Surapprentissage bénin et noyaux quantiques

Joachim TOMASI[1,2] Sandrine ANTHOINE[2] Hachem KADRI[1]

[1]Laboratoire d'Informatique et des Systèmes, 163 avenue de Luminy, 13288 Marseille Cedex 09, France

[2]Institut de Mathématiques de Marseille, 3 place Victor Hugo, 13331 Marseille Cedex 3, France

Résumé – Cet article aborde le défi de la généralisation dans l'apprentissage automatique quantique (QML) avec les noyaux quantiques. Bien que ces derniers offrent un potentiel d'expressivité élevé en exploitant des espaces de Hilbert de grande dimension, ils souffrent fréquemment d'une concentration exponentielle des valeurs du noyau, conduisant à un surapprentissage et à une mauvaise généralisation. Pour pallier cette limitation, nous nous inspirons du phénomène de surapprentissage bénin récemment observé en apprentissage classique. Nous introduisons un nouveau cadre pour la construction de noyaux quantiques : le noyau quantique local-global. Cette approche combine un noyau quantique local, basé sur des mesures de sous-systèmes quantiques, et un noyau quantique global, issu de mesures de l'ensemble du système. Grâce à des expériences numériques, nous montrons que le noyau local-global présente un comportement de surapprentissage bénin, interpolant les données d'entraînement tout en maintenant de bonnes performances de généralisation. Nos résultats soulignent l'efficacité de cette stratégie pour améliorer la conception de noyaux quantiques performants.

Abstract – This paper addresses the challenge of generalization in quantum machine learning (QML) with quantum kernels. While these kernels offer high expressivity by exploiting large-dimensional Hilbert spaces, they frequently suffer from exponential concentration of kernel values, leading to overfitting and poor generalization. To overcome this limitation, we draw inspiration from the phenomenon of benign overfitting recently observed in classical learning. We introduce a novel framework for constructing quantum kernels: the local-global quantum kernel. This innovative approach combines a local quantum kernel, based on measurements of quantum subsystems, and a global quantum kernel, derived from measurements of the entire system. Through numerical experiments, we show that the local-global kernel exhibits benign overfitting, interpolating the training data while maintaining good generalization performance. Our results highlight the effectiveness of this strategy for improving the design of high-performing quantum kernels.

1 Introduction

Parmi les différentes approches de l'apprentissage automatique quantique (QML), les noyaux quantiques projettent des données classiques dans des états quantiques, correspondant à des espaces de Hilbert de haute dimension, dans le but d'exploiter les avantages computationnels potentiels des dispositifs quantiques [5, 11, 12]. Cependant, en grande dimension, les noyaux quantiques tendent à produire des matrices de noyau proches de la matrice identité [7]. Ce comportement limite leur capacité à extraire des corrélations significatives entre les points de données, donnant des modèles favorisant un surapprentissage et ayant de mauvaises performances de généralisation.

Pour permettre un meilleure généralisation, des méthodes d'ajustement de la largeur de bande des noyaux quantiques et de réduction de la dimensionnalité de l'espace des caractéristiques quantiques ont été proposées [3, 6, 7]. Notre approche adopte une voie différente : elle s'inspire des résultats récents en apprentissage automatique sur les modèles surparamétrés qui parviennent à généraliser même quand ils sont entraînés de façon à (quasi) annuler le risque empirique [13]. Dans notre travail, nous étendons ces travaux en proposant un cadre pour les noyaux quantiques qui favorise intrinsèquement le surapprentissage bénin [1]. En nous inspirant du concept de noyaux "*spiky-smooth*" [4], nous proposons un noyau quantique *local-global* construit comme une somme pondérée de deux noyaux quantiques. Une composante est dérivée d'une mesure locale (c'est-à-dire de faible dimension par rapport à l'espace d'intégration quantique complet) qui imite le comportement du noyau lisse, tandis que l'autre provient d'une mesure globale de l'ensemble de l'espace des caractéristiques quantiques, capturant la composante "spiky". Ce noyau quantique local-global conduit à une bonne généralisation tout en interpolant les données d'apprentissage, conduisant à un surapprentissage bénin en apprentissage automatique quantique.

Notations En notation quantique *bra-ket*, les vecteurs colonne (appelés *kets*) sont notés $|\psi\rangle$, tandis que leurs duaux, les vecteurs ligne (*bras*), s'écrivent $\langle\psi| = |\psi\rangle^\dagger$, où $\dagger$ représente l'adjoint (transposée conjuguée). Le produit scalaire entre deux états $|\psi\rangle$ et $|\phi\rangle$ se note $\langle\psi|\phi\rangle$, et leur produit tensoriel s'exprime sous les formes $|\psi\phi\rangle$, $|\psi\rangle|\phi\rangle$ ou $|\psi\rangle \otimes |\phi\rangle$. Un produit tensoriel répété t-fois s'abrège en $|\psi\rangle^{\otimes t}$ ou $|\psi^t\rangle$. Les états purs sont décrits par des matrices de densité $\rho = |\psi\rangle\langle\psi|$, où chaque état de base $|i\rangle$ ($i \in \{0, 1, \ldots, 2^n - 1\}$) correspond à un produit tensoriel d'états de qubits individuels. Les observables quantiques, représentés par des opérateurs hermitiens O, ont une valeur espérée donnée par $\langle O\rangle_\rho = \mathrm{Tr}(\rho O)$, l'indice ρ étant souvent omis lorsque le contexte est clair. Pour une introduction complète à l'informatique quantique, les lecteurs peuvent se référer à [9].

2 Contexte

Nous décrivons dans cette section les principaux travaux sur le surapprentissage bénin et sur les noyaux quantiques qui ont conduit à cet article.

2.1 Surapprentissage bénin

Les réseaux neuronaux surparamétrés défient le compromis biais-variance classique en généralisant efficacement malgré une interpolation des données d'apprentissage. Cette observation a motivé une reconsidération des mécanismes de généralisation en apprentissage automatique [13]. Ce phénomène, peut être interprété comme du *surapprentissage bénin*, permet aux modèles de mémoriser le bruit (surapprend), sans dégrader (bénin) leurs performances en généralisation [1]. Il se distingue des régimes de surapoprentissage tempérés ou catastrophiques, où le surapprentissage nuit aux prédictions [8].

Le surapprentissage bénin en régression linéaire ou par noyaux peut être expliqué grâce à une décomposition "simple-plus-spiky" de l'interpolant de norme minimale [2]. Dans cette décomposition, la composante "simple" capture le signal principal dans les données, tandis que la composante "spiky" interpole localement le bruit sans affecter significativement la performance de la prédiction.

Considérons le modèle linéaire $y_i = \langle \theta^*, x_i \rangle + \epsilon_i$, où θ^* est le vecteur de paramètres et ϵ_i est le bruit. Dans le régime surparamétré ($d \gg n$), l'interpolant de norme minimale est donné par $\hat{\theta} = X^\top (XX^\top)^+ y$, où X est la matrice de données. Nous pouvons toujours choisir un $l \in \mathbb{N}$ pour diviser les caractéristiques et les paramètres comme $x = [x_{\leq l}, x_{>l}]$ et $\hat{\theta} = [\hat{\theta}_{\leq l}, \hat{\theta}_{>l}]$, de sorte que la fonction de prédiction et la matrice de covariance se décomposent en :

$$\hat{f}(x) = \langle \hat{\theta}_{\leq l}, x_{\leq l} \rangle + \langle \hat{\theta}_{>l}, x_{>l} \rangle$$

et

$$XX^\top = X_{\leq l} X_{\leq l}^\top + X_{>l} X_{>l}^\top.$$

Si le terme $X_{>l} X_{>l}^\top$ a un spectre suffisamment plat pour un certain l, nous pouvons l'approximer par ρI_n, de sorte que $XX^\top \approx X_{\leq l} X_{\leq l}^\top + \rho I_n$. Sous cette approximation, l'estimateur pour les l premières composantes ressemble à la solution de la régression ridge :

$$\hat{\theta}_{\leq l} \approx \arg\min_{\theta \in \mathbb{R}^l} \|X_{\leq l}\theta - y\|_2^2 + \rho \|\theta\|_2^2.$$

Ici, la composante simple, $\hat{\theta}_{\leq l}$, capture le signal principal, tandis que la composante "spiky", $\langle \hat{\theta}_{>l}, x_{>l} \rangle$, tient compte de l'interpolation du bruit résiduel.

Une décomposition similaire s'applique à la régression par noyaux. La matrice du noyau peut être exprimée comme $K = K_{\leq l} + K_{>l} \approx K_{\leq l} + \rho I$, séparant l'interpolant de norme minimale calculé dans l'espace de Hilbert à noyau reproduisant (RKHS) en une décomposition "simple plus spiky" [2, 8, 10]. Récemment, [4] a proposé une règle de prédiction basée sur un noyau "spiky-smooth" de la forme $k_{\rho,\gamma}(x, z) = \tilde{k}(x, z) + \rho \hat{k}_\gamma(x, z)$, où $\tilde{k}$ est un noyau universel lisse et $\hat{k}_\gamma$ est un noyau "spiky", permettant un surapprentissage bénin.

2.2 Noyaux quantiques

Les algorithmes quantiques requièrent l'encodage des données classiques x_i en états quantiques $|\phi_{x_i}\rangle$. Cette transformation définit alors une représentation explicite des caractéristiques dans un espace de Hilbert définissant les états quantiques [5, 12]. Considérons une représentation quantique qui encode un point de données x dans un état quantique représenté par la matrice de densité ρ_x. Un noyau quantique $k(x, z)$ est défini comme le produit scalaire de Hilbert-Schmidt entre les états ρ_x et ρ_z, donné par:

$$k(x, z) = \text{Tr}[\rho_x \rho_z]. \tag{1}$$

Lorsque la donnée x est encodée dans un état quantique pur de la forme $x \to \rho_x = |\phi_x\rangle\langle\phi_x|$, le noyau quantique en (1) se simplifie à:

$$k(x, z) = |\langle \phi_x | \phi_z \rangle|^2. \tag{2}$$

Ceci correspond à la mesure de la fidélité entre les états quantiques $|\phi_x\rangle$ et $|\phi_z\rangle$, faisant de ce noyau le noyau de fidélité quantique. Cependant, en grande dimension (c.a.d., quand le nombre de qubits utilisés pour l'encodage augmente), ce type de noyaux donne des matrices noyaux proche de la matrice Identité, entraînant un surapprentissage et une mauvaise généralisation. Si des approches comme l'introduction de largeur de bande pour les noyaux quantiques [3] ou la réduction de dimension de l'état [6, 7] permettent d'atténuer ce phénomène, nous montrons que le surapprentissage bénin peut, au contraire, favoriser une meilleure généralisation des noyaux quantiques.

3 Surapprentissage bénin avec le noyau quantique local-global

Dans cette section, nous présentons de nouvelles contributions qui étendent la littérature existante. Nous concevons un nouveau cadre tenant compte du phénomène de surapprentissage bénin pour construire des noyaux quantiques capables d'atteindre une bonne généralisation. Notre cadre est basé sur la notion de noyau quantique local-global, qui peut être considéré comme un analogue quantique du noyau classique "spiky-smooth" [4].

3.1 Définition du noyau quantique local-global

Définition 1 (noyau quantique local-global). Soit $U(x)$ un opérateur unitaire agissant sur t qubits. Définissons les états quantiques locaux et globaux par:

$$\rho_L^x = U(x) \left(L_s \otimes \frac{1}{2^{t-s}} I_{t-s} \right) U^\dagger(x),\ \rho_G^x = U(x) G_t U^\dagger(x),$$

où L_s est un état quantique pur, c'est-à-dire un projecteur de rang un, de $s < t$ qubits et G_t est un état quantique pur de t qubits. Les noyaux quantiques locaux et globaux correspondants sont définis, respectivement, comme suit: $k_L(x, z) = \text{Tr}[\rho_L^x \rho_L^z]$ et $k_G(x, z) = \text{Tr}[\rho_G^x \rho_G^z]$. Le noyau quantique local-global est donné par la somme pondérée:

$$k_{LG}(x, z) = \lambda_L k_L(x, z) + \lambda_G k_G(x, z), \tag{3}$$

où λ_L et λ_G sont des poids scalaires.

Le terme "*local-global*" provient des termes, L_s qui agit comme un projecteur de rang un local et G_t qui agit comme un projecteur de rang un global. Plus précisément, le noyau local est dérivé d'une mesure d'un sous-ensemble de s qubits, tandis que le noyau global est obtenu par une mesure de l'ensemble des t qubits. L'intuition clé derrière notre construction est quand t augmente, la matrice noyau générée par le noyau

global k_G devrait tendre vers la matrice identité. Ceci découle du fait que les produits scalaires entre les états quantiques devraient s'annuler quand le nombre de qubits augmente. En revanche, le noyau local k_L, dérivé d'une mesure locale, devrait produire une matrice de noyau plus lisse avec des éléments hors diagonale plus importants. Suivant l'idée des noyaux "spiky-smooth", le noyau quantique local-global est conçu pour capturer le signal principal grâce à sa composante locale tout en permettant l'interpolation localisée du bruit via sa partie globale, conduisant ainsi à un surapprentissage bénin.

3.2 Encodage global séparable

Pour simplifier l'analyse, nous supposons que le projecteur global G_t et l'unitaire d'encodage $U(x)$ admettent une factorisation séparable en q fois de s-qubits:

$$G_t = L_s^{\otimes q}, \quad U(x) = V_s(x)^{\otimes q}, \quad t = qs,$$

où chaque $V_s(x)$ agit sur s qubits. Nous appelons ce schéma Encodage Global Séparable. Dans ce contexte, la matrice de densité réduite de ρ_L^x s'écrit comme:

$$\tilde{\rho}_L^x := \mathrm{Tr}_{s+1:t}[\rho_L^x] = V_s(x) L_s V_s^\dagger(x),$$

où $\mathrm{Tr}_{s+1:t}$ est l'opération de trace partielle [9] sur les qubits $s+1$ à t. Le noyau local est alors exprimé par:

$$\begin{aligned} k_L(x,z) &:= \mathrm{Tr}[\rho_L^x \rho_L^z] \\ &= \mathrm{Tr}[(\tilde{\rho}_L^x \otimes \frac{1}{2^{t-s}} I_{t-s})(\tilde{\rho}_L^z \otimes \frac{1}{2^{t-s}} I_{t-s})] \\ &= \frac{1}{2^{2(t-s)}} \mathrm{Tr}[\tilde{\rho}_L^x \tilde{\rho}_L^z], \end{aligned}$$

et le noyau global est donné par:

$$\begin{aligned} k_G(x,z) &:= \mathrm{Tr}[\rho_G^x \rho_G^z] \\ &= \mathrm{Tr}[(V_s(x) L_s V_s^\dagger(x) V_s(z) L_s V_s^\dagger(z))^{\otimes q}] \\ &= \mathrm{Tr}[V_s(x) L_s V_s^\dagger(x) V_s(z) L_s V_s^\dagger(z)]^q = \mathrm{Tr}[\tilde{\rho}_L^x \tilde{\rho}_L^z]^q. \end{aligned}$$

Ainsi, le noyau local-global se réduit à :

$$k_{LG}(x,z) = \tilde{\lambda}_L k(x,z) + \lambda_G k(x,z)^q, \tag{4}$$

où $k(x,z) = \mathrm{Tr}[\tilde{\rho}_L^x \tilde{\rho}_L^z]$ et $\tilde{\lambda}_L = \frac{\lambda_L}{2^{2(t-s)}}$. Le noyau quantique local-global inclut un paramètre q, qui détermine le degré du noyau et sert de paramètre de réglage pour contrôler la largeur de bande de la composante globale. La Figure 1 illustre cette idée. Le noyau local-global suit de près la composante locale mais s'en écarte dans des régions étroites spécifiques. Quand le paramètre q augmente, ces déviations deviennent plus localisées, renforçant l'analogie avec les noyaux "spiky-smooth".

4 Résultats expérimentaux

Dans cette section, nous présentons des simulations numériques montrant le surapprentissage bénin du noyau quantique local-global. Le noyau local-global utilisé dans nos expériences est défini comme $k_{LG}(x,z) = k_c(x,z) + \rho k_c(x,z)^q$, ce qui correspond au noyau de l'équation Eq(4) où $\tilde{\lambda}_L = 1$,

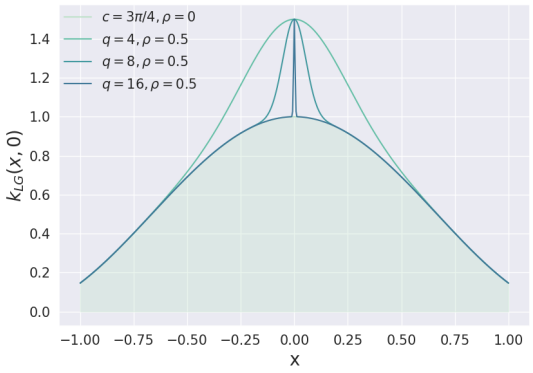

Figure 1: Noyaux local-global en dimension $d = 1$, avec $k_{LG}(x,z) = \cos^2\left(\frac{c(x-z)}{2}\right) + \rho\cos^{2q}\left(\frac{c(x-z)}{2}\right)$, où $c = \frac{3\pi}{4}$, $\rho = 0.5$ et $q = 4, 8, 16$. Comme nous le verrons dans la section 4, le noyau $k(x,z) = \cos^2\left(\frac{c(x-z)}{2}\right)$ est le noyau de fidélité quantique obtenu par l'encodage quantique d'angle.

$\lambda_G = \rho$ et k_c défini comme dans [3] par l'équation (6). La représentation quantique des données est défini comme suit :

$$U_c(x)|0^d\rangle = \bigotimes_{j=1}^{d} \left[\cos\left(\frac{cx_j}{2}\right)|0\rangle + i\sin\left(\frac{cx_j}{2}\right)|1\rangle\right], \tag{5}$$

et le noyau quantique résultant est alors :

$$\begin{aligned} k_c(x,z) &= |\langle 0^d | U_c^\dagger(z) U_c(x) | 0^d \rangle|^2 \\ &= \prod_{j=1}^{d} \cos^2\left(\frac{c(x_j - z_j)}{2}\right). \end{aligned} \tag{6}$$

Nous générons $n = 8$ échantillons selon $y = f(x) + \epsilon$, avec $\epsilon \sim \mathcal{N}(0, 0.5)$ et $x \sim \mathcal{U}([-0.75, 0.75])$. La fonction cible f est définie dans l'espace de Hilbet à noyau reproduisant comme $f(x) = 1 + \sum_{i=1}^{5} k_c(x, w_i)$, avec $c = \frac{3\pi}{4}$ et $\{w_i\}_{i=1}^5$ tirés de $\mathcal{U}([-0.75, 0.75])$. La Figure 2 illustre la régression par moindres carrés (sans régularisation) avec le noyau quantique local-global. Sans la composante globale, k_c n'est pas assez d'expressif pour interpoler les données d'entraînement. En ajoutant la composante global, l'interpolation des données devient possible, menant au surapprentissage. Celui-ci varie avec le degré q du noyau : catastrophique pour $q = 4$, tempéré pour $q = 8$ et bénin pour $q = 16$. Quand q augmente, le modèle sans régularisation se rapproche du régresseur avec noyau local et une régularisation ridge avec un paramètre ρ, indiquant que la composante globale joue un rôle de régularisation implicite lorsque q est suffisamment grand. Cette observation est cohérente avec les résultats de [4].

Nous menons également des expériences avec une dimension plus grande d. L'ensemble de données comprend $n = 200$ échantillons iid $x_i \sim \mathcal{U}([-1, 1]^d)$, avec $d = 20$. Les étiquettes sont générées selon le modèle : $y = \sum_{j=1}^{d} \cos(0.01\pi x^{(j)}) + \epsilon$, où $\epsilon \sim \mathcal{N}(0, 0.5)$. Le Tableau 1 compare les performances de la régression Ridge et sans Ridge avec le noyau quantique local, ainsi que la régression sans Ridge utilisant le noyau quantique local-global. Les résultats montrent que le noyau quantique local-global résout efficacement le problème de généralisation des noyaux quantiques, permettant un surapprentissage bénin.

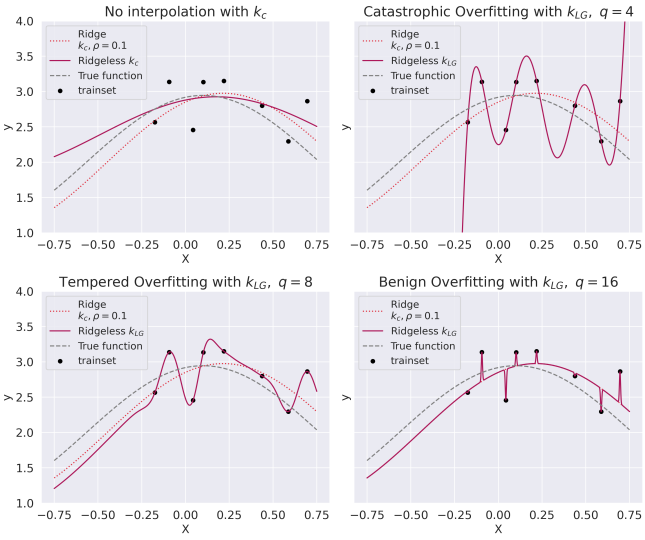

Figure 2: Comportement de surapprentissage des noyaux local-global $k_{LG}(x,z) = k_c(x,z) + \rho k_c(x,z)^q$ basé sur le noyau $k_c(x,z) = \cos^2\left(\frac{c(x-z)}{2}\right)$, en fonction du degré q, avec $\rho = 0.1$. Le noyau quantique local k_c sans la composante globale n'est pas suffisamment expressif pour surapprendre les données d'apprentissage (en haut à gauche). Le surapprentissage est catastrophique pour $q = 4$ (en haut à droite), tempéré pour $q = 8$ (en bas à gauche) et bénin pour $q = 16$ (en bas à droite).

5 Conclusion

Nous avons introduit une nouvelle approche de construction de noyaux quantiques, appelée noyaux '*local-global*', qui combine des composantes locales et globales pour permettre un surapprentissage bénin dans l'apprentissage automatique quantique. En exploitant l'encodage global séparable, nous offrons un mécanisme simple pour contrôler la largeur de bande de la composante globale du noyau. Nos résultats empiriques démontrent que l'ajustement de la composante globale favorise le surapprentissage bénin. Bien que cette approche soit conçue pour favoriser le surapprentissage bénin, elle présente également une stratégie prometteuse pour la construction de noyaux quantiques efficaces.

Table 1: Comparaison des performances, en termes d'erreur quadratique moyenne (MSE), de la régression ridge et sans Ridge avec le noyau quantique local k_c par rapport à la régression sans Ridge utilisant le noyau quantique local-global k_{LG}. $k_{LG}(x,z) = k_c(x,z) + \rho k_c(x,z)^q$, où $k_c(x,z) = \prod_{j=1}^{d} \cos^2\left(\frac{c(x_j - z_j)}{2}\right)$ avec $c = \frac{\pi}{20}$.

Type de Noyau	ρ	q	Train MSE	Test MSE
Local (Ridge 0.07)	0.0	–	2.44e-01	0.293
Local	0.0	–	2.42e-29	40.464
Local-Global	0.07	3	1.94e-23	0.844
Local-Global	0.07	5	2.38e-26	0.396
Local-Global	0.07	7	1.09e-27	0.286

References

[1] Peter L BARTLETT, Philip M LONG, Gábor LUGOSI et Alexander TSIGLER : Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

[2] Peter L BARTLETT, Andrea MONTANARI et Alexander RAKHLIN : Deep learning: a statistical viewpoint. *Acta numerica*, 2021.

[3] Abdulkadir CANATAR, Evan PETERS, Cengiz PEHLEVAN, Stefan M WILD et Ruslan SHAYDULIN : Bandwidth enables generalization in quantum kernel models. *arXiv preprint arXiv:2206.06686*, 2022.

[4] Moritz HAAS, David HOLZMÜLLER, Ulrike LUXBURG et Ingo STEINWART : Mind the spikes: Benign overfitting of kernels and neural networks in fixed dimension. *Advances in Neural Information Processing Systems*, 2024.

[5] Vojtěch HAVLÍČEK, Antonio D CÓRCOLES, Kristan TEMME, Aram W HARROW, Abhinav KANDALA, Jerry M CHOW et Jay M GAMBETTA : Supervised learning with quantum-enhanced feature spaces. *Nature*, 2019.

[6] Hsin-Yuan HUANG, Michael BROUGHTON, Masoud MOHSENI, Ryan BABBUSH, Sergio BOIXO, Hartmut NEVEN et Jarrod R MCCLEAN : Power of data in quantum machine learning. *Nature communications*, 2021.

[7] Jonas KÜBLER, Simon BUCHHOLZ et Bernhard SCHÖLKOPF : The inductive bias of quantum kernels. *Advances in Neural Information Processing Systems*, 2021.

[8] Neil MALLINAR, James SIMON, Amirhesam ABEDSOLTAN, Parthe PANDIT, Misha BELKIN et Preetum NAKKIRAN : Benign, tempered, or catastrophic: Toward a refined taxonomy of overfitting. *Advances in Neural Information Processing Systems*, 2022.

[9] Michael A. NIELSEN et Isaac L. CHUANG : *Quantum Computation and Quantum Information*. Cambridge University Press, 2010.

[10] Evan PETERS et Maria SCHULD : Generalization despite overfitting in quantum machine learning models. *Quantum*, 2023.

[11] Pablo RODRIGUEZ-GRASA, Yue BAN et Mikel SANZ : Training embedding quantum kernels with data reuploading quantum neural networks. *arXiv preprint arXiv:2401.04642*, 2024.

[12] Maria SCHULD et Nathan KILLORAN : Quantum machine learning in feature Hilbert spaces. *Physical review letters*, 2019.

[13] Chiyuan ZHANG, Samy BENGIO, Moritz HARDT, Benjamin RECHT et Oriol VINYALS : Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 2021.