

Towards provably efficient quantum algorithms for large-scale machine-learning models

Junyu Liu,^{1,2,3,4,5,6} Minzhao Liu,^{7,8} Jin-Peng Liu,^{9,10,11} Ziyu Ye,² Yuri Alexeev,^{8,2,3} Jens Eisert,¹² and Liang Jiang^{1,3}

¹*Pritzker School of Molecular Engineering, The University of Chicago, Chicago, IL 60637, USA*

²*Department of Computer Science, The University of Chicago, Chicago, IL 60637, USA*

³*Chicago Quantum Exchange, Chicago, IL 60637, USA*

⁴*Kadanoff Center for Theoretical Physics, The University of Chicago, Chicago, IL 60637, USA*

⁵*qBraid Co., Chicago, IL 60615, USA*

⁶*SeQure, Chicago, IL 60615, USA*

⁷*Department of Physics, The University of Chicago, Chicago, IL 60637, USA*

⁸*Computational Science Division, Argonne National Laboratory, Lemont, IL 60439, USA*

⁹*Simons Institute for the Theory of Computing, University of California, Berkeley, CA 94720, USA*

¹⁰*Department of Mathematics, University of California, Berkeley, CA 94720, USA*

¹¹*Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

¹²*Dahlem Center for Complex Quantum Systems, Free University Berlin, Berlin, 14195, Germany*

Large machine learning models are revolutionary technologies of artificial intelligence whose bottlenecks include huge computational expenses, power, and time used both in the pre-training and fine-tuning process. In this work, we show that fault-tolerant quantum computing could possibly provide provably efficient resolutions for generic (stochastic) gradient descent algorithms, scaling as $\mathcal{O}(T^2 \times \text{polylog}(n))$, where n is the size of the models and T is the number of iterations in the training, as long as the models are both sufficiently dissipative and sparse, with small learning rates. Based on earlier efficient quantum algorithms for dissipative differential equations, we find and prove that similar algorithms work for (stochastic) gradient descent, the primary algorithm for machine learning. In practice, we benchmark instances of large machine learning models from 7 million to 103 million parameters. We find that, in the context of sparse training, a quantum enhancement is possible at the early stage of learning after model pruning, motivating a sparse parameter download and re-upload scheme. Our work shows solidly that fault-tolerant quantum algorithms could potentially contribute to most state-of-the-art, large-scale machine-learning problems.

It is widely believed that large-scale machine learning might be one of the most revolutionary technologies benefiting society [1], including already important breakthroughs in digital arts [2], conversation like GPT-3 [3, 4], and mathematical problem solving [5]. However, training such models with considerable parameters is costly and has high carbon emissions. For instance, twelve million dollars and over five-hundred tons of CO₂ equivalent emissions have been produced to train GPT-3 [6]. Thus, on the one hand, it is important to make large-scale machine-learning models (like large language models, LLM) sustainable and efficient.

On the other hand, machine learning might possibly be one of the flag applications of quantum technology. Running machine learning algorithms on quantum devices, implementing readings of so-called *quantum machine learning*, is widely seen as a potentially very fruitful application of quantum algorithms [7]. Despite rapid development and significant progress, current quantum machine learning algorithms feature substantial limitations both in theory and practice. First, practical applica-

tions of quantum machine learning algorithms for near-term devices are often lacking theoretical grounds that guarantee or at least plausibly suggest to outperform their classical counterparts. Second, for fault-tolerant settings of quantum machine learning problems [8–16], rigorous super-polynomial quantum speedups can actually be proven [17–19] for highly structured problems. That said, these prescriptions are arguably still far from real state-of-the-art applications of classical machine learning. Some of them are primarily using quantum states as training data instead of classical data, which can be—highly encouraging as these approaches are—argued to be not the currently most important classical machine learning application [18, 20–23]. Efforts need to be made to extend our understanding of quantum machine learning, in the sense that we have to understand how they could have theoretical guarantees and how they could solve timely and natural problems, at least in principle, of classical machine learning. For instance, they should relate to scalable and sustainable natural problems in large-scale machine-learning.

In this work, we take significant steps in this direction by designing end-to-end quantum machine learning algorithms that are expected to be timely for the current machine learning community and that are to an extent equipped with guarantees. Based on a typical large-scale (classical) machine-learning process (see Fig. 1 for an illustration), we find that after a significant number of neural network training parameters have been pruned (sparse training) [24–27] and the classical training parameters compiled to a quantum computer, we suggest to find a quantum enhancement at the early state of training before the error grows exponentially. At its heart, the quantum algorithm part of the work includes suitable modifications of the quantum algorithm [28] for solving differential equations to running (stochastic) gradient descent algorithms—presumably the primary classical machine learning algorithm—into a quantum processor after linearization. The expectation of a possible quantum enhancement is rooted in an application of a variant of the so-called *Harrow-Hassidim-Lloyd* (HHL) algorithm [29], an efficient quantum algorithm for sparse matrix inversion that solves the problem within $\mathcal{O}(\log n)$ time for suitably conditioned $n \times n$ sparse matrices. We find that our algorithm can solve large-scale model-dimension- n machine learning problems in $\mathcal{O}(\text{polylog}(n) \times T)$ or $\mathcal{O}(\text{polylog}(n) \times T^2)$

time, where T is the number of iterations. The scaling in n outperforms the scaling of any classical algorithms we know of. However, for a given machine learning problem with required performances, there is no guarantee that our hybrid quantum-classical algorithm will necessarily outperform all other conceivable classical algorithms for related, but different tasks (for instance, for algorithms that are not gradient-based). Thus, our result gives, to the best of our knowledge, rise to a potential substantial quantum speedup or enhancement of particular classical algorithms, instead of a quantum advantage over the entire problem class.

It is important to stress that the above algorithm has a number of requirements that do admit a quantum enhancement. First, both the machine learning model and the weight vectors have to be sufficiently *sparse*, which will ensure a fast interface between classical and quantum processors (this requirement could be relaxed in the presence of *quantum random access memory* (QRAM) [30], a fast uploader towards quantum data, but we stress that this is *not* required and there are no hidden resources in our scheme). Second, the model has to be sufficiently *dissipative*. For dissipative systems, the linearization error is well controlled, ensuring that the HHL algorithm can obtain reliable results even with non-linear machine learning models. We find dissipation happens generically in the early training process of large-scale machine learning.

We corroborate the intuition developed here by a number of theorems, as well as extensive numerical experiments. The formal definition of dissipation, sparsity, and quantum speedups are rigorously proven in the supplementary material. Informal readings of the main theorems are presented in Section I, while solid numerical evidence up to 103 million training parameters are provided in Section II. Finally, a conclusion providing also an outlook will be provided in Section III.

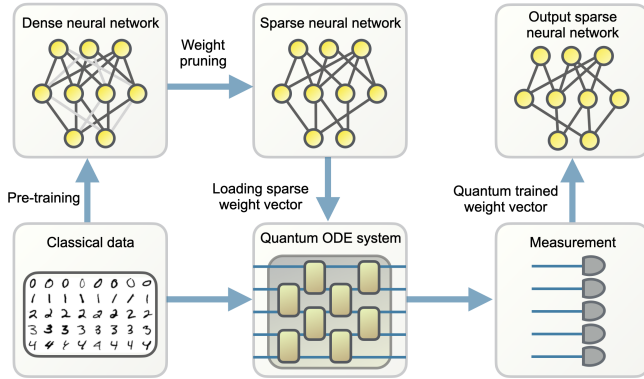


Figure 1. A possible learning process in large-scale models, which might use sparse training, whose early stage in learning might admit possible quantum enhancement. A dense neural network is pre-trained classically. The neural network weights are then pruned and only a small fraction is preserved. A quantum ordinary difference equation system that corresponds to the sparse training dynamics is created using the sparse network and the training data. To allow quantum enhancement, the system must be sparse and dissipative. Measurement on the solution state is performed to obtain final training parameters, used to construct a trained classical sparse neural network.

I. THEOREMS

In this section, we will lay out the informally formulated main theorems that are established in this work. Details can be found in the supplementary material.

Theorem 1 (Informal). *For a sparse machine learning model with model size n , running T iterations, with the algorithm being fully dissipative with small learning rates (whose formal definition is given in the Appendix), there is a quantum algorithm that runs in*

$$\mathcal{O}\left(T \times \text{poly}\left(\log n, \frac{1}{\epsilon}\right)\right) \quad (1)$$

time with precision $\epsilon > 0$. The sparsity condition also ensures the efficiency of uploading and downloading quantum states towards classical processors.

Theorem 2 (Informal). *For a sparse machine learning model with model size n , running in T iterations, and the algorithm being almost dissipative with small learning rates (whose formal definition is given in the Appendix), then there is a quantum algorithm runs in*

$$\mathcal{O}\left(T^2 \times \text{poly}\left(\log n, \frac{1}{\epsilon}\right)\right) \quad (2)$$

time with precision $\epsilon > 0$. The sparsity condition also ensures the efficiency of uploading and downloading quantum states towards classical processors.

In these expressions, $m := \log_2(n)$ takes the role of a system size of the quantum system. First, we describe the problem we are trying to solve. A machine learning model is defined partially by a function \mathcal{L}_A , called the *loss function*, as a function of weight vector (variational angle) $\theta \in \mathbb{R}^n = (\theta_\mu)$, and the *input training set* \mathcal{A} . The weight vector has n components if an n -dimensional model. The task is to minimize the function \mathcal{L}_A by adjusting θ making use of T iterations.

The presumably most widely utilized algorithm in machine learning is called (stochastic) gradient descent. Starting from the initial weight vector $\theta(t=0)$, we implement the following ordinary differential equation from $t=0$ to $t=T$ with small, positive learning rate η ,

$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \frac{d\mathcal{L}_A}{d\theta_\mu} \Big|_{\theta(t)}. \quad (3)$$

Variants of the gradient descent algorithms also include adding random noise $\xi_\mu(t)$ in each step, so-called stochastic gradient descent algorithms. One can show that in many cases, at the end of training, $\theta_\mu(t=T)$ can make the loss function $\mathcal{L}_A(\theta(t=T))$ sufficiently small.

The quantum algorithm with the promised efficiency in Theorem 1 and Theorem 2 described in the following.

- Our starting point of the algorithm is given by a initial weight vector, $\theta(0)$, the maximal number of iterations T , and the machine learning architecture \mathcal{L}_A , with model size n .

- In a first step, we use so-called *quantum Carleman linearization* introduced in Ref. [28], to linearize the model \mathcal{L}_A with the matrix M (see the supplemental material for more details).
- Then, we need to upload the sparse weight vector $\theta(0)$ as a state vector in quantum devices, using tools of Ref. [31] or alternatively more sophisticated and at the same time challenging architectures like *quantum random access memory* (QRAM) [30].
- Then, in a further step, we use a variant of the HHL solver that has been introduced in Ref. [29] and the supplementary material, to solve the state vector at the end $t = T$.
- Finally, we exploit tomographic methods described in, for instance, Refs. [22, 32] and the supplementary material, to obtain the classical model parameters $\theta(T)$.

II. NUMERICAL ANALYSIS

In this part of our work, we focus on providing numerical evidence of a potential quantum enhancement for large-scale machine-learning models. Commercial large-scale LLMs like GPT-3 can have $\mathcal{O}(100)$ billion parameters and even more, which is challenging as a starting point due to its tremendous computational costs. Instead, here we provide examples of classification and computer vision machine learning models, which are relatively small compared to language models used in industry. Our computational resources allow us to achieve the scale up to $\mathcal{O}(100)$ million, which is both practically minded and reachable. We expect that LLMs and other models will feature a similar behavior to those examples since our algorithm works in general as a replacement for stochastic gradient descent.

Thus, in order to provide evidence of the functioning of our quantum algorithm in the context of practically minded machine learning, we perform numerical experiments on a state-of-the-art machine vision architectures, namely the so-called *ResNet*, to tentatively outline schemes with a potential quantum enhancement. First, we study a model with 7 million trainable parameters trained to distinguish images of 100 classes [33]. We first pre-train the neural network, use the largest 10% of learned parameters for initialization, and use the quantum ODE system to obtain a sparse output model. We record the Hessian spectra during sparse training, allowing us to track the evolution of an error bound related quantity, given by

$$\frac{1}{N_c} \int_{-\infty}^{-0.4} \rho(a) |(1+a)^t| da + \frac{1}{N_c} \int_{0.4}^{\infty} \rho(a) |(1+a)^t| da, \quad (4)$$

where ρ is the eigenvalue density, a is the negative of Hessian eigenvalues, and N_c is the renormalization constant implicitly defined by

$$\frac{1}{N_c} \int_{-\infty}^{-0.4} \rho(a) da + \frac{1}{N_c} \int_{0.4}^{\infty} \rho(a) da = 1. \quad (5)$$

This error proxy discards small magnitude Hessian eigenvalues because they are close to 0, extremely abundant, and renders the error proxy stationary.

More dissipative systems have more positive Hessian eigenvalues, more negative a , and a better behaved error proxy. Specifically, the dissipative nature of the training dynamics initially leads to a reduction in this error proxy, which then gets overtaken by divergent modes and leads to an exponentially increasing error bound as shown in Fig. 2 (b). This motivates us to download the quantum trained model parameters sparsely and re-upload to the quantum computer to continue training every 100 steps. The effect of this procedure is that the exponentially increasing error restarts at 0 after re-uploading, with the side effect of Hessian broadening and accuracy reduction as shown in Fig. 2.

There is another strategy assuming the existence of QRAM. To combat the effect of Hessian broadening on the error proxy, we train the model classically for 10 steps after download before re-uploading of the new dense parameters, during which no training error is accrued. Although classical training has a cost linear in n , it is a small fraction of the entire training process. The accuracy dips immediately after download improves as training progresses, so our quantum training scheme is capable of producing useful sparse models. Finally, we ex-

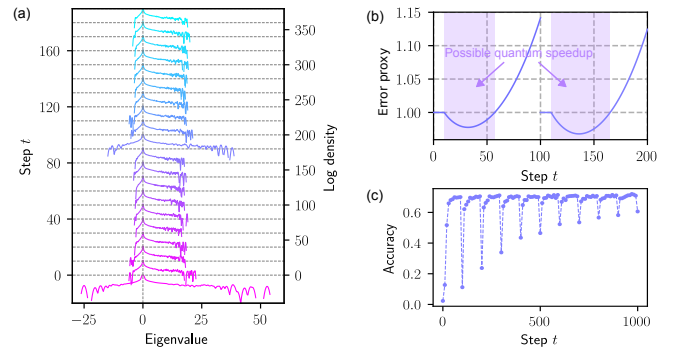


Figure 2. (a) - (c) Numerical results on ResNet as a function of step. Each step corresponds to a step of stochastic gradient descent based on the derivatives of the loss computed from 2048 randomly selected training samples. (a) ResNet Hessian spectra during training. (b) Estimated error proxy during training. (c) Training accuracy evolution for ResNet.

amine the Hessian of a 103 million parameter ResNet. We start with a pre-trained model and prune 90% of the parameters. Due to the immense computational cost of computing Hessian for a large machine learning model (a relatively large-scale model for computational vision based on our computational resources), we only benchmark the Hessian spectra to provide evidences of dissipation and potential quantum enhancements. Fig. 3 shows the initial Hessian, which clearly shows the dominance of dissipative modes over divergent modes similar to the 7 million parameter model. Since the Hessian improves with training for the 7 million parameter model, we believe this is evidence that the 103 million parameter model will have similarly manageable error growth.

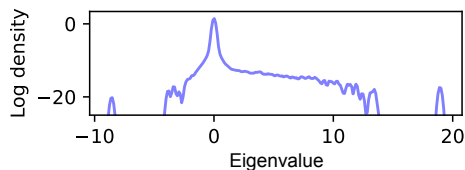


Figure 3. Hessian of the pruned 103 million parameter model immediately after pruning without any additional training.

III. CONCLUSION AND OUTLOOK

In our work, we have provided quantum algorithm strategies that are presumably helpful for solving the (stochastic) gradient descent dynamics for large-scale classical machine learning models, like LLMs such as GPT-3. We identify certain precisely stated dissipative and sparse regimes of the model where quantum devices could meaningfully contribute, providing an end-to-end HHL-type quantum algorithm application that could outperform known classical algorithms. The observation that an efficient classical algorithm for efficiently solving all instances of non-linear dissipative differential equations would imply an efficient classical algorithm for any problem that can be solved efficiently by a quantum computer can be seen as a strong argument that our algorithm is implausible to be de-quantized by classical proposals along the lines of Ref. [34]. Frankly, the core thesis of this work is that a main application of quantum computers may be in the *training of classical neural networks*.

Indeed, we claim that our algorithm might significantly increase the scalability and sustainability of classical large-scale machine-learning models and provide evidence for our claims numerically up to 103 million training parameters. Our work provides solid theoretical guarantees and intersections with state-of-the-art classical machine learning research. It sharply deviates from the mindset of variational quantum algorithms, and instead aims at augmenting classical machine learning by a key quantum step that constitutes a bottleneck for the classical training. In a way, it can be seen as adding flesh to the expectation that quantum formulations of neural networks may lead to new computational tools [35].

Our work is expected to open up several potential directions in the field of quantum machine learning where one can reasonably hope for algorithmic improvements. In the supplemental material, we hint at a number of potentially

particularly fruitful directions for future research. In short, they include the development of an alternative, time-dependent version during gradient descent trajectories, the identification of better formal criteria for dissipation, work on connections to diffusion models in classical machine learning and LLMs [36], theoretical improvements on the truncated HHL algorithms, and the identification of mechanisms of possible quantum speedups beyond notions of dissipation. We hope that this work can provide some stimulus for this type of research.

Acknowledgments

We thank Senrui Chen, Vedran Dunjko, Aram Harrow, Robert Huang, Daliang Li, John Preskill, Jonah Sherman, Umesh Vazirani, Carl Vondrick, Han Zheng, Sisi Zhou and Peter Zoller for many valuable discussions. This research has been supported by and has used the resources of the Argonne Leadership Computing Facility, which is a U.S. Department of Energy (DOE) Office of Science User Facility supported under Contract DE-AC02-06CH11357. J. L. is supported in part by International Business Machines (IBM) Quantum through the Chicago Quantum Exchange, and the Pritzker School of Molecular Engineering at the University of Chicago through AFOSR MURI (FA9550-21-1-0209). M. L. acknowledges support from DOE Q-NEXT. J.-P. L. acknowledges the support by the NSF (grant CCF-1813814, PHY-1818914), an NSF QISE-NET triplet award (DMR-1747426), an NSF QLCI program (OMA-2016245), a Simons Foundation award (No. 825053), and the Simons Quantum Postdoctoral Fellowship. Y. A. acknowledges support from the U.S. Department of Energy, Office of Science, under contract DE-AC02-06CH11357 at Argonne National Laboratory and Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0068. J. E. acknowledges funding of the BMBF (Hybrid), the BMWK (PlanQK, EniQma), the Munich Quantum Valley (K-8), the QuantERA (HQCC), the DFG (The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689), CRC 183), and the Einstein Research Foundation (Einstein Research Unit on Quantum Devices). L. J. acknowledges support from the the ARO(W911NF-23-1-0077), ARO MURI (W911NF-21-1-0325), AFOSR MURI (FA9550-19-1-0399, FA9550-21-1-0209), AFRL (FA8649-21-P-0781), DoE Q-NEXT, NSF (OMA-1936118, ERC-1941583, OMA-2137642), NTT Research, and the Packard Foundation (2020-71479).

-
- [1] LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
 - [2] Johnson, K. OpenAI debuts DALL-E for generating images from text. *VentureBeat* (2021).
 - [3] Brown, T. *et al.* Language models are few-shot learners. *Adv. Neur. Inf. Process. Sys.* **33**, 1877–1901 (2020). arXiv:2005.14165.
 - [4] Roose, K. The brilliance and weirdness of ChatGPT. *The New York Times* (2022).
 - [5] Lewkowycz, A. *et al.* Solving quantitative reasoning problems with language models (2022). arXiv:2206.14858.
 - [6] Patterson, D. *et al.* Carbon emissions and large neural network training (2021). arXiv:2104.10350.
 - [7] Biamonte, J. *et al.* Quantum machine learning. *Nature* **549**, 195–202 (2017). arXiv:1611.09347.
 - [8] Peruzzo, A. *et al.* A variational eigenvalue solver on a photonic quantum processor. *Nature Comm.* **5**, 1–7 (2014). arXiv:1304.3061.

- [9] McClean, J. R., Romero, J., Babbush, R. & Aspuru-Guzik, A. The theory of variational hybrid quantum-classical algorithms. *New J. Phys.* **18**, 023023 (2016). arXiv:1509.04279.
- [10] McArdle, S., Endo, S., Aspuru-Guzik, A., Benjamin, S. C. & Yuan, X. Quantum computational chemistry. *Rev. Mod. Phys.* **92**, 015003 (2020). arXiv:1808.10402.
- [11] Cerezo, M. *et al.* Variational quantum algorithms. *Nature Rev. Phys.* 1–20 (2021). arXiv:2012.09265.
- [12] Farhi, E., Goldstone, J. & Gutmann, S. A quantum approximate optimization algorithm (2014). arXiv:1411.4028.
- [13] Ebadi, S. *et al.* Quantum optimization of maximum independent set using Rydberg atom arrays. *Science* **376**, 1209–1215 (2022). arXiv:2202.09372.
- [14] Farhi, E. & Neven, H. Classification with quantum neural networks on near term processors (2018). arXiv:1802.06002.
- [15] Havlíček, V. *et al.* Supervised learning with quantum-enhanced feature spaces. *Nature* **567**, 209–212 (2019). arXiv:1804.11326.
- [16] Liu, J., Wilde, F., Mele, A. A., Jiang, L. & Eisert, J. Noise can be helpful for variational quantum algorithms (2022). arXiv:2210.06723.
- [17] Sweke, R., Seifert, J.-P., Hangleiter, D. & Eisert, J. On the quantum versus classical learnability of discrete distributions. *Quantum* **5**, 417 (2021). arXiv:2007.14451.
- [18] Liu, Y., Arunachalam, S. & Temme, K. A rigorous and robust quantum speed-up in supervised machine learning. *Nature Phys.* **17**, 1–5 (2021). arXiv:2010.02174.
- [19] Pirnay, N., Sweke, R., Eisert, J. & Seifert, J.-P. A super-polynomial quantum-classical separation for density modelling. *Phys. Rev. A* **107**, 042416 (2023). arXiv:2210.14936.
- [20] Lloyd, S., Mohseni, M. & Rebentrost, P. Quantum principal component analysis. *Nature Phys.* **10**, 631–633 (2014). arXiv:1307.0401.
- [21] Huang, H.-Y. *et al.* Power of data in quantum machine learning. *Nature Comm.* **12**, 2631 (2021). arXiv:2011.01938.
- [22] Huang, H.-Y., Kueng, R. & Preskill, J. Predicting many properties of a quantum system from very few measurements. *Nature Phys.* **16**, 1050–1057 (2020). arXiv:2002.08953.
- [23] Huang, H.-Y. *et al.* Quantum advantage in learning from experiments. *Science* **376**, 1182–1186 (2022). arXiv:2112.00778.
- [24] Hoeffler, T., Alistarh, D., Ben-Nun, T., Dryden, N. & Peste, A. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *J. Mach. Learn. Res.* **22**, 1–124 (2021). arXiv:2102.00554.
- [25] Lee, N., Ajanthan, T. & Torr, P. Snip: Single-shot network pruning based on connection sensitivity. In *International Conference on Learning Representations* (2019).
- [26] Wang, C., Zhang, G. & Grosse, R. Picking winning tickets before training by preserving gradient flow. In *International Conference on Learning Representations* (2020).
- [27] Tanaka, H., Kunin, D., Yamins, D. L. & Ganguli, S. Pruning neural networks without any data by iteratively conserving synaptic flow. *Adv. Neur. Inf. Process. Sys.* **33**, 6377–6389 (2020). arXiv:2006.05467.
- [28] Liu, J.-P. *et al.* Efficient quantum algorithm for dissipative nonlinear differential equations. *Proc. Natl. Ac. Sc.* **118**, e2026805118 (2021). arXiv:2011.03185.
- [29] Harrow, A. W., Hassidim, A. & Lloyd, S. Quantum algorithm for linear systems of equations. *Phys. Rev. Lett.* **103**, 150502 (2009). arXiv:0811.3171.
- [30] Giovannetti, V., Lloyd, S. & Maccone, L. Quantum random access memory. *Phys. Rev. Lett.* **100**, 160501 (2008). arXiv:0708.1879.
- [31] Gleinig, N. & Hoeffler, T. An efficient algorithm for sparse quantum state preparation. In *2021 58th ACM/IEEE Design Automation Conference (DAC)*, 433–438 (IEEE, 2021).
- [32] Guță, M., Kahn, J., Kueng, R. & Tropp, J. A. Fast state tomography with optimal error bounds. *J. Phys. A* **53**, 204001 (2020). arXiv:1809.11162.
- [33] Krizhevsky, A., Nair, V. & Hinton, G. CIFAR-100 (Canadian Institute for Advanced Research) (2009).
- [34] Tang, E. Quantum-inspired classical algorithms for principal component analysis and supervised clustering (2018). arXiv:1811.00414.
- [35] Bondesan, R. & Welling, M. The Hinton in your neural network: a quantum field theory view of deep learning (2021). arXiv:2103.04913.
- [36] Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neur. Inf. Proc. Sys.* **33**, 6840–6851 (2020). arXiv:2006.11239.

Supplemental material: Towards provably efficient quantum algorithms for large-scale machine learning models

Junyu Liu,^{1,2,3,4,5,6} Minzhao Liu,^{7,8} Jin-Peng Liu,^{9,10,11} Ziyu Ye,² Yuri Alexeev,^{8,2,3} Jens Eisert,¹² and Liang Jiang^{1,3}

¹*Pritzker School of Molecular Engineering, The University of Chicago, Chicago, IL 60637, USA*

²*Department of Computer Science, The University of Chicago, Chicago, IL 60637, USA*

³*Chicago Quantum Exchange, Chicago, IL 60637, USA*

⁴*Kadanoff Center for Theoretical Physics, The University of Chicago, Chicago, IL 60637, USA*

⁵*qBraid Co., Chicago, IL 60615, USA*

⁶*SeQure, Chicago, IL 60615, USA*

⁷*Department of Physics, The University of Chicago, Chicago, IL 60637, USA*

⁸*Computational Science Division, Argonne National Laboratory, Lemont, IL 60439, USA*

⁹*Simons Institute for the Theory of Computing, University of California, Berkeley, CA 94720, USA*

¹⁰*Department of Mathematics, University of California, Berkeley, CA 94720, USA*

¹¹*Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

¹²*Dahlem Center for Complex Quantum Systems, Free University Berlin, Berlin, 14195, Germany*

I. MOTIVATION

A. Quantum machine learning

Quantum computing is considered one of the most compelling alternatives of von Neumann architectures in the post-Moore era [1]. Making use of quantum features such as superposition and entanglement along the computation, quantum computing is expected to outperform its classical counterparts for many specific tasks, including factoring [2], database search [3], the simulation of complex quantum systems [4], sampling [5–7], and tasks of linear algebra [8]. Those algorithms might be useful in multiple areas, while applications in quantum machine learning might combine several of them at the same time. *Quantum machine learning* [9] utilizes quantum devices to run suitably modified machine learning algorithms. It has been argued that it may have the potential to lead to dramatically faster algorithms compared to what classical machine learning algorithms can provide in certain scenarios. However, substantial limitations exist for currently known quantum machine learning models.

- There are many studies about hybrid quantum-classical implementations of quantum machine learning, so-called *variational quantum algorithms* [10]. Relevant algorithms of such a type have shown potential applications in quantum computational chemistry [10–13], in combinatorial optimization [14, 15], and data-driven learning tasks [16, 17]. However, there is increasing evidence that such variational algorithms run on near-term quantum devices [11] are challenged by noise beyond logarithmic depth [18–21] (even though in some instances, noise can even be of some help [22]). Those algorithms might still be implementable on fault-tolerant quantum computing devices, but it is to date unclear whether and to what extent they are generically expected to have a quantum advantage compared to classical ones.
- Full-scale quantum machine learning algorithms have at the same time been established in the fault-tolerant setting, including steps of principle component analysis [23] (which has been largely de-quantized in the meantime [24]), quantum kernel methods [25, 26], or notions of quantum machine learning of quantum physical states and systems [27, 28]. Those algorithms are promising when fault-tolerant quantum devices will eventually be available, and some may even be equipped with theoretical guarantees. For example, it has been shown that quantum computers fare better in notions of distribution learning [29, 30] and applications of kernel methods to achieve robust quantum speed-ups in supervised machine learning [25]. However, to be fair, those algorithms are still quite far away from the realm of applicability of state-of-the-art classical machine learning applications. To start with, most data in common applications arising in everyday life are classical, not quantum. So if quantum computing requires quantum data to have a speedup, the applicability in our current world, especially from commercial perspectives, seems narrow. In other instances, the data have to be highly structured. It is to date not clear if those theoretical results are helpful for the general classical machine learning community at the moment. For instance, Ref. [25] presents an algorithm showing a robust quantum advantage in notions of machine learning, but it requires constructing a problem in the discrete logarithmic scenario, and there is no immediate state-of-the-art classical machine learning usage for such structured distributions as far as the authors know. Similar in spirit, theoretical contributions about other quantum machine learning works [31]—despite them contributing substantially to our conceptual and mathematical understanding—often provide information-theoretic bounds using statistical learning theory which might miss intricacies of practical real-world applications [32].

The algorithm designed in this work manages to overcome several of the above issues. Compared to existing literature, our work can be seen as contributing the following improvements.

- Instead of decoupling from applications of real-world classical machine learning, our algorithm is relatively practical and expected to be beneficial to the state-of-the-art classical machine learning community, touching upon one of the most important tasks of the current interest: large-scale machine learning models, including *large language models* (LLM) [33–35]. In order to validate our theory, our work also contains a practically minded simulations of around 100 million classical parameters, which is, as far as we know, a record for the quantum computing literature at the moment.
- The performance of our algorithm is, at least in several aspects, amenable to rigorous guarantees. Our algorithm is based on a provably efficient quantum algorithm, the HHL [8] algorithm, which promises to feature exponentially faster performance in sparse matrix inversions over known classical counterparts. Based on the HHL algorithm and Carleman linearization introduced for solving *dissipative ordinary differential equations* (ODEs) [36], we constructively show that there is an algorithm with poly-logarithmic performance in the number of model parameters and linear or quadratic dependence in the number of iterations when the training is sparse, dissipative, and close to a local minimum of the cost function. Moreover, in the more realistic scenarios, we also have solid theoretical bounds on the performance, which are supported by our numerical simulations.
- In the sparse training setup, our work could be applicable to classical data as well and is not relying on QRAM [37], while in the case of dense training, QRAM is required to ensure fast uploading, and tomographic methods could be a natural pruning approach for further classical machine learning steps. Under the assumptions of sparse training [38–41], we could create efficient interfaces between classical and quantum processors. Moreover, our algorithm only requires that interface to be established *once* for totally T iterations, which maximally reduces the burden of transmissions between classical and quantum worlds. Since our inputs and outputs are sparse in the quantum device, we do not necessarily need access to QRAM, which might be challenging to build in the large-scale and fault-tolerant regime [42, 43]. On the other hand, in the dense training task, QRAM could be used to ensure efficient uploading from the classical processors to quantum devices, and in the example of Section II D, the tomographic algorithm could read r components with largest norms with high probability, thus could serve a pruning method for dense or sparse training in the classical processor in the follow-up steps naturally.
- Our algorithm can be seen as a *quantum-classical hybrid algorithm*, but this is meant in a very different way than what is commonly considered in variational quantum algorithms. The quantum algorithm introduced here tackles a key step in otherwise classical machine learning algorithms.

One of the critical concepts in our work is sparse training, which makes the quantum implementation feasible and practical. We will summarize the sparse training perspectives in Section I B.

B. Making large-scale machine learning sustainable

Large-scale deep learning models have been shown to be an efficient method for enhancing performance across various tasks [33–35]. Those models are widely used to interact with humans [44], to solve mathematical problems [45], in the digital arts [46]. After all, they might after all be revolutionary for the entire society. Large-scale models, especially LLMs, are beyond reasonable doubt, some of most impressive and cutting-edge technologies in the world. However, training such models with considerable numbers of parameters is costly and can have high carbon emissions. For instance, twelve millions dollars and over five-hundred tons of CO₂ equivalent emissions are being produced to train GPT-3 [47]. It is hard to imagine that, in order to build such a machine, we have to use the equivalent of the annual electricity energy consumption of a small city. If one were to vote for the specific areas where quantum computing devices could contribute, LLMs might be the most promising playgrounds.

In our work, we make a substantial step towards such a topic, where quantum and classical computing might both help in the regime of *sparse training*, namely, training an otherwise classical network with most parameters being zero or very close to zero. Recently, sparse training has emerged as a promising direction to reduce the training cost and improve the inference efficiency [38]. One may say that the future of deep learning is sparse: By inducing sparsity in neural networks, we are able to train models at scale with unprecedented efficiency, paving the way for the next-generation artificial intelligence architectures. Sparse training is also a field where multiple research efforts have intersections, including sparse fine-tuning [48, 49], sparse-scratch [39–41], different pruning techniques and the techniques built on the *lottery ticket hypothesis* [50, 51]. One of the intuitive reasons for the advantages of sparse training is apparent: The task itself only fundamentally requires a few data and the current artificial neural networks are highly over-parametrized. In principle, the parameters required for given tasks should be fixed, hence in large-scale tasks, such parameters should be set to zero (see Ref. [52] for a comprehensive discussion).

Appealingly, sparsity is also one of the requirements of the HHL algorithm to obtain an exponential improvement over classical algorithms [52]. Moreover, the sparsity is also required for establishing interfaces between classical and quantum devices: The uploading step of dense quantum states necessitates realizations of QRAM, whereas downloading dense quantum states necessitates readings of tomography, which requiring significant resources for generic dense quantum states but efficient for sparse states. See more discussions around Section II D and Section V A.

Fig. 1 illustrates a typical process that elucidates how modern large-scale machine learning models work. The early stage involves pre-training whose data inputs and data processing are typically dense (which might be still sparse after the technology improves). In those processes, since they are more data-dependent, it will be harder to think about how quantum technologies could possibly help. The second process is the actual training where quantum technologies might contribute more substantially, where also another key factor of our quantum algorithm comes into play: dissipation.

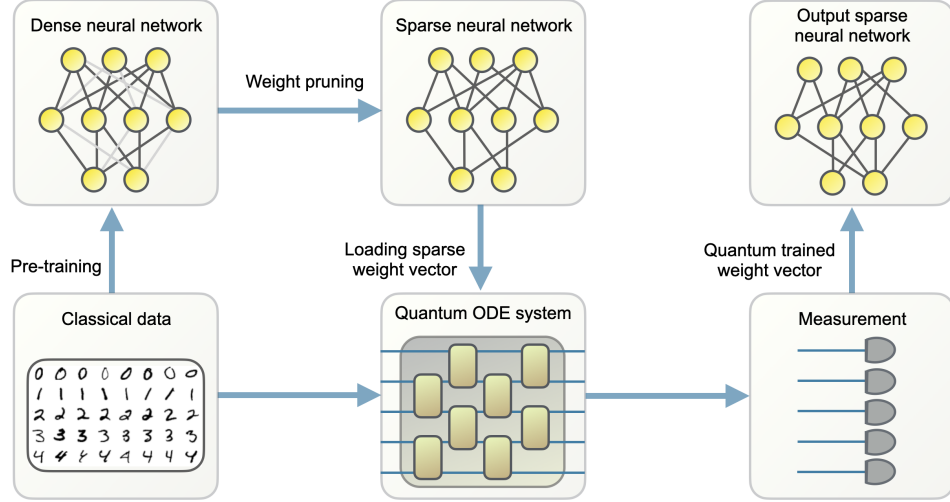


FIG. 1. A possible learning processes in large-scale models. A dense neural network is pre-trained classically. The neural network weights are then pruned and only a small fraction is preserved. A quantum ordinary difference equation system that corresponds to the sparse training dynamics is created using the sparse weight parameters and the training data. To allow for a quantum enhancement, the system must be sparse and dissipative. Measurement on the solution state is performed to obtain the final trained weight parameters, which is used to construct a trained classical sparse neural network.

When applying the ODE algorithm related to Ref. [36], dissipation is an essential condition. In fact, the algorithm presented in Ref. [36] employs linearization techniques to convert a system non-linear ODEs (in our case, the stochastic gradient descent equations in machine learning) into a sparse linear system problem to which the HHL algorithm applies. If the system is dissipative, the linearization error is dissipative as well and thus exponentially decaying, making it possible to linearize the system in a controllable fashion. In practical machine learning algorithms that are able to generalize efficiently (the generalization regime [52]), we find that the actual training process undergoes two phases: First, the neural network is actually learning non-trivial and significant examples. The system becomes dissipative, and equivalently, sensitive during the phase. Later, when the system accumulates enough data, the model becomes less dissipative.

In our quantum algorithm, we quantify the level of dissipation by the positivity of Hessian matrices during gradient descent (more precisely, *quantum Carleman matrices* (QCM), and the above phenomena might be intuitively explained from well-established information bottleneck theory [52–54]. When the system is actually learning knowledge, the system is more sensitive, and the trace of the Fisher information matrix tends to be higher (see Ref. [52]). When the learning is effectively ending, the Fisher information is approaching zero. Because the Hessian matrix is a second-order approximation of the Fisher matrix, the above observation also holds for Hessian matrices. Thus, we identify here important and specific circumstances where quantum algorithms could meaningfully contribute: This is when the system is both sparse and dissipative. Then we are likely to be able to employ quantum algorithms to accelerate large-scale machine learning models which are super-polynomially more efficient than known classical algorithms for this task. This regime is likely to occur in the core learning process of sparse training.

C. Theorems

In this subsection, we will primarily state our theorems of the main text and present the respective proofs in full.

Theorem I.1 (First main theorem). *For a sparse machine learning model with model size n , running T iterations, with the*

algorithm being fully dissipative (see Equation (II.1)), there is a quantum algorithm that runs in

$$O\left(sqT \times \text{poly}\left(m, \frac{2}{\log(1/\|\bar{u}\|)} \log \frac{3 \sum_{\ell=2}^q (\ell-1) \|F_\ell\|}{\|A\|}, \frac{1}{\epsilon}\right)\right) \quad (\text{I.1})$$

time with error $\epsilon > 0$, with s , $\|\bar{u}\|$, $\|\bar{F}_\ell\|$, q , $\|A\|$ are model-related parameters, and $m := \log_2(n)$ is the number of qubits. The algorithm also assumes that the learning rate is small, such that the condition in V.29 is satisfied.

Moreover, if we assume that the output weight vectors are r -sparse, the tomographic cost is $O(m^2 r^3 / \epsilon^2)$, ignoring log factors to bring the quantum states to the classical devices. The algorithm ignores the state preparation cost to prepare the initial weight vector in the quantum device, which is also computationally efficient for sparse training.

Theorem I.2 (Second main theorem). *For a machine learning model with model size n , running T iterations, with the algorithm being almost dissipative (see Equation (II.2)), there is a quantum algorithm that runs in*

$$O\left(s^2 q^2 T^2 \times \text{poly}\left(m, \frac{2}{\log(1/\|\bar{u}\|)} \log \frac{3 \sum_{\ell=2}^q (\ell-1) \|F_\ell\|}{\|A\|}, \frac{1}{\epsilon}\right)\right) \quad (\text{I.2})$$

time with error $\epsilon > 0$, with s , $\|\bar{u}\|$, $\|\bar{F}_\ell\|$, q , $\|A\|$ are model-related parameters, under the assumption of Theorem II.4, and $m = \log_2(n)$ is the number of qubits. The algorithm also assumes that the learning rate is small, such that the condition in V.29 is satisfied.

Moreover, if we assume that the output weight vectors are r -sparse, the tomographic cost is $O(m^2 r^3 / \epsilon^2)$ ignoring log factors to bring the quantum states to the classical devices. The algorithm ignores the state preparation cost to prepare the initial weight vector in the quantum device, which is also computationally efficient for sparse training.

For an algorithm for the efficient uploading of sparse quantum states, see Ref. [55], and for further discussion and details, see Section V A.

D. Algorithm description

In this subsection, we describe the actual quantum algorithm. A machine learning model is defined partially by a function $\mathcal{L}_{\mathcal{A}}$, called the *loss function*, as a function of weight vector (variational angle) $\theta \in \mathbb{R}^n$ with components θ_μ , and the *input training set* \mathcal{A} . In instances, we will also write θ when we emphasize that it is a vector. The weight vector has n components for an n -dimensional model. The task is to minimize the function $\mathcal{L}_{\mathcal{A}}$ by adjusting θ with T iterations. The most widely utilized algorithm in machine learning is called (*stochastic*) *gradient descent*. Starting from the initial weight vector $\theta(t=0)$, we implement the following ordinary differential equation from $t=0$ to $t=T$,

$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \left. \frac{d\mathcal{L}_{\mathcal{A}}}{d\theta_\mu} \right|_{\theta(t)}. \quad (\text{I.3})$$

Variants of the gradient descent algorithms also include adding random noise $t \mapsto \xi_\mu(t)$ in each step, so-called *stochastic gradient descent algorithms*. One can show that in many cases, at the end of training, $\theta_\mu(t=T)$ could make the loss function $\mathcal{L}_{\mathcal{A}}(\theta(t=T))$ sufficiently small. In our work, we propose the following quantum algorithm.

Algorithm 1: Efficient quantum (stochastic) gradient descent algorithm

Input: An initial weight vector $\theta(0)$, maximal number of iterations T , machine learning architecture $\mathcal{L}_{\mathcal{A}}$ with size n .

Output: A terminal weight vector $\theta(T)$.

1Linearize the model $\mathcal{L}_{\mathcal{A}}$ following Section II with a sparse matrix M .

2Upload the sparse weight vector $\theta(0)$ as a state vector in the quantum device, using the methods of Ref. [55] or more sophisticated architectures like QRAM.

3Perform the quantum linear system solver described in Section II C once to produce the quantum state vector at the end $t=T$.

4Perform state tomography described in Section II D to output the classical model parameter $\theta(T)$.

Result: $\theta(T)$.

As long as the conditions described in Section I E are satisfied, the overall complexity of the efficient quantum (stochastic) gradient descent algorithm scales poly-logarithmically in n and linearly or quadratic in T .

E. Assumptions on the machine learning setup

In this part, we briefly list the assumptions made in the setup.

- *Sparsity of the model.* First, the sparsity of the matrix F_ℓ used in the matrix A (see Section II for details) should be a constant when n grows. This is depending on the machine learning model architectures. Common models like ResNet networks [56, 57], or *multilayer perceptron* (MLP) models with bounded-polynomial activation functions would satisfy the requirement.
- *Sparsity of the weight vectors.* The sparsity of the weight vectors is satisfied naturally in sparse training, where the sparsity of the weight vectors stay as a constant and do not scale with the size of the model.
- *Dissipative assumptions.* The model should be sufficiently dissipative, such that either the eigenvalues of A are all non-negative, or the number of positive eigenvalues of A is much more than the number of negative eigenvalues (see Section II and Section III for detailed theoretical and experimental discussions).
- *Complexity of uploading and downloading quantum states.* The tomographic procedure is efficient due to the assumption of sparse training (see the algorithm described by Section II D).

II. EFFICIENT QUANTUM ALGORITHMS FOR DISSIPATIVE DIFFERENTIAL EQUATIONS

The idea of Ref. [36] can be directly generalized to differential equations. Roughly speaking, Ref. [36] primarily makes use of Carleman linearization, basically transforming non-linear ODEs into a set of linear ODEs. Then one can discretize ODEs as differential equations. The error in this procedure arises from the following two sources: the linearization error from Carleman linearization, and the discretization error for solving ODEs. If we are interested in the differential equations, the discretization error should not be considered. Other perspectives of the proof will be very similar to the pure ODE case. The results of Ref. [36] can be generalized from the quadratic case to higher-degree cases [58–60]. There are alternative linearization approaches that can be implemented on quantum computers [61–69].

A. Quantum Carleman linearization and error bounds

In this subsection, we turn to discussing the specific procedure of *quantum Carleman linearization* [36]. As an example, let us consider the following q -th degree polynomial differential equation,

$$\begin{aligned} \delta u &= u(t+1) - u(t) = \sum_{\ell=0}^q F_\ell u^{\otimes \ell}(t) \\ &= F_q u^{\otimes q}(t) + \dots + F_2 u^{\otimes 2}(t) + F_1 u(t) + F_0, \\ u(0) &= u_{\text{in}}. \end{aligned} \tag{II.1}$$

We have,

$$\begin{aligned} F_0 &\in \mathbb{R}^{n \times 1}, \\ u &= (u_1, \dots, u_n) \in \mathbb{R}^n, F_1 \in \mathbb{R}^{n \times n}, \\ u^{\otimes 2} &= (u_1^2, u_1 u_2, \dots, u_1 u_n, u_2 u_1, \dots, u_n u_{n-1}, u_n^2) \in \mathbb{R}^{n^2}, F_2 \in \mathbb{R}^{n \times n^2}, \\ u^{\otimes 3} &= (u_1^3, u_1^2 u_2, \dots) \in \mathbb{R}^{n^3}, F_3 \in \mathbb{R}^{n \times n^3}, \\ &\dots \\ u^{\otimes q} &= (u_1^q, u_1^{q-1} u_2, \dots) \in \mathbb{R}^{n^q}, F_q \in \mathbb{R}^{n \times n^q}. \end{aligned} \tag{II.2}$$

Here, we assume that $F_{q,\dots,2,1}$ does not change over time, while $F_0 = F_0(t)$ might depend on time. We further assume that all F_q are at most s -sparse. Moreover, one can define

$$\begin{aligned} A_{j+q-1}^j &:= F_q \otimes I^{\otimes j-1} + I \otimes F_q \otimes I^{\otimes j-2} + \dots + I^{\otimes j-1} \otimes F_q, \\ &\dots \\ A_{j+1}^j &:= F_2 \otimes I^{\otimes j-1} + I \otimes F_2 \otimes I^{\otimes j-2} + \dots + I^{\otimes j-1} \otimes F_2, \\ A_j^j &:= F_1 \otimes I^{\otimes j-1} + I \otimes F_1 \otimes I^{\otimes j-2} + \dots + I^{\otimes j-1} \otimes F_1, \\ A_{j-1}^j &:= F_0 \otimes I^{\otimes j-1} + I \otimes F_0 \otimes I^{\otimes j-2} + \dots + I^{\otimes j-1} \otimes F_0. \end{aligned} \quad (\text{II.3})$$

We can now formulate the so-called quantum Carleman linearization (there is another contribution due to the discrete nature of the dynamics. We will discuss it in Section V B. The contribution is expected to be ignored if the learning rate is small.)

$$\delta \begin{pmatrix} u \\ u^{\otimes 2} \\ u^{\otimes 3} \\ \vdots \\ u^{\otimes(N-1)} \\ u^{\otimes N} \\ \vdots \end{pmatrix} = \begin{pmatrix} A_1^1 & A_2^1 & A_3^1 & \dots & A_q^1 & & \\ A_1^2 & A_2^2 & A_3^2 & \dots & A_q^2 & A_{q+1}^2 & \\ & A_2^3 & A_3^3 & A_4^3 & \dots & & \\ & & \ddots & \ddots & \ddots & & \\ & & & A_{N-2}^{N-1} & A_{N-1}^{N-1} & A_N^{N-1} & \dots \\ & & & & A_{N-1}^N & A_N^N & \ddots \\ & & & & & \ddots & \ddots \end{pmatrix} \begin{pmatrix} u \\ u^{\otimes 2} \\ u^{\otimes 3} \\ \vdots \\ u^{\otimes(N-1)} \\ u^{\otimes N} \\ \vdots \end{pmatrix} + \begin{pmatrix} F_0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \end{pmatrix}. \quad (\text{II.4})$$

Based on these expressions, we can construct the linear system truncated with the parameter N

$$\delta \hat{y} = A \hat{y} + b, \quad \hat{y}(0) = \hat{y}_{\text{in}}, \quad (\text{II.5})$$

which is

$$\delta \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \vdots \\ \hat{y}_{N-1} \\ \hat{y}_N \end{pmatrix} = \begin{pmatrix} A_1^1 & A_2^1 & \dots & A_q^1 & 0 & 0 \\ A_1^2 & A_2^2 & \dots & A_q^2 & A_{q+1}^2 & 0 \\ 0 & A_2^3 & A_3^3 & A_4^3 & 0 & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & A_{N-2}^{N-1} & A_{N-1}^{N-1} & A_N^{N-1} \\ 0 & 0 & 0 & 0 & A_{N-1}^N & A_N^N \end{pmatrix} \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \vdots \\ \hat{y}_{N-1} \\ \hat{y}_N \end{pmatrix} + \begin{pmatrix} F_0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}. \quad (\text{II.6})$$

Here, we specify

$$\begin{aligned} \hat{y}_j &:= u^{\otimes j} \in \mathbb{R}^{n^j}, \\ \hat{y}_{\text{in}} &:= (u_{\text{in}}, u_{\text{in}}^{\otimes 2}, \dots, u_{\text{in}}^{\otimes N}). \end{aligned} \quad (\text{II.7})$$

Now we define the error of the linearization process as,

$$\eta_j(t) := u^{\otimes j}(t) - \hat{y}_j(t). \quad (\text{II.8})$$

One can show that it will satisfy

$$\delta \eta = A \eta + \hat{b}, \quad \eta(0) = 0, \quad (\text{II.9})$$

which is

$$\delta \begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \vdots \\ \eta_{N-1} \\ \eta_N \end{pmatrix} = \begin{pmatrix} A_1^1 & A_2^1 & \dots & A_q^1 & 0 & 0 \\ A_1^2 & A_2^2 & \dots & A_q^2 & A_{q+1}^2 & 0 \\ 0 & A_2^3 & A_3^3 & A_4^3 & 0 & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & A_{N-2}^{N-1} & A_{N-1}^{N-1} & A_N^{N-1} \\ 0 & 0 & 0 & 0 & A_{N-1}^N & A_N^N \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \vdots \\ \eta_{N-1} \\ \eta_N \end{pmatrix} + \begin{pmatrix} 0 \\ \vdots \\ A_{N+1}^{N-q+2} u^{\otimes(N+1)} \\ \vdots \\ \sum_{\ell=1}^{q-2} A_{N+\ell}^{N-1} u^{\otimes(N+\ell)} \\ \sum_{\ell=1}^{q-1} A_{N+\ell}^N u^{\otimes(N+\ell)} \end{pmatrix}. \quad (\text{II.10})$$

The term $t \mapsto F_0(t)$ could be understood as the noise term in the stochastic gradient descent. The solution of the iterative relation is given by

$$\eta(t+1) = \left(\prod_{i=t}^0 (1 + A(i)) \right) \eta(0) + \sum_{i=0}^t \left(\left(\prod_{j=l}^{i+1} (1 + A(j)) \right) \hat{b}(i) \right). \quad (\text{II.11})$$

Thus

$$\|\eta(t+1)\| \leq \left\| \left(\prod_{i=t}^0 (1 + A(i)) \right) \eta(0) \right\| + \left\| \sum_{i=0}^t \left(\left(\prod_{j=t}^{i+1} 1 + A(j) \right) \hat{b}(i) \right) \right\|. \quad (\text{II.12})$$

We call the matrix $A(t)$ here the *quantum Carleman matrix* (QCM). The dimension of the QCM is given by

$$\Delta \equiv n + n^2 + \dots + n^N = \frac{n^{N+1} - n}{n - 1} = O(n^N). \quad (\text{II.13})$$

Assuming that $A(t)$ has Δ eigenvalues denoted by $a_\mu(t)$, $\mu \in [\Delta]$.

First, we assume that $A(t)$ does not change over time, $A(t) = A$, which corresponds to a time-independent F_0 . More generally, we use the index notation $\mu, \nu, \dots \in [\Delta]$. So we have

$$\eta_\mu(t+1) = \sum_{\mu_1, \mu_2} (1 + a_{\mu_1})^{t+1} S_{\mu, \mu_1}^\dagger S_{\mu_1, \mu_2} \eta_{\mu_2}(0) + \sum_{\mu_1, \mu_2} \left((1 + a_{\mu_1})^{t+1} - 1 \right) a_{\mu_1}^{-1} S_{\mu, \mu_1}^\dagger S_{\mu_1, \mu_2} \hat{b}_{\mu_2}. \quad (\text{II.14})$$

Here, the matrix S is diagonalizing the QCM as

$$\sum_{\mu_1} S_{\mu, \mu_1}^\dagger a_{\mu_1} S_{\mu_1, \nu} = A_{\mu, \nu}. \quad (\text{II.15})$$

Thus, we have

$$\begin{aligned} \left(\left(\prod_{i=t}^0 (1 + A) \right) \eta(0) \right)_\mu &= \left((1 + A)^{t+1} \eta(0) \right)_\mu = \sum_{\mu_1, \mu_2} (1 + a_{\mu_1})^{t+1} S_{\mu, \mu_1}^\dagger S_{\mu_1, \mu_2} \eta_{\mu_2}(0). \\ \left(\sum_{i=0}^t \left(\left(\prod_{j=t}^{i+1} 1 + A \right) \hat{b}(i) \right) \right)_\mu &= \left(\sum_{i=0}^t \left((1 + A)^{t-i} \hat{b}(i) \right) \right)_\mu \\ &= \sum_{i=0, \mu_1, \mu_2}^t \left(S_{\mu, \mu_1}^\dagger (1 + a_{\mu_1})^{t-i} S_{\mu_1, \mu_2} \hat{b}_{\mu_2}(i) \right). \end{aligned} \quad (\text{II.16})$$

Making use of the sub-multiplicativity of the operator norm and by bounding the right hand side by a largest element, one can bound the second term by

$$\left\| \sum_{i=0}^t \left(\left(\prod_{j=t}^{i+1} 1 + A \right) \hat{b}(i) \right) \right\| \leq \sum_{i=0}^t \|(1 + A)\|^{t-i} \|\hat{b}(i)\|. \quad (\text{II.17})$$

Moreover, this term can be put into a more refined form, since

$$\sum_{i=0, \mu_1, \mu_2}^t \left(S_{\mu, \mu_1}^\dagger (1 + a_{\mu_1})^{t-i} S_{\mu_1, \mu_2} \hat{b}_{\mu_2}(i) \right) = \sum_{\mu_1, \mu_2} \left(S_{\mu, \mu_1}^\dagger S_{\mu_1, \mu_2} \sum_{i=0}^t (1 + a_{\mu_1})^{t-i} \hat{b}_{\mu_2}(i) \right). \quad (\text{II.18})$$

Assuming that all b_μ are upper bounded as

$$|\hat{b}_{\mu_2}(i)| \leq \bar{b}_{\mu_2} \quad (\text{II.19})$$

for all i , we find

$$\begin{aligned} \left| \sum_{i=0, \mu_1, \mu_2}^t \left(S_{\mu, \mu_1}^\dagger (1 + a_{\mu_1})^{t-i} S_{\mu_1, \mu_2} \hat{b}_{\mu_2}(i) \right) \right| &\leq \sum_{\mu_1, \mu_2} \left(|S_{\mu, \mu_1}^\dagger S_{\mu_1, \mu_2} \bar{b}_{\mu_2}| \left| \sum_{i=0}^t (1 + a_{\mu_1})^{t-i} \right| \right) \\ &= \sum_{\mu_1, \mu_2} \left(|S_{\mu, \mu_1}^\dagger S_{\mu_1, \mu_2} \bar{b}_{\mu_2}| \left| \frac{(a_{\mu_1} + 1)^{t+1} - 1}{a_{\mu_1}} \right| \right). \end{aligned} \quad (\text{II.20})$$

At the same time, we can also write

$$|\eta_\mu(t+1)| \leq \sum_\nu |(1+a_\nu)^{t+1}| Q_{\mu,\nu} + C_\mu, \quad (\text{II.21})$$

where

$$\begin{aligned} Q_{\mu,\nu} &= \left| S_{\mu,\nu}^\dagger \sum_{\mu_1} S_{\nu,\mu_1} \eta_{\mu_1}(0) \right| + |a_\nu^{-1}| \left| S_{\mu,\nu}^\dagger \sum_{\mu_1} S_{\nu,\mu_1} \bar{b}_{\mu_1} \right| = |a_\nu^{-1}| \left| S_{\mu,\nu}^\dagger \sum_{\mu_1} S_{\nu,\mu_1} \bar{b}_{\mu_1} \right|, \\ C_\mu &= \left| \sum_{\mu_2,\nu} a_\nu^{-1} S_{\mu,\nu}^\dagger S_{\nu,\mu_2} \bar{b}_{\mu_2} \right|. \end{aligned} \quad (\text{II.22})$$

Since the matrix Q and the vector C are time independent, the specific time dependence is completely depending on the eigenvalues of A . In the most extreme case, all eigenvalues of A satisfy

$$|1 + a_\mu| < 1 \quad (\text{II.23})$$

then we get a bounded η , in that

$$\|\eta(t)\| \leq \|Q\| + \|C\| \quad (\text{II.24})$$

and we also have

$$\begin{aligned} \|C\| &\leq \|A\|^{-1} \|\bar{b}\|, \\ \|Q\| &\leq \|A\|^{-1} \|\bar{b}\|. \end{aligned} \quad (\text{II.25})$$

Now we have the following theorem.

Theorem II.1 (Fully dissipative). *If $\|1 + A\| < 1$ and $\|\hat{b}(t)\|$ is bounded by $\|\bar{b}\| < \infty$, then we have*

$$\|\eta(T)\| \leq \|Q\| + \|C\| \quad (\text{II.26})$$

and

$$\begin{aligned} \|Q\| &\leq \|A\|^{-1} \|\bar{b}\|, \\ \|C\| &\leq \|A\|^{-1} \|\bar{b}\|, \end{aligned} \quad (\text{II.27})$$

where T is the total number of iterations.

Moreover, we can formulate a probabilistic version. We assume that there are Δ_+ eigenvalues a_μ satisfying $|1 + a_\mu| < 1$. Without loss of generality, we assume that

$$\begin{aligned} |1 + a_\nu| &< 1 : 1 \leq \nu \leq \Delta_+, \\ |1 + a_\nu| &\geq 1 : \nu > \Delta_+. \end{aligned} \quad (\text{II.28})$$

Moreover, we define $\Delta_- := \Delta - \Delta_+$, and we write

$$\begin{aligned} [\Delta_+] &:= \{1 \leq \nu \leq \Delta_+\}, \\ [\Delta_-] &:= \{\nu > \Delta_+\}. \end{aligned} \quad (\text{II.29})$$

Thus, we have

$$\sum_\nu |(1+a_\nu)^{t+1}| Q_{\mu,\nu} + C_\mu = \sum_{\nu_+ \in [\Delta_+]} |(1+a_{\nu_+})^{t+1}| Q_{\mu,\nu_+} + \sum_{\nu_- \in [\Delta_-]} |(1+a_{\nu_-})^{t+1}| Q_{\mu,\nu_-} + C_\mu. \quad (\text{II.30})$$

Then, the exponential convergence will still be approximately valid, if

$$\sum_{\nu_+ \in [\Delta_+]} |(1+a_{\nu_+})^{t+1}| Q_{\mu,\nu_+} \gg \sum_{\nu_- \in [\Delta_-]} |(1+a_{\nu_-})^{t+1}| Q_{\mu,\nu_-}. \quad (\text{II.31})$$

This condition means that, weighted by $Q_{\mu,\nu}$, if the contribution of diverging modes in QCM (eigenvalues satisfying $|1 + a_\nu| \geq 1$) is significantly smaller than the contribution of the dissipative modes (eigenvalues satisfying $|1 + a_\nu| < 1$), then we still get the convergent error η .

Let us give a simple example at this point. Assuming that we only have two typical a_μ , a_+ as a dissipative mode, and a_- as a divergent mode. Furthermore, assuming that $Q_{\mu,\nu}$ are equal for all components. We have

$$\Delta_+ |(1 + a_+)^{t+1}| \gg \Delta_- |(1 + a_-)^{t+1}| \quad (\text{II.32})$$

which will imply

$$\left| \frac{1 + a_+}{1 + a_-} \right|^{t+1} \gg \frac{\Delta_-}{\Delta_+}. \quad (\text{II.33})$$

Replacing $t + 1$ with the total number of iterations T , we get a constraint for the time T , that

$$T \ll \frac{\log \frac{\Delta_-}{\Delta_+}}{\log \left| \frac{1+a_+}{1+a_-} \right|} \quad (\text{II.34})$$

has to hold. This translates to

$$\Delta_+ \gg \frac{\Delta}{\left| \frac{1+a_+}{1+a_-} \right|^T + 1}. \quad (\text{II.35})$$

This calculation gives a precise condition: we have choose the number of dissipative modes to be much larger than the number of diverging modes in the QCM. One can easily generalize the above argument to different modes. for the μ component, one could derive the condition for the total number of iteration T_μ , as

$$T_\mu \ll \frac{\left(\sum_{\nu_- \in [\Delta_-]} \log |Q_{\mu,\nu_-}| - \sum_{\nu_+ \in [\Delta_+]} \log |Q_{\mu,\nu_+}| \right)}{\left(\sum_{\nu_+ \in [\Delta_+]} \log |(1 + a_{\nu_+})| - \sum_{\nu_- \in [\Delta_-]} \log |(1 + a_{\nu_-})| \right)}. \quad (\text{II.36})$$

Now we care mostly about η_1 , which is the scalar difference between the original non-linear equation solution and the linearized version, so we can take $\mu = 1$.

Finally, in terms of mathematical statement, we could replace \gg as $>$. Thus, since bounds could be directly established with $>$. Thus, we arrive at the following claim.

Theorem II.2 (Almost dissipative). *If $\|\hat{b}(t)\|$ is bounded by $\|\bar{b}\| < \infty$, and,*

$$\sum_{\nu_+ \in [\Delta_+]} |(1 + a_{\nu_+})^T| Q_{\mu,\nu_+} > \sum_{\nu_- \in [\Delta_-]} |(1 + a_{\nu_-})^T| Q_{\mu,\nu_-}. \quad (\text{II.37})$$

Then, we have

$$\|\eta(T)\| \leq 2 \|Q\| + \|C\| \quad (\text{II.38})$$

and

$$\begin{aligned} \|Q\| &\leq \|A\|^{-1} \|\bar{b}\|, \\ \|C\| &\leq \|A\|^{-1} \|\bar{b}\|. \end{aligned} \quad (\text{II.39})$$

We now turn to discussing the time-dependent cases, which correspond to settings of stochastic gradient descent. Similarly as before, we have the diagonalization

$$\sum_{\mu_1} S_{\mu,\mu_1}^\dagger(i) a_{\mu_1}(i) S_{\mu_1,\nu}(i) = A_{\mu,\nu}(i) \quad (\text{II.40})$$

and we find

$$\left(\left(\prod_{i=t}^0 (1 + A(i)) \right) \eta(0) \right)_{\mu} = \sum_{\nu} \sum_{\alpha_t, \alpha_{t-1}, \dots, \alpha_1, \alpha_0} f_{\alpha_t, \alpha_{t-1}, \dots, \alpha_1, \alpha_0}^{\mu, \nu}(t; 0) (1 + a_{\alpha_t}(t)) (1 + a_{\alpha_{t-1}}(t-1)) \dots (1 + a_{\alpha_0}(0)) \eta_{\nu}(0) \quad (\text{II.41})$$

where

$$f_{\alpha_t, \alpha_{t-1}, \dots, \alpha_1, \alpha_0}^{\mu, \nu}(t; 0) := \sum_{\mu_t, \mu_{t-1}, \dots, \mu_1} S_{\mu \alpha_t}^{\dagger}(t) S_{\alpha_t, \mu_t}(t) S_{\mu_t, \alpha_{t-1}}^{\dagger}(t-1) S_{\alpha_{t-1}, \mu_{t-1}}(t-1) \dots S_{\mu_2, \alpha_1}^{\dagger}(1) S_{\alpha_1, \mu_1}(1) S_{\mu_1, \alpha_0}^{\dagger}(0) S_{\alpha_0, \nu}(0). \quad (\text{II.42})$$

Furthermore, we find

$$\left(\sum_{i=0}^t \left(\prod_{j=i}^{i+1} 1 + A(j) \right) \hat{b}(i) \right)_{\mu} = \sum_{i=0}^t \sum_{\nu} \sum_{\alpha_t, \alpha_{t-1}, \dots, \alpha_{i+2}, \alpha_{i+1}} f_{\alpha_t, \alpha_{t-1}, \dots, \alpha_{i+2}, \alpha_{i+1}}^{\mu, \nu}(t; i+1) \times (1 + a_{\alpha_t}(t)) (1 + a_{\alpha_{t-1}}(t-1)) \dots (1 + a_{\alpha_{i+1}}(i+1)) \hat{b}_{\nu}(i). \quad (\text{II.43})$$

Thus, the structure is similar as before, if we have sufficiently small numbers of divergent modes, and much more dissipative modes, we will have similar bounds to guarantee that η is decaying and gives us the exponentially converging errors.

One might still worry about how realistic this approach is for practically useful machine learning problems. An assumption we could consider is to assume a distribution of the QCM eigenspectra. For simplicity, we assume that a_{ν} is independently following the normal distribution $\mathcal{N}(\bar{a}, \sigma)$ (with the measure $D a_{\nu}$). So we get

$$\begin{aligned} & \left(\sum_{\nu} \mathcal{Q}_{\mu, \nu} \int D a_{\nu} (1 + a_{\nu})^{t+1} \right) + C_{\mu} \\ &= (-i)^{t+1} 2^{\frac{t+1}{2}} \sigma^{t+1} U \left(\frac{1}{2}(-t-1), \frac{1}{2}, -\frac{(1+\bar{a})^2}{2\sigma^2} \right) \left(\sum_{\nu} \mathcal{Q}_{\mu, \nu} \right) + C_{\mu}. \end{aligned} \quad (\text{II.44})$$

Here, U is the U -confluent hypergeometric function. If $\bar{a} > 0$ (the divergent case), the formula will quickly be dominated by the exponential divergence for large t ,

$$\begin{aligned} & \left(\sum_{\nu} \mathcal{Q}_{\mu, \nu} \int D a_{\nu} (1 + a_{\nu})^{t+1} \right) + C_{\mu} \\ &\approx \left((1 + \bar{a})^{t+1} + \frac{t(t+1)}{2} (1 + \bar{a})^{t-1} \sigma^2 + \dots \right) \left(\sum_{\nu} \mathcal{Q}_{\mu, \nu} \right) + C_{\mu}. \end{aligned} \quad (\text{II.45})$$

If $\bar{a} < 0$ (the dissipative case), the value of σ matters to see if the error is bounded or not. If σ is large, we might still get divergence. If σ is small enough, the divergence will not show up since it is dominated by the exponential decay $(1 + \bar{a})^{t+1}$. For a given value of \bar{a} and σ , one can plot the time dependence from the confluent hypergeometric function to estimate the error. For instance, one can evaluate

$$\Omega(t, \bar{a}, \sigma) := \left| \frac{g(t, \bar{a}, \sigma)}{g(0, \bar{a}, \sigma)} \right| \quad (\text{II.46})$$

to check the convergence/divergence properties, where

$$g(t, \bar{a}, \sigma) := (-i)^{t+1} 2^{\frac{t+1}{2}} \sigma^{t+1} U \left(\frac{1}{2}(-t-1), \frac{1}{2}, -\frac{(1+\bar{a})^2}{2\sigma^2} \right). \quad (\text{II.47})$$

A particularly interesting example is where $\bar{a} = 1$. In the case where $\sigma < 1$, one can estimate,

$$g(t, \bar{a} = 1, \sigma) \approx 1 + \frac{t(t+1)}{2} \sigma^2 + \dots \quad (\text{II.48})$$

The expression of g is now dominated by the quadratic term. Thus, the error is quadratic in time $t + 1$, which is much more mild than exponential divergences.

The mathematics is easier when we consider the independent uniform distribution with the range $[\bar{a} - \sigma_a, \bar{a} + \sigma_a]$ (note that now σ_a is not exactly the standard deviation). We have

$$\begin{aligned} & \left(\sum_{\nu} Q_{\mu, \nu} \int D a_{\nu} (1 + a_{\nu})^{t+1} \right) + C_{\mu} \\ & \approx \left(\frac{(1 + \bar{a} + \sigma_a)^{t+2} - (1 + \bar{a} - \sigma_a)^{t+2}}{2\sigma_a(t+2)} \right) \left(\sum_{\nu} Q_{\mu, \nu} \right) + C_{\mu}. \end{aligned} \quad (\text{II.49})$$

Thus, one can get exponential convergence if $-1 < \bar{a} + \sigma_a < 0$. One can use those criteria to numerically check the property of the bounds. Finally, we estimate the bound on $\|\bar{b}\|$. We can express this as

$$\|\hat{b}(t)\| \leq \sum_{\ell'=1}^{q-1} \sum_{\ell=1}^{q-\ell'} \left\| A_{N+\ell}^{N-\ell'+1} u^{\otimes(N+\ell)} \right\|. \quad (\text{II.50})$$

Using

$$\left\| A_{N+\ell}^{N-\ell'+1} \right\| \leq (N - \ell' + 1) \|F_{\ell+\ell'}\|, \quad (\text{II.51})$$

we get

$$\begin{aligned} \|\hat{b}(t)\| & \leq \sum_{\ell'=1}^{q-1} \sum_{\ell=1}^{q-\ell'} (N - \ell' + 1) \|F_{\ell+\ell'}\| \|u\|^{N+\ell} \leq \|u\|^{N+1} \sum_{\ell'=1}^{q-1} \sum_{\ell=1}^{q-\ell'} (N - \ell' + 1) \|F_{\ell+\ell'}\| \\ & = \|u\|^{N+1} \sum_{\ell=2}^q \|F_{\ell}\| \left((\ell-1)N - \frac{(\ell-1)(\ell-2)}{2} \right) \\ & \leq N \|u\|^{N+1} \sum_{\ell=2}^q (\ell-1) \|F_{\ell}\|, \end{aligned} \quad (\text{II.52})$$

with $N > q$. Moreover, we can assume that

$$\|u(t)\| \leq \|\bar{u}\| < \infty. \quad (\text{II.53})$$

Namely, the condition is that $t \mapsto u(t)$ is bounded in time. By scaling variables, one could without loss of generality, assume that

$$0 < \|\bar{u}\| < 1. \quad (\text{II.54})$$

So we set

$$\|\bar{b}\| := N \|\bar{u}\|^{N+1} \sum_{\ell=2}^q (\ell-1) \|F_{\ell}\| \quad (\text{II.55})$$

and we have

$$\begin{aligned} \|A\|^{-1} \|\bar{b}\| & = N \|A\|^{-1} \|\bar{u}\|^{N+1} \sum_{\ell=2}^q (\ell-1) \|F_{\ell}\|, \\ \|Q\| & \leq N \|A\|^{-1} \|\bar{u}\|^{N+1} \sum_{\ell=2}^q (\ell-1) \|F_{\ell}\|, \\ \|C\| & \leq N \|A\|^{-1} \|\bar{u}\|^{N+1} \sum_{\ell=2}^q (\ell-1) \|F_{\ell}\|. \end{aligned} \quad (\text{II.56})$$

So in all cases, we find

$$\|\eta(T)\| \leq 3N\|A\|^{-1}\|\bar{u}\|^{N+1} \sum_{\ell=2}^q (\ell-1) \|F_\ell\|. \quad (\text{II.57})$$

Therefore, if we assume that the linearization error is mostly ϵ_{le} , we get

$$3N\|A\|^{-1}\|\bar{u}\|^{N+1} \sum_{\ell=2}^q (\ell-1) \|F_\ell\| = \epsilon_{\text{le}}. \quad (\text{II.58})$$

For this reason, there exists an N such that

$$N\|\bar{u}\|^{N+1} \leq \|\bar{u}\|^{\frac{1}{2}N}. \quad (\text{II.59})$$

So we obtain

$$N \geq \frac{2}{\log(1/\|\bar{u}\|)} \log \frac{3 \sum_{\ell=2}^q (\ell-1) \|F_\ell\|}{\|A\| \epsilon_{\text{le}}}, \quad (\text{II.60})$$

for a sufficiently small $\epsilon_{\text{le}} > 0$.

B. The QCM spectra

1. Dominance from Hessian matrices

In this part, we will derive how the QCM spectra could be related to the Hessian in the machine learning problems. For simplicity, let us firstly consider

$$F_0 = F_{\ell \geq 2} = 0. \quad (\text{II.61})$$

Now, assuming that for

$$\text{spec}(F_1) = (f_a)_{a=1}^n, \quad (\text{II.62})$$

we have

$$\begin{aligned} \text{spec}(A_j^j) &= (f_{a_1} + f_{a_2} + \dots + f_{a_j})_{a_1, a_2, \dots, a_j=1}^n, \\ \dim(A_j^j) &= n^j. \end{aligned} \quad (\text{II.63})$$

Thus we can identify the spectra of the QCM to be

$$\begin{aligned} \text{spec}(A) &= \bigcup_{j=1}^N (f_{a_1} + f_{a_2} + \dots + f_{a_j})_{a_1, a_2, \dots, a_j=1}^n, \\ \dim(A) &= \frac{n^{N+1} - n}{n - 1} = \Delta. \end{aligned} \quad (\text{II.64})$$

We can now bring these findings into the context of actual machine learning models. For a machine learning model with trainable parameters θ_a and the loss function \mathcal{L}_A , we have (see Section III for further information, and \supset indicates that we are calculating the first order term in the Taylor expansion)

$$\begin{aligned} \delta\theta_a &= -\eta \frac{\partial \mathcal{L}_A}{\partial \theta_a} \supset -\eta \sum_b \frac{\partial^2 \mathcal{L}_A}{\partial \theta_a \partial \theta_b} \theta_b \\ -\eta \frac{\partial^2 \mathcal{L}_A}{\partial \theta_a \partial \theta_b} &= (F_1)_{a,b} \\ -\eta \times \text{spec} \left(\frac{\partial^2 \mathcal{L}_A}{\partial \theta_a \partial \theta_b} \right) &= -\eta \times \text{spec}(H) = \text{spec}(F_1). \end{aligned} \quad (\text{II.65})$$

Thus, more positive Hessian eigenvalues reflect more dissipative modes in QCM.

2. Perturbation theory

We here put the findings into the context of perturbation theory. We take the case where

$$\|F_{0,\ell \geq 2}\| \ll \|F_1\| \quad (\text{II.66})$$

In this case, one can use perturbation theory in quantum mechanics to compute the shift of the QCM spectra from the perturbation of $F_{0,\ell \geq 2}$. We assume that

$$\begin{aligned} F_{0,\ell \geq 2} &= g_{0,\ell \geq 2} \hat{F}_{0,\ell \geq 2}, \\ 0 < g_{0,\ell \geq 2} &\ll 1. \end{aligned} \quad (\text{II.67})$$

We discuss notions of perturbation theory for such cases. For simplicity, we assume that f_a s are all different numbers, and there is no extra degeneracy except the permutations,

$$\text{spec}(A_j^j) \supset \bigcup_{\sigma; a} \left(f_{\sigma(a_1)} + f_{\sigma(a_2)} + \dots + f_{\sigma(a_j)} \right)_{a_1, a_2, \dots, a_j=1}^n. \quad (\text{II.68})$$

Here, we combine the degenerate eigenvalues and their corresponding eigenspaces induced by the permutation σ of the list a_1, a_2, \dots, a_j . We give examples about non-degenerate and degenerate perturbations as the following.

- *Non-degenerate perturbation theory:* Consider the eigenvalues

$$\text{spec}(A_j^j) \supset (j \times f_a)_{a=1}^n. \quad (\text{II.69})$$

Those eigenvalues are non-degenerate if those eigenvalues of A have no extra degeneracy and are generically chosen. In this case, the minimal corrections are the second-order terms in perturbation theory $O(g^2)$. We denote the eigenvalue as

$$a_{j,\mu_0} = j \times f_a \quad (\text{II.70})$$

so that the eigenvalue perturbation is given by

$$\begin{aligned} a_{j,\mu_0} &\mapsto a_{j,\mu_0} + g_0^2 \sum_{a_{j-1,\mu} \neq a_{j,\mu_0}} \frac{|\langle j, \mu_0 | \hat{A}_{j-1}^j | j-1, \mu \rangle|^2}{a_{j-1,\mu} - a_{j,\mu_0}} \\ &+ \sum_{\ell=1}^{q-1} g_{\ell+1}^2 \sum_{a_{j+\ell,\mu} \neq a_{j,\mu_0}} \frac{|\langle j, \mu_0 | \hat{A}_{j+\ell}^j | j+\ell, \mu \rangle|^2}{a_{j+\ell,\mu} - a_{j,\mu_0}}. \end{aligned} \quad (\text{II.71})$$

Here,

$$\begin{aligned} \hat{A}_{j+q-1}^j &= \hat{F}_q \otimes I^{\otimes j-1} + I \otimes \hat{F}_q \otimes I^{\otimes j-2} + \dots + I^{\otimes j-1} \otimes \hat{F}_q, \\ &\dots \\ \hat{A}_{j+1}^j &= \hat{F}_2 \otimes I^{\otimes j-1} + I \otimes \hat{F}_2 \otimes I^{\otimes j-2} + \dots + I^{\otimes j-1} \otimes \hat{F}_2, \\ \hat{A}_{j-1}^j &= \hat{F}_0 \otimes I^{\otimes j-1} + I \otimes \hat{F}_0 \otimes I^{\otimes j-2} + \dots + I^{\otimes j-1} \otimes \hat{F}_0, \end{aligned} \quad (\text{II.72})$$

and the sum is taken over all the eigenvalues of A_{j-1}^{j-1} and $A_{j+\ell}^{j+\ell}$ for $1 \leq \ell \leq q-1$, denoted by $a_{j-1,\mu}$ and $a_{j+\ell,\mu}$. The eigenvectors are given by $|j-1, \mu\rangle$ and $|j+\ell, \mu\rangle$. Similarly, the eigenvector corresponding to the eigenvalue a_{j,μ_0} is $|j, \mu_0\rangle$. We have

$$\begin{aligned} \dim\{|j-1, \mu\rangle\} &= n^{j-1}, \\ \dim\{|j, \mu_0\rangle\} &= n^j, \\ \dim\{|j+\ell, \mu\rangle\} &= n^{j+\ell} \end{aligned} \quad (\text{II.73})$$

and $A_{j-1}^j \in \mathbb{R}^{n^j \times n^{j-1}}$, $A_{j+\ell}^j \in \mathbb{R}^{n^j \times n^{j+\ell}}$. Thus, the inner product is well-defined.

• *Degenerate perturbation theory:* We consider the eigenvalue

$$\text{spec} \left(A_j^j \right) \supset \left(a_{j,\sigma} \equiv f_{\sigma(a_1)} + f_{\sigma(a_2)} + \dots + f_{\sigma(a_j)} \right)_\sigma \quad (\text{II.74})$$

with

$$1 \leq a_1 < a_2 \dots < a_j \leq n. \quad (\text{II.75})$$

Here, σ is a permutation among a_1, a_2, \dots, a_j . Those eigenvalues are degenerate, with respective degeneracy $j!$, since there are $j!$ permutations in total. Thus, for a given j , all $a_{j,\sigma}$ are equal, where we could denote as a_j . In this case, the degenerate perturbation theory gives a possible eigenvalue splitting with respect to σ ,

$$\begin{aligned} a_j \mapsto a_j + g_0^2 \sum_{a_{j-1}, \mu \neq a_{j,\sigma}} \frac{\left| \langle j, \sigma | \hat{A}_{j-1}^j | j-1, \mu \rangle \right|^2}{a_{j-1, \mu} - a_{j, \sigma}} \\ + \sum_{\ell=1}^{q-1} g_{\ell+1}^2 \sum_{a_{j+\ell}, \mu \neq a_{j,\sigma}} \frac{\left| \langle j, \sigma | \hat{A}_{j+\ell}^j | j+\ell, \mu \rangle \right|^2}{a_{j+\ell, \mu} - a_{j, \sigma}}. \end{aligned} \quad (\text{II.76})$$

Here, $|j, \sigma\rangle$ is the eigenvector assigned with the permutation σ in the eigenspace with the eigenvalue a_j . Thus we get at most $j!$ different $a_{\sigma,j}$ s in the second-order degenerate perturbation theory, where we have $j!$ eigenvalue splitting at most.

C. Quantum linear system solvers

Until now, we have established how to obtain a linearized differential equation. We now aim at approximately solving this equation by making use of a quantum computer. Given the initial condition described by a quantum state, we aim to provide a quantum-state description of the solution given the terminal time T . Quantum computers can identify the solution of a N_h -dimensional linear system in Hilbert space. Originally proposed by Harrow, Hassidim, and Lloyd [8], quantum linear system algorithms [8, 70–75] have been developed to provide a quantum state encoding the solution with complexity $\text{poly}(\log N_h)$. Building on this body of work, generalizations to solve linear ordinary and partial differential equations [76–80] have been proposed.

Theorem II.3 (Quantum linear system algorithm [75]). *Let $Mx = b$ be a system of linear equations, where M is an s_M -sparse $N_h \times N_h$ matrix with condition number $\kappa := \|M\| \|M^{-1}\|$. Given a sparse matrix oracle for M and an oracle for $|b\rangle$, there exists a quantum algorithm which produces the state vector $|M^{-1}b\rangle$ within error $\epsilon_{\text{HHL}} > 0$ using gate complexity*

$$O(s_M \kappa \cdot \text{poly} \log(N_h / \epsilon_{\text{HHL}})). \quad (\text{II.77})$$

We consider the linearized differential equation

$$\delta \hat{y} = \hat{y}(t+1) - \hat{y}(t) = A \hat{y}(t) + b, \quad \hat{y}(0) = \hat{y}_{\text{in}}, \quad (\text{II.78})$$

with iterations $t \in [T+1]_0 := \{0, 1, \dots, T\}$, which we rewrite as one linear system $MX = B$ in the form

$$\begin{pmatrix} I & & & & \\ -(I+A) & I & & & \\ & \ddots & \ddots & & \\ & & -(I+A) & I & \\ & & & -(I+A) & I \end{pmatrix} \begin{pmatrix} \hat{y}(0) \\ \hat{y}(1) \\ \vdots \\ \hat{y}(T-1) \\ \hat{y}(T) \end{pmatrix} = \begin{pmatrix} \hat{y}_{\text{in}} \\ b \\ \vdots \\ b \\ b \end{pmatrix}. \quad (\text{II.79})$$

M^{-1} can actually be written as

$$\begin{pmatrix} I & & & & \\ I+A & I & & & \\ (I+A)^2 & I+A & I & & \\ \dots & \dots & \dots & \dots & \\ (I+A)^T & \dots & (I+A)^2 & I+A & I \end{pmatrix}, \quad (\text{II.80})$$

such that we find the upper bounds

$$\|M^{-1}\| \leq \|I\| + \|I + A\| + \dots + \|(I + A)^T\|. \quad (\text{II.81})$$

As discussed in Theorem II.1, if all eigenvalues of A satisfy

$$|1 + a_\mu| < 1, \quad (\text{II.82})$$

or, in other words, $\|I + A\| < 1$ holds true, then

$$\kappa = \|M\| \|M^{-1}\| \leq (1 + \|I + A\|) \left(\sum_{t=0}^T \|I + A\|^t \right) = 2(T + 1). \quad (\text{II.83})$$

Given a sparse matrix oracle, the cost of encoding an s -sparse $N_h \times N_h$ matrix A is found to be $O(sq)$. The upper bound on the condition number shows that $\kappa = 2(T + 1)$, which indicates the gate complexity of solving the desired linear system is $O(sqT \cdot \text{poly}(\log(N_h/\epsilon_{\text{HHL}})))$.

Furthermore, we can ask what happens when we have divergent modes? Here, we notice that the HHL algorithm can be implemented only on the well-conditioned eigenspaces by adding auxiliary qubits (see Ref. [8], discussion section), which could be likely applicable to the adiabatic method [75] as well. In the context of the original work [8] when solving $MX = b$, we set

$$|b\rangle := \sum_j \beta_j |u_j\rangle \quad (\text{II.84})$$

where $|u_j\rangle$ are normalized eigenvectors of M . The algorithm will give

$$|x\rangle = \left(\sum_j |\lambda_j^{-1} \beta_j|^2 \right)^{-1/2} \sum_j \lambda_j^{-1} \beta_j |u_j\rangle \quad (\text{II.85})$$

up to the given precision. Instead, one can add auxiliary qubits such that the output is

$$\begin{aligned} |\tilde{x}\rangle &= \left(\sum_{j:|\lambda_j| \geq 1/\kappa_a} |\lambda_j^{-1} \beta_j|^2 + \sum_{j:|\lambda_j| < 1/\kappa_a} |\beta_j|^2 \right)^{-1/2} \\ &\times \left(\sum_{j:|\lambda_j| \geq 1/\kappa_a} \lambda_j^{-1} \beta_j |u_j\rangle |\text{well}\rangle + \sum_{j:|\lambda_j| < 1/\kappa_a} \beta_j |u_j\rangle |\text{ill}\rangle \right). \end{aligned} \quad (\text{II.86})$$

Here, $\kappa_a > 0$ is an effective positive number and may not necessarily be the condition number of M . When $\|M\| = 1$, and $\kappa_a = \kappa$, the algorithm effectively returns to Theorem II.3 regardless of those auxiliary qubits. The auxiliary states $|\text{well}\rangle$ and $|\text{ill}\rangle$ will be post-selected at the end of the algorithm. If we set a general κ_a , the error on the state vector $|x\rangle$ can be estimated by the fidelity

$$\langle x | \tilde{x} \rangle = \frac{\sum_{j:|\lambda_j| \geq 1/\kappa_a} |\lambda_j^{-1} \beta_j|^2}{\left(\sum_j |\lambda_j^{-1} \beta_j|^2 \right)^{1/2} \left(\sum_{j:|\lambda_j| \geq 1/\kappa_a} |\lambda_j^{-1} \beta_j|^2 + \sum_{j:|\lambda_j| < 1/\kappa_a} |\beta_j|^2 \right)^{1/2}}. \quad (\text{II.87})$$

In an actual machine learning process, the weights and biases might depend highly randomly. A good model to reflect this randomness is to assume that $|b\rangle$ is a Haar-random state vector. Namely, $|b\rangle = U_{\text{Haar}} |0\rangle$ where U_{Haar} is distributing uniformly in the whole unitary group (see Ref. [81]). We denote $\rho_0 = |0\rangle\langle 0|$ and

$$\begin{aligned} \Lambda^{-1} &= \text{diag}_j(\lambda_j^{-1}), \\ \bar{\Lambda}^{-1} &= \text{diag}_{j:|\lambda_j| \geq 1/\kappa_a}(\lambda_j^{-1}) \oplus \text{diag}_{j:|\lambda_j| < 1/\kappa_a}(0), \end{aligned} \quad (\text{II.88})$$

where Λ^{-1} is the diagonal matrix of all eigenvalue inversions of M , and $\bar{\Lambda}^{-1}$ is the *pruned* version of Λ^{-1} . This is the step of a sparse truncation: Here all eigenvalues satisfying $|\lambda_j| \geq 1/\kappa_a$ are included, while other eigenvalues satisfying $|\lambda_j| < 1/\kappa_a$ are excluded and set to zero, leading to a sparse setting. We have

$$\begin{aligned} \int dU_{\text{Haar}} \sum_{j: |\lambda_j| \geq 1/\kappa_a} |\lambda_j^{-1} \beta_j|^2 &= \int dU_{\text{Haar}} \langle 0 | U_{\text{Haar}}^\dagger \bar{\Lambda}^{-1\dagger} \bar{\Lambda}^{-1} U_{\text{Haar}} | 0 \rangle \\ &= \int dU_{\text{Haar}} \text{Tr}(U_{\text{Haar}}^\dagger \bar{\Lambda}^{-1\dagger} \bar{\Lambda}^{-1} U_{\text{Haar}} \rho_0) = \frac{1}{\Delta(T+1)} \text{Tr}(\bar{\Lambda}^{-1\dagger} \bar{\Lambda}^{-1}). \end{aligned} \quad (\text{II.89})$$

Since the expression of $\langle x | \tilde{x} \rangle$ only cares about the inner product that involves a pair of U_{Haar} and its Hermitian conjugate, based on the *Page theorem* [81, 82], it is equivalently assuming that $|b\rangle$ is a superposition of all computational basis states. Similarly, we have

$$\int dU_{\text{Haar}} \sum_j |\lambda_j^{-1} \beta_j|^2 = \frac{1}{\Delta(T+1)} \text{Tr}(\Lambda^{-1\dagger} \Lambda^{-1}), \quad (\text{II.90})$$

$$\begin{aligned} \int dU_{\text{Haar}} \sum_{j: |\lambda_j| \geq 1/\kappa_a} |\lambda_j^{-1} \beta_j|^2 &+ \int dU_{\text{Haar}} \sum_{j: |\lambda_j| < 1/\kappa_a} |\beta_j|^2 \\ &= \frac{1}{\Delta(T+1)} \text{Tr}(\bar{\Lambda}^{-1\dagger} \bar{\Lambda}^{-1}) + \frac{\#(j: |\lambda_j| < 1/\kappa_a)}{\Delta(T+1)}, \end{aligned} \quad (\text{II.91})$$

where $\#(j: |\lambda_j| < 1/\kappa_a)$ refers to the number of eigenvalues that are smaller than the cutoff $1/\kappa_a$.

We can use the above formulas to estimate the inner product (II.87). Note that we average over the numerator, the denominator separately, inside or outside the square root separately. This is the difference between the annealed average and its opposite [81, 83], where the error is likely negligible in the limit of large $(T+1)\Delta$, the dimension of the Hilbert space. We can specify the situation in the machine learning case, where we have Δ_+ dissipative modes with eigenvalues λ_+ . We can simply take $\kappa_a = 2(T+1)$ as the upper bound, such that the divergent modes are ill-conditioned. In this case we estimate the fidelity as

$$\begin{aligned} \langle x | \tilde{x} \rangle &= \frac{\Delta_+ \sum_{t=0}^T (1+a_+)^t}{((T+1)\Delta)^{1/2} \left(\Delta_+ \sum_{t=0}^T (1+a_+)^t + (T+1)\Delta_- \right)^{1/2}} \\ &= \frac{\Delta_+ \frac{(a_++1)^{T+1}-1}{a_+}}{((T+1)\Delta)^{1/2} \left(\Delta_+ \frac{(a_++1)^{T+1}-1}{a_+} + (T+1)\Delta_- \right)^{1/2}}. \end{aligned} \quad (\text{II.92})$$

In the limit where $a_+ \rightarrow 0^-$, the formula is simply

$$\langle x | \tilde{x} \rangle = \frac{\Delta_+}{\Delta}, \quad (\text{II.93})$$

which indicates that only a constant number of errors will be introduced if we set κ_a linear in T , where the constant is referring to the ratio of dissipative modes. This observation is independent of the number of iterations T or $(T+1)$.

Finally, we claim that filtering out ill-conditioned eigenvalues might cause additional overhead on the algorithm. The naive usage of the auxiliary qubits in Ref. [8] introduces extra $O(\kappa_a^2)$ gates. Further works about this filtering algorithm require rigorous improvements on auxiliary qubits and post-selection using amplitude amplification and adapting the above framework to the adiabatic setting [75] which we leave for future research.

Theorem II.4 (Truncated quantum linear system algorithm). *Let $Mx = b$ be a system of linear equations, where M is an s_M -sparse $N_h \times N_h$ matrix. We define the effective condition number $\kappa_a > 0$ that filters out all eigenvalues of M smaller than $1/\kappa_a$, and define the truncated solution $|\tilde{x}\rangle$ in Equation (II.86). Given a sparse matrix oracle for M and an oracle for $|b\rangle$, there exists a quantum algorithm which produces the state vector $|\tilde{x}\rangle$ within error $\epsilon_{\text{HHL}} > 0$ using a gate complexity of*

$$O(s_M^2 \kappa_a^2 \text{polylog}(N_h/\epsilon_{\text{HHL}})). \quad (\text{II.94})$$

According to the analysis in Eqs. (II.92) and (II.93), it suffices to choose the effective condition number as $\kappa_a = 2(T+1)$ so that the fidelity of $|\tilde{x}\rangle$ to $|x\rangle$ is close to 1. The gate complexity of solving the desired linear system is $O(s^2 q^2 T^2 \text{polylog}(N_h/\epsilon_{\text{HHL}}))$. Overall, combining with the error bound in differential equations, we get the gate complexity

$$O\left(sqT \times \text{poly}(\log n, \frac{2}{\log(1/\|\bar{u}\|)} \log \frac{3 \sum_{\ell=2}^q (\ell-1) \|F_\ell\|}{\|A\|_{\epsilon_{\text{le}}}}, \frac{1}{\epsilon_{\text{HHL}}})\right) \quad (\text{II.95})$$

for $\|1 + A\| < 1$, and

$$O\left(s^2 q^2 T^2 \times \text{poly}(\log n, \frac{2}{\log(1/\|\bar{u}\|)} \log \frac{3 \sum_{\ell=2}^q (\ell-1) \|F_\ell\|}{\|A\| \epsilon_{\text{le}}}, \frac{1}{\epsilon_{\text{HHL}}})\right) \quad (\text{II.96})$$

for dissipative QCM.

D. Tomographic recovery and sparsity

In this subsection, we briefly discuss notions of tomographic recovery and what role they take in sparse training. The requirement of sparsity for the implementation of our quantum algorithms has two components: The sparsity of weight and bias vectors θ on the one hand and the sparsity of the QCM matrix A on the other hand. The sparsity of the QCM matrix originates from the machine learning architecture and Taylor expansion, while the sparsity of weight matrices comes from the nature of sparse training. Here, we primarily discuss the sparsity of the output weight vector, where we denote by r the number of non-zero entries of a vector with Hilbert space dimension N . The vectors to be recovered tomographically in this step are of the form

$$|x\rangle = \sum_{\alpha=1}^r v_\alpha |p_\alpha\rangle, \quad (\text{II.97})$$

normalized as

$$\sum_{\alpha=1}^r |v_\alpha|^2 = 1, \quad (\text{II.98})$$

where each $|p_\alpha\rangle$ is a computational basis vector of a system of $m = \log_2(n)$ qubits. That is to say, the state is r -sparse in the sense that there are at most r non-vanishing coefficients. So in a first step, it has to be determined which computational basis vectors feature and hence which coefficients are non-zero, as the sparsity pattern in this sense that is not known. This problem can be identified as an instance of the *coupon collector's problem* (see, for instance, Ref. [84]). The algorithm in this step goes as follows. One performs projective measurements in the computational Pauli-Z basis and records observed patterns. This procedure is repeated until all r different patterns are observed. The original coupon collector's problem is recovered in this setting exactly in the situation of the state vector $|x\rangle = r^{-1/2} \sum_{\alpha=1}^r |p_\alpha\rangle$, as an equal superposition, as then each vector reflecting a different pattern is observed with the same probability. For this problem, the expected number of measurements is known to scale as $O(r \log r)$. For the state vector as specified in Equation (II.97), the weights are different for each pattern. For this weighted coupon collector's problem, the expected number of steps until all patterns have been observed is also known [84]: The expected number of repetitions is given by

$$\int_0^{+\infty} \left(1 - \prod_{\alpha=1}^r (1 - e^{-|v_\alpha|^2 y})\right) dy. \quad (\text{II.99})$$

It takes a moment of thought that for equal probability for each probability one arrives at the minimum expectation value. There is a subtlety that is worth commenting upon that relates to small non-vanishing weights. If weights converge to zero, the expectation values of observing all patterns including those of the small weights clearly diverges. In this situation, one can consider only the significant weights above a given threshold $\delta > 0$, which will lead to an efficient algorithm for an approximation up to weights larger than δ , as a detailed analysis shows. Overall, again, this step can be performed making use of $O(r \log r)$ many measurements.

After this step has been taken of identifying the at most r relevant computational basis vectors, corresponding to a r -sparse weight state, we can pursue a variant of state tomography to extract the full pure quantum state including phases. We suggest to do so by implementing a variant of the *shadow estimation scheme* of Ref. [27] making use of random quantum circuits involving the n -qubit Clifford group. For the physical implementation of this, $O(m^2)$ entangling gates are necessary. After implementing random Clifford circuits, one performs computational basis measurements, giving rise to outcomes o reflecting computational basis vectors $|o\rangle$ of the m qubit system. For the input $|x\rangle \langle x|$, the output of this random quantum channel is obtained as

$$\mathcal{M}(|x\rangle \langle x|) = \mathbb{E} \left(U^\dagger |o\rangle \langle o| U \right), \quad (\text{II.100})$$

where $U \in C_m$ is from the m -qubit Clifford group. To compute the inverse as

$$|x\rangle \langle x| = \mathbb{E} \left(\mathcal{M}^{-1}(U^\dagger |o\rangle \langle o| U) \right), \quad (\text{II.101})$$

one has to compute overlaps of computational basis vectors and stabilizer states obtained from applying Clifford circuits to computational basis vectors. Such overlaps can be classically efficiently computed with an effort polynomial in m . The inverse map is found to be

$$\mathcal{M}^{-1}(X) = (2^m + 1)X - I. \quad (\text{II.102})$$

For a state vector of the form as in Equation (II.97), one can take samples of random Clifford circuits applied to $|x\rangle$, and sample from the output distribution obtained from computational basis measurements, creating a classical shadow. From these data, one can infer about the elements v_α for $\alpha = 1, \dots, r$, requiring an overall effort of $\mathcal{O}(m^2 r^3 / \epsilon^2)$ in the random measurement procedure, for a recovery with the 1-norm error $\epsilon > 0$, ignoring logarithmic factors. It also involves a polynomial classical effort in m , due to computing overlaps of stabilizer states and computational basis states [85]. Alternative methods are also conceivable, e.g., based on compressing circuits and the methods presented in Ref. [86] based on projected least squares.

In the actual machine learning process, the ratio of pruned parameters in sparse training $\delta_{\text{tr}}(n)$ is scaling with the size of the model n . In this case, the weight sparsity is given by $r = n(1 - \delta_{\text{tr}}(n))$. For this to be feasible, we expect that r is a constant in n . This requires that

$$\delta_{\text{tr}}(n) = 1 - \mathcal{O}\left(\frac{1}{n}\right). \quad (\text{II.103})$$

This will guarantee that the cost of our algorithm will have a poly-logarithmic dependence in n .

Finally, we wish to mention that in the case of dense input vectors, the tomographic algorithm will read the r vector components with the largest norms, and this could be interpreted as a natural pruning method for further classical processes. In this case, we are assuming a generic dense training approach, and the input dense weight vectors can be uploaded via QRAM. The downloading process could still be efficient if r is not scaling with n , and this could be interpreted as a hybrid quantum-classical pruning method.

III. MACHINE LEARNING AND NUMERICAL ASSESSMENT

In this section, we present substantial numerical evidence regarding the feasibility of accelerating machine learning with our proposed quantum algorithm. First, we formulate the learning dynamics of a *multilayer perceptron* (MLP) machine learning model as an ODE system and make some error scaling statements. Second, we numerically analyze the convergence characteristics of the quantum algorithm for large vision models using their Hessian spectra. This is because computing the original QCM and diagonalize it for large models is numerically intractable. Lastly, we present numerical evidence supporting the use of the Hessian spectra for convergence analysis.

A. Training dynamics of MLPs as ODE systems

To analytically understand neural networks, we consider the MLP, a simple neural network architecture (see Ref. [87]). The definition of its elements is given by

$$\begin{aligned} z_i^{(1)}(\mathbf{x}) &:= b_i^{(1)} + \sum_{j=1}^{n_0} W_{i,j}^{(1)} x_j, \\ \text{for } i &= 1, \dots, n_1, \\ z_i^{(\ell+1)}(\mathbf{x}) &:= b_i^{(\ell+1)} + \sum_{j=1}^{n_\ell} W_{i,j}^{(\ell+1)} \sigma\left(z_j^{(\ell)}(\mathbf{x})\right), \\ \text{for } i &= 1, \dots, n_{\ell+1}; \quad \ell = 1, \dots, L-1, \end{aligned} \quad (\text{III.1})$$

where \mathbf{x} is the input data vector, $\mathbf{W}^{(\ell)}$ and $\mathbf{b}^{(\ell)}$ are the ℓ -th layer trainable weights and biases, $\mathbf{z}^{(\ell)}$ is the ℓ -th layer preactivation, n_ℓ is the ℓ -th layer width, and σ is a non-linear activation function. In vector notation,

$$\mathbf{z}^{(\ell+1)}(\mathbf{x}) := \mathbf{b}^{(\ell+1)} + \mathbf{W}^{(\ell+1)} \sigma\left(\mathbf{z}^{(\ell)}(\mathbf{x})\right), \quad (\text{III.2})$$

where σ is applied element-wise. During training of the neural network, parameters update according to the gradient descent algorithm on the *mean squared error* (MSE) loss

$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \frac{d\mathcal{L}_{\mathcal{A}}}{d\theta_\mu} \Big|_{\theta(t)}, \quad (\text{III.3})$$

$$\mathcal{L}_{\mathcal{A}} := \frac{1}{2} \sum_{\alpha \in \mathcal{A}} |z^{(L)}(\mathbf{x}_\alpha; \boldsymbol{\theta}) - \mathbf{y}_\alpha|^2, \quad (\text{III.4})$$

where $\alpha \in \mathcal{A}$ are elements in the training set \mathcal{A} , \mathbf{y} are the labels (referred to as ground truths), $\mathbf{z}^{(L)}$ are the final predictions from the MLP model, η is the training rate, and $\boldsymbol{\theta}$ is the vectorized combination of all weights and biases. In the language of ODE, $u \mapsto \boldsymbol{\theta}$ and

$$\sum_{\ell=0}^q F_\ell u^{\otimes \ell}(t) \mapsto -\eta \frac{d\mathcal{L}_{\mathcal{A}}}{d\theta_\mu} \Big|_{\theta(t)}.$$

Therefore, we need to express $\eta \frac{d\mathcal{L}_{\mathcal{A}}}{d\theta_\mu} \Big|_{\theta(t)}$ as a polynomial in $\boldsymbol{\theta}$. We have

$$\frac{d\mathcal{L}_{\mathcal{A}}}{d\theta_\mu} = \sum_{\alpha \in \mathcal{A}} \left[\mathbf{z}^{(L)}(\mathbf{x}_\alpha) \cdot \frac{d\mathbf{z}^{(L)}(\mathbf{x}_\alpha)}{d\theta_\mu} - \mathbf{y}_\alpha \cdot \frac{d\mathbf{z}^{(L)}(\mathbf{x}_\alpha)}{d\theta_\mu} \right]. \quad (\text{III.5})$$

Since \mathbf{x}, \mathbf{y} are fixed, $\mathbf{z}^{(\ell)}$ are already functions of θ , what remains to be shown is how to express the derivatives as functions of $\boldsymbol{\theta}$ as well. We find

$$\frac{dz_i^{(L)}(\mathbf{x}_\alpha)}{db_j^{(\ell)}} = \frac{dz_i^{(L)}(\mathbf{x}_\alpha)}{dz_j^{(\ell)}(\mathbf{x}_\alpha)}, \quad \frac{dz_i^{(L)}(\mathbf{x}_\alpha)}{dW_{jk}^{(\ell)}} = \frac{dz_i^{(L)}(\mathbf{x}_\alpha)}{dz_j^{(\ell)}(\mathbf{x}_\alpha)} \sigma \left(z_k^{(\ell-1)}(\mathbf{x}_\alpha) \right), \quad (\text{III.6})$$

$$\begin{aligned} \frac{dz_i^{(L)}(\mathbf{x}_\alpha)}{dz_j^{(\ell)}(\mathbf{x}_\alpha)} &= \sum_{k_{\ell+1}, k_{\ell+2}, \dots, k_{L-1}}^{n_{\ell+1}, n_{\ell+2}, \dots, n_{L-1}} W_{i,j}^{(L)} W_{k_{L-1}j}^{(L-1)} \dots W_{k_{\ell+2}j}^{(\ell+2)} W_{k_{\ell+1}j}^{(\ell+1)} \\ &\quad \times \sigma \left(z_j^{(L-1)}(\mathbf{x}_\alpha) \right)' \sigma \left(z_j^{(L-2)}(\mathbf{x}_\alpha) \right)' \dots \sigma \left(z_j^{(\ell+1)}(\mathbf{x}_\alpha) \right)' \sigma \left(z_j^{(\ell)}(\mathbf{x}_\alpha) \right)'. \end{aligned} \quad (\text{III.7})$$

We now have increments of $\boldsymbol{\theta}$ in terms of $\boldsymbol{\theta}$ (and training data), and have obtained a non-linear ODE system. The highest order term of the ODE corresponds to the terms derived from $d\mathbf{z}^{(L)}/dW_{j,k}^{(1)}$ and $d\mathbf{z}^{(L)}/db_j^{(1)}$. Assuming that the non-linearity σ is an order s polynomial, then $\sigma^{(1)}$ is an order s polynomial in θ_μ , $\mathbf{z}^{(2)}$ has order $s+1$, $\sigma^{(2)}$ has order $s(s+1)$, and $\sigma^{(\ell)}$ has order $\sum_{m=1}^{\ell} s^m = s(s^\ell - 1)/(s - 1)$. From Equation (III A), $d\mathbf{z}_{i;\alpha}^{(L)}/dz_{j;\alpha}^{(1)}$ is of order

$$O((L-1) + \sum_{\ell=1}^{L-1} \left(\frac{s(s^\ell - 1)}{s - 1} - 1 \right)). \quad (\text{III.8})$$

Since \mathbf{z} is one order higher than $\sigma^{(L-1)}$, Equation (III A) leads to the following result.

Theorem III.1 (Bound to the order). *For an L layer MLP with activation functions of order s , the ODE $d\theta_\mu/dt = d\mathcal{L}_{\mathcal{A}}/d\theta_\mu$ has order*

$$O \left(\left(\frac{s(s^{L-1} - 1)}{s - 1} + 1 \right) + \left[(L-1) + \sum_{\ell=1}^{L-1} \left(\frac{s(s^\ell - 1)}{s - 1} - 1 \right) \right] \right). \quad (\text{III.9})$$

For large s , is ODE is approximately of order $q = s^{L-1}$ in terms of θ_μ .

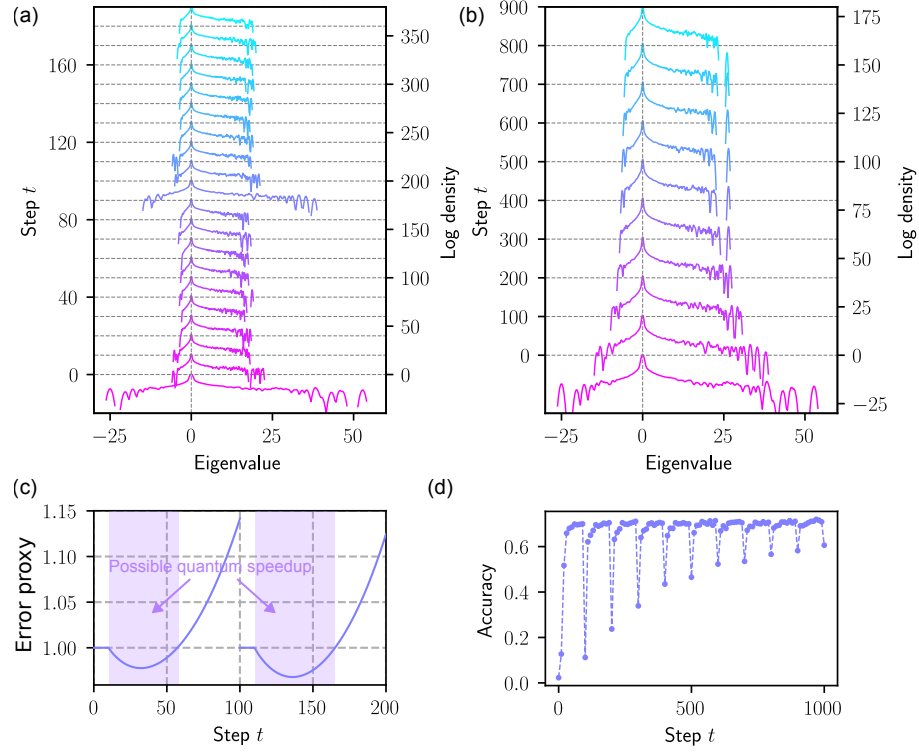


FIG. 2. Numerical results on ResNet as a function of step. Each step corresponds to a step of stochastic gradient descent based on the derivatives of the loss computed from 2048 randomly selected training samples. (a) Hessian spectra of the ResNet during the first 200 steps of training shown for every 10 steps. The neural network is pruned with ratio 0.9 every 100 steps, which corresponds to the sudden broadening of the Hessian spectra. (b) Hessian spectra shown for every 100 steps where spectra broadening take place due to pruning. (c) Estimate of error proxy. The first 10 steps after pruning are trained classically, hence incurring no error. (d) Neural network accuracy during training.

B. Numerical analysis on sparse ResNet for vision

In order to understand the potential of our approach in accelerating training of state-of-the-art large neural networks, we study the training dynamics of the *ResNet* [88] applied to the *CIFAR-100* [89] data set. *CIFAR-100* is a set of 60000 images in 100 categories, and the neural network is trained to classify given images. We first train a dense *ResNet*, where all parameters are kept as normal, with depth 32 and widen factor 4. The parameter count for the dense model is 7.451.044. After that, 90% of the parameters are being pruned (set to 0), and a new model is trained. We allow pruned parameters to evolve to non-zero values so that we might obtain a different set of parameters as sparse weights. Further, since the quantum training algorithm deviates from the ideal training trajectory after a number of steps, we download sparse weights (10% of all parameters in this case) every 100 steps from the quantum ODE system and reupload the sparse weights to continue quantum training. We train the *ResNet* and examine the Hessian spectra at different steps as shown in Fig. 2 (a) and (b), where each step reflects an update using gradient estimated from 2048 randomly selected training samples with a learning rate of 0.01. We observe in Fig. 2 (a) that the Hessian spectra immediately after sparse weights download are broad, but rapidly concentrates and becomes dissipative within 10 steps of training. We also show the Hessian spectra immediately after sparse weights download for the entire training process. The spectra broadening effect of downloads becomes less significant as training progresses. The entire training process contains more dissipative modes than divergent modes as can be seen in the Hessian spectra.

We further quantify the impact of the Hessian spectra on the convergence of the linearized solver. Equation (II.21) gives us a bound on the linearization error. Assuming $Q_{\mu,\nu}$ are the same for all μ, ν , we test this criterion by examining

$$\text{error proxy} = \frac{1}{\Delta} \sum_{\nu} |(1 + a_{\nu})^t|, \quad (\text{III.10})$$

where a_{ν} s are the negatives of the Hessian eigenvalues. If all eigenvalues are exactly 1, neither dissipative or divergent, the quantity stays exactly 1. If the divergent contributions outweigh the dissipative contributions, the quantity is greater than 1. We numerically compute the density of Hessian eigenvalues given by $\rho(a)$, and we have $\int_{-\infty}^{\infty} \rho(a) da = 1$. Assuming the Hessian

spectrum stays constant, the error quantity can be estimated as

$$\text{error proxy} = \int_{-\infty}^{\infty} \rho(a)|(1+a)^t|da. \quad (\text{III.11})$$

We estimate the error proxy by sampling the eigenvalue densities at 256 points. Further, since eigenvalues close to 0 are extremely abundant and lead to stationary error proxy, we remove eigenvalues with magnitudes less than 0.4 and renormalize the densities. More explicitly, we define the error proxy as

$$\text{error proxy} = \frac{1}{N_c} \int_{-\infty}^{-0.4} \rho(a)|(1+a)^t|da + \frac{1}{N_c} \int_{0.4}^{\infty} \rho(a)|(1+a)^t|da \quad (\text{III.12})$$

where

$$\frac{1}{N_c} \int_{-\infty}^{-0.4} \rho(a)da + \frac{1}{N_c} \int_{0.4}^{\infty} \rho(a)da = 1. \quad (\text{III.13})$$

Since the Hessian spectra are broad immediately after download, the error proxy tends to diverge. Therefore, we consider the training scenario where the first 10 steps after download are trained classically without error, and the new dense weights are uploaded to the quantum ODE system with QRAM. As can be seen in Fig. 2 (a), the Hessian spectra after 10 steps following download are already concentrated and mainly dissipative, which results in less rapidly growing linearization error.

Fig. 2 (c) shows the estimated error proxy. The error proxy stays exactly at 1 for 10 steps after download because the model is trained classically and has no linearization error. Afterwards, we observe that the error proxy initially decreases due to the dominance of dissipative modes. However, as time increases, the exponentially growing contribution from divergent modes eventually takes over, leading to an apparently exploding error. We download the weights before the error becomes too unreasonable, and resumes the classical training and weights reuploading scheme. Since the Hessian spectra for larger step numbers have less divergent contributions as indicated in Fig. 2 (b), we can be fairly certain that the error proxy will be even better managed in later stages of training.

Finally, we examine the generalization ability of the model following this training scheme. Fig. 2 (d) shows that the test set accuracy drops after download, but the drop becomes less significant as training progresses. Our quantum ODE training scheme, therefore, produces useful sparse machine learning models.

In addition, we examine the Hessian of a 103 million parameter ResNet. We start with a pre-trained model and prune 90% of the parameters. Due to the immense computational cost of computing Hessian for a large machine learning model (a relatively large-scale model for computational vision based on our computational resources), we only benchmark the Hessian spectra to provide evidences of dissipation and potential quantum enhancements. Fig. 3 shows the initial Hessian, which clearly shows the dominance of dissipative modes over divergent modes similar to the 7 million parameter model. Since the Hessian improves with training for the 7 million parameter model, we believe this is evidence that the 103 million parameter model will have similarly manageable error growth.

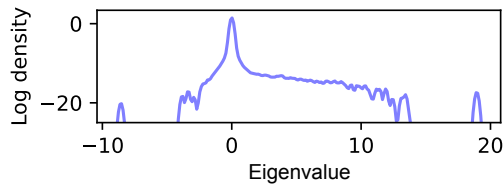


FIG. 3. Hessian of the pruned 103 million parameter model immediately after pruning without any additional training.

C. Numerical verification of the effect of Hessian on convergence

In this subsection, we provide further evidence that dissipative Hessians actually lead to better convergence. We have simulated systems of quadratic ODEs, where the F_1 matrix has a tunable percentage of positive eigenvalues, and the F_2 matrix entries are randomly sampled from a uniform distribution $U(-0.015, 0.015)$. To generate random F_1 matrices, we first generate positive eigenvalues from $U(0, 1)$ and negative eigenvalues from $U(-0.01, 0)$, form F_1 in the diagonal form, and transform its basis using a randomly generated orthogonal matrix. F_0 is 0 for all entries.

First, we provide evidence for the suppression of error in time and the QCM order N in the case that all modes are dissipative in Fig. 4(a), which shows the error of a randomly chosen parameter. If all modes are dissipative, an exponential reduction in the

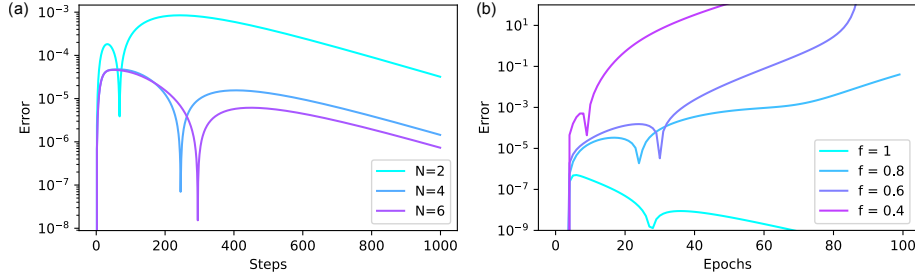


FIG. 4. Numerical evidence for the effect of the Hessian on the approximation error. (a) Error scaling for different QCM orders N for a Hessian with all positive eigenvalues. (b) Error scaling for different fractions f of positive Hessian eigenvalues.

error as time increases is observed due to the exponentially vanishing error bound. Second, for different fractions f of positive Hessian eigenvalues, the error of a randomly chosen parameter for $N = 4$ is shown in Fig. 4(b). For $f < 1$, divergent modes take over and eventually lead to exponentially growing errors in the limit of large times. However, reducing the number of divergent modes significantly suppresses the error, and extends the time period within which the error remains controlled.

IV. FUTURE IMPROVEMENTS

Our algorithm, as an end-to-end application of the HHL algorithm to the context of quantum machine learning, seems to open up several novel research directions. In this section, we hint at some future improvements building on our algorithm that may eventually lead to more effective enhancements of classical machine learning tasks using quantum technologies.

- *Time-dependent version during gradient descent trajectories:* In our work, we have defined the QCMs by Taylor expanding the machine learning models around specific starting points. For models beyond the scale-invariant activation functions, it might, however, be favorable to improve the previous analysis towards a time-dependent picture of generic matrix elements in the QCM matrices. The current strategy we have laid out in Section III is to expand the model around different points of parameters during the gradient descent trajectories. A re-formulation of the existing theory towards fully time-dependent QCM might make our theory more solid and complete.
- *Better criteria for dissipation:* In our current work, we define the contributions of different signs of the eigenvalues of QCM to be the primary criteria of dissipation. These, however, are hard to evaluate in large-scale models. Instead, in practice, we might want to count the contributions from the diagonal elements, or, alternatively, the Hessian eigenvalues as criteria when aiming at identifying regimes of possible quantum advantages. The question arises whether there are any better criteria that are both more solid and more computationally efficient.
- *Connections to diffusion models in classical machine learning:* Our work highlights the observation that dissipation is important to obtain such an efficient quantum enhancement. The celebrated diffusion models used in the current LLM [90] are also dissipative systems. It might be interesting to explore their connections.
- *Theoretical improvements on the truncated HHL algorithms:* In Theorem II.4, we propose a truncated version of the HHL algorithm. However, improvements were plausible if one could extend the ideas from Ref. [75] to such a truncated version, which will likely reduce the overhead related to the sparsity and the condition number from $O(s^2 \kappa_a^2)$ down to $O(s \kappa_a)$.
- *Quantum enhancement beyond dissipation:* It is demonstrated in Ref. [36] that there is numerical evidence that the ODE system may not need to feature full dissipation to obtain a quantum speedup. It might be interesting to explore alternative methods to avoid the condition of dissipation and still provide efficient quantum algorithms.
- *Classical counterparts:* In the sparse training process, it might be possible that for certain sparse training schemes with sufficient few non-zero entries in the weight vectors, efficient classical algorithms might also be created to restrict the calculations inside the r -dimensional subspaces over the whole weight vector space. It might be interesting also to consider more sophisticated hybrid quantum-classical combinations to ensure maximal performances with low overheads.
- *Large learning rate:* In our theory we assume that the learning rate is small enough and the update of weight vectors is slow. This is a usual setting in machine learning problems. However, the learning rate could be large in principle for specific applications [91, 92]. We give an explicit solution about the Carleman linearization in the discrete case in Section V B. We leave the studies about how the discreteness will affect the performances of machine learning problems in large learning rates for future works.

V. OTHER COMMENTS

A. Aaronson's criteria

In this section, we finally briefly comment on our quantum algorithms under the view of Ref. [93]. We regard our quantum algorithm as an end-to-end provable application of the HHL algorithm under certain conditions. Ref. [93] summarizes four perspectives on applying HHL, solving $M|x\rangle = |b\rangle$ and getting $|x\rangle$, into machine learning problems with solid guarantees. We will address how our algorithm takes into consideration about all four conditions:

- *Initial state preparation (resolved by either QRAM or sparse training)*: Applying the HHL algorithm requires the efficient implementations of the initial state. In general, QRAM [37] might be required to encode non-sparse classical data into the quantum computer. QRAM as a possible efficient uploading oracle architecture has been extensively studied in Refs. [42, 94–96]. It is practically challenging to realize such a machine in practice [43], although it is realized that polynomially controllable error scaling is possible in certain devices [94]. In our work under the sparse training assumption, the input state is sparse due to the assumption of sparse training, and an efficient algorithm has been constructed without resorting to QRAM. For instance, Ref. [55] proposes an algorithm for loading an r -sparse quantum state for use in machine learning models with size n , using $O(r^2 \log r \log n)$ classical time, and $O(r \log n)$ quantum circuit depth (which only contains single qubit gates and CNOT gates). Thus, under the sparse training assumption, we are able to avoid using QRAM altogether. On the other hand, in the dense training case where input weight vectors are dense, our analysis is still valid when QRAM is available.
- *The sparsity of the inverted matrix (resolved by machine learning model architectures)*: HHL requires the sparsity of M to guarantee the efficiency of the algorithm. In our setup, this issue is related to machine learning architectures. If the machine learning model is polynomial (for instance, the ReLu architecture model), the matrix M is guaranteed to be sparse. Other architectures are possible as well to make sure the sparsity of M , as long as there are reasonable controls on the Taylor truncation about the weight parameters.
- *Condition number of the inverted matrix (resolved by the dissipative assumptions)*: The HHL algorithm also requires a polynomial dependence on the condition number of the matrix M to maintain efficiency. In our setup, the condition number is related to our assumption of dissipation. If our QCM matrix A satisfies $\|1 + A\| < 1$, it is entirely dissipative, and the condition number is linearly bounded by the number of iterations T . If not, Theorem II.2 and Theorem II.4 will bound the efficiency of the algorithm at early times of training, as has been shown in Section III.
- *The tomographic overhead of the output state vector (resolved by sparse training)*: HHL algorithm provides $|x\rangle$ as a quantum state vector as an output, and tomographic efforts are needed to bring the quantum data back to the classical devices. This issue could be resolved by our assumptions of sparse training, as has been discussed in Section II D. For a r -sparse quantum state, the tomographic recovery need $O(m^2 r^3 / \epsilon^2)$ additional resource with error $\epsilon > 0$ ignoring logarithmic factors, which is computationally efficient. On the other hand, if we are under the dense training assumption, the tomographic recovery could still work to read r components of the weight vectors with the largest norms, and this is effectively a pruning method toward further training steps in the classical processors.

Overall, our algorithm is considered an end-to-end HHL algorithm applications under our assumptions that provides guarantees satisfying the requirements of the HHL algorithm and offers efficient interfaces between classical and quantum processors.

The potentially exponential quantum enhancement we suggest invites efforts for training general large-scale machine learning models, rather than specific machine learning tasks such as those considered in Refs. [25, 28, 29]. The capability of deep neural network models is far from being well understood, and they may not provide the most efficient resolution for certain tasks due to the over-parameterization. However, the computational overhead in terms of parameterizations n is generally unavoidable and costly in practice. Actually, it seems highly unlikely that any classical algorithm for non-linear dissipative differential equations (and training dynamics) can run in time complexity poly-logarithmic in n . If we consider the continuous-time limit of the differential equations as a system of differential equations, and let the dissipation and non-linearity approach zero, then asymptotically we might have a system of linear differential equations with no dissipation. The problem of efficiently simulating non-dissipative linear differential equations is BQP-hard even when the dynamics is entirely unitary, generated by Hamiltonian evolution [97]. In other words, an efficient classical algorithm for non-linear dissipative differential equations would imply a classical algorithms for any problem that can be solved efficiently by a quantum computer, which is considered implausible. A similar argument refers to Section 7 of Ref. [36]. This can be seen as an argument that the de-quantization by classical algorithms along the lines of Ref. [98] of the presented quantum algorithm seems highly implausible, even though to assess this matter could give rise to a fruitful research question.

B. Discrete contributions from Carleman linearization

Here we will give a more detail explanation on Equation II.4.
Say that we have a differential equation,

$$du = f(u) , \quad (\text{V.1})$$

where d is an infinitesimal differential operator. The Taylor expansion tells,

$$f(u) = f(0) + \sum_{\mu} (\partial_{\mu} f) u_{\mu} + \sum_{\mu_1 \mu_2} \frac{1}{2} (\partial_{\mu_1 \mu_2} f) u_{\mu_1 \mu_2} + \dots \quad (\text{V.2})$$

There is an alternative way of using this, which is,

$$f(u) = \sum_{k=0}^{\Lambda_{\infty}} \frac{1}{k!} (\partial_{\otimes(k)} f) u^{\otimes(k)} , \quad (\text{V.3})$$

where,

$$\partial_{\otimes(k)} f = \text{vec}(\partial_{\mu_1 \mu_2 \dots \mu_k} f) , \quad (\text{V.4})$$

and we use the notation where $\Lambda_{\infty} = +\infty$. Specifically, when we are dealing with the following differential equation,

$$du = F_q u^{\otimes(q)}(t) + \dots + F_2 u^{\otimes(2)}(t) + F_1 u(t) + F_0 . \quad (\text{V.5})$$

Here, F_0 could be either time-dependent or time-independent. Then we can identify,

$$\frac{1}{k!} (\partial_{\otimes(k)} f) = F_k . \quad (\text{V.6})$$

Using the language of the Kronecker products, we could try to understand Carleman linearization. Let us start with a simple example,

$$d(u^{\otimes(2)}) = d(u \otimes u) = d(u) \otimes u + u \otimes d(u) . \quad (\text{V.7})$$

One could compute each term as,

$$\begin{aligned} d(u) \otimes u &= \sum_{k=0}^{\Lambda_{\infty}} \left(\text{vec}(F_k) u^{\otimes(k)} \right) \otimes u = \sum_{k=0}^{\Lambda_{\infty}} \left(\text{vec}(F_k \otimes I) \left(u^{\otimes(k+1)} \right) \right) , \\ u \otimes d(u) &= \sum_{k=0}^{\Lambda_{\infty}} u \otimes \left(\text{vec}(F_k) u^{\otimes(k)} \right) = \sum_{k=0}^{\Lambda_{\infty}} \left(\text{vec}(I \otimes F_k) \left(u^{\otimes(k+1)} \right) \right) . \end{aligned} \quad (\text{V.8})$$

In general, we get,

$$d(u^{\otimes(i)}) = \sum_{k=0}^{\Lambda_{\infty}-i+1} G_{i,k} u^{\otimes(k+i-1)} , \quad (\text{V.9})$$

where,

$$G_{i,k} = \text{vec} \left(\sum_{\ell=0}^{i-1} I^{\otimes(\ell)} \otimes F_k \otimes I^{\otimes(i-1-\ell)} \right) . \quad (\text{V.10})$$

Now, instead, let us estimate the difference between the discrete difference δu and the differential du . Starting again with the example we have before, we get,

$$\begin{aligned} \delta(u^{\otimes(2)}) &= \delta(u \otimes u) = u(t+1) \otimes u(t+1) - u(t) \otimes u(t) \\ &= u(t+1) \otimes (u(t+1) - u(t)) + (u(t+1) - u(t)) \otimes u(t) \\ &= u(t+1) \otimes \delta u(t) + \delta u(t) \otimes u(t) . \end{aligned} \quad (\text{V.11})$$

For each term, we have,

$$\delta u(t) \otimes u(t) = \sum_{k=0}^{\Lambda_{\infty}} \left(\text{vec}(F_k) u^{\otimes(k)}(t) \right) \otimes u(t) = \sum_{k=0}^{\Lambda_{\infty}} \left(\text{vec}(F_k \otimes I) \left(u^{[0,k+1]}(t) \right) \right), \quad (\text{V.12})$$

$$u(t+1) \otimes \delta u(t) = \sum_{k=0}^{\Lambda_{\infty}} u(t+1) \otimes \left(\text{vec}(F_k) u^{\otimes(k)}(t) \right) = \sum_{k=0}^{\Lambda_{\infty}} \left(\text{vec}(I \otimes F_k) \left(u^{[1,k]}(t) \right) \right). \quad (\text{V.13})$$

Here, we use the notation,

$$u^{[i,j]}(t) \equiv u^{\otimes(i)}(t+1) \otimes u^{\otimes(j)}(t). \quad (\text{V.14})$$

Thus, in general we get,

$$\delta \left(u^{\otimes(i)} \right) = \sum_{k=0, \ell=0}^{\Lambda_{\infty}-i+1, i-1} g_{k,\ell}^i u^{[\ell, k+i-1-\ell]}, \quad (\text{V.15})$$

where,

$$g_{k,\ell}^i = \text{vec} \left(I^{\otimes(\ell)} \otimes F_k \otimes I^{\otimes(i-1-\ell)} \right). \quad (\text{V.16})$$

Note that we have the identity,

$$G_{i,k} = \sum_{\ell=0}^{i-1} g_{k,\ell}^i. \quad (\text{V.17})$$

For differential equations, we could try to derive the recursion relation for the linearization error. Using the notation in the main text, we denote,

$$\eta_j(t) = u^{\otimes(j)}(t) - \hat{y}_j(t). \quad (\text{V.18})$$

With the Taylor expansion truncation at the order N , we have,

$$\begin{aligned} d\eta_j(t) &= du^{\otimes(j)}(t) - d\hat{y}_j(t) \\ &= \sum_{k=0}^{\Lambda_{\infty}-j+1} G_{i,k} u^{\otimes(k+j-1)} - \sum_{k=0}^{N-j+1} G_{i,k} \hat{y}_{k+j-1} \\ &= \sum_{k=0}^{N-j+1} G_{i,k} \eta_{k+j-1} + \sum_{k=N-j+2}^q G_{i,k} u^{\otimes(k+j-1)}, \end{aligned} \quad (\text{V.19})$$

which is exactly Equation II.10 in the matrix form. Here, we use the fact that $G_{i,k} = 0$ if $k > q$. On the other hand, in the discrete case, we have,

$$\begin{aligned} \delta\eta_j(t) &= \delta u^{\otimes(j)}(t) - \delta\hat{y}_j(t) \\ &= \sum_{k=0, \ell=0}^{\Lambda_{\infty}-j+1, j-1} g_{k,\ell}^j u^{[\ell, k+j-1-\ell]} - \sum_{k=0}^{N-j+1} G_{i,k} \hat{y}_{k+j-1} \\ &= \sum_{k=0, \ell=0}^{\Lambda_{\infty}-j+1, j-1} g_{k,\ell}^j u^{[\ell, k+j-1-\ell]} - \sum_{k=0}^{\Lambda_{\infty}-j+1} G_{i,k} u^{\otimes(k+j-1)} + \sum_{k=0}^{\Lambda_{\infty}-j+1} G_{i,k} u^{\otimes(k+j-1)} - \sum_{k=0}^{N-j+1} G_{i,k} \hat{y}_{k+j-1} \\ &= \sum_{k=0, \ell=0}^{\Lambda_{\infty}-j+1, j-1} g_{k,\ell}^j u^{[\ell, k+j-1-\ell]} - \sum_{k=0}^{\Lambda_{\infty}-j+1} G_{i,k} u^{\otimes(k+j-1)} + \sum_{k=0}^{N-j+1} G_{i,k} \eta_{k+j-1} + \sum_{k=N-j+2}^q G_{i,k} u^{\otimes(k+j-1)}. \end{aligned} \quad (\text{V.20})$$

Thus, we could define the extra piece between du and δu as the extra contribution from discretization,

$$\delta\zeta_j(t) \equiv \sum_{k=0, \ell=0}^{\Lambda_{\infty}-j+1, j-1} g_{k,\ell}^j u^{[\ell, k+j-1-\ell]} - \sum_{k=0}^{\Lambda_{\infty}-j+1} G_{i,k} u^{\otimes(k+j-1)}. \quad (\text{V.21})$$

In fact, we have,

$$\begin{aligned}\delta\zeta_j(t) &= \sum_{k=0, \ell=0}^{\Lambda_\infty-j+1, j-1} g_{k, \ell}^j \left(u^{[\ell, k+j-1-\ell]} - u^{\otimes(k+j-1)} \right) \\ &= \sum_{k=0, \ell=1}^{\Lambda_\infty-j+1, j-1} g_{k, \ell}^j \delta u^{\otimes(\ell)} \otimes u^{\otimes(k+j-1-\ell)}.\end{aligned}\quad (\text{V.22})$$

$\delta\zeta_j(t)$ is $O(\eta^2)$ in learning rate η , since $g = O(\eta)$ and $\delta u = O(\eta)$. The condition for making $\delta\zeta_j(t)$ ignored is that,

$$\left| \sum_{k=0, \ell=1}^{\Lambda_\infty-j+1, j-1} g_{k, \ell}^j \delta u^{\otimes(\ell)} \otimes u^{\otimes(k+j-1-\ell)} \right| \ll \left| \delta u^{\otimes(j)} \right|, \quad (\text{V.23})$$

which is,

$$\left| \sum_{k=0, \ell=1, k'=0}^{\Lambda_\infty-j+1, j-1, \Lambda_\infty-\ell+1} g_{k, \ell}^j G_{i, k'} u^{\otimes(k+k'+j-2)} \right| \ll \left| \sum_{k'=0}^{\Lambda_\infty-j+1} G_{i, k'} u^{\otimes(k'+j-1)} \right|. \quad (\text{V.24})$$

in the limit where η is small, such that we could do perturbative expansion.

There is another more direct condition. We start from,

$$\delta u^{\otimes(i)} = \sum_{k_1=0, \ell_1=0}^{\Lambda_\infty-i+1, i-1} g_{k_1, \ell_1}^i u^{\otimes(k_1+i-1)} + \sum_{k_1=0, \ell_1=1}^{\Lambda_\infty-i+1, i-1} g_{k_1, \ell_1}^i \delta u^{\otimes(\ell_1)} \otimes u^{\otimes(k_1+i-1-\ell_1)}, \quad (\text{V.25})$$

and we also have,

$$\delta u^{\otimes(\ell_1)} = \sum_{k_2=0, \ell_2=0}^{\Lambda_\infty-\ell_1+1, \ell_1-1} g_{k_2, \ell_2}^{\ell_1} u^{\otimes(k_2+\ell_1-1)} + \sum_{k_2=0, \ell_2=1}^{\Lambda_\infty-\ell_1+1, \ell_1-1} g_{k_2, \ell_2}^{\ell_1} \delta u^{\otimes(\ell_2)} \otimes u^{\otimes(k_2+\ell_1-1-\ell_2)}. \quad (\text{V.26})$$

Using this recursion relation multiple times, we get,

$$\begin{aligned}\delta u^{\otimes(i)} &= \sum_{k_1=0, \ell_1=0}^{\Lambda_\infty-i+1, i-1} g_{k_1, \ell_1}^i u^{\otimes(k_1+i-1)} + \sum_{\ell_1=1, \ell_2=0, K=k_1, k_1=0}^{i-1, \ell_1-1, 2\Lambda_\infty-i-\ell_1+2, \Lambda_\infty-i+1} g_{k_1, \ell_1}^i g_{K-k_1, \ell_2}^{\ell_1} u^{\otimes(K+i-2)} \\ &+ \dots + \sum_{\ell_1=1, \ell_2=1, \dots, \ell_i=1, \ell_i=0}^{i-1, \ell_1-1, \ell_{i-2}-1, \ell_{i-1}-1, i\Lambda_\infty-\ell_1-\ell_2-\dots-\ell_{i-1}, \Lambda_\infty-i+1, \Lambda_\infty-\ell_1+1, \dots, \Lambda_\infty-\ell_{i-1}+1} g_{k_1, \ell_1}^i g_{k_2, \ell_2}^{\ell_1} g_{k_3, \ell_3}^{\ell_2} \dots g_{k_{i-1}, \ell_{i-1}}^{\ell_{i-2}} g_{K-k_1-\dots-k_{i-1}, \ell_i}^{\ell_{i-1}} u^{\otimes K}.\end{aligned}\quad (\text{V.27})$$

Note that since $g = O(\eta)$, we schematically could understand the equation as,

$$\delta u^{\otimes(i)} = O(\eta) + O(\eta^2) + \dots + O(\eta^i). \quad (\text{V.28})$$

This equation could be understood as the exact solution of the Carleman linearization in the discrete case. The linearization error η vector, could be written as $\delta\eta = A'\eta + \hat{b}'$, where A' and \hat{b}' will get reduced to A and \hat{b} in Equation II.9 when the learning rate η is small.

In fact, we know that the discretization contribution could be ignored if,

$$\left| \sum_{k_1=0, \ell_1=0}^{\Lambda_\infty-i+1, i-1} g_{k_1, \ell_1}^i u^{\otimes(k_1+i-1)} \right| \gg \left| \begin{aligned} &\sum_{\ell_1=1, \ell_2=0, K=k_1, k_1=0}^{i-1, \ell_1-1, 2\Lambda_\infty-i-\ell_1+2, \Lambda_\infty-i+1} g_{k_1, \ell_1}^i g_{K-k_1, \ell_2}^{\ell_1} u^{\otimes(K+i-2)} \\ &+ \dots + \sum_{\ell_1=1, \ell_2=1, \dots, \ell_i=1, \ell_i=0}^{i-1, \ell_1-1, \ell_{i-2}-1, \ell_{i-1}-1, i\Lambda_\infty-\ell_1-\ell_2-\dots-\ell_{i-1}, \Lambda_\infty-i+1, \Lambda_\infty-\ell_1+1, \dots, \Lambda_\infty-\ell_{i-1}+1} g_{k_1, \ell_1}^i g_{k_2, \ell_2}^{\ell_1} g_{k_3, \ell_3}^{\ell_2} \dots g_{k_{i-1}, \ell_{i-1}}^{\ell_{i-2}} g_{K-k_1-\dots-k_{i-1}, \ell_i}^{\ell_{i-1}} u^{\otimes K} \end{aligned} \right|. \quad (\text{V.29})$$

Factorizing u , one could make the inequality purely a condition about the tensor g , namely F_k . Thus, the condition in Equation V.29 is a quantitative measure on how small the learning rate η is required to be.

C. Alternative methods

The main text discusses the brute-force linearization of the original differential equations. Alternatively, one could possibly make use of some transformations before linearization to simplify the linearization procedure, and even of transformations of a non-dissipative regime into a dissipative one. We discuss the following concrete example from the toy model discussed in Ref. [91] and restrict our analysis in the ODE limit. We consider a data-set with one-dimensional inputs. We write the network function with one hidden layer and linear activations as

$$f(x) = n^{-1/2} v^T u, \quad (\text{V.30})$$

where n is the width, and $u, v \in \mathbb{R}^n$ are the model parameters. We denote the loss function as $L = f^2/2$. The gradient descent equations are

$$\begin{aligned} u_{t+1} &= u_t - \eta n^{-1/2} f_t v_t, \\ v_{t+1} &= v_t - \eta n^{-1/2} f_t u_t, \end{aligned} \quad (\text{V.31})$$

for the t -th iteration. Taking the limit of $(u_{t+1} - u_t)/\eta \rightarrow \dot{u}$ and $(v_{t+1} - v_t)/\eta \rightarrow \dot{v}$, we obtain the ODE system

$$\begin{aligned} \dot{u} &= -n^{-1/2} f v = -n^{-1} (v^T u) v, \\ \dot{v} &= -n^{-1/2} f u = -n^{-1} (v^T u) v. \end{aligned} \quad (\text{V.32})$$

The gradient descent equations Equation (V.32) can be written in terms of the so-called *neural tangent kernel* (NTK). We denote the primary kernel eigenvalue as

$$\lambda = n^{-1} (\|u\|_2^2 + \|v\|_2^2). \quad (\text{V.33})$$

Then the gradient descent equations can then be expressed as

$$\begin{aligned} f_{t+1} &= f_t - \eta \lambda_t f_t + n^{-1} \eta^2 f_t^3, \\ \lambda_{t+1} &= \lambda_t - 4n^{-1} \eta f_t^2 + n^{-1} \eta^2 \lambda_t f_t^2. \end{aligned} \quad (\text{V.34})$$

Taking the limit of $(f_{t+1} - f_t)/\eta \rightarrow \dot{f}$ and $(\lambda_{t+1} - \lambda_t)/\eta \rightarrow \dot{\lambda}$, we obtain the ODE system

$$\begin{aligned} \dot{f} &= -\lambda f + n^{-1} \eta f^3, \\ \dot{\lambda} &= -4n^{-1} f^2 + n^{-1} \eta \lambda f^2. \end{aligned} \quad (\text{V.35})$$

We are now in the position to describe the quantum algorithm. Given an initial quantum state vector $u(0), v(0)$ proportional to $u(0), v(0) \in \mathbb{R}^{2n}$, we perform the quantum Carleman linearization algorithm as the main text for the non-linear ODE in Equation (V.32). One can also regard f as an independent variable to simplify the non-linearities in Equation (V.32). We observe that the NTK dynamics Equation (V.35) is dissipative. Hence, f and λ can be computed classically (i.e., regard f as a known time-varying parameter, and thus the quadratic Equation (V.32) is degenerated to a linear ODE), or quantumly (i.e., simulate the quantum state vector $|f(t)\rangle$ coupled with $|u(t)\rangle, |v(t)\rangle$). Another alternative is to simulate $f v$ and $f u$ in Equation (V.32) in a quantum fashion, such that Equation (V.32) is always linear for the multi-layer case. An advantage is that the dynamics of $f \dot{f}$ is also dissipative as

$$\begin{aligned} \frac{d}{dt}(f u) &= -\lambda(f u) + n^{-1} \eta f^3 u - n^{-1/2} f^2 v, \\ \frac{d}{dt}(f v) &= -\lambda(f v) + n^{-1} \eta f^3 v - n^{-1/2} f^2 u. \end{aligned} \quad (\text{V.36})$$

According to the results of Refs. [36, 58] and the main text, the dissipation in Equation (V.35), in particular the negative definite kernel $-\lambda$, is beneficial for the convergence of the linearization algorithm. In general, we expect the linearization is efficient whenever u, v is nearly located in the flat local minima of f . If that works, there is an exponential enhancement in n .

-
- [1] J. Preskill, “Lecture notes for physics 229: Quantum information and computation,” *California Institute of Technology* **16** (1998) 1–8.
 - [2] P. W. Shor, “Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer,” *SIAM Rev.* **41** (1999) 303–332. arXiv:quant-ph/9508027.
 - [3] L. K. Grover, “A fast quantum mechanical algorithm for database search,” in *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pp. 212–219. 1996. arXiv:quant-ph/9605043.
 - [4] S. Lloyd, “Universal quantum simulators,” *Science* **273** (1996) 1073–1078.
 - [5] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. Brandao, D. A. Buell, *et al.*, “Quantum supremacy using a programmable superconducting processor,” *Nature* **574** (2019) 505–510.
 - [6] 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, *et al.*, “Quantum computational advantage using photons,” *Science* **370** (2020) 1460–1463. arXiv:2012.01625.
 - [7] D. Hangleiter and J. Eisert, “Computational advantage of quantum random sampling,” 2022. arXiv:2206.04079.
 - [8] A. W. Harrow, A. Hassidim, and S. Lloyd, “Quantum algorithm for linear systems of equations,” *Phys. Rev. Lett.* **103** (2009) 150502. arXiv:0811.3171.
 - [9] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, “Quantum machine learning,” *Nature* **549** (2017) 195–202. arXiv:1611.09347.
 - [10] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, *et al.*, “Variational quantum algorithms,” *Nature Rev. Phys.* (2021) 1–20. arXiv:2012.09265.
 - [11] 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 Comm.* **5** (2014) 1–7. arXiv:1304.3061.
 - [12] J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, “The theory of variational hybrid quantum-classical algorithms,” *New J. Phys.* **18** (2016) 023023. arXiv:1509.04279.
 - [13] S. McArdle, S. Endo, A. Aspuru-Guzik, S. C. Benjamin, and X. Yuan, “Quantum computational chemistry,” *Rev. Mod. Phys.* **92** (2020) 015003. arXiv:1808.10402.
 - [14] E. Farhi, J. Goldstone, and S. Gutmann, “A quantum approximate optimization algorithm,” 2014. arXiv:1411.4028.
 - [15] S. Ebadi, A. Keesling, M. Cain, T. T. Wang, H. Levine, D. Bluvstein, G. Semeghini, A. Omran, J.-G. Liu, R. Samajdar, *et al.*, “Quantum optimization of maximum independent set using Rydberg atom arrays,” *Science* **376** (2022) 1209–1215. arXiv:2202.09372.
 - [16] E. Farhi and H. Neven, “Classification with quantum neural networks on near term processors,” 2018. arXiv:1802.06002.
 - [17] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, “Supervised learning with quantum-enhanced feature spaces,” *Nature* **567** (2019) 209–212. arXiv:1804.11326.
 - [18] D. Stilck Franca and R. García-Patrón, “Limitations of optimization algorithms on noisy quantum devices,” *Nature Phys.* **17** (2020) 1221.
 - [19] G. González-García, R. Trivedi, and J. I. Cirac, “Error propagation in NISQ devices for solving classical optimization problems,” 2022. arXiv:2203.15632.
 - [20] A. Deshpande, P. Niroula, O. Shtanko, A. V. Gorshkov, B. Fefferman, and M. J. Gullans, “Tight bounds on the convergence of noisy random circuits to the uniform distribution,” *PRX Quantum* **3** (2022) 040329.
 - [21] Y. Quek, D. S. França, S. Khatri, J. J. Meyer, and J. Eisert, “Exponentially tighter bounds on limitations of quantum error mitigation,” 2022. arXiv:2210.11505.
 - [22] J. Liu, F. Wilde, A. A. Mele, L. Jiang, and J. Eisert, “Noise can be helpful for variational quantum algorithms,” 2022. arXiv:2210.06723.
 - [23] S. Lloyd, M. Mohseni, and P. Rebentrost, “Quantum principal component analysis,” *Nature Phys.* **10** (2014) 631–633. arXiv:1307.0401.
 - [24] E. Tang, “Quantum principal component analysis only achieves an exponential speedup because of its state preparation assumptions,” *Phys. Rev. Lett.* **127** (2021) 060503.
 - [25] Y. Liu, S. Arunachalam, and K. Temme, “A rigorous and robust quantum speed-up in supervised machine learning,” *Nature Phys.* **17** (2021) 1–5. arXiv:2010.02174.
 - [26] 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 Comm.* **12** (2021) 2631. arXiv:2011.01938.
 - [27] H.-Y. Huang, R. Kueng, and J. Preskill, “Predicting many properties of a quantum system from very few measurements,” *Nature Phys.* **16** (2020) 1050–1057. arXiv:2002.08953.
 - [28] H.-Y. Huang, M. Broughton, J. Cotler, S. Chen, J. Li, M. Mohseni, H. Neven, R. Babbush, R. Kueng, J. Preskill, *et al.*, “Quantum advantage in learning from experiments,” *Science* **376** (2022) 1182–1186. arXiv:2112.00778.
 - [29] R. Sweke, J.-P. Seifert, D. Hangleiter, and J. Eisert, “On the quantum versus classical learnability of discrete distributions,” *Quantum* **5** (2021) 417. arXiv:2007.14451.
 - [30] N. Pirnay, R. Sweke, J. Eisert, and J.-P. Seifert, “A super-polynomial quantum-classical separation for density modelling,” *Phys. Rev. A* **107** (2023) 042416. arXiv:2210.14936.
 - [31] H.-Y. Huang, R. Kueng, and J. Preskill, “Information-theoretic bounds on quantum advantage in machine learning,” *Phys. Rev. Lett.* **126** (2021) 190505. arXiv:2101.02464.
 - [32] Similar situations exist for the classical machine learning community as well. In the history of deep learning [99], some of the most important inventions, like ResNet [57], transformers [100], large language models [34], have first been discovered by large-scale numerical experiments instead of by statistical learning theory. In large-scale machine learning tasks, it is often hard to theoretically investigate the detailed behavior of the model, while heuristic theories and numerical experiments might provide important guidance

[101].

- [33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2018. arXiv:1810.04805.
- [34] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, “Language models are few-shot learners,” *Adv. Neur. Inf. Process Sys.* **33** (2020) 1877–1901. arXiv:2005.14165.
- [35] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, *et al.*, “Palm: Scaling language modeling with pathways,” 2022. arXiv:2204.02311.
- [36] J.-P. Liu, H. Ø. Kolden, H. K. Krovi, N. F. Loureiro, K. Trivisa, and A. M. Childs, “Efficient quantum algorithm for dissipative nonlinear differential equations,” *Proc. Natl. Ac. Sc.* **118** (2021) e2026805118. arXiv:2011.03185.
- [37] V. Giovannetti, S. Lloyd, and L. Maccone, “Quantum random access memory,” *Phys. Rev. Lett.* **100** (2008) 160501. arXiv:0708.1879.
- [38] T. Hoeffler, D. Alistarh, T. Ben-Nun, N. Dryden, and A. Peste, “Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks,” *J. Mach. Learn. Res.* **22** (2021) 1–124. arXiv:2102.00554.
- [39] N. Lee, T. Ajanthan, and P. Torr, “Snip: Single-shot network pruning based on connection sensitivity,” in *International Conference on Learning Representations*. 2019.
- [40] C. Wang, G. Zhang, and R. Grosse, “Picking winning tickets before training by preserving gradient flow,” in *International Conference on Learning Representations*. 2020.
- [41] H. Tanaka, D. Kunin, D. L. Yamins, and S. Ganguli, “Pruning neural networks without any data by iteratively conserving synaptic flow,” *Adv. Neur. Inf. Process Sys.* **33** (2020) 6377–6389. arXiv:2006.05467.
- [42] C. T. Hann, C.-L. Zou, Y. Zhang, Y. Chu, R. J. Schoelkopf, S. M. Girvin, and L. Jiang, “Hardware-efficient quantum random access memory with hybrid quantum acoustic systems,” *Phys. Rev. Lett.* **123** (2019) 250501.
- [43] O. Di Matteo, V. Gheorghiu, and M. Mosca, “Fault-tolerant resource estimation of quantum random-access memories,” *IEEE Tras. Quant. Eng.* **1** (2020) 1–13. arXiv:1902.01329.
- [44] K. Roose, “The brilliance and weirdness of ChatGPT,” *The New York Times* (2022) .
- [45] A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. Ramasesh, A. Slone, C. Anil, I. Schlag, T. Gutman-Solo, *et al.*, “Solving quantitative reasoning problems with language models,” 2022. arXiv:2206.14858.
- [46] K. Johnson, “OpenAI debuts DALL-E for generating images from text,” *VentureBeat* (2021) .
- [47] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean, “Carbon emissions and large neural network training,” 2021. arXiv:2104.10350.
- [48] U. Evci, F. Pedregosa, A. Gomez, and E. Elsen, “The difficulty of training sparse neural networks,” in *Identifying and Understanding Deep Learning Phenomena Workshop, International Conference on Machine Learning*. 2019. arXiv:1906.10732.
- [49] J. Frankle, G. K. Dziugaite, D. Roy, and M. Carbin, “Pruning neural networks at initialization: Why are we missing the mark?,” in *International Conference on Learning Representations*. 2021.
- [50] J. Frankle and M. Carbin, “The lottery ticket hypothesis: Finding sparse, trainable neural networks,” in *International Conference on Learning Representations*. 2019.
- [51] H. You, C. Li, P. Xu, Y. Fu, Y. Wang, X. Chen, R. G. Baraniuk, Z. Wang, and Y. Lin, “Drawing early-bird tickets: Towards more efficient training of deep networks,” 2019. arXiv:1909.11957.
- [52] Z. Ye, “Generalization and memorization in sparse neural networks.”
<https://github.com/ZIYU-DEEP/Generalization-and-Memorization-in-Sparse-Training>.
- [53] R. Shwartz-Ziv and N. Tishby, “Opening the black box of deep neural networks via information,” 2017. arXiv:1703.00810.
- [54] A. Achille, G. Paolini, and S. Soatto, “Where is the information in a deep neural network?,” 2019. arXiv:1905.12213.
- [55] N. Gleinig and T. Hoeffler, “An efficient algorithm for sparse quantum state preparation,” in *2021 58th ACM/IEEE Design Automation Conference (DAC)*, pp. 433–438, IEEE. 2021.
- [56] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034. 2015. arXiv:1502.01852.
- [57] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778. 2016.
- [58] D. An, D. Fang, S. Jordan, J.-P. Liu, G. H. Low, and J. Wang, “Efficient quantum algorithm for nonlinear reaction-diffusion equations and energy estimation,” arXiv:2205.01141.
- [59] H. Krovi, “Improved quantum algorithms for linear and nonlinear differential equations,” 2022. arXiv:2202.01054.
- [60] A. Surana, A. Gnanasekaran, and T. Sahai, “Carleman linearization based efficient quantum algorithm for higher order polynomial differential equations,” 2022. arXiv:2212.10775.
- [61] I. Joseph, “Koopman-von Neumann approach to quantum simulation of nonlinear classical dynamics,” *Phys. Rev. Res.* **2** (2020) 043102. arXiv:2003.09980.
- [62] I. Y. Dodin and E. A. Startsev, “On applications of quantum computing to plasma simulations,” *Phys. Plasmas* **28** (2021) 092101. arXiv:2005.14369.
- [63] S. Lloyd, G. De Palma, C. Gokler, B. Kiani, Z.-W. Liu, M. Marvian, F. Tennie, and T. Palmer, “Quantum algorithm for nonlinear differential equations,” 2020. arXiv:2011.06571.
- [64] A. Engel, G. Smith, and S. E. Parker, “Linear embedding of nonlinear dynamical systems and prospects for efficient quantum algorithms,” *Phys. Plasmas* **28** (2021) 062305. arXiv:2012.06681.
- [65] C. Xue, Y.-C. Wu, and G.-P. Guo, “Quantum homotopy perturbation method for nonlinear dissipative ordinary differential equations,” *New J. Phys.* **23** (2021) 123035. arXiv:2111.07486.
- [66] Y. T. Lin, R. B. Lowrie, D. Aslangil, Y. Subaşı, and A. T. Sornborger, “Koopman von Neumann mechanics and the Koopman representation: A perspective on solving nonlinear dynamical systems with quantum computers,” 2022. arXiv:2202.02188.

- [67] S. Jin and N. Liu, “Quantum algorithms for computing observables of nonlinear partial differential equations,” 2022. arXiv:2202.07834.
- [68] S. Jin, N. Liu, and Y. Yu, “Time complexity analysis of quantum algorithms via linear representations for nonlinear ordinary and partial differential equations,” 2022. arXiv:2209.08478.
- [69] S. Jin, N. Liu, and Y. Yu, “Quantum simulation of partial differential equations via schrodingerisation,” 2022. arXiv:2212.13969.
- [70] A. M. Childs, R. Kothari, and R. D. Somma, “Quantum algorithm for systems of linear equations with exponentially improved dependence on precision,” *SIAM J. Comp.* **46** (2017) 1920–1950. arXiv:1511.02306.
- [71] A. Gilyén, Y. Su, G. H. Low, and N. Wiebe, “Quantum singular value transformation and beyond: exponential improvements for quantum matrix arithmetics,” in *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pp. 193–204. 2019. arXiv:1806.01838.
- [72] Y. Subaşı, R. D. Somma, and D. Orsucci, “Quantum algorithms for systems of linear equations inspired by adiabatic quantum computing,” *Phys. Rev. Lett.* **122** (2019) 060504. arXiv:1805.10549.
- [73] D. An and L. Lin, “Quantum linear system solver based on time-optimal adiabatic quantum computing and quantum approximate optimization algorithm,” *ACM Trans. Quant. Comp.* **3** (2022) 1–28. arXiv:1909.05500.
- [74] L. Lin and Y. Tong, “Optimal polynomial based quantum eigenstate filtering with application to solving quantum linear systems,” *Quantum* **4** (2020) 361. arXiv:1910.14596.
- [75] P. Costa, D. An, Y. R. Sanders, Y. Su, R. Babbush, and D. W. Berry, “Optimal scaling quantum linear systems solver via discrete adiabatic theorem,” 2021. arXiv:2111.08152.
- [76] D. W. Berry, “High-order quantum algorithm for solving linear differential equations,” *J. Phys. A* **47** (2014) 105301. arXiv:1010.2745.
- [77] D. W. Berry, A. M. Childs, A. Ostrander, and G. Wang, “Quantum algorithm for linear differential equations with exponentially improved dependence on precision,” *Commun. Math. Phys.* **356** (2017) 1057–1081. arXiv:1701.03684.
- [78] A. M. Childs and J.-P. Liu, “Quantum spectral methods for differential equations,” *Commun. Math. Phys.* **375** (2020) 1427–1457. arXiv:1911.00961.
- [79] A. M. Childs, J.-P. Liu, and A. Ostrander, “High-precision quantum algorithms for partial differential equations,” *Quantum* **5** (2021) 574. arXiv:2002.07868.
- [80] D. An, J.-P. Liu, D. Wang, and Q. Zhao, “A theory of quantum differential equation solvers: limitations and fast-forwarding,” 2022. arXiv:2211.05246.
- [81] D. A. Roberts and B. Yoshida, “Chaos and complexity by design,” *JHEP* **04** (2017) 121.
- [82] D. N. Page, “Average entropy of a subsystem,” *Phys. Rev. Lett.* **71** (1993) 1291–1294.
- [83] J. Cotler, N. Hunter-Jones, J. Liu, and B. Yoshida, “Chaos, complexity, and random matrices,” *JHEP* **11** (2017) 048.
- [84] M. Mitzenmacher and E. Upfal, *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge University Press, 2017.
- [85] We specifically thank Robert Huang for his clarifications.
- [86] M. Guţă, J. Kahn, R. Kueng, and J. A. Tropp, “Fast state tomography with optimal error bounds,” *J. Phys. A* **53** (2020) 204001. arXiv:1809.11162.
- [87] D. A. Roberts, S. Yaida, and B. Hanin, “The principles of deep learning theory,”. arXiv:2106.10165.
- [88] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778. 2016. arXiv:1512.03385.
- [89] A. Krizhevsky, V. Nair, and G. Hinton, “CIFAR-100 (Canadian Institute for Advanced Research),” (2009) .
- [90] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Adv. Neur. Inf. Proc. Sys.* **33** (2020) 6840–6851. arXiv:2006.11239.
- [91] A. Lewkowycz, Y. Bahri, E. Dyer, J. Sohl-Dickstein, and G. Gur-Ari, “The large learning rate phase of deep learning: the catapult mechanism,” 2020. arXiv:2003.02218.
- [92] D. Meltzer and J. Liu, “Catapult dynamics and phase transitions in quadratic nets,” *arXiv preprint arXiv:2301.07737* (2023) .
- [93] S. Aaronson, “Read the fine print,” *Nature Phys.* **11** (2015) 291–293.
- [94] C. T. Hann, G. Lee, S. Girvin, and L. Jiang, “Resilience of quantum random access memory to generic noise,” *PRX Quant.* **2** (2021) 020311. arXiv:2012.05340.
- [95] K. C. Chen, W. Dai, C. Errando-Herranz, S. Lloyd, and D. Englund, “Scalable and high-fidelity quantum random access memory in spin-photon networks,” *PRX Quantum* **2** (2021) 030319. arXiv:2103.07623.
- [96] L. Bugalho, E. Z. Cruzeiro, K. C. Chen, W. Dai, D. Englund, and Y. Omar, “Resource-efficient simulation of noisy quantum circuits and application to network-enabled QRAM optimization,” 2022. arXiv:2210.13494.
- [97] K. G. H. Vollbrecht and J. I. Cirac, “Quantum simulators, continuous-time automata, and translationally invariant systems,” *Phys. Rev. Lett.* **100** (2008) 010501.
- [98] E. Tang, “Quantum-inspired classical algorithms for principal component analysis and supervised clustering,”. arXiv:1811.00414.
- [99] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature* **521** (2015) 436–444.
- [100] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Adv. Neur. Inf. Proc. Sys.* **30** (2017) . arXiv:1706.03762.
- [101] M. Schuld and N. Killoran, “Is quantum advantage the right goal for quantum machine learning?,” 2022. arXiv:2203.01340.