Supplemental Materials: Quantum Continual Learning Overcoming Catastrophic Forgetting

Wenjie Jiang,[1] Zhide Lu,[1] and Dong-Ling Deng[1,2]

[1]*Center for Quantum Information, IIIS, Tsinghua University, Beijing 100084, People's Republic of China*
[2]*Shanghai Qi Zhi Institute, 41th Floor, AI Tower, No. 701 Yunjin Road, Xuhui District, Shanghai 200232, China*

I. THE SETTING

Our numerical simulations are based on the open source package Yao.jl [1]. To illustrate the catastrophic forgetting phenomena, we randomly initialize an eight-qubit variational quantum circuit (as shown in Fig. S1) as the ansatz for our quantum classifier, in which those rotation angles are variational parameters updated in the training process and unchanged in the inference process, and the CNOT gates is necessary to entangle all qubits since entanglement in quantum circuits is a key resource for potential quantum advantages. This variational architecture is hardware-efficient [2] and is capable to achieve satisfactory performances for our classification tasks (see Fig. S2). Besides, this architecture does not take advantages of the specific structure information of datasets. We remark that the EWC method does not rely on the specific variational ansatz chosen in this manuscript. In principle, the same strategy can be utilized for other variational models after corresponding modifications, such as quantum conventional neural networks [3], recurrent quantum neural networks [4], etc., as long as there exist common solutions for the sequentially coming tasks. Here, we choose this hardware-efficient for several conveniences. First, this ansatz is universal. That is to say, it can represent any unitary to any precision in the limit of long depth, which means that theoretically we can find common solutions for the tasks learned continually. Second, this ansatz is hardware-efficient and experiment-friendly [2]. So, it is possible to accomplish quantum continual learning via the EWC method with this ansatz in the future. Third, this

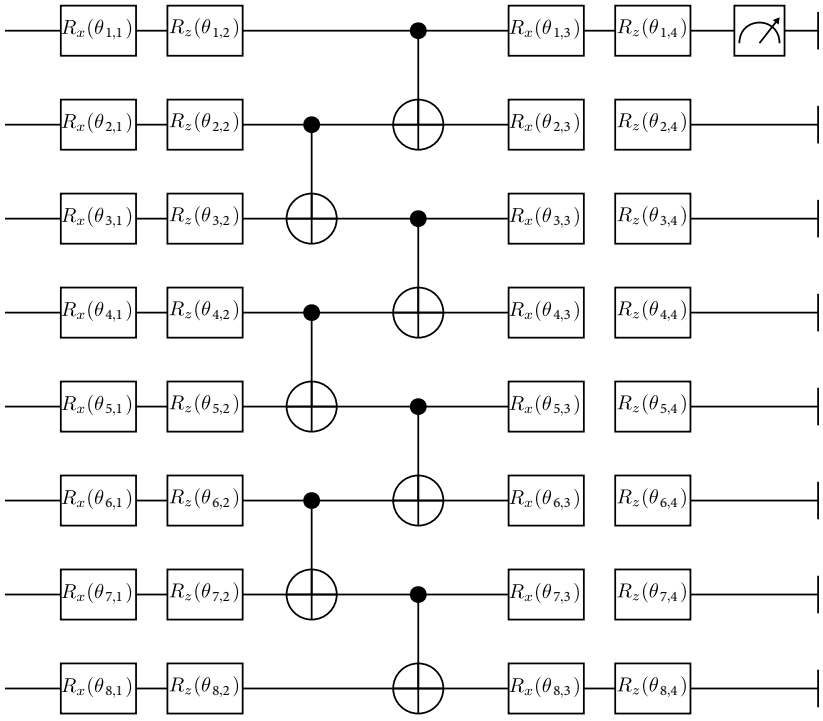

FIG. S1. A single layer of the variational ansatz of our quantum classifier. This is a single layer of the ansatz. All single qubit gates in this ansatz are rotation gates ($R_x(\theta) = e^{i\theta/2X}$ and $R_z(\theta) = e^{i\theta/2Z}$). Those rotation angles $\theta_{i,j}$ (i indicates the i-th qubit and j indicates the j-th parameter of this qubit) in the rotation gates are the variational parameters of the quantum classifier and will be updated during the training process. Those CNOT gates are adapted here to introduce necessary entanglement among different qubits. We measure the first qubit and treat its output as the classification result of this quantum classifier. Our quantum classifier used in numerical simulations consists of ten repeated layers.

ansatz contains only unitary gates in the bulk layers and the measurement gates in the last layer, which provides conveniences for both the numerical simulation and the experimental implementation. Despite those conveniences, there are also some caveats for this architecture. One of them is that, this ansatz does not take advantages of the information for the specific tasks. For some specific tasks, other carefully-designed architectures may achieve better performances. Besides, this ansatz may encounter the trainability problems when the depth of the actual circuit grows, such as the barren plateau problem. Those problems can make the training process very challenging, even the solutions exist in principle. In our simulations, the size of the learning tasks is limit. So, we can relatively easily simulate the learning process and the inference process using this vanilla ansatz. When the size of the tasks grow, those caveats discussed above should be seriously considered.

All data encountered in our numerical simulations consists of 256 features and can be represented by eight qubits using amplitude encoding. For the original MNIST hand-written digit images, those 28×28-pixel images [5] are reduced to 16×16-pixel images (see Fig. S2(a)), so that we can simulate this quantum learning process with moderate classical computational resources. Then, we randomly choose a permutation of the 256 pixels and apply it for all images, which produces a new dataset consisting of pixel-permuted images (see Fig. S2(b)). For time-of-flight (TOF) images, we diagonalize the Hamiltonian of quantum anomalous Hall effect with an open boundary condition and calculate the atomic density distributions with different spin bases for the lower band in momentum space to obtain input data. We vary the strength of the spin-orbit coupling and the strength of the on-site Zeeman interaction in both the topological and topologically trivial regions to generate several thousand data samples (see Fig. S2(c)). For the symmetry protected topological state (SPT), we consider the model involving eight spins and exactly diagonalize its Hamiltonian to obtain the ground state which can be naturally represented using eight qubits (see Fig. S2(d)). In this work, we use amplitude encoding to convert the data of our classification tasks into the input quantum states for the quantum classifier.

The process of sequential learning is divided into different phases and our quantum classifier are trained with only one specific dataset in each training phase. For example, to illustrate the catastrophic forgetting phenomena, we first use the randomly initialized quantum classifier to learn to classify original MNIST images. After a satisfactory performance is obtained, this classifier are trained to distinguish permuted MINIST images. The results of different learning phases are shown in the main text, where the forgetting phenomena is revealed. As for continual learning via EWC method, the Fisher information matrix for each task is computed after the corresponding training phases and is stored for those following training phases.

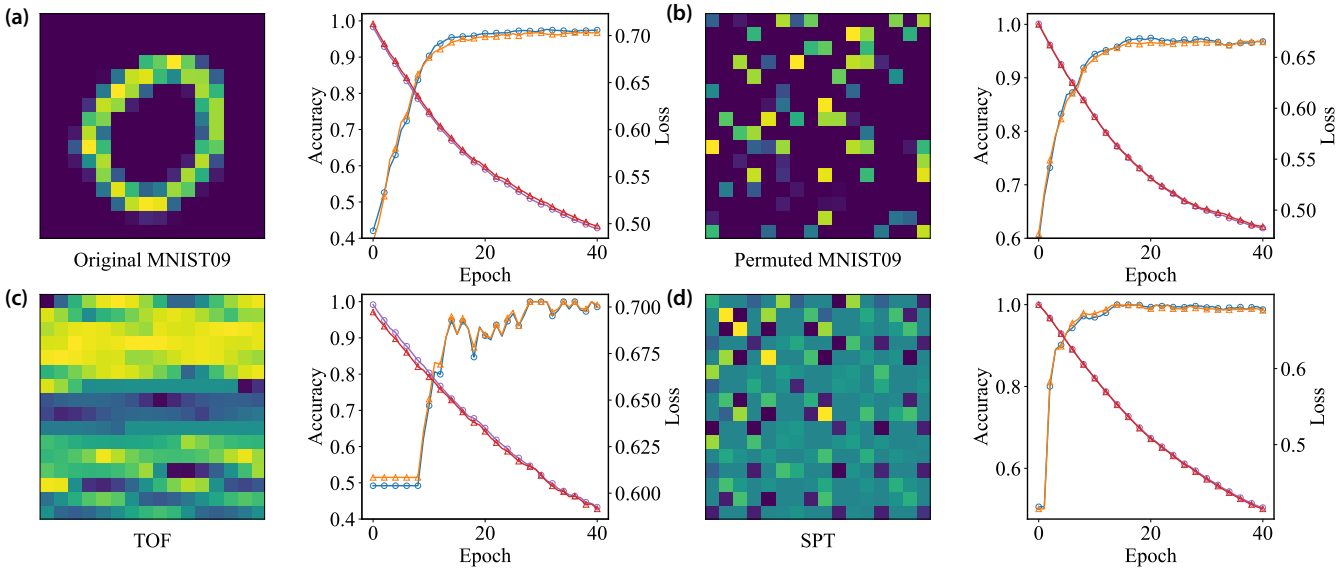

FIG. S2. Results of learning single task. Here we show the classification performances of our quantum classifier on tasks used in our simulations of quantum continual learning and we also plot a sample image of each task: (a) the original MNIST image and its learning performance; (b) the permuted MNIST image and its learning performance; (c) the time-of-flight image and its learning performance; (d) the symmetry protected topological state and its learning performance.

II. ELASTIC WEIGHT CONSOLIDATION

From a high-level perspective, overcoming catastrophic forgetting in quantum continual learning requires protecting the learned knowledge of those previous tasks, as well as learning the new-coming knowledge of following tasks [6, 7]. So our quantum learning model should have enough capacity to store those information. Besides, appropriate management of model's capacity is required to achieve quantum continual learning in practice. EWC method offers a practical method to do the capacity management: it estimates the necessary capacity for previous tasks and refreshes the rest part which contains rare information about those previously trained tasks. To do this, EWC method evaluates the importance of each variational parameter in the quantum classifier and only allows significant twist for those relatively unimportant ones.

We then give a detailed mathematical derivation of EWC method. For simplicity, we concern the two-task scenario here and use the similar philosophy to explicitly write down the result for the multi-task scenario. From the perspective of maximum likelihood estimation [8], we explore all possibilities of parameters $\boldsymbol{\theta}$ of the quantum classifier to maximize the likelihood function $p(\boldsymbol{\theta}|\Sigma)$, where $\Sigma = \Sigma_A + \Sigma_B$ is the total dataset (Σ_A and Σ_B are datasets for task A and task B respectively and we assume that these two tasks are independent to each other). So we have expression

$$\begin{aligned}\log p(\boldsymbol{\theta}|\Sigma) &= \log\left(\frac{p(\Sigma_B|\Sigma_A,\boldsymbol{\theta})p(\Sigma_A,\boldsymbol{\theta})}{p(\Sigma_A,\Sigma_B)}\right) = \log\left(p(\Sigma_B|\boldsymbol{\theta})\cdot\frac{p(\boldsymbol{\theta},\Sigma_A)}{p(\Sigma_A)}\cdot\frac{1}{p(\Sigma_B)}\right)\\ &= \log p(\Sigma_B|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}|\Sigma_A) - \log p(\Sigma_B),\end{aligned}$$

where the first and third equation use the Bayes' rule and the second equation uses the independence condition. As shown in the main text, we have Taylor Series for the second term:

$$\log p(\boldsymbol{\theta}|\Sigma_A) = \log p(\boldsymbol{\theta}_A^*|\Sigma_A) + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_A^*)^T H_{\boldsymbol{\theta}_A^*}(\boldsymbol{\theta}-\boldsymbol{\theta}_A^*).$$

It is worthwhile to mention that from the perspective of parameter estimation [9], this treatment means that we sample parameters from a multivariate normal distribution:

$$p(\boldsymbol{\theta}|\Sigma_A) \propto \mathcal{N}\left(\boldsymbol{\theta}_A^*, H_{\boldsymbol{\theta}_A^*}^{-1}\right), \tag{S1}$$

where the optimal solution $\boldsymbol{\theta}_A^*$ for task A is the mean value of this normal distribution and $H_{\boldsymbol{\theta}_A^*}^{-1}$ is the precision matrix ($\left(H_{\boldsymbol{\theta}_A^*}\right)_{ij} = \frac{\partial}{\partial\theta_i\partial\theta_j}\log p(\boldsymbol{\theta}|\Sigma_A)|_{\boldsymbol{\theta}_A^*}$ is the Hessian matrix at the optimal solution $\boldsymbol{\theta}_A^*$ for task A and is equal to the minus of the Fisher information matrix F under some specific conditions [10]). We can rewrite the quadratic term using the Fisher information matrix and absorb it into the likelihood function of sequential tasks. This leads to the loss function for the second task in our scenario:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}_B(\boldsymbol{\theta}) + \lambda\cdot(\boldsymbol{\theta}-\boldsymbol{\theta}_A^*)^T F_{\boldsymbol{\theta}_A^*}(\boldsymbol{\theta}-\boldsymbol{\theta}_A^*).$$

To reduce the potential storage and computation overhead for those possible large quantum models, we use the diagonal elements of the Fisher matrix as the weights of variational parameters and neglect those off-diagonal entries, which will be discussed later. Thus, we could add the regularization term shown in the main text to the loss function of the second task in order to maximize the likelihood function of joint tasks.

For continual learning of more than two tasks, we can compute the regularization term for each trained task and add them together to overcome catastrophic forgetting:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}_0(\boldsymbol{\theta}) + \lambda_A\sum_i F_i^{(A)}\cdot\left(\theta_i-\theta_{A,i}^*\right)^2 + \lambda_{A,B}\sum_i F_i^{(A,B)}\cdot\left(\theta_i-\theta_{A,B,i}^*\right)^2 + ...,$$

where $\mathcal{L}_0(\boldsymbol{\theta})$ is the original loss function for current task given current parameters $\boldsymbol{\theta}$, $F_i^{(A)}$ is the i-th diagonal element of the Fisher information matrix at the optimal point $\boldsymbol{\theta}_A^*$ for previous task A, λ_A is a hyper-parameter controlling the strength of this EWC restriction and so on.

Here are some discussions about the independence assumption in the theoretical analysis above. In the main text, we adapt the EWC method to continually learn the original MNIST dataset and the permuted MNIST dataset, which share the similar underlying structure. As described before, to generate the permuted MNIST dataset, we randomly choose a permutation of pixels and apply it on every images in the original MNIST dataset. This might introduce inner correlations into the joint dataset, and violate the independence assumption. Nevertheless, from the numerical simulation, we find that the EWC method also functions well in this case. We can interpret this result from the following discussion. The possibly chosen permutations are exponentially many, and the information that the permuted dataset carries about the original MNIST dataset is exponentially rare. This means

that the independence assumption is only slightly violated, and the theoretical analysis is nearly valid in practice. Besides, the independence assumption is a simplification in theoretical analysis. From a more general perspective, the dependance among different datasets can even be beneficial for continually learning them. Imagine that we need to find a public local minimum in the loss landscape for several tasks. It is a natural conjecture that the more similar those tasks are, the more possible public solutions there will be. In those cases, dependance among different tasks provides advantages for continual learning. We remark that things can be more interesting in quantum continual learning scenario. Despite those possible advantages coming from classical statistical correlations among datasets, the quantum entanglement can also be introduced into quantum datasets. The benefits of quantum entanglement in quantum machine learning is discovered in [11], and we can conjecture that it can benefit the quantum continual learning as well.

III. REASONS FOR NEGLECTING OFF-DIAGONAL ELEMENTS

In our numerical simulations, the quantum classifier consists of 248 variational parameters, in which computing and storing the full Fisher matrix is not very hard. Nevertheless, if the number of parameters gets larger and larger to match the exponentially growing dimensionality of the Hilbert space, computing and storing its full Fisher matrix can be quite challenging. From a more practical perspective, we use the diagonal elements of the Fisher matrix which can be estimated by the first order derivative [12]:

$$\left(F_{\boldsymbol{\theta}_A^*}\right)_{ij} = \mathbb{E}\left(-\left.\frac{\partial^2 L}{\partial\theta_i\partial\theta_j}\right|_{\boldsymbol{\theta}_A^*}\right) = \mathbb{E}\left(\left.\left(\frac{\partial L}{\partial\theta_i}\right)\left(\frac{\partial L}{\partial\theta_j}\right)^T\right|_{\boldsymbol{\theta}_A^*}\right) \tag{S2}$$

To compare the learning result of using the diagonal elements of the Fisher matrix and that of using the full Fisher matrix, we train our quantum classifier using the original MNIST images and the permuted MNIST images sequentially. In this simulation, the diagonal elements of the Fisher matrix and the full Fisher matrix are adapted as the metric to quantify the derivative distance in the parameter space respectively. The results in Fig. S3 shows that the performances of both metric choices are at the same level. We remark here that in consideration of the summation of those off-diagonal elements, we manually lower down the strength parameter λ in the simulation of using the full Fisher matrix. The similar performances between those two learning scenarios indicate that neglecting those off-diagonal elements in the Fisher matrix has no significant influence on the results of quantum continual learning. Thus, we use the diagonal elements as our distance metric in all other numerical simulations.

IV. MORE NUMERICAL RESULTS

In this section, we give more results of continual learning. Performances of learning single tasks are shown in Fig. S2 and one sample image of each dataset is plotted. Those results indicate that our quantum classifier is capable to achieve satisfactory performances on those chosen classification tasks.

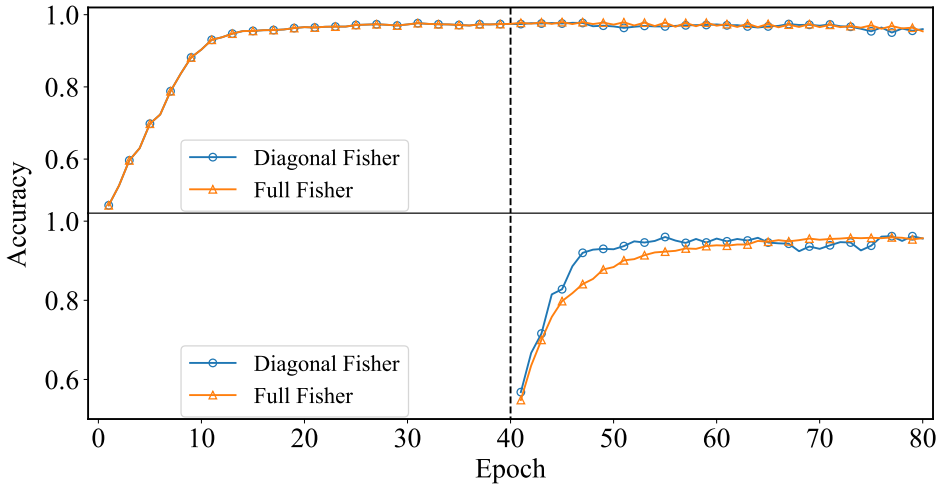

FIG. S3. Comparison between the learning result of using diagonal elements of the Fisher matrix and that of using the full Fisher matrix. We train our quantum classifier using the original MNIST dataset and then adapt two kinds of regularization terms to train this classifier using the new-coming permuted MNIST dataset. The learning settings for both cases are exactly the same except the strength parameter λ.

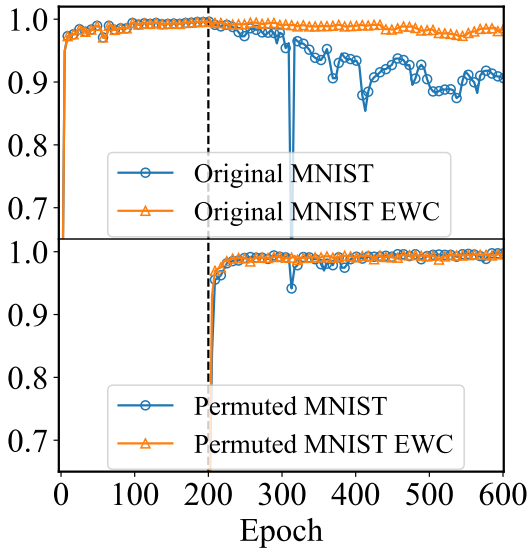

FIG. S4. Illustration of continual learning using a classical neural network. The learning model is the fully-connected neural network with a hidden layer containing 16 hidden neurons, and the activation function is the Rectified Linear Unit (ReLU) function.

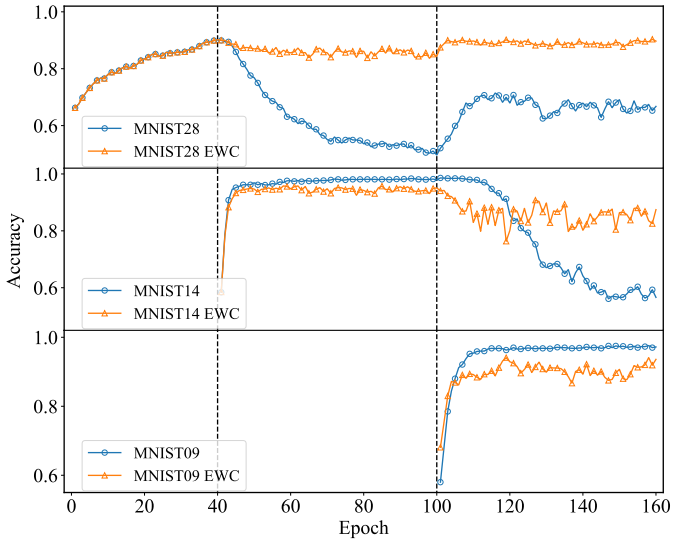

FIG. S5. Illustration of quantum continual learning of classifying different MNIST images. Learning curves of three related tasks: classifying digit 2 and digit 8, classifying digit 1 and digit 4, and classifying digit 0 and digit 9.

In the main text, we show results of quantum continual learning. For completeness, we also use classical neural networks to learn two classical tasks (classifying the original MNIST images and the permuted MNIST images, see Fig. S4). In this simulation, we use a fully-connected neural network to continually learn the MNIST images of digit 0 and 9 and their randomly permuted images via the EWC method. The comparison between those similar results shows that even for the totally classical tasks, the variational circuit based quantum classier can achieve comparable performance in the terms of continual learning.

In the main text, we also show that quantum continual learning of two-task case can be accomplished when those two problems are similar or dissimilar to each other. As a complementary example, we also simulate the quantum continual learning of two related problems. We use MNIST images of different digits to construct several classification tasks and find that the continual learning of this kind of tasks can also be accomplished (see Fig. S5).

We group MNIST hand-written images of different digits to construct several binary classification tasks and use them to train our quantum classifier. For multi-task cases, we choose three pairs of digits and use our quantum classifier to classify their hand-written images. We first train our quantum classifier using images of digit 2 and images of digit 8, which ends with a high classification performance ($> 90\%$). Then, we train this quantum classifier to identify digit 1 and digit 4. In the favor of EWC

method, our quantum classifier behaves reasonably well at both tasks after the second training phase. Sequentially, we train this circuit to classify digit 0 and digit 9, and find that our quantum classifier can perform relatively well in all three different classification tasks after those training processes.

We also notice that in the continual learning scenario, the performance of our quantum classifier on each task has a slight reduction compared with that in the single task learning scenario. Intuitively, this is caused by an inevitable small deviation from the optimal solution of a single task to the optimal solution of the joint task.

---

[1] X.-Z. Luo, J.-G. Liu, P. Zhang, and L. Wang, Quantum **4**, 341 (2020).
[2] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, Nature **549**, 242 (2017).
[3] I. Cong, S. Choi, and M. D. Lukin, Nat. Phys. **15**, 1273 (2019).
[4] J. Bausch, in *Advances in Neural Information Processing Systems*, Vol. 33 (Curran Associates, Inc., 2020) pp. 1368–1379.
[5] Y. LeCun, C. Cortes, and C. Burges, MNIST handwritten digit database (1998).
[6] D. Rao, F. Visin, A. A. Rusu, Y. W. Teh, R. Pascanu, and R. Hadsell, arXiv:1910.14481 (2019).
[7] R. Aljundi, arXiv:1910.02718 (2019).
[8] W. A. Scott, J. Stat. Comput. Simul. **72**, 599 (2002).
[9] Precise and accurate estimation, in *Parameter Estimation for Scientists and Engineers* (John Wiley, Sons, Ltd, 2007) Chap. 5, pp. 99–162.
[10] A. Ly, M. Marsman, J. Verhagen, R. Grasman, and E.-J. Wagenmakers, arXiv:1705.01064 (2017).
[11] K. Sharma, M. Cerezo, Z. Holmes, L. Cincio, A. Sornborger, and P. J. Coles, Phys. Rev. Lett. **128**, 070501 (2022).
[12] Fundamentals of statistical signal processing, in *Blind Equalization and System Identification: Batch Processing Algorithms, Performance and Applications* (Springer London, London, 2006) pp. 83–182.