal{E}(f) = \mathcal{R}^*(f) - \mathcal{R}(f)\,. \tag{G.15}$$

Statistical learning theory provides probabilistic upper-bounds on the generalization error depending on the type of task at hand and on the specific model used to tackle it. In order to find such an upper bound for a binary classification tasks, it is convenient to consider a relaxed version of the risk, the η-margin-risk $\mathcal{R}_\eta(f)$ defined in the main text. The upper bound on the generalization properties involves the empirical Rademacher complexity of a class of trial functions $\mathcal{H}$ with respect to the training sample $\mathcal{S}$. It is defined as:

$$\mathfrak{R}_{\mathcal{S}}\left(\mathcal{H}\right) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{H}}\frac{1}{N_{\text{train}}}\sum_{i=1}^{N_{\text{train}}}\sigma_i f(\boldsymbol{x}_i)\right] \tag{G.16}$$

where $\boldsymbol{\sigma}$ is a vector of Rademacher variables that are discrete, independent and identically distributed following a uniform law over $\{-1, 1\}$. The Rademacher complexity measures the ability of a hypothesis class $\mathcal{H}$ to fit noise, and as such it is a measure of the expressivity of $\mathcal{H}$. We now give a upper-bound on the generalization error (Theorem 5.8 in [186]):

Theorem II.1. *Let $\mathcal{H}$ be a set of trial functions and $\eta > 0$. Then $\forall\delta > 0$, with probability at least $1-\delta$, we have $\forall f \in \mathcal{H}$:*

$$\mathcal{R}^*(f) \leq \mathcal{R}_\eta(f) + \frac{2}{\eta}\mathfrak{R}_{\mathcal{S}}(\mathcal{H}) + 3\sqrt{\frac{\log(\frac{2}{\delta})}{2N_{\text{train}}}}$$

This upper-bound can be specialized to the case of kernel methods where the hypothesis class is the RKHS of a kernel k. In this the Rademacher complexity is upper bounded by a quantity that depends only on the trace of the empirical kernel matrix $\mathbf{K}$ (Theorem 6.12 in [186]):

Theorem II.2. *Let $\mathcal{H}$ by the RKHS associated to a given kernel k. For $\Lambda \geq 0$ consider the set of hypothesis functions $\mathcal{H}_\Lambda = \{f : x \mapsto \boldsymbol{w}^T \delta\boldsymbol{\phi}(x), \quad \|\boldsymbol{w}\|^2 \Lambda \leq 1\} \subseteq \mathcal{H}$. Then we have:*

$$\mathfrak{R}_{\mathcal{S}}\left(\mathcal{H}_\Lambda\right) \leq \frac{1}{N_{\text{train}}} \sqrt{\frac{\text{Tr}\left[\mathbf{K}\right]}{\Lambda}} \,.$$

Injecting this result in the previous upper bound, we get the desired result. In particular, using the centered noisy quantum kernel k_c and Eq. (G.7), we get Eq. (2.35).

III Time-multiplexing and model expressivity

The maximal class of trial functions $\mathcal{H}_{\text{full}}$ (see Sec. II) associated to a given embedding is obtained by performing a complete tomography of the embedded quantum states $\hat{\rho}(\boldsymbol{x})$ right after the end of the encoding procedure. The system evolution after time τ according to a Lindblad master equation with a constant Hamiltonian and disspator for a given duration δt_m can be expressed into a set of Kraus operators $\{\hat{W}_i\}$ [137] satisfying:

$$\sum_i \hat{W}_i^\dagger \hat{W}_i = \hat{\mathbb{1}} \,. \tag{G.17}$$

The evolved density matrices $\hat{\rho}(\boldsymbol{x}; \delta t_m)$ are given by:

$$\hat{\rho}(\boldsymbol{x}; \delta t_m) = \sum_i \hat{W}_i \hat{\rho}(\boldsymbol{x}) \hat{W}_i^\dagger \,. \tag{G.18}$$

In the Heisenberg picture, the observables evolve in time following an adjoint master equation [137]. Hence, we can see the non-unitary evolution of the open quantum system as a simple change in the set of observables that are measured on the state $\hat{\rho}(\boldsymbol{x})$. Suppose that we want to measure observables from the orthonormal basis introduced in section **??** after the previous evolution. We define the $(P+1) \times (P+1)$ matrix $\boldsymbol{\Xi}$ whose elements are:

$$\Xi_{kl} = \text{Tr}\left[\sum_i \hat{W}_i^\dagger \hat{O}_k \hat{W}_i \hat{O}_l\right] \,. \tag{G.19}$$

The measurement at time $\tau + \delta t_m$ of the observable $\hat{O}_l$ can now be expressed using the decomposition in Eq. (2.41) and the elements of $\boldsymbol{\Xi}$ as:

$$\text{Tr}\left[\hat{\rho}(\boldsymbol{x}; \delta t_m) \hat{O}_l\right] = \frac{1}{2^N} \left(\Xi_{0l} + \sum_k \text{Tr}\left[\hat{\rho}(\boldsymbol{x}) \hat{O}_k\right] \Xi_{kl}\right) \,. \tag{G.20}$$

Thus the embedding map $\boldsymbol{\phi}(\boldsymbol{x})$ is transformed by the non-unitary evolution during δt_m and becomes:

$$\boldsymbol{\phi}(\boldsymbol{x}; \delta t_m) = \boldsymbol{\Xi} \boldsymbol{\phi}(\boldsymbol{x}) \,. \tag{G.21}$$

Assuming we only make measurements on a subset of the basis $\{\hat{O}_i\}$, then we can write for the feature vector:

$$\boldsymbol{\phi}(\boldsymbol{x}; \delta t_m) = \mathbf{D} \boldsymbol{\Xi} \boldsymbol{\phi}(\boldsymbol{x}) \,, \tag{G.22}$$

where $\boldsymbol{D}$ is a diagonal $(P+1)\times(P+1)$ matrix whose diagonal entries i are 1 if $\hat{O}_i$ is measured and 0 otherwise. When we repeat the measurements at different times, we can stack the previous vectors at each time steps. For N_{rep} repetitions, we denote $\boldsymbol{\Lambda}$ the $N_{\text{rep}}(P+1)\times(P+1)$ matrix of the form:

$$\boldsymbol{\Lambda} = \begin{pmatrix} \mathbf{D\Xi} \\ \mathbf{D\Xi}^2 \\ \vdots \\ \mathbf{D\Xi}^{N_{\text{rep}}} \end{pmatrix} . \tag{G.23}$$

The final vector reads:

$$\tilde{\boldsymbol{\phi}}(\boldsymbol{x}) = \boldsymbol{\Lambda}\boldsymbol{\phi}(\boldsymbol{x}) . \tag{G.24}$$

Hence, by performing repeated measurements in-between non-unitary evolutions amount to performing a restricted number of measurements on the encoded states $\hat{\rho}(\boldsymbol{x})$ at time τ. This implies that the time-multiplexing decoding lowers the model expressivity. The difference between the models obtained from the full tomography and the time-multiplexing decoding is encapsulated in the matrix $\boldsymbol{\Lambda}$.

H Examples of Clifford approximant circuits

In this appendix we provide a sample of Clifford approximant circuits for the estimation of $\mathbb{E}_{\boldsymbol{\theta}}[C(\boldsymbol{\theta})]$ and $\mathbb{E}_{\boldsymbol{\theta}}[C(\boldsymbol{\theta})^2]$ for the simple circuit depicted in Fig. H.1. The generalisation to Clifford approximants for other quantities, such as the expectation of the squared gradient, can be derived from that example as it suffices to introduce the adequate Clifford gates to the fixed layers to obtain the right estimators (see Sec. I.1 and I.4). This circuit acts on three qubits and is composed of two layers of rotations that are alternated with fixed two-qubits Control-Z gates. To obtain a first order approximant for these circuits it suffices to randomly replace each rotation by either the identity gate (a wire) or the Pauli gate corresponding to the direction of the concerned rotation gate. Three examples of first order Clifford approximant are represented in Fig. H.2. The second order approximant are derived by first mapping each rotation along X or Y to a rotation along Z, making use of the identities $\hat{X} = \hat{H}^\dagger \hat{Z} \hat{H}$ and $\hat{Y} = (\hat{S}\hat{H})\hat{Z}(\hat{S}\hat{H})^\dagger$ where $\hat{H}, \hat{S}$ are respectively the Hadamard and phase gates. As a result we get the ansatz with layers of Z rotations alternated with fixed layers composed of Clifford gates represented on Fig. H.3. This circuit is then doubled vertically to give a circuit acting on six qubits. Finally, each pairs of rotations sharing the same angle is randomly replaced by two single-qubit gates according to the scheme of Fig. 3.2.

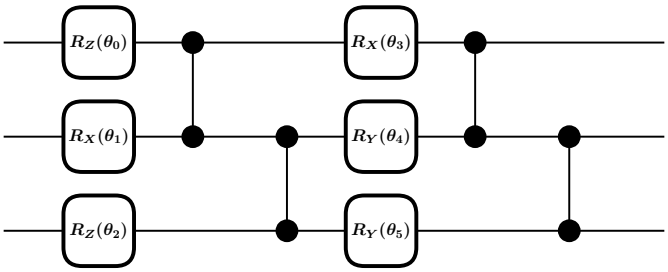


Figure H.1: Initial variational circuit with random rotation angles.

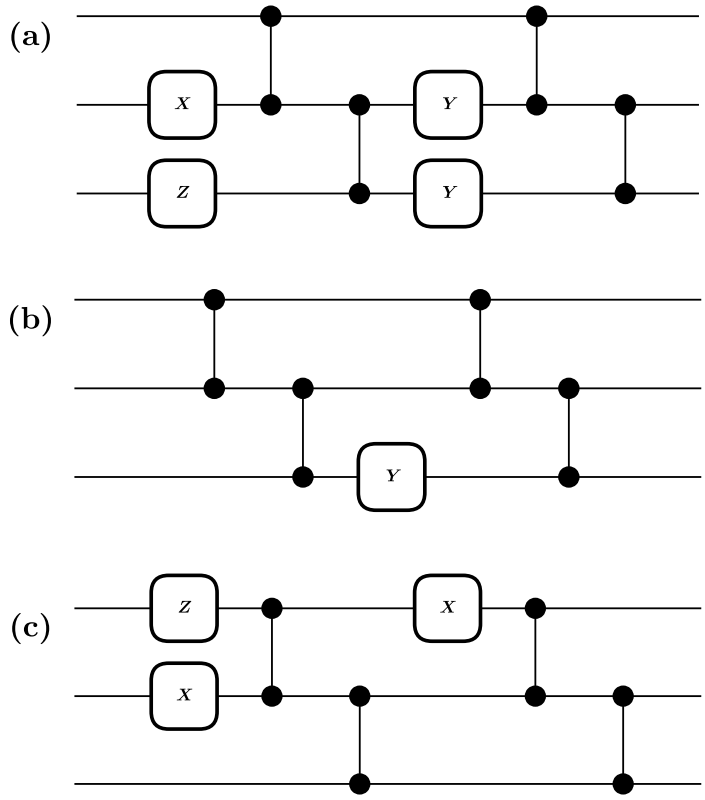


Figure H.2: Examples of first-order Clifford approximant circuits for the ansatz of Fig. H.1. Assuming the probability distribution of the angles is even, we replace each rotation by a Clifford gate that is sampled according to Eq. (3.36).

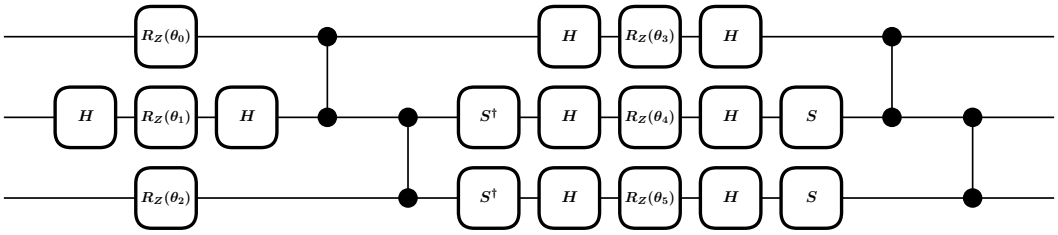
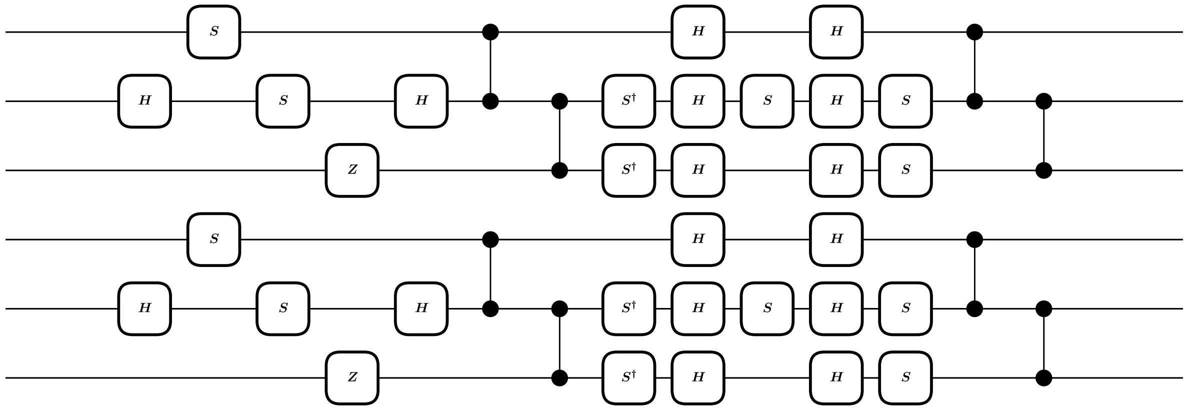
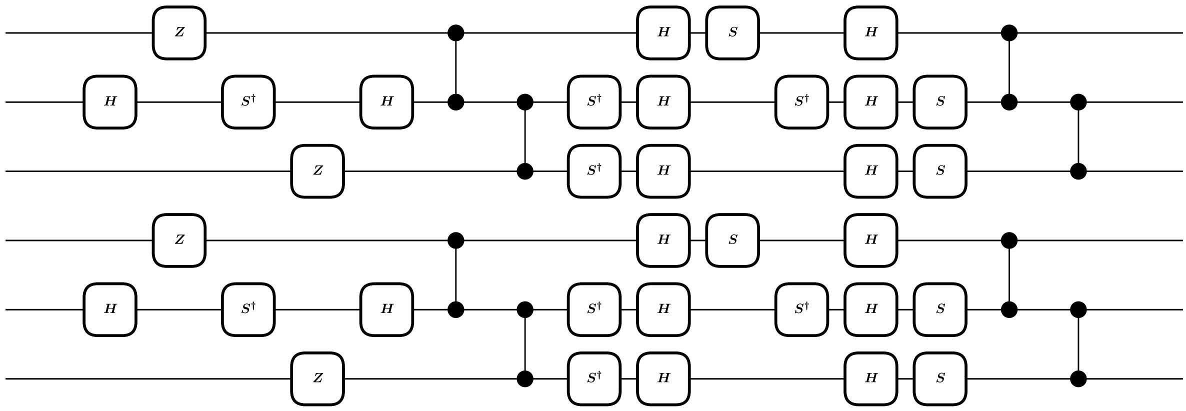


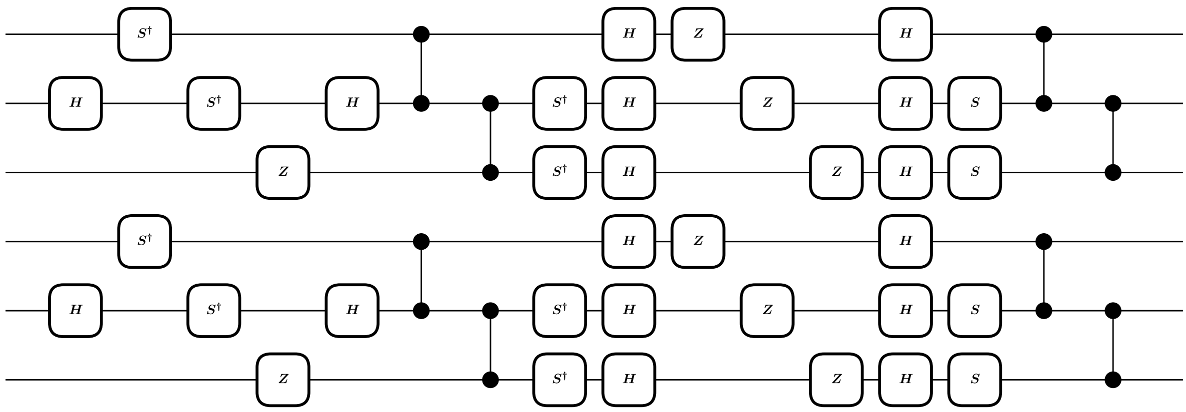
Figure H.3: Equivalent form of the initial circuit with Z-rotations only.



(a)



(b)



(c)

Figure H.4: Examples of second-order Clifford approximant circuits for the ansatz of Fig. H.1.

I Résumé substantiel

Cette thèse est consacrée à l'exploration de l'interface entre l'informatique quantique et l'apprentissage automatique, en mettant l'accent sur les effets du bruit et de la décohérence. Elle s'articule autour de deux axes de recherche principaux :

- l'utilisation de systèmes quantiques ouverts pour la réalisation de tâches classiques de reconnaissance de formes, au travers d'algorithmes d'apprentissage automatique hybrides classiques-quantiques (chapitre 2) ;

- l'estimation de la capacité d'entrainement d'algorithmes variationnels quantiques grâce à un schéma de simulation efficace basé sur un phénomène de décohérence artificielle et sur le théorème de Gottesman-Knill (chapitre 3).

Le chapitre 1 contient l'essentiel des bases théoriques sur lesquelles sont construits les résultats présentés dans la suite du texte. En particulier, les outils nécessaires à la description des systèmes quantiques ouverts y sont réintroduits et illustrés sur deux exemples classiques de systèmes à deux niveaux sujets à de la décohérence. L'état d'un système quantique à un temps t est représenté par un opérateur densité $\hat{\rho}(t)$. Pour un système fermé, l'évolution du système est donnée par l'équation de Liouville [137], qui s'écrit dans le système d'unité naturelles ($\hbar = 1$) :

$$\frac{\mathrm{d}\hat{\rho}}{\mathrm{d}t} = -\mathrm{i}\left[\hat{H}, \hat{\rho}\right] , \tag{I.1}$$

avec $\hat{H}$ le hamiltonien du système. Dans un premier temps, on considère un système à deux niveaux, ou *qubit*, soumis à une excitation classique aléatoire $\xi(t)$ et dont le hamiltonien s'écrit

$$\hat{H}(t) = -\frac{1}{2}\omega_0\hat{\sigma}_z - \frac{1}{2}\xi(t)\hat{\sigma}_z \, . \tag{I.2}$$

On suppose par ailleurs que l'excitation $\xi(t)$ est donnée par un processus stochastique gaussien. À chaque réalisation de ξ correspond un état $\hat{\rho}_\xi$ du système. En écrivant $\hat{\rho} = \mathbb{E}_\xi\left[\hat{\rho}_\xi\right]$ l'état du système moyenné sur les réalisations de l'excitation, on montre que l'évolution moyenne du système est donnée par l'équation maîtresse

$$\frac{\mathrm{d}\hat{\rho}(t)}{\mathrm{d}t} = -\mathrm{i}\left[\hat{H}_0, \hat{\rho}\right] + \gamma\mathcal{D}\left[\hat{\sigma}_z\right]\left(\hat{\rho}(t)\right) , \tag{I.3}$$

où l'on a introduit $\hat{H}_0 = \frac{\omega_0}{2}\hat{\sigma}_z$, le dissipateur $\mathcal{D}\left[\hat{A}\right]$ d'un opérateur $\hat{A}$ défini par

$$\mathcal{D}\left[\hat{A}\right](\hat{\rho}) = \hat{A}\hat{\rho}\hat{A}^\dagger - \frac{1}{2}\left(\hat{A}^\dagger\hat{A}\hat{\rho} + \hat{\rho}\hat{A}^\dagger\hat{A}\right) , \tag{I.4}$$

ainsi que le taux de dissipation γ dont l'expression exacte dépend des caractéristiques de l'excitation $\xi(t)$. On considère ensuite un modèle analogue où l'excitation aléatoire classique est remplacée par une interaction avec un environnement constitué d'un ensemble de modes bosoniques dans un état thermique. L'évolution du système global est dictée par l'équation (I.1) et le hamiltonien

$$\hat{H} = \frac{\omega_0}{2}\hat{\sigma}_z + \sum_{k\in\mathbb{N}} \omega_k \hat{b}_k^\dagger \hat{b}_k + \sum_{k\in\mathbb{N}} \hat{\sigma}_z \left(g_k \hat{b}_k^\dagger + g_k^* \hat{b}_k\right) , \tag{I.5}$$

où $\hat{b}_k^\dagger$ et $\hat{b}_k$ sont respectivement les opérateurs création et annihilation du mode k. L'état du qubit est obtenu en prenant la trace de l'opérateur densité du système global par rapport aux degrés de liberté de l'environnement. Cela revient à considérer la version quantique de la moyenne sur les réalisations du modèle précédent. À nouveau, on montre que dans ces conditions l'évolution du qubit est donnée par l'équation (I.3). Cette équation est un exemple d'équation maîtresse de type Lindblad, de la forme

$$\frac{\mathrm{d}\hat{\rho}}{\mathrm{d}t} = \mathcal{L}(t)(\hat{\rho}) = -\mathrm{i}\left[\hat{H}(t), \hat{\rho}(t)\right] + \sum_k \gamma_k(t)\mathcal{D}\left[\hat{A}_k(t)\right](\hat{\rho}) , \tag{I.6}$$

où les $\hat{A}_k$ sont des opérateurs de saut quantique et γ_k les taux de dissipation associés. Plus généralement, il est possible de montrer que, sous certaines conditions sur les interactions entre un système et son environnement, l'évolution d'un système quantique ouvert peut toujours s'écrire sous cette forme [137]. Par ailleurs, l'évolution temporelle d'un système quantique peut être décrite en terme d'applications associant l'opérateur densité du système à un temps t_1 à celui à un temps ultérieur $t_2 > t_1$. Afin qu'elle représente une transformation physique acceptable, une telle application doit être complètement positive et préserver la trace des opérateurs, auquel cas elle appartient à la classe plus large des *canaux de communication quantiques* [141]. Dans les exemples précédents, l'évolution entre le temps $t = 0$ et le temps $t > 0$ est donnée (dans la représentation d'interaction, voir l'annexe A) par le *canal de décohérence* [142]

$$\hat{\rho}(t) = \Lambda_{(t,0)}(\hat{\rho}_0) = \frac{1}{2}\left(1 + e^{-\gamma t}\right)\hat{\rho}_0 + \frac{1}{2}\left(1 - e^{-\gamma t}\right)\hat{\sigma}_z\hat{\rho}_0\hat{\sigma}_z , \tag{I.7}$$

avec $\hat{\rho}_0 = \hat{\rho}(0)$.

Le chapitre 2 est consacré à l'étude de l'utilisation de systèmes quantiques ouverts pour la réalisation de tâches classiques de reconnaissance de formes dans un contexte d'apprentissage supervisé. Plus précisément, on s'intéresse aux modèles de machines à noyaux quantiques utilisant des systèmes dissipatifs. Les méthodes d'apprentissage supervisé visent à approximer une relation causale entre les éléments $\boldsymbol{x}$ d'un ensemble de données d'entrée $\mathcal{X}$ et les éléments cibles y d'un ensemble d'arrivée $\mathcal{Y}$, en se basant sur un jeu d'entrainement $\mathcal{S} = \{(\boldsymbol{x}_i, y_i) \,|\, i \in [\![1, N_{\text{train}}]\!]\}$. En supposant que la relation entre les

entrées et les cibles est donnée par une fonction inconnue f telle que $y = f(\boldsymbol{x})$, on cherche à approximer f par une fonction $f_{\boldsymbol{w}}$ dépendant d'un ensemble de paramètres $\boldsymbol{w}$. On considère ici des modèles de calcul par réservoir présentant une structure d'encodage-décodage. Dans un premier temps, une donnée d'entrée $\boldsymbol{x}$ est encodée de façon non-linéaire dans l'état $\hat{\rho}(\boldsymbol{x})$ d'un système quantique ouvert, que l'on appelle *réservoir*. On mesure ensuite un ensemble d'observables du système $\left\{\hat{O}_1, \ldots, \hat{O}_M\right\}$, dont les valeurs moyennes forment un vecteur de caractéristiques

$$\boldsymbol{\phi}(\boldsymbol{x}) = \left(1, \mathrm{Tr}\left[\hat{\rho}(\boldsymbol{x})\hat{O}_1\right], \ldots, \mathrm{Tr}\left[\hat{\rho}(\boldsymbol{x})\hat{O}_M\right]\right) . \tag{I.8}$$

Finalement, l'information sur l'élément cible y est décodée en considérant une combinaison linéaire des résultats des mesures, et on définit :

$$f_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) . \tag{I.9}$$

Les poids $\boldsymbol{w}$ de la combinaison linéaire sont entrainés à partir des données d'entrainement et d'une fonction de perte associée à la tâche à accomplir. Cette approche de décodage, caractéristique du calcul par réservoir, a l'avantage de permettre un entrainement peu coûteux, voire analytique. D'autre part, l'étape d'encodage dans la dynamique du réservoir et les mesures subséquentes génèrent un ensemble de transformations non linéaires des données d'entrées. Ces transformations ne sont pas entrainées, ce qui évite d'avoir à contrôler précisément la dynamique système. Les modèles considérés appartiennent à l'ensemble plus large des machine à noyaux. En particulier, à l'ensemble des observables mesurées est associée la fonction noyau

$$k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \boldsymbol{\phi}(\boldsymbol{x}_1)^T \boldsymbol{\phi}(\boldsymbol{x}_2) . \tag{I.10}$$

Dans le cas où $\left\{\hat{O}_1, \ldots, \hat{O}_M\right\}$ forme une base de l'ensemble des observables du système, le noyau se reformule [171]

$$k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathrm{Tr}\left[\hat{\rho}(\boldsymbol{x}_1)\hat{\rho}(\boldsymbol{x}_2)\right] . \tag{I.11}$$

La fonction noyau caractérise l'ensemble des transformations qui peuvent être obtenue à partir d'une méthode d'encodage donnée et des mesures des observables considérées. La théorie des fonctions noyaux fournit des outils permettant d'analyser l'expressivité et la capacité de généralisation des modèles étudiés [186, 253]. Dans ce cadre, nous évaluons l'impact du bruit et de la décohérence affectant le système physique sous-jacent sur les performances des modèles. En particulier, nous montrons que le bruit physique peut aider à lutter contre le phénomène de sur-ajustement, agissant ainsi comme un processus de régularisation. Ces résultats sont confirmés par une étude numérique, où l'on considère des réservoirs composés de chaines de qubits couplés et dissipatifs évoluant selon une équation de Lindblad. L'encodage est réalisé par une excitation temporelle des qubits dont l'amplitude dépend des données d'entrée à traiter. L'ensemble des résultats théoriques et numériques présentés dans ce chapitre sont issus de notre publication [α].

Le chapitre 3 présente les résultats de notre publication [γ]. On y introduit une méthode permettant l'évaluation efficace de l'entrainabilité d'une large classe d'algorithmes

quantiques variationnels. Ces algorithmes quantiques polyvalents sont inspirés des méthodes d'apprentissage automatique, et sont des candidats prometteurs à un avantage quantiques pour des tâches variées, allant de l'optimisation combinatoire à la chimie quantique [95]. Ils reposent sur l'utilisation de circuits quantiques variationnels dont les paramètres sont optimisés par des techniques de descente de gradient. En général, la tâche à accomplir est présentée sous la forme d'une minimisation d'une énergie moyenne associée à un hamiltonien $\hat{H}$ agissant sur un système de n qubits. On cherche alors un état $|\phi\rangle$ du système de qubits minimisant l'énergie moyenne $\langle\hat{H}\rangle_\phi = \langle\phi|\,\hat{H}\,|\phi\rangle$. Pour cela, on utilise un circuit variationnel définissant une transformation unitaire paramétrée $\hat{U}(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathbb{R}^M$. En supposant que les qubits sont dans un état initial $|\phi_0\rangle$, on obtient en sortie du circuit un état variationnel

$$|\phi(\boldsymbol{\theta})\rangle = \hat{U}(\boldsymbol{\theta})\,|\phi_0\rangle\ . \tag{I.12}$$

On définit alors la fonction de coût à minimiser :

$$\begin{aligned} C(\boldsymbol{\theta}) &= \langle\phi(\boldsymbol{\theta})|\,\hat{H}\,|\phi(\boldsymbol{\theta})\rangle \\ &= \langle\phi_0|\,\hat{U}(\boldsymbol{\theta})^\dagger \hat{H}\hat{U}(\boldsymbol{\theta})\,|\phi_0\rangle\ . \end{aligned} \tag{I.13}$$

On s'intéresse ici aux circuits variationnels présentant une structure par couches, composées de transformations fixes $\hat{W}_i$ et de rotations à un seul qubit $\hat{U}_i(\theta_i)$ de générateurs $P_i \in \left\{\hat{X}, \hat{Y}, \hat{Z}\right\}$, telle que

$$\hat{U}(\boldsymbol{\theta}) = \prod_{i=1}^{M} \hat{U}_i(\theta_i)\hat{W}_i\,, \qquad \hat{U}_i(\theta_i) = e^{-\mathrm{i}\frac{\theta_i}{2}\hat{P}_i}\,. \tag{I.14}$$

Les angles des rotations sont entrainés selon un algorithme de descente de gradient. Dans le cas le plus simple, le vecteur $\boldsymbol{\theta}^{k+1}$ des angles de rotation à l'itération $k+1$ est donné par

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k - \eta\nabla_{\boldsymbol{\theta}} C(\boldsymbol{\theta}^k)\,. \tag{I.15}$$

Les composantes du gradient sont obtenues grâce à la *règle des paramètres translatés*, selon laquelle

$$\partial_k C(\boldsymbol{\theta}) = \frac{1}{2}\left(C(\boldsymbol{\theta} + \frac{\pi}{2}\boldsymbol{e}_k) - C(\boldsymbol{\theta} - \frac{\pi}{2}\boldsymbol{e}_k)\right), \tag{I.16}$$

où l'on note $\partial_k C(\boldsymbol{\theta}) = \frac{\partial C}{\theta_k}(\boldsymbol{\theta})$ et $\boldsymbol{e}_k$ le vecteur de la base canonique associé à la composante k. Lors de l'exécution d'un tel algorithme sur une plateforme de calcul quantique, les quantités apparaissant dans le terme de droite de l'équation (I.16) sont mesurées sur le système quantique considéré. La précision du gradient obtenu dépend donc des limitations expérimentales, et notamment du nombre de répétition de chaque mesure. Or, les algorithmes quantiques variationnels souffrent parfois d'un phénomène de disparition du gradient évoluant exponentiellement avec le nombre de qubits, connu dans la littérature sous le nom de *barren plateaus* [237]. La présence de barren plateaus rend le coût d'une évaluation précise du gradient prohibitif, ce qui empêche d'entrainer efficacement le modèle. Pour un circuit variationnel donné, les angles initiaux sont choisi aléatoirement selon une loi de probabilité fixée $\mathbb{P}$. On dit qu'un circuit souffre de barren plateaus s'il existe un $\alpha > 0$ tel que

$$\mathbb{P}(|\partial_k C| > \epsilon) \leq \mathcal{O}(\exp(-\alpha n))\,. \tag{I.17}$$

Dans de nombreux cas, la valeur moyenne de $C(\boldsymbol{\theta})$ est nulle, par exemple lorsque les composantes de $\boldsymbol{\theta}$ sont indépendantes et de loi uniforme sur $[0, 2\pi]$. L'inégalité de Bienaymé-Tchebytchev pour $\epsilon > 0$ donne alors

$$\mathbb{P}(|\partial_k C| > \epsilon) \leq \mathrm{Var}\left[\partial_k C\right] / \epsilon^2 \,. \tag{I.18}$$

Ainsi, un système dont la variance du gradient disparait exponentiellement avec le nombre de qubits n présentera des barren plateaus. Réciproquement, une large variance du gradient garantit des fluctuations importantes du gradient initial. La variance du gradient apparait donc comme une mesure adéquate de l'existence de barren plateaus. On propose ici une méthode permettant d'estimer efficacement cette variance sur un ordinateur classique, et ainsi de garantir l'entrainabilité d'un circuit variationnel donnée. Cette méthode est basée sur l'utilisation de circuits quantiques de Clifford, dont la transformation unitaire associée peut-être simulée classiquement avec une complexité polynomiale en n, d'après le théorème de Gottesman-Knill [42, 142]. On considère d'abord le canal quantique $\mathcal{E}$ qui représente la transformation moyenne d'un état initial $\hat{\rho}$ pour un choix aléatoire de $\boldsymbol{\theta}$:

$$\begin{aligned} \mathcal{E}\left(\hat{\rho}\right) &= \mathbb{E}_{\boldsymbol{\theta}}\left[\hat{U}(\boldsymbol{\theta})\hat{\rho}\hat{U}(\boldsymbol{\theta})^{\dagger}\right] \\ &= \int_{[0,2\pi]^M} \hat{U}(\boldsymbol{\theta})\hat{\rho}\hat{U}^{\dagger}(\boldsymbol{\theta})\mathbb{P}(\mathrm{d}\boldsymbol{\theta}) \,. \end{aligned} \tag{I.19}$$

On montre que, pour un circuit variationnel composé de transformations de Clifford fixées et de rotations d'angles aléatoires, sous certaines conditions sur la distribution $\mathbb{P}$ des angles, le canal quantique $\mathcal{E}$ peut se réécrire comme une somme convexe de canaux de Clifford

$$\mathcal{E}\left(\hat{\rho}\right) = \sum_i p_i \hat{C}_i \hat{\rho} \hat{C}_i^{\dagger} \,, \tag{I.20}$$

avec $p_i \geq 0$, $\sum_i p_i = 1$, et $\hat{C}_i$ des transformations de Clifford. Ce résultat est déduit de considérations sur le cas d'une rotation d'axe Z et d'angle aléatoire θ agissant sur un unique qubit. Par exemple, en supposant que θ est distribué uniformément dans $[0, 2\pi]$, il est facile de montrer que

$$\mathbb{E}_{\theta}\left[\hat{R}_Z(\theta)\hat{\rho}\hat{R}_Z^{\dagger}(\theta)\right] = \frac{1}{2}\hat{\rho} + \frac{1}{2}\hat{Z}\hat{\rho}\hat{Z} \,, \tag{I.21}$$

où $\hat{R}_Z(\theta) = e^{-\mathrm{i}\frac{\theta}{2}\hat{Z}}$. Ainsi, dans ce cas simple, le canal quantique moyen obtenu est un canal de décohérence. Puisque les transformations $\hat{Z}$ et l'identité sont des transformations de Clifford, ce résultat est bien cohérent avec l'équation (I.20). En généralisant ce raisonnement à des circuits à n et à des moyennes quadratiques, on obtient une méthode qui permet d'estimer des quantités telles que l'énergie moyenne ou la variance du gradient avec un algorithme classique de complexité polynomiale. D'un point de vue physique, ce résultat est la conséquence d'une décohérence artificiellement induite par le choix aléatoire des angles de rotations. Finalement, notre méthode pourrait permettre d'optimiser l'architecture de circuits variationnels quantiques tout en garantissant leur entrainabilité, notamment au travers d'algorithmes d'apprentissage automatique.

Bibliography

[α] V. Heyraud, Z. Li, Z. Denis, A. Le Boité and C. Ciuti, "Noisy quantum kernel machines", Physical Review A **106**, 052421 (2022).

[γ] V. Heyraud, Z. Li, K. Donatella, A. Le Boité and C. Ciuti, "Efficient estimation of trainability for variational quantum circuits", 10.48550/arXiv.2302.04649 (2023).

[β] Z. Li, V. Heyraud, K. Donatella, Z. Denis and C. Ciuti, "Machine learning via relativity-inspired quantum dynamics", Physical Review A **106**, 032413 (2022).

[1] A. Aspect, "14 From Einstein, Bohr, Schrödinger to Bell and Feynman: a New Quantum Revolution?", in *14 From Einstein, Bohr, Schrödinger to Bell and Feynman: a New Quantum Revolution?* (EDP Sciences, 2021), pp. 407–434.

[2] J. Bardeen and W. H. Brattain, "The Transistor, A Semi-Conductor Triode", Physical Review **74**, 230–231 (1948).

[3] W. Shockley, "The Theory of p-n Junctions in Semiconductors and p-n Junction Transistors", Bell System Technical Journal **28**, 435–489 (1949).

[4] A. Kastler, "Quelques suggestions concernant la production optique et la détection optique d'une inégalité de population des niveaux de quantifigation spatiale des atomes. Application à l'expérience de Stern et Gerlach et à la résonance magnétique", J. Phys. Radium **11**, 255 (1950).

[5] A. Einstein, B. Podolsky and N. Rosen, "Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?", Physical Review **47**, 777–780 (1935).

[6] J. S. Bell, "On the Einstein Podolsky Rosen paradox", Physics Physique Fizika **1**, 195–200 (1964).

[7] J. S. BELL, "On the Problem of Hidden Variables in Quantum Mechanics", Reviews of Modern Physics **38**, 447–452 (1966).

[8] J. F. Clauser, M. A. Horne, A. Shimony and R. A. Holt, "Proposed Experiment to Test Local Hidden-Variable Theories", Physical Review Letters **23**, 880–884 (1969).

[9] J. F. Clauser and A. Shimony, "Bell's theorem. Experimental tests and implications", Reports on Progress in Physics **41**, 1881 (1978).

[10] A. Aspect, P. Grangier and G. Roger, "Experimental Tests of Realistic Local Theories via Bell's Theorem", Physical Review Letters **47**, 460–463 (1981).

[11] A. Aspect, J. Dalibard and G. Roger, "Experimental Test of Bell's Inequalities Using Time-Varying Analyzers", Physical Review Letters **49**, 1804–1807 (1982).

[12] J. P. Gordon, "Quantum Effects in Communications Systems", Proceedings of the IRE **50**, 1898–1908 (1962).

[13] P. Benioff, "Quantum Mechanical Models of Turing Machines That Dissipate No Energy", Physical Review Letters **48**, 1581–1585 (1982).

[14] P. Benioff, "Quantum mechanical hamiltonian models of turing machines", Journal of Statistical Physics **29**, 515–546 (1982).

[15] R. P. Feynman, "Simulating physics with computers", International Journal of Theoretical Physics **21**, 467–488 (1982).

[16] Yu. I. Manin, *Vychislimoe i nevychislimoe* (Sov. radio, 1980).

[17] D. Deutsch and R. Penrose, "Quantum theory, the Church–Turing principle and the universal quantum computer", Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences **400**, 97–117 (1997).

[18] E. Bernstein and U. Vazirani, "Quantum complexity theory", in Proceedings of the twenty-fifth annual ACM symposium on Theory of Computing, STOC '93 (1993), pp. 11–20.

[19] D. Simon, "On the power of quantum computation", in Proceedings 35th Annual Symposium on Foundations of Computer Science (1994), pp. 116–123.

[20] S. Lloyd, "Universal Quantum Simulators", Science **273**, 1073–1078 (1996).

[21] L. K. Grover, "A fast quantum mechanical algorithm for database search", in Proceedings of the twenty-eighth annual ACM symposium on Theory of Computing, STOC '96 (1996), pp. 212–219.

[22] P. Shor, "Algorithms for quantum computation: discrete logarithms and factoring", in Proceedings 35th Annual Symposium on Foundations of Computer Science (1994), pp. 124–134.

[23] A. G. J. MacFarlane, J. P. Dowling and G. J. Milburn, "Quantum technology: the second quantum revolution", Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences **361**, 1655–1674 (2003).

[24] A. Acín, I. Bloch, H. Buhrman, T. Calarco, C. Eichler, J. Eisert, D. Esteve, N. Gisin, S. J. Glaser, F. Jelezko, S. Kuhr, M. Lewenstein, M. F. Riedel, P. O. Schmidt, R. Thew, A. Wallraff, I. Walmsley and F. K. Wilhelm, "The quantum technologies roadmap: a European community view", New Journal of Physics **20**, 080201 (2018).

[25] L. Gyongyosi and S. Imre, "A Survey on quantum computing technology", Computer Science Review **31**, 51–71 (2019).

[26] A. Montanaro, "Quantum algorithms: an overview", npj Quantum Information **2**, 1–8 (2016).

[27] W. H. Zurek, "Decoherence, einselection, and the quantum origins of the classical", Reviews of Modern Physics **75**, 715–775 (2003).

[28] D. Loss and D. P. DiVincenzo, "Quantum computation with quantum dots", Physical Review A **57**, 120–126 (1998).

[29] C. Kloeffel and D. Loss, "Prospects for Spin-Based Quantum Computing in Quantum Dots", Annual Review of Condensed Matter Physics **4**, 51–81 (2013).

[30] J. L. O'Brien, "Optical Quantum Computing", Science **318**, 1567–1570 (2007).

[31] P. Kok, W. J. Munro, K. Nemoto, T. C. Ralph, J. P. Dowling and G. J. Milburn, "Linear optical quantum computing with photonic qubits", Reviews of Modern Physics **79**, 135–174 (2007).

[32] W. S. Warren, "The Usefulness of NMR Quantum Computing", Science **277**, 1688–1690 (1997).

[33] J. A. Jones and M. Mosca, "Implementation of a quantum algorithm on a nuclear magnetic resonance quantum computer", The Journal of Chemical Physics **109**, 1648–1653 (1998).

[34] J. I. Cirac and P. Zoller, "Quantum Computations with Cold Trapped Ions", Physical Review Letters **74**, 4091–4094 (1995).

[35] H. Häffner, C. F. Roos and R. Blatt, "Quantum computing with trapped ions", Physics Reports **469**, 155–203 (2008).

[36] R. Blatt and C. F. Roos, "Quantum simulations with trapped ions", Nature Physics **8**, 277–284 (2012).

[37] J. Q. You and F. Nori, "Superconducting Circuits and Quantum Information", Physics Today **58**, 42–47 (2005).

[38] G. Wendin, "Quantum information processing with superconducting circuits: a review", Reports on Progress in Physics **80**, 106001 (2017).

[39] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. S. L. Brandao, D. A. Buell, B. Burkett, Y. Chen, Z. Chen, B. Chiaro, R. Collins, W. Courtney, A. Dunsworth, E. Farhi, B. Foxen, A. Fowler, C. Gidney, M. Giustina, R. Graff, K. Guerin, S. Habegger, M. P. Harrigan, M. J. Hartmann, A. Ho, M. Hoffmann, T. Huang, T. S. Humble, S. V. Isakov, E. Jeffrey, Z. Jiang, D. Kafri, K. Kechedzhi, J. Kelly, P. V. Klimov, S. Knysh, A. Korotkov, F. Kostritsa, D. Landhuis, M. Lindmark, E. Lucero, D. Lyakh, S. Mandrà, J. R. McClean, M. McEwen, A. Megrant, X. Mi, K. Michielsen, M. Mohseni, J. Mutus, O. Naaman, M. Neeley, C. Neill, M. Y. Niu, E. Ostby, A. Petukhov, J. C. Platt, C. Quintana, E. G. Rieffel, P. Roushan, N. C. Rubin, D. Sank, K. J. Satzinger, V. Smelyanskiy, K. J. Sung, M. D. Trevithick, A. Vainsencher, B. Villalonga, T. White, Z. J. Yao, P. Yeh, A. Zalcman, H. Neven and J. M. Martinis, "Quantum supremacy using a programmable superconducting processor", Nature **574**, 505–510 (2019).

[40] H.-S. Zhong, H. Wang, Y.-H. Deng, M.-C. Chen, L.-C. Peng, Y.-H. Luo, J. Qin, D. Wu, X. Ding, Y. Hu, P. Hu, X.-Y. Yang, W.-J. Zhang, H. Li, Y. Li, X. Jiang, L. Gan, G. Yang, L. You, Z. Wang, L. Li, N.-L. Liu, C.-Y. Lu and J.-W. Pan, "Quantum computational advantage using photons", Science **370**, 1460–1463 (2020).

[41] Y. Kim, A. Eddins, S. Anand, K. X. Wei, E. van den Berg, S. Rosenblatt, H. Nayfeh, Y. Wu, M. Zaletel, K. Temme and A. Kandala, "Evidence for the utility of quantum computing before fault tolerance", Nature **618**, 500–505 (2023).

[42] D. Gottesman, *The Heisenberg Representation of Quantum Computers*, 1998.

[43] S. D. Bartlett, B. C. Sanders, S. L. Braunstein and K. Nemoto, "Efficient Classical Simulation of Continuous Variable Quantum Information Processes", Physical Review Letters **88**, 097904 (2002).

[44] G. Vidal, "Efficient Classical Simulation of Slightly Entangled Quantum Computations", Physical Review Letters **91**, 147902 (2003).

[45] S. Aaronson and D. Gottesman, "Improved simulation of stabilizer circuits", Physical Review A **70**, 052328 (2004).

[46] Y.-Y. Shi, L.-M. Duan and G. Vidal, "Classical simulation of quantum many-body systems with a tree tensor network", Physical Review A **74**, 022320 (2006).

[47] F. Verstraete, V. Murg and J. Cirac, "Matrix product states, projected entangled pair states, and variational renormalization group methods for quantum spin systems", Advances in Physics **57**, 143–224 (2008).

[48] A. Mari and J. Eisert, "Positive Wigner Functions Render Classical Simulation of Quantum Computation Efficient", Physical Review Letters **109**, 230503 (2012).

[49] S. Bravyi and D. Gosset, "Improved Classical Simulation of Quantum Circuits Dominated by Clifford Gates", Physical Review Letters **116**, 250501 (2016).

[50] G. Carleo and M. Troyer, "Solving the quantum many-body problem with artificial neural networks", Science **355**, 602–606 (2017).

[51] S. Bravyi, D. Gosset and R. Movassagh, "Classical algorithms for quantum mean values", Nature Physics **17**, 337–341 (2021).

[52] M. Medvidović and G. Carleo, "Classical variational simulation of the Quantum Approximate Optimization Algorithm", npj Quantum Information **7**, 1–7 (2021).

[53] S. Bravyi, D. Gosset and Y. Liu, "How to simulate quantum measurement without computing marginals", Physical Review Letters **128**, 220503 (2022).

[54] Y. Zhou, E. M. Stoudenmire and X. Waintal, "What Limits the Simulation of Quantum Computers?", Physical Review X **10**, 041038 (2020).

[55] J. Tindall, M. Fishman, M. Stoudenmire and D. Sels, *Efficient tensor network simulation of IBM's kicked Ising experiment*, 2023.

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

[57] W. G. Unruh, "Maintaining coherence in quantum computers", Physical Review A **51**, 992–997 (1995).

[58] B. M. Terhal, "Quantum error correction for quantum memories", Reviews of Modern Physics **87**, 307–346 (2015).

[59] J. Roffe, "Quantum error correction: an introductory guide", Contemporary Physics **60**, 226–245 (2019).

[60] P. Campagne-Ibarcq, A. Eickbusch, S. Touzard, E. Zalys-Geller, N. E. Frattini, V. V. Sivak, P. Reinhold, S. Puri, S. Shankar, R. J. Schoelkopf, L. Frunzio, M. Mirrahimi and M. H. Devoret, "Quantum error correction of a qubit encoded in grid states of an oscillator", Nature **584**, 368–372 (2020).

[61] Y. Zhao, Y. Ye, H.-L. Huang, Y. Zhang, D. Wu, H. Guan, Q. Zhu, Z. Wei, T. He, S. Cao, F. Chen, T.-H. Chung, H. Deng, D. Fan, M. Gong, C. Guo, S. Guo, L. Han, N. Li, S. Li, Y. Li, F. Liang, J. Lin, H. Qian, H. Rong, H. Su, L. Sun, S. Wang, Y. Wu, Y. Xu, C. Ying, J. Yu, C. Zha, K. Zhang, Y.-H. Huo, C.-Y. Lu, C.-Z. Peng, X. Zhu and J.-W. Pan, "Realization of an Error-Correcting Surface Code with Superconducting Qubits", Physical Review Letters **129**, 030501 (2022).

[62] K. Bharti, A. Cervera-Lierta, T. H. Kyaw, T. Haug, S. Alperin-Lea, A. Anand, M. Degroote, H. Heimonen, J. S. Kottmann, T. Menke, W.-K. Mok, S. Sim, L.-C. Kwek and A. Aspuru-Guzik, "Noisy intermediate-scale quantum algorithms", Reviews of Modern Physics **94**, 015004 (2022).

[63] A. L. Fradkov, "Early History of Machine Learning", IFAC-PapersOnLine, 21st IFAC World Congress **53**, 1385–1390 (2020).

[64] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning", Nature **521**, 436–444 (2015).

[65] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*, edited by F. Bach, Adaptive Computation and Machine Learning Series (MIT Press, Cambridge, MA, USA, 2016).

[66] J. Schulman, F. Wolski, P. Dhariwal, A. Radford and O. Klimov, *Proximal Policy Optimization Algorithms*, 2017.

[67] R. S. Sutton and A. G. Barto, *Reinforcement Learning, second edition: An Introduction* (MIT Press, 2018).

[68] A. Bochkovskiy, C.-Y. Wang and H.-Y. M. Liao, *YOLOv4: Optimal Speed and Accuracy of Object Detection*, 2020.

[69] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell and S. Xie, *A ConvNet for the 2020s*, 2022.

[70] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis", in Advances in Neural Information Processing Systems, Vol. 34 (2021), pp. 8780–8794.

[71] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon and B. Poole, *Score-Based Generative Modeling through Stochastic Differential Equations*, 2021.

[72] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu and M. Chen, *Hierarchical Text-Conditional Image Generation with CLIP Latents*, 2022.

[73] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, J. Ho, D. J. Fleet and M. Norouzi, "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding", Advances in Neural Information Processing Systems **35**, 36479–36494 (2022).

[74] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan and B. Guo, "Vector Quantized Diffusion Model for Text-to-Image Synthesis", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022), pp. 10696–10706.

[75] F.-A. Croitoru, V. Hondru, R. T. Ionescu and M. Shah, "Diffusion Models in Vision: A Survey", IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–20 (2023).

[76] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei, *Language Models are Few-Shot Learners*, 2020.

[77] E. M. Bender, T. Gebru, A. McMillan-Major and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21 (2021), pp. 610–623.

[78] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto and L. Zdeborová, "Machine learning and the physical sciences", Reviews of Modern Physics **91**, 045002 (2019).

[79] J. Degrave, F. Felici, J. Buchli, M. Neunert, B. Tracey, F. Carpanese, T. Ewalds, R. Hafner, A. Abdolmaleki, D. de las Casas, C. Donner, L. Fritz, C. Galperti, A. Huber, J. Keeling, M. Tsimpoukelli, J. Kay, A. Merle, J.-M. Moret, S. Noury, F. Pesamosca, D. Pfau, O. Sauter, C. Sommariva, S. Coda, B. Duval, A. Fasoli, P. Kohli, K. Kavukcuoglu, D. Hassabis and M. Riedmiller, "Magnetic control of tokamak plasmas through deep reinforcement learning", Nature **602**, 414–419 (2022).

[80] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, "Highly accurate protein structure prediction with AlphaFold", Nature **596**, 583–589 (2021).

[81] V. Dunjko and H. J. Briegel, "Machine learning & artificial intelligence in the quantum domain: a review of recent progress", Reports on Progress in Physics **81**, 074001 (2018).

[82] M. Bukov, A. G. R. Day, D. Sels, P. Weinberg, A. Polkovnikov and P. Mehta, "Reinforcement Learning in Different Phases of Quantum Control", Physical Review X **8**, 031086 (2018).

[83] T. Fösel, P. Tighineanu, T. Weiss and F. Marquardt, "Reinforcement Learning with Neural Networks for Quantum Feedback", Physical Review X **8**, 031084 (2018).

[84] P. Andreasson, J. Johansson, S. Liljestrand and M. Granath, "Quantum error correction for the toric code using deep reinforcement learning", Quantum **3**, 183 (2019).

[85] H. P. Nautrup, N. Delfosse, V. Dunjko, H. J. Briegel and N. Friis, "Optimizing Quantum Error Correction Codes with Reinforcement Learning", Quantum **3**, 215 (2019).

[86] M. Y. Niu, S. Boixo, V. N. Smelyanskiy and H. Neven, "Universal quantum control through deep reinforcement learning", npj Quantum Information **5**, 1–8 (2019).

[87] M. M. Wauters, E. Panizon, G. B. Mbeng and G. E. Santoro, "Reinforcement-learning-assisted quantum optimization", Physical Review Research **2**, 033446 (2020).

[88] J. Yao, M. Bukov and L. Lin, *Policy Gradient based Quantum Approximate Optimization Algorithm*, 2020.

[89] Y.-H. Zhang, P.-L. Zheng, Y. Zhang and D.-L. Deng, "Topological Quantum Compiling with Reinforcement Learning", Physical Review Letters **125**, 170501 (2020).

[90] Z. He, L. Li, S. Zheng, Y. Li and H. Situ, "Variational quantum compiling with double Q-learning", New Journal of Physics **23**, 033002 (2021).

[91] M. Ostaszewski, L. M. Trenkwalder, W. Masarczyk, E. Scerri and V. Dunjko, "Reinforcement learning for optimization of variational quantum circuit architectures", in Advances in Neural Information Processing Systems, Vol. 34 (2021), pp. 18182–18194.

[92] D. A. Herrera-Martí, "Policy Gradient Approach to Compilation of Variational Quantum Circuits", Quantum **6**, 797 (2022).

[93] V. V. Sivak, A. Eickbusch, H. Liu, B. Royer, I. Tsioutsios and M. H. Devoret, "Model-Free Quantum Control with Reinforcement Learning", Physical Review X **12**, 011059 (2022).

[94] 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).

[95] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio and P. J. Coles, "Variational quantum algorithms", Nature Reviews Physics **3**, 625–644 (2021).

[96] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe and S. Lloyd, "Quantum machine learning", Nature **549**, 195–202 (2017).

[97] C. Ciliberto, M. Herbster, A. D. Ialongo, M. Pontil, A. Rocchetto, S. Severini and L. Wossnig, "Quantum machine learning: a classical perspective", Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences **474**, 20170551 (2018).

[98] M. Cerezo, G. Verdon, H.-Y. Huang, L. Cincio and P. J. Coles, "Challenges and opportunities in quantum machine learning", Nature Computational Science **2**, 567–576 (2022).

[99] H.-Y. Huang, M. Broughton, M. Mohseni, R. Babbush, S. Boixo, H. Neven and J. R. McClean, "Power of data in quantum machine learning", Nature Communications **12**, 2631 (2021).

[100] A. Abbas, D. Sutter, C. Zoufal, A. Lucchi, A. Figalli and S. Woerner, "The power of quantum neural networks", Nature Computational Science **1**, 403–409 (2021).

[101] M. Schuld, R. Sweke and J. J. Meyer, "Effect of data encoding on the expressive power of variational quantum-machine-learning models", Physical Review A **103**, 032430 (2021).

[102] M. Schuld and F. Petruccione, "Quantum Models as Kernel Methods", in *Machine Learning with Quantum Computers*, edited by M. Schuld and F. Petruccione, Quantum Science and Technology (Springer International Publishing, Cham, 2021), pp. 217–245.

[103] J. Kübler, S. Buchholz and B. Schölkopf, "The Inductive Bias of Quantum Kernels", in Advances in Neural Information Processing Systems, Vol. 34 (2021), pp. 12661–12673.

[104] T. Hofmann, B. Schölkopf and A. J. Smola, "Kernel methods in machine learning", The Annals of Statistics **36**, 1171–1220 (2008).

[105] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer Science & Business Media, 2013).

[106] S. Y. Kung, *Kernel Methods and Machine Learning* (Cambridge University Press, 2014).

[107] D. Anguita, S. Ridella, F. Rivieccio and R. Zunino, "Quantum optimization for training support vector machines", Neural Networks, Advances in Neural Networks Research: IJCNN '03 **16**, 763–770 (2003).

[108] P. Rebentrost, M. Mohseni and S. Lloyd, "Quantum Support Vector Machine for Big Data Classification", Physical Review Letters **113**, 130503 (2014).

[109] M. Schuld, M. Fingerhuth and F. Petruccione, "Implementing a distance-based classifier with a quantum interference circuit", Europhysics Letters **119**, 60002 (2017).

[110] C. Blank, D. K. Park, J.-K. K. Rhee and F. Petruccione, "Quantum classifier with tailored quantum kernel", npj Quantum Information **6**, 1–7 (2020).

[111] K. Bartkiewicz, C. Gneiting, A. Černoch, K. Jiráková, K. Lemr and F. Nori, "Experimental kernel-based quantum machine learning in finite feature space", Scientific Reports **10**, 12356 (2020).

[112] D. K. Park, C. Blank and F. Petruccione, "The theory of the quantum kernel-based binary classifier", Physics Letters A **384**, 126422 (2020).

[113] X. Wang, Y. Du, Y. Luo and D. Tao, "Towards understanding the power of quantum kernels in the NISQ era", Quantum **5**, 531 (2021).

[114] M. Schuld, *Supervised quantum machine learning models are kernel methods*, 2021.

[115] T. Hubregtsen, D. Wierichs, E. Gil-Fuster, P.-J. H. S. Derks, P. K. Faehrmann and J. J. Meyer, "Training quantum embedding kernels on near-term quantum computers", Physical Review A **106**, 042431 (2022).

[116] R. Shaydulin and S. M. Wild, "Importance of kernel bandwidth in quantum machine learning", Physical Review A **106**, 042407 (2022).

[117] S. Thanasilp, S. Wang, M. Cerezo and Z. Holmes, *Exponential concentration and untrainability in quantum kernel methods*, 2022.

[118] G. Tanaka, T. Yamane, J. B. Héroux, R. Nakane, N. Kanazawa, S. Takeda, H. Numata, D. Nakano and A. Hirose, "Recent advances in physical reservoir computing: A review", Neural Networks **115**, 100–123 (2019).

[119] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush and H. Neven, "Barren plateaus in quantum neural network training landscapes", Nature Communications **9**, 4812 (2018).

[120] S. Wang, E. Fontana, M. Cerezo, K. Sharma, A. Sone, L. Cincio and P. J. Coles, "Noise-induced barren plateaus in variational quantum algorithms", Nature Communications **12**, 6961 (2021).

[121] M. Cerezo, A. Sone, T. Volkoff, L. Cincio and P. J. Coles, "Cost function dependent barren plateaus in shallow parametrized quantum circuits", Nature Communications **12**, 1791 (2021).

[122] C. Ortiz Marrero, M. Kieferová and N. Wiebe, "Entanglement-Induced Barren Plateaus", PRX Quantum **2**, 040316 (2021).

[123] Z. Holmes, K. Sharma, M. Cerezo and P. J. Coles, "Connecting Ansatz Expressibility to Gradient Magnitudes and Barren Plateaus", PRX Quantum **3**, 010313 (2022).

[124] R. Durrett, *Probability: Theory and Examples* (Cambridge University Press, 2019).

[125] C. Cohen-Tannoudji, B. Diu and F. Laloe, *Mécanique Quantique - Tome 1: Nouvelle édition* (EDP Sciences, 2018).

[126] J. B. Conway, *A Course in Functional Analysis* (Springer, 2019).

[127] J. von Neumann, *Mathematische Grundlagen der Quantenmechanik* (Springer, Berlin, Heidelberg, 1996).

[128] E. Nelson, *Dynamical Theories of Brownian Motion* (Princeton University Press, 2001).

[129] A. Holevo, "Statistical structure of quantum theory and hidden variables", in *Probabilistic and Statistical Aspects of Quantum Theory*, edited by A. Holevo, Publications of the Scuola Normale Superiore (Edizioni della Normale, Pisa, 2011), pp. 265–303.

[130] A. Fine, "Hidden Variables, Joint Probability, and the Bell Inequalities", Physical Review Letters **48**, 291–295 (1982).

[131] M. M. Wolf, D. Perez-Garcia and C. Fernandez, "Measurements Incompatible in Quantum Theory Cannot Be Measured Jointly in Any Other No-Signaling Theory", Physical Review Letters **103**, 230402 (2009).

[132] N. Stevens and P. Busch, "Steering, incompatibility, and Bell-inequality violations in a class of probabilistic theories", Physical Review A **89**, 022123 (2014).

[133] W. Heisenberg, "Über den anschaulichen inhalt der quantentheoretischen kinematik und mechanik", Zeitschrift für Physik **43**, 172–198 (1927).

[134] L. E. BALLENTINE, "The Statistical Interpretation of Quantum Mechanics", Reviews of Modern Physics **42**, 358–381 (1970).

[135] M. Ozawa, "Universally valid reformulation of the Heisenberg uncertainty principle on noise and disturbance in measurement", Physical Review A **67**, 042105 (2003).

[136] L. A. Rozema, A. Darabi, D. H. Mahler, A. Hayat, Y. Soudagar and A. M. Steinberg, "Violation of Heisenberg's Measurement-Disturbance Relationship by Weak Measurements", Physical Review Letters **109**, 100404 (2012).

[137] H.-P. Breuer and F. Petruccione, *The Theory of Open Quantum Systems* (Oxford University Press, Oxford, 2007).

[138] H. M. Wiseman and G. J. Milburn, *Quantum Measurement and Control* (Cambridge University Press, 2010).

[139] P. Busch, "Unsharp reality and joint measurements for spin observables", Physical Review D **33**, 2253–2261 (1986).

[140] M. Reed, *Methods of Modern Mathematical Physics: Functional Analysis* (Elsevier, 2012).

[141] J. Watrous, *The Theory of Quantum Information* (Cambridge University Press, 2018).

[142] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information: 10th Anniversary Edition* (Cambridge University Press, 2010).

[143] M. B. Plenio, "Logarithmic Negativity: A Full Entanglement Monotone That is not Convex", Physical Review Letters **95**, 090503 (2005).

[144] G. Vidal and R. F. Werner, "Computable measure of entanglement", Physical Review A **65**, 032314 (2002).

[145] K. Mølmer, Y. Castin and J. Dalibard, "Monte Carlo wave-function method in quantum optics", JOSA B **10**, 524–538 (1993).

[146] H.-P. Breuer, "Genuine quantum trajectories for non-Markovian processes", Physical Review A **70**, 012106 (2004).

[147] J. Piilo, S. Maniscalco, K. Härkönen and K.-A. Suominen, "Non-Markovian Quantum Jumps", Physical Review Letters **100**, 180402 (2008).

[148] J. Preskill, "Lecture notes for Ph219/CS219: Quantum Information - Chapter 3", 65 (2018).

[149] L. Isserlis, "On a Formula for the Product-Moment Coefficient of any Order of a Normal Frequency Distribution in any Number of Variables", Biometrika **12**, 134–139 (1918).

[150] Á. Rivas and S. F. Huelga, *Open Quantum Systems: An Introduction* (Springer Science & Business Media, 2011).

[151] V. Gorini, A. Kossakowski and E. C. G. Sudarshan, "Completely positive dynamical semigroups of N-level systems", Journal of Mathematical Physics **17**, 821–825 (1976).

[152] G. Lindblad, "On the generators of quantum dynamical semigroups", Communications in Mathematical Physics **48**, 119–130 (1976).

[153] D. Manzano, "A short introduction to the Lindblad master equation", AIP Advances **10**, 025106 (2020).

[154] H.-P. Breuer, E.-M. Laine and J. Piilo, "Measure for the Degree of Non-Markovian Behavior of Quantum Processes in Open Systems", Physical Review Letters **103**, 210401 (2009).

[155] Á. Rivas, S. F. Huelga and M. B. Plenio, "Entanglement and Non-Markovianity of Quantum Evolutions", Physical Review Letters **105**, 050403 (2010).

[156] D. Chruściński, A. Kossakowski and Á. Rivas, "Measures of non-Markovianity: Divisibility versus backflow of information", Physical Review A **83**, 052128 (2011).

[157] D. Chruściński and S. Maniscalco, "Degree of Non-Markovianity of Quantum Evolution", Physical Review Letters **112**, 120404 (2014).

[158] M. J. W. Hall, J. D. Cresser, L. Li and E. Andersson, "Canonical form of master equations and characterization of non-Markovianity", Physical Review A **89**, 042120 (2014).

[159] L. Li, M. J. W. Hall and H. M. Wiseman, "Concepts of quantum non-Markovianity: A hierarchy", Physics Reports, Concepts of Quantum Non-Markovianity: A Hierarchy **759**, 1–51 (2018).

[160] U. Shrikant and P. Mandayam, "Quantum non-Markovianity: Overview and recent developments", Frontiers in Quantum Science and Technology **2**, 1134583 (2023).

[161] G. M. Palma, K.-a. Suominen and A. Ekert, "Quantum computers and dissipation", Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences **452**, 567–584 (1997).

[162] M. Brune, E. Hagley, J. Dreyer, X. Maître, A. Maali, C. Wunderlich, J. M. Raimond and S. Haroche, "Observing the Progressive Decoherence of the "Meter" in a Quantum Measurement", Physical Review Letters **77**, 4887–4890 (1996).

[163] W. H. Zurek, "Pointer basis of quantum apparatus: Into what mixture does the wave packet collapse?", Physical Review D **24**, 1516–1525 (1981).

[164] E. Strubell, A. Ganesh and A. McCallum, "Energy and policy considerations for deep learning in NLP", in Proceedings of the 57th annual meeting of the association for computational linguistics (2019), pp. 3645–3650.

[165] G.-B. Huang, Q.-Y. Zhu and C.-K. Siew, "Extreme learning machine: Theory and applications", Neurocomputing, Neural Networks **70**, 489–501 (2006).

[166] A. Opala, S. Ghosh, T. C. Liew and M. Matuszewski, "Neuromorphic Computing in Ginzburg-Landau Polariton-Lattice Systems", Physical Review Applied **11**, 064029 (2019).

[167] Z. Denis, I. Favero and C. Ciuti, "Photonic Kernel Machine Learning for Ultrafast Spectral Analysis", Physical Review Applied **17**, 034077 (2022).

[168] D. Ballarini, A. Gianfrate, R. Panico, A. Opala, S. Ghosh, L. Dominici, V. Ardizzone, M. De Giorgi, G. Lerario, G. Gigli, T. C. H. Liew, M. Matuszewski and D. Sanvitto, "Polaritonic Neuromorphic Computing Outperforms Linear Classifiers", Nano Letters **20**, 3506–3512 (2020).

[169] D. Pierangeli, G. Marcucci and C. Conti, "Photonic extreme learning machine by free-space optical propagation", Photonics Research **9**, 1446–1454 (2021).

[170] A. Jacot, F. Gabriel and C. Hongler, "Neural Tangent Kernel: Convergence and Generalization in Neural Networks", in Advances in Neural Information Processing Systems, Vol. 31 (2018).

[171] M. Schuld and N. Killoran, "Quantum Machine Learning in Feature Hilbert Spaces", Physical Review Letters **122**, 040504 (2019).

[172] S. Lloyd, M. Schuld, A. Ijaz, J. Izaac and N. Killoran, "Quantum embeddings for machine learning", `10.48550/arXiv.2001.03622` (2020).

[173] T. Kusumoto, K. Mitarai, K. Fujii, M. Kitagawa and M. Negoro, "Experimental quantum kernel trick with nuclear spins in a solid", npj Quantum Information **7**, 1–7 (2021).

[174] Y. Liu, S. Arunachalam and K. Temme, "A rigorous and robust quantum speed-up in supervised machine learning", Nature Physics **17**, 1013–1017 (2021).

[175] Y. Wu, B. Wu, J. Wang and X. Yuan, "Provable Advantage in Quantum Phase Learning via Quantum Kernel Alphatron", arXiv:2111.07553 [quant-ph] (2021).

[176] K. Fujii and K. Nakajima, "Harnessing Disordered-Ensemble Quantum Dynamics for Machine Learning", Physical Review Applied **8**, 024030 (2017).

[177] H. Xu, T. Krisnanda, W. Verstraelen, T. C. H. Liew and S. Ghosh, "Superpolynomial quantum enhancement in polaritonic neuromorphic computing", Physical Review B **103**, 195302 (2021).

[178] J. Suykens and J. Vandewalle, "Least Squares Support Vector Machine Classifiers", Neural Processing Letters **9**, 293–300 (1999).

[179] S. Jerbi, L. J. Fiderer, H. Poulsen Nautrup, J. M. Kübler, H. J. Briegel and V. Dunjko, "Quantum machine learning beyond kernel methods", Nature Communications **14**, 517 (2023).

[180] V. I. Paulsen and M. Raghupathi, *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*, Cambridge Studies in Advanced Mathematics (Cambridge University Press, Cambridge, 2016).

[181] B. Bordelon, A. Canatar and C. Pehlevan, "Spectrum Dependent Learning Curves in Kernel Regression and Wide Neural Networks", in Proceedings of the 37th International Conference on Machine Learning (2020), pp. 1024–1034.

[182] A. Canatar, B. Bordelon and C. Pehlevan, "Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks", Nature Communications **12**, 2914 (2021).

[183] K. Nakaji, H. Tezuka and N. Yamamoto, "Quantum-enhanced neural networks in the neural tangent kernel framework", arXiv:2109.03786 [quant-ph] (2021).

[184] T. Hastie and J. Zhu, "Comment: [Support vector machines with applications]", Statistical Science **21**, 352–357 (2006).

[185] N. Cristianini, J. Kandola, A. Elisseeff and J. Shawe-Taylor, "On Kernel Target Alignment", in *Innovations in Machine Learning: Theory and Applications*, edited by D. E. Holmes and L. C. Jain, Studies in Fuzziness and Soft Computing (Springer, Berlin, Heidelberg, 2006), pp. 205–256.

[186] M. Mohri, A. Rostamizadeh and A. Talwalkar, *Foundations of Machine Learning, second edition* (MIT Press, 2018).

[187] L. Banchi, J. Pereira and S. Pirandola, "Generalization in Quantum Machine Learning: A Quantum Information Standpoint", PRX Quantum **2**, 040321 (2021).

[188] A. Czerwinski, "Quantum state tomography with informationally complete POVMs generated in the time domain", Quantum Information Processing **20**, 105 (2021).

[189] A. Di Lorenzo, "Sequential Measurement of Conjugate Variables as an Alternative Quantum State Tomography", Physical Review Letters **110**, 010404 (2013).

[190] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik and J. L. O'Brien, "A variational eigenvalue solver on a photonic quantum processor", Nature Communications **5**, 4213 (2014).

[191] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow and J. M. Gambetta, "Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets", Nature **549**, 242–246 (2017).

[192] GOOGLE AI QUANTUM AND COLLABORATORS, F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, S. Boixo, M. Broughton, B. B. Buckley, D. A. Buell, B. Burkett, N. Bushnell, Y. Chen, Z. Chen, B. Chiaro, R. Collins, W. Courtney, S. Demura, A. Dunsworth, E. Farhi, A. Fowler, B. Foxen, C. Gidney, M. Giustina, R. Graff, S. Habegger, M. P. Harrigan, A. Ho, S. Hong, T. Huang, W. J. Huggins, L. Ioffe, S. V. Isakov, E. Jeffrey, Z. Jiang, C. Jones, D. Kafri, K. Kechedzhi, J. Kelly, S. Kim, P. V. Klimov, A. Korotkov, F. Kostritsa, D. Landhuis, P. Laptev, M. Lindmark, E. Lucero, O. Martin, J. M. Martinis, J. R. McClean, M. McEwen, A. Megrant, X. Mi, M. Mohseni, W. Mruczkiewicz, J. Mutus, O. Naaman, M. Neeley, C. Neill, H. Neven, M. Y. Niu, T. E. O'Brien, E. Ostby, A. Petukhov, H. Putterman, C. Quintana, P. Roushan, N. C. Rubin, D. Sank, K. J. Satzinger, V. Smelyanskiy, D. Strain, K. J. Sung, M. Szalay, T. Y. Takeshita, A. Vainsencher, T. White, N. Wiebe, Z. J. Yao, P. Yeh and A. Zalcman, "Hartree-Fock on a superconducting qubit quantum computer", Science **369**, 1084–1089 (2020).

[193] E. Farhi, J. Goldstone and S. Gutmann, *A Quantum Approximate Optimization Algorithm*, 2014.

[194] N. Lacroix, C. Hellings, C. K. Andersen, A. Di Paolo, A. Remm, S. Lazar, S. Krinner, G. J. Norris, M. Gabureac, J. Heinsoo, A. Blais, C. Eichler and A. Wallraff, "Improving the Performance of Deep Quantum Optimization Algorithms with Continuous Gate Sets", PRX Quantum **1**, 020304 (2020).

[195] M. P. Harrigan, K. J. Sung, M. Neeley, K. J. Satzinger, F. Arute, K. Arya, J. Atalaya, J. C. Bardin, R. Barends, S. Boixo, M. Broughton, B. B. Buckley, D. A. Buell, B. Burkett, N. Bushnell, Y. Chen, Z. Chen, Ben Chiaro, R. Collins, W. Courtney, S. Demura, A. Dunsworth, D. Eppens, A. Fowler, B. Foxen, C. Gidney, M. Giustina, R. Graff, S. Habegger, A. Ho, S. Hong, T. Huang, L. B. Ioffe, S. V. Isakov, E. Jeffrey, Z. Jiang, C. Jones, D. Kafri, K. Kechedzhi, J. Kelly, S. Kim, P. V. Klimov, A. N. Korotkov, F. Kostritsa, D. Landhuis, P. Laptev, M. Lindmark, M. Leib, O. Martin, J. M. Martinis, J. R. McClean, M. McEwen, A. Megrant, X. Mi, M. Mohseni, W. Mruczkiewicz, J. Mutus, O. Naaman, C. Neill, F. Neukart, M. Y. Niu, T. E. O'Brien, B. O'Gorman, E. Ostby, A. Petukhov, H. Putterman, C. Quintana, P. Roushan, N. C. Rubin, D. Sank, A. Skolik, V. Smelyanskiy, D. Strain, M. Streif, M. Szalay, A. Vainsencher, T. White, Z. J. Yao, P. Yeh, A. Zalcman, L. Zhou, H. Neven, D. Bacon, E. Lucero, E. Farhi and R. Babbush, "Quantum approximate optimization of non-planar graph problems on a planar superconducting processor", Nature Physics **17**, 332–336 (2021).

[196] M. Schuld, I. Sinayskiy and F. Petruccione, "The quest for a Quantum Neural Network", Quantum Information Processing **13**, 2567–2586 (2014).

[197] D. Marković and J. Grollier, "Quantum neuromorphic computing", Applied Physics Letters **117**, 150501 (2020).

[198] M. Benedetti, E. Lloyd, S. Sack and M. Fiorentini, "Parameterized quantum circuits as machine learning models", Quantum Science and Technology **4**, 043001 (2019).

[199] E. Farhi and H. Neven, "Classification with Quantum Neural Networks on Near Term Processors", arXiv:1802.06002 [quant-ph] (2018).

[200] A. V. Uvarov and J. D. Biamonte, "On barren plateaus and cost function locality in variational quantum algorithms", Journal of Physics A: Mathematical and Theoretical **54**, 245301 (2021).

[201] T. L. Patti, K. Najafi, X. Gao and S. F. Yelin, "Entanglement devised barren plateau mitigation", Physical Review Research **3**, 033090 (2021).

[202] R. Wiersema, C. Zhou, J. F. Carrasquilla and Y. B. Kim, *Measurement-induced entanglement phase transitions in variational quantum circuits*, 2021.

[203] J. Kim and Y. Oz, "Entanglement Diagnostics for Efficient Quantum Computation", Journal of Statistical Mechanics: Theory and Experiment **2022**, 073101 (2022).

[204] J. Kim and Y. Oz, "Quantum energy landscape and circuit optimization", Physical Review A **106**, 052424 (2022).

[205] J. Kim, Y. Oz and D. Rosa, *Quantum Chaos and Circuit Parameter Optimization*, 2022.

[206] S. H. Sack, R. A. Medina, A. A. Michailidis, R. Kueng and M. Serbyn, "Avoiding Barren Plateaus Using Classical Shadows", PRX Quantum **3**, 020365 (2022).

[207] L. Friedrich and J. Maziero, "Avoiding barren plateaus with classical deep neural networks", Physical Review A **106**, 042433 (2022).

[208] E. Grant, L. Wossnig, M. Ostaszewski and M. Benedetti, "An initialization strategy for addressing barren plateaus in parametrized quantum circuits", Quantum **3**, 214 (2019).

[209] H.-Y. Liu, Z.-Y. Chen, T.-P. Sun, Y.-C. Wu, Y.-J. Han and G.-P. Guo, *Mitigating Barren Plateaus with Transfer-learning-inspired Parameter Initializations*, 2022.

[210] K. Mitarai, Y. Suzuki, W. Mizukami, Y. O. Nakagawa and K. Fujii, "Quadratic Clifford expansion for efficient benchmarking and initialization of variational quantum algorithms", Physical Review Research **4**, 033012 (2022).

[211] G. S. Ravi, P. Gokhale, Y. Ding, W. M. Kirby, K. N. Smith, J. M. Baker, P. J. Love, H. Hoffmann, K. R. Brown and F. T. Chong, *CAFQA: A classical simulation bootstrap for variational quantum algorithms*, 2022.

[212] M. H. Cheng, K. E. Khosla, C. N. Self, M. Lin, B. X. Li, A. C. Medina and M. S. Kim, *Clifford Circuit Initialisation for Variational Quantum Algorithms*, 2022.

[213] J. Dborin, F. Barratt, V. Wimalaweera, L. Wright and A. G. Green, "Matrix product state pre-training for quantum machine learning", Quantum Science and Technology **7**, 035014 (2022).

[214] A. Pesah, M. Cerezo, S. Wang, T. Volkoff, A. T. Sornborger and P. J. Coles, "Absence of Barren Plateaus in Quantum Convolutional Neural Networks", Physical Review X **11**, 041011 (2021).

[215] L. Schatzki, M. Larocca, Q. T. Nguyen, F. Sauvage and M. Cerezo, *Theoretical Guarantees for Permutation-Equivariant Quantum Neural Networks*, 2022.

[216] Z. Holmes, A. Arrasmith, B. Yan, P. J. Coles, A. Albrecht and A. T. Sornborger, "Barren Plateaus Preclude Learning Scramblers", Physical Review Letters **126**, 190501 (2021).

[217] A. Arrasmith, M. Cerezo, P. Czarnik, L. Cincio and P. J. Coles, "Effect of barren plateaus on gradient-free optimization", Quantum **5**, 558 (2021).

[218] A. Arrasmith, Z. Holmes, M. Cerezo and P. J. Coles, "Equivalence of quantum barren plateaus to cost concentration and narrow gorges", Quantum Science and Technology **7**, 045015 (2022).

[219] S. Wang, P. Czarnik, A. Arrasmith, M. Cerezo, L. Cincio and P. J. Coles, *Can Error Mitigation Improve Trainability of Noisy Variational Quantum Algorithms?*, 2021.

[220] Y. Du, T. Huang, S. You, M.-H. Hsieh and D. Tao, "Quantum circuit architecture search for variational quantum algorithms", npj Quantum Information **8**, 1–8 (2022).

[221] K. Sharma, M. Cerezo, L. Cincio and P. J. Coles, "Trainability of Dissipative Perceptron-Based Quantum Neural Networks", Physical Review Letters **128**, 180505 (2022).

[222] K. Mitarai, M. Negoro, M. Kitagawa and K. Fujii, "Quantum Circuit Learning", Physical Review A **98**, 032309 (2018).

[223] D. A. Roberts and B. Yoshida, "Chaos and complexity by design", Journal of High Energy Physics **2017**, 121 (2017).

[224] S. Sim, P. D. Johnson and A. Aspuru-Guzik, "Expressibility and Entangling Capability of Parameterized Quantum Circuits for Hybrid Quantum-Classical Algorithms", Advanced Quantum Technologies **2**, 1900070 (2019).

[225] A. W. Harrow and R. A. Low, "Random Quantum Circuits are Approximate 2-designs", Communications in Mathematical Physics **291**, 257–302 (2009).

[226] F. G. S. L. Brandão, A. W. Harrow and M. Horodecki, "Local Random Quantum Circuits are Approximate Polynomial-Designs", Communications in Mathematical Physics **346**, 397–434 (2016).

[227] J. Haferkamp, "Random quantum circuits are approximate unitary t-designs in depth $O\left(nt\{5+o(1)\}\right)$", Quantum **6**, 795 (2022).

[228] P. Mujal, R. Martínez-Peña, J. Nokkala, J. García-Beni, G. L. Giorgi, M. C. Soriano and R. Zambrini, "Opportunities in Quantum Reservoir Computing and Extreme Learning Machines", Advanced Quantum Technologies **4**, 2100027 (2021).

[229] G. Marcucci, D. Pierangeli and C. Conti, "Theory of Neuromorphic Computing by Waves: Machine Learning by Rogue Waves, Dispersive Shocks, and Solitons", Physical Review Letters **125**, 093901 (2020).

[230] M. Cerezo and P. J. Coles, "Higher order derivatives of quantum neural networks with barren plateaus", Quantum Science and Technology **6**, 035006 (2021).

[231] J. Liu, Z. Lin and L. Jiang, *Laziness, Barren Plateau, and Noise in Machine Learning*, 2022.

[232] L. Bittel and M. Kliesch, "Training Variational Quantum Algorithms Is NP-Hard", Physical Review Letters **127**, 120502 (2021).

[233] E. R. Anschuetz and B. T. Kiani, "Beyond Barren Plateaus: Quantum Variational Algorithms Are Swamped With Traps", Nature Communications **13**, 7760 (2022).

[234] C. Zhao and X.-S. Gao, "Analyzing the barren plateau phenomenon in training quantum neural networks with the ZX-calculus", Quantum **5**, 466 (2021).

[235] R. S. Bennink, E. M. Ferragut, T. S. Humble, J. A. Laska, J. J. Nutaro, M. G. Pleszkoch and R. C. Pooser, "Unbiased simulation of near-Clifford quantum circuits", Physical Review A **95**, 062337 (2017).

[236] C. Piveteau, D. Sutter and S. Woerner, "Quasiprobability decompositions with reduced sampling overhead", npj Quantum Information **8**, 1–9 (2022).

[237] J. R. McClean, J. Romero, R. Babbush and A. Aspuru-Guzik, "The theory of variational hybrid quantum-classical algorithms", New Journal of Physics **18**, 023023 (2016).

[238] J. R. Seddon and E. T. Campbell, "Quantifying magic for multi-qubit operations", Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences **475**, 20190251 (2019).

[239] J. R. Seddon, B. Regula, H. Pashayan, Y. Ouyang and E. T. Campbell, "Quantifying Quantum Speedups: Improved Classical Simulation From Tighter Magic Monotones", PRX Quantum **2**, 010345 (2021).

[240] C. McDiarmid, "On the method of bounded differences", in *Surveys in Combinatorics, 1989: Invited Papers at the Twelfth British Combinatorial Conference*, edited by J. Siemons, London Mathematical Society Lecture Note Series (Cambridge University Press, Cambridge, 1989), pp. 148–188.

[241] C. Ferrie and J. Emerson, "Framed Hilbert space: hanging the quasi-probability pictures of quantum theory", New Journal of Physics **11**, 063040 (2009).

[242] H. Pashayan, J. J. Wallman and S. D. Bartlett, "Estimating Outcome Probabilities of Quantum Circuits Using Quasiprobabilities", Physical Review Letters **115**, 070501 (2015).

[243] V. Veitch, C. Ferrie, D. Gross and J. Emerson, "Negative quasi-probability as a resource for quantum computation", New Journal of Physics **14**, 113011 (2012).

[244] R. Raussendorf, J. Bermejo-Vega, E. Tyhurst, C. Okay and M. Zurel, "Phase-space-simulation method for quantum computation with magic states on qubits", Physical Review A **101**, 012350 (2020).

[245] R. Raussendorf, J. Bermejo-Vega, E. Tyhurst, C. Okay and M. Zurel, "Erratum: Phase-space-simulation method for quantum computation with magic states on qubits [Phys. Rev. A 101, 012350 (2020)]", Physical Review A **105**, 039902 (2022).

[246] H.-Y. Huang, R. Kueng and J. Preskill, "Predicting many properties of a quantum system from very few measurements", Nature Physics **16**, 1050–1057 (2020).

[247] H.-Y. Huang, R. Kueng, G. Torlai, V. V. Albert and J. Preskill, "Provably efficient machine learning for quantum many-body problems", Science **377**, eabk3333 (2022).

[248] D. E. Koh and S. Grewal, "Classical Shadows With Noise", Quantum **6**, 776 (2022).

[249] A. Elben, S. T. Flammia, H.-Y. Huang, R. Kueng, J. Preskill, B. Vermersch and P. Zoller, "The randomized measurement toolbox", Nature Reviews Physics **5**, 9–24 (2023).

[250] C. Benedetti, M. G. A. Paris and S. Maniscalco, "Non-Markovianity of colored noisy channels", Physical Review A **89**, 012114 (2014).

[251] K. E. Cahill and R. J. Glauber, "Ordered Expansions in Boson Amplitude Operators", Physical Review **177**, 1857–1881 (1969).

[252] K. P. Murphy, *Machine Learning: A Probabilistic Perspective* (MIT Press, 2012).

[253] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* (MIT Press, 2002).

[254] C. Williams and M. Seeger, "Using the Nyström Method to Speed up Kernel Machines", in Advances in neural information processing systems 13 (2001), pp. 682–688.