Reinforcement Learning with Neural Networks for Quantum Multiple Hypothesis Testing

Sarah Brandsen*, Kevin D. Stubbs‡, and Henry D. Pfister†‡
*Department of Physics, Duke University, Durham, North Carolina 27708, USA
†Department of Electrical and Computer Engineering, Duke University, Durham, North Carolina 27708, USA
‡Department of Mathematics, Duke University, Durham, North Carolina 27708, USA
Email: {sarah.brandsen, kevin.stubbs, henry.pfister}@duke.edu

***Abstract*—Reinforcement learning with neural networks (RLNN) has recently demonstrated great promise for many problems, including some problems in quantum information theory. In this work, we apply reinforcement learning to quantum hypothesis testing, where one designs measurements that can distinguish between multiple quantum states $\{\rho_j\}|_{j=1}^m$ while minimizing the error probability. Although the Helstrom measurement is known to be optimal when there are $m = 2$ states, the general problem of finding a minimal-error measurement is challenging. Additionally, in the case where the candidate states correspond to a quantum system with many qubit subsystems, implementing the optimal measurement on the entire system may be impractical. In this work, we develop locally-adaptive measurement strategies that are experimentally feasible in the sense that only one quantum subsystem is measured in each round. RLNN is used to find the optimal measurement protocol for arbitrary sets of tensor product quantum states. Numerical results for the network performance are shown. In special cases, the neural network testing-policy achieves the same probability of success as the optimal collective measurement.**

***Index Terms*—quantum state discrimination, LOCC, neural networks, reinforcement leraning, quantum hypothesis testing, min-entropy**

I. Introduction

The task of quantum hypothesis testing consists of finding the optimal quantum measurement $\{\Pi_j\}|_{j=1}^m$ to distinguish between m candidate states $\{\rho_j\}_{j=1}^m$ with prior probabilities $\{q_j\}_{j=1}^m$. This can be applied to discrimination between quantum coherent states [1] and also to the task of decoding one of m codewords that has been sent through a known noisy quantum channel [2], [3]. One important example of locally adaptive multiple hypothesis testing protocols is the Dolinar receiver, which uses an adaptive measurement scheme to distinguish between m optical signals [4].

In the special case where $m = 2$, the optimal measurement is called the Helstrom measurement [5] and is given by the eigenvectors of the matrix $M \triangleq q_1\rho_1 - q_2\rho_2$. For the general case, the optimal measurement is challenging to compute. Although an analytic solution for the optimal measurement is not known in general, semidefinite programming techniques have been used to approach the problem of finding the minimal-error measurement and computing the optimal success probability [6], [7].

When the candidate states are high-dimensional (corresponding to a quantum system composed of many qubit subsystems), it may be experimentally difficult to implement operations on all subsystems at once. Thus, we also focus on finding optimal (or near-optimal) approaches that include the experimentally desirable property of locality, where only a single subsystem is measured in each round. We know that dynamic programming can be used to recursively find the optimal local approach [8]. However, even in the simplest case where $m = 2$, the complexity grows like $O(2^n nQ)$, where n is the number of qubit subsystems and Q is the number of different local measurements considered [9].

A powerful alternative tool for developing optimal adaptive protocols is reinforcement learning with neural networks (RLNN). In this process, an agent learns an optimized protocol through repeated interaction with an environment. RLNN has been successfully applied to problems in quantum information theory such as generating error-correcting sequences [10]. While RLNN was introduced more than 20 years ago [11], [12], interest in these methods was recently rekindled by its remarkable success for atari games [13], [14]. This motivates our use of RLNN to find optimal locally-adaptive measurement protocols.

We provide preliminary results for the network performance and demonstrate that the network can achieve optimal performance in some special cases. Additionally, we prove by counterexample that, in contrast to the binary case [9], [15], locally-adaptive greedy strategies are not necessarily optimal when all candidate states are pure. While we demonstrate that the neural network is able to find a better local strategy, we also conjecture that the gap between the optimal local strategy and the optimal collective strategy persists.

Next, we compare the neural network performance to the collective "Pretty Good Measurement" (PGM) [16] for general tensor product quantum states, and demonstrate that, for a large number of candidate states ($m = 5$), the neural network exceeds the PGM success probability in most trials. Use of the PGM as a benchmark is motivated by its optimality in several important cases, including sufficiently symmetric candidate states [17]–[19], and its relation to a least-squares approximation [20]. Finally, we compare the neural network performance to a semidefinite programming (SDP) technique outlined in [21] in order to upper bound the gap between the optimal local and optimal collective strategy for more general cases.

II. Reinforcement Learning

Each round of the reinforcement learning process involves an agent choosing one action from an allowed action space, implementing the action, and receiving a reward from the environment. For a Markov decision process, the agent can eventually learn to choose actions according to an optimal policy that maximizes the expected future reward.

In the context of state discrimination, the environment is a parameterized measurement protocol for the quantum system of interest. The action space (denoted by $\mathcal{A}$) is the set of allowed quantum measurements and the reward equals the indicator function of success at the end of the process (i.e., $+1$ for successful discrimination once all measurements are completed and 0 otherwise). Denote by s_t the state of the environment just before round t and let n be the total number of rounds. The agent's policy, $\pi_\theta(a_t|s_t)$, is parameterized by θ and equals the probability of selecting action $a_t \in \mathcal{A}$ in round t conditioned on the state s_t of the environment.

We train the agent using the proximal policy optimization (PPO) algorithm [22]. This algorithm optimizes the advantage function

$$A^\pi(s_t,a_t) = \sum_{\ell=t}^{n} \gamma^{\ell-t}\Big(\mathbb{E}_{\pi_\theta}[r(s_\ell,a_\ell)\big|s_t,a_t] - \mathbb{E}_{\pi_\theta}[r(s_\ell,a_\ell)\big|s_t]\Big),$$

where $r(s,a)$ is the average reward for taking action a in state s and γ is a discount factor. Given a current state s_t, the advantage function compares the expected reward of choosing action a_t to the expected average reward for the policy $\pi_\theta(a_t|s_t)$. PPO optimizes a modified advantage function to prevent rapid changes in the policy π_θ. The modified objective function is given by:

$$L(\theta) \triangleq \mathbb{E}_t\Big[\min\Big(R_t(\theta)A^\pi(s_t,a_t), \text{clip}(R_t(\theta),1-\epsilon,1+\epsilon)A^\pi(s_t,a_t)\Big)\Big]$$

where the ratio function $R_t(\theta) \triangleq \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ is a measure of the change in the policy, ϵ is a hyperparameter, and

$$\text{clip}(x,\text{min},\text{max}) \triangleq \begin{cases} x & \text{if min} \le x \le \text{max} \\ \text{min} & \text{if } x < \text{min} \\ \text{max} & \text{otherwise.} \end{cases}$$

The final objective function also has terms that allow the neural network to estimate the value of each state. For these details, we refer the interested reader to [22]. Results are generated using the default PPO algorithm from the RLlib package included in Ray version 0.7.6 [23], [24]. Only the learning rate is tuned. The relevant hyperparameters are the clipping parameter $\epsilon = 0.3$, the discount factor $\gamma = 0.99$, and the learning rate $\text{lr} = 5 \times 10^{-5}$.

The neural network is fully connected and has an input layer with $m + n - 1$ neurons which take the state s as the input. The input layer feeds into two parallel sets of subnetworks (each with two hidden layers of 256 neurons, tanh activation functions, and each with their own linear output layers.) The output layer of the first subnetwork consists of a single neuron, and computes an estimate for the value a state. The output layer of the second subnetwork has nQ neurons (i.e. the number of allowed actions) and computes the policy, $\pi(a|s)$. See https://github.com/SarahBrandsen/RLNN-QSD for numerical results and the source code used to obtain them.

III. Discrimination Task and Structure

We consider the task of deriving the minimum-error adaptive measurement protocol to distinguish between m tensor-product quantum states $\{\rho_j\}|_{j=1}^m$ with prior probability vector $\boldsymbol{q}$ where $q_j = \Pr(\rho = \rho_j)$. Since each candidate state is assumed to be a tensor product of n subsystems, it can be written as

$$\rho_j = \bigotimes_{k=1}^{n} \rho_j^{(k)},$$

where $\rho_j^{(k)}$ is a qubit density matrix for all $j \in [1,...,m]$ and all $k \in [1,...,n]$. Thus, the quantum system ρ is composed of n unentangled qubit subsystems.

We build an OpenAI gym environment [25] capable of implementing local measurement algorithms. In each round, the algorithm chooses the next subsystem j to measure as well as which measurement to implement. Since all subsystems are assumed to be qubits, re-measuring an already measured subsystem is non-informative. To speed convergence, we introduce a penalty of -0.3 from the environment if the agent attempts re-measurement. Each element of the action space is a pair $(\hat{\Pi}, k)$, indicating that measurement $\hat{\Pi}$ is to be implemented on subsystem k. The set of allowed quantum measurements consists of binary real qubit POVMs spaced on the Bloch sphere according to quantization parameter Q, or $\{\hat{\Pi}_Q(\ell)\}|_{\ell=1}^Q$, where

$$\hat{\Pi}_Q(\ell) \triangleq \left\{ \begin{pmatrix} (\frac{\ell}{Q})^2 & \frac{\ell}{Q}\sqrt{(1-(\frac{\ell}{Q})^2)} \\ \frac{\ell}{Q}\sqrt{(1-(\frac{\ell}{Q})^2)} & 1-(\frac{\ell}{Q})^2 \end{pmatrix}, \begin{pmatrix} 1-(\frac{\ell}{Q})^2 & -\frac{\ell}{Q}\sqrt{(1-(\frac{\ell}{Q})^2)} \\ -\frac{\ell}{Q}\sqrt{(1-(\frac{\ell}{Q})^2)} & (\frac{\ell}{Q})^2 \end{pmatrix} \right\}$$

and $Q = 20$ (unless otherwise specified). After all subsystems are measured, the network receives a reward of 1 if the correct state has the largest posterior probability. That is, if $\rho = \rho_{j^*}$ for

$$j^* = \text{argmax}_j(p_j(\boldsymbol{q},\boldsymbol{d})),$$

where $\boldsymbol{d}$ is the vector containing all previous measurement results and $\boldsymbol{p}(\boldsymbol{q},\boldsymbol{d})$ is the updated probability after $\text{len}(\boldsymbol{d})$ rounds given the starting prior $\mathbf{q}$.

Given the candidate state set, all remaining information about the environment can be found from the updated prior vector given the vector of previous measurement results $\boldsymbol{d}$ and a length-n vector $\boldsymbol{v}$ where $v_k = 1$ if subsystem k has already been measured and 0 else. Thus, the environment before each

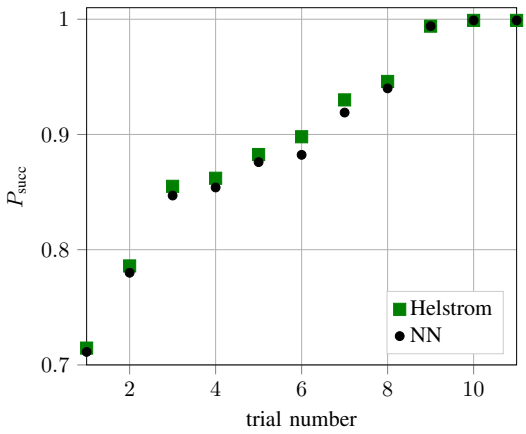

Fig. 1. Probability of success for the optimal RLNN policy after 1000 training iterations vs. the collective Helstrom measurement for tensor-products of pure states when $m = 2$, $n = 3$. The neural network approximates the Helstrom success probability, with quantization error from the action space contributing to the difference.

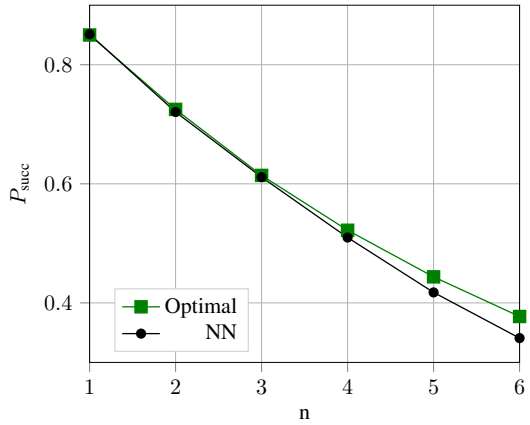

Fig. 2. Performance of the RLNN policy after 1000 training iterations vs. the optimal success probability as a function of the number of subsystems n. The RLNN approach has decreasing accuracy as the number of subsystems n increases.

round can be represented as $s \triangleq (\boldsymbol{v}, p(\boldsymbol{q}, \boldsymbol{d}))$. The episode is terminated when all subsystems have been measured, or equivalently when $\boldsymbol{v}$ is the all-ones vector.

IV. Numerical Results for RL with PPO

For an initial test, we consider the special case of binary discrimination (i.e. $m = 2$) between tensor products of pure states such that $\rho_j^{(k)} = |\psi_j^{(k)}\rangle \langle\psi_j^{(k)}|$ for all $k \in \{1, ..., n\}$ and $j \in \{1, 2\}$. In this case, the optimal collective (Helstrom) success probability can be achieved through locally-adaptive strategies [9], [15]. We randomly generate eleven trials with $n = 3$ and order the trials according to increasing distinguishability measured by the Helstrom success probability. For each trial, we compare the Helstrom success probability with the neural network best performance, as shown in Fig. 1. The neural network comes close to the Helstrom success probability in each case, and we believe the gap is mainly due to action space quantization.

An additional case where locally adaptive protocols are strictly optimal has been found by Sasaki et. al in [26]. Consider a set of states $\mathcal{S}_1 \triangleq \{\rho_j\}|_{j=1}^m$ and associated probabilities $\{q_j\}|_{j=1}^m$. Suppose the known optimal POVM for these is $\{\Pi_j\}|_{j=1}^m$. The set of n-subsystem product states generated by $\mathcal{S}$ can be written as

$$\mathcal{S}_n \triangleq \left\{ \bigotimes_{j=1}^{n} \rho_{i_j} \;\middle|\; i \in \{1, ..., m\}^n \right\},$$

with corresponding probabilities defined as $q_{i_1 \ldots i_n} \triangleq q_{i_1} \times ... \times q_{i_n}$. Then, the optimal POVM candidate state set $\mathcal{S}_n$ has elements that can be written in tensor product form as:

$$\Pi_{i_1 \ldots i_n} = \bigotimes_{j=1}^{n} \Pi_{i_j}.$$

It immediately follows that the optimal probability of success can be achieved using locally-adaptive protocols. The same proof idea can also be used to extend that result to a broader class of candidate states. That is, consider n sets of states $\{\rho_j(k)\}|_{j=1}^m$ for $k \in \{1, .., n\}$ with associated probabilities $\{q_j\}|_{j=1}^m$. Let the optimal POVM for the k-th state set be $\{\Pi_j^{(k)}\}_{j=1}^m$. The state set $\mathcal{S}_n \triangleq \left\{ \bigotimes_{j=1}^n \rho_{i_j}(i_j) \;\middle|\; i \in \{1, ..., m\}^n \right\}$ with corresponding prior probabilities $q_{i_1 \ldots i_n} \triangleq q_{i_1} \times ... \times q_{i_n}$ then has an optimal POVM with elements

$$\Pi_{i_1 \ldots i_n} = \bigotimes_{j=1}^{n} \Pi_{i_j}^{(j)}.$$

Thus, locally-optimal strategies can also achieve the optimal collective success probability for this extended class of problems. This provides a useful test of the neural network performance. We take the initial state set to be $\{\rho_1, \rho_2\}$, where the states are initially orthogonal depolarized states with depolarizing parameter $\gamma = 0.3$. Since the optimal local POVM belongs to the allowed action set, there should be no quantization loss. We train the neural network for 1000 iterations (using a custom learning rate schedule where the learning rate starts at 5.5×10^{-5} and decays by 0.95 every 10 iterations), and compare the neural network performance after training to the optimal success probability. For the case of relatively small subsystems $n \leq 4$, the neural network attains or approximately attains the exact success probability. As the subsystem number grows, the accuracy decreases, as depicted in Fig. 2.

Finally, we evaluate the RLNN performance for some tensor-product quantum states where the exact locally-optimal success probability is unknown. In these cases, we use the "Pretty Good Measurement" (PGM) as a benchmark. For a given state set $\{\rho_j\}|_{j=1}^m$ and prior vector $\boldsymbol{q}$, elements of the PGM are given as:

$$\Pi_j^{\text{PGM}} \triangleq \left(\sum_{\ell=1}^{m} q_\ell \rho_\ell \right)^{-1/2} q_j \rho_j \left(\sum_{\ell=1}^{m} q_\ell \rho_\ell \right)^{-1/2}.$$

The PGM is optimal in several cases, such as the symmetric case with equiprobable states $\rho_j = U^{j-1} \rho_1 \left(U^\dagger\right)^{j-1}$ and

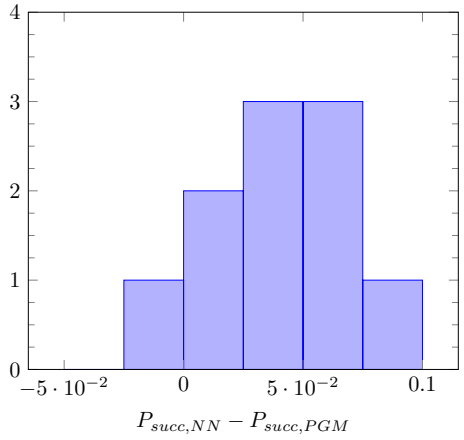

Fig. 3. Histogram of difference in the success probability the RLNN policy given 1000 training iterations and the PGM when $n = 3$ and $m = 5$. The NN comes close to or outperforms the PGM in all trials.

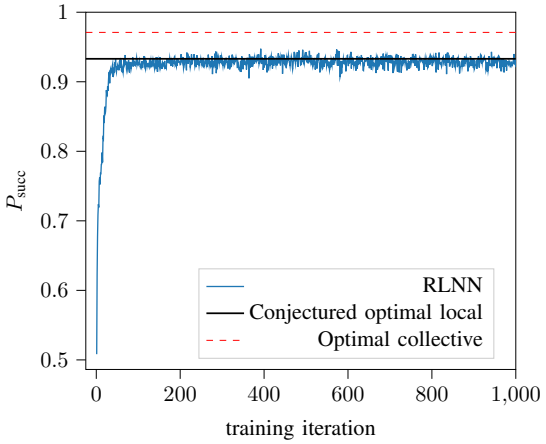

Fig. 4. Training curve (solid blue line) over the first 1000 iterations. The final success probability after 2000 iterations is $P_{\text{succ}} \approx 0.93 > P_{\text{succ, lg}}$. This approaches the conjectured optimal success probability for local strategies of $P_{\text{succ}} = 0.93$, represented by the dashed black line. However, there remains a gap between the neural network performance and the best collective measurement success probability $P_{\text{succ, opt}} \approx 0.97$.

$U^m = \mathbb{I}$ [17] or the case where the square root of the Gram matrix for weighted states $\{q_j\rho_j\}$ has equal diagonal elements [26], [27]. This motivates the use of the PGM as a benchmark. Additionally, we observe that the PGM can be achieved adaptively if $\sum_{j=1}^{m} q_j\rho_j$ can be written as a tensor product of m qubit density matrices.

We generate ten trials, where each trial has a randomly chosen (mixed) state set with $m = 5$ candidate states and $n = 3$ subsystems. In most cases the RLNN success probability $P_{\text{succ, NN}}$ is significantly greater than the success probability of the PGM, $P_{\text{succ, PGM}}$ (in the case where the number of candidate states is reduced to $m = 3$, the NN no longer outperforms the PGM [28]). A histogram of the success probability difference is shown in Fig. 3.

V. Pure Tensor Product Candidate States

In the special case where $m = 2$, it has been shown [9], [15] that locally-greedy algorithms are optimal for distinguishing between pure tensor product states. This, however is not the case when $m > 2$.

Theorem V.1. *Denote by $P_{\text{succ,lg}}(\{\rho_j\}, \boldsymbol{q})$ the probability of success when following a locally greedy strategy, and likewise denote by $P_{\text{succ,opt}}$ the success probability for the optimal collective measurement on the full quantum system. For $m > 2$, there exists at least one set of pure tensor product states $\{\rho_j\}|_{j=1}^m$ with starting prior $\boldsymbol{q}$ such that $P_{\text{succ,lg}}(\{\rho_j\}, \boldsymbol{q}) < P_{\text{succ,opt}}(\{\rho_j\}, \boldsymbol{q})$.*

Proof- Consider as an example the case where $n = 2$ and the candidate state set is symmetric with

$$\begin{aligned}\rho_j &\triangleq \left(U^j |0\rangle \langle 0| (U^j)^\dagger\right)^{\otimes 2} \\ &= (U \otimes U)^j |00\rangle \langle 00| \left((U \otimes U)^j\right)^\dagger,\end{aligned}$$

where

$$U \triangleq \begin{pmatrix} \cos(\frac{2\pi}{3}) & -\sin(\frac{2\pi}{3}) \\ \sin(\frac{2\pi}{3}) & \cos(\frac{2\pi}{3}) \end{pmatrix}$$

and $\boldsymbol{q} = [1/3, 1/3, 1/3]$. Since $(U \otimes U)^m = \mathbb{I}$, the PGM is optimal [17]. The corresponding probability of success is then $P_{\text{succ, opt}} = \frac{1}{6}(3 + 2\sqrt{2}) \approx 0.971$. Since both subsystems are identical copies, there is no need to consider the order of measurement.

The unique locally optimal measurement for the first subsystem is the local PGM, that is, the PGM for states $\{U^j |0\rangle \langle 0| (U^j)^\dagger\}|_{j=1}^3$ with $\boldsymbol{q} = [1/3, 1/3, 1/3]$. The updated probability vector after measuring the first subsystem and obtaining measurement outcome d_1 will be $p_j(d_1) = \frac{2}{3}\delta_{j,d_1} + \frac{1}{6}(1 - \delta_{j,d_1})$. From [27], [29], [30], we may verify that a locally optimal measurement for the second subsystem given the new prior is the PGM for states $\{U^j |0\rangle \langle 0| (U^j)^\dagger\}|_{j=1}^m$ with probabilities $\tilde{p}_j(d_1) = \frac{37}{39}\delta_{j,d_1} + \frac{1}{39}(1 - \delta_{j,d_1})$ because they satisfy the sufficient condition,

$$p_j(d_1) = \frac{C}{\langle 0| (U^j)^\dagger \left(\sum_k \tilde{p}_k(d_1) U^j |0\rangle \langle 0| (U^j)^\dagger\right)^{-\frac{1}{2}} U^j |0\rangle}$$

for normalization constant C. Note that, for the last subsystem, any locally-optimal measurement will yield the same final success probability, so it is not necessary to verify uniqueness of this measurement. The resulting success probability is $P_{\text{succ, lg}} = \frac{4}{5}$, hence $P_{\text{succ, lg}} < P_{\text{succ, opt}}$. □

1) RLNN finds a better strategy: Given the demonstrated reliable performance of our neural network when the number of subsystems is small, aim to find whether the neural network can outperform the locally-greedy strategy for the above example. Given the symmetry of the candidate states, we extend the set of allowed quantum measurements to include SIC POVMs of the form:

$$\hat{\Pi}(\ell) \triangleq \left\{ R\left(\frac{\ell\pi}{(Q-1)}\right) |0\rangle \langle 0| R^\dagger\left(\frac{\ell\pi}{(Q-1)}\right),\right.$$

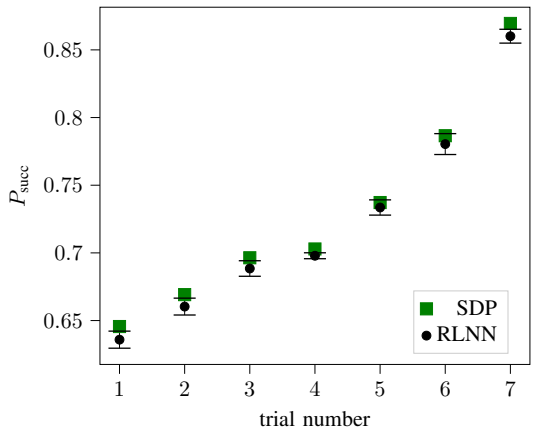

Fig. 5. Probability of success for SDP and RLNN when $m = 2$ and $n = 3$. For each trial, the RLNN success probability is computed by separately training the neural network five times with 2000 iterations each. The error bars represent the standard deviation in the final success probability over the five independent trainings. In all trials, the gap between local and non-local measurements is small.

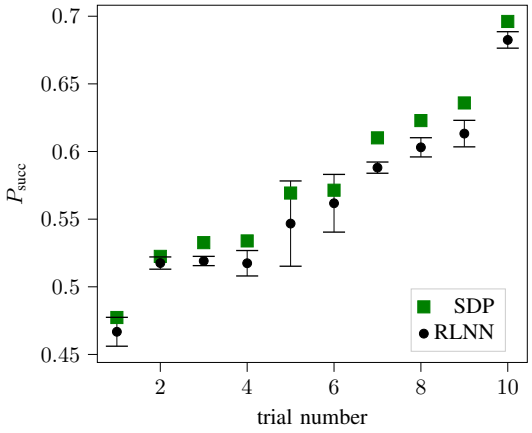

Fig. 6. Probability of success for SDP and RLNN after 2000 training iterations when $m = 3$, $n = 3$. For each trial, the RLNN success probability is computed by separately training the neural network five times with 2000 iterations each. The error bars represent the standard deviation in the final success probability over the five independent trainings. Compared to the case where $m = 2$, there is greater variation in training results and a larger gap between local and non-local measurements.

$$R\left(\frac{\ell\pi}{(Q-1)}+\frac{2\pi}{3}\right)|0\rangle\langle 0|R^{\dagger}\left(\frac{\ell\pi}{(Q-1)}+\frac{2\pi}{3}\right),$$
$$R\left(\frac{\ell\pi}{(Q-1)}+\frac{4\pi}{3}\right)|0\rangle\langle 0|R^{\dagger}\left(\frac{\ell\pi}{(Q-1)}+\frac{4\pi}{3}\right)\Big\}$$

for $\ell \in [0, ..., Q-1]$ for $Q = 12$. This is combined with the set of binary projective measurements for $Q = 12$ to form the full measurement set.

The probability of success is plotted as a function of the training iteration in Fig. 4. The training curve indicates rapid learning and a final success probability of $P_{\text{succ}} = 0.928$, which represents a significant improvement over the locally greedy approach. This suggests that the optimal method is to implement $\hat{\Pi}(Q-1)$ on the first subsystem, which corresponds to a rotated SIC POVM where each measurement outcome is orthogonal to at least one candidate state. Then the second subsystem is measured according to the Helstrom measurement for two remaining candidate states, with $P_{\text{succ}} \approx 0.93$. Despite this improvement over the locally greedy method, there remains a gap between the best local approach we know and the optimal collective measurement.

VI. Gap between Locally Optimal Algorithm and Collective Measurement

Finally, we use RLNN to estimate the gap between the best locally adaptive algorithm and the optimal collective (non-local) measurement in more general cases where the best locally adaptive algorithm is not otherwise known. The probability of success for the optimal collective measurement is closely approximated via semidefinite programming (SDP), as introduced in [21].

The simulation setup for a given m and n is as follows: for each trial, we randomly generate pure tensor product candidate states and then apply depolarizing noise with a randomly chosen noise parameter. The RLNN algorithm is independently trained 5 times over 2000 iterations, and the average final success probability is compared (with error bars) to the optimal collective success probability found via SDP. Results are plotted for $m = 2$, $n = 3$ in Figure 5 and for $m = 3$, $n = 3$ in Figure 6, and indicate that the gap between local and collective measurements increases with m.

VII. Conclusion

We apply RLNN to devise near-optimal locally-adaptive measurement schemes for multiple state discrimination. We provide preliminary results for the neural network performance in cases where the locally-adaptive probability of success is known, and show that the network can achieve good performance when the total number of subsystems to be measured is small. We show by counterexample that, unlike in the binary case, adaptive locally greedy algorithms no longer achieve the optimal collective success probability when all candidate states are pure tensor product states. We also use RLNN to find an improved, (less-greedy) locally-adaptive protocol and observe that the gap between the optimal collective success probability appears to persist for all locally-adaptive algorithms. Finally, we use RLNN to estimate the gap between the optimal local and optimal collective strategies in more general cases.

Acknowledgment

The authors would like to thank Narayanan Rengaswamy for helpful discussions. The work of Brandsen and Pfister was supported in part by the National Science Foundation (NSF) under Grant No. 1908730 and 1910571. Any opinions, findings, conclusions, and recommendations expressed in this material are those of the authors and do not necessarily reflect the views of these sponsors.

References

[1] A. Ferdinand, M. DiMario, and F. Becerra, "Multi-state discrimination below the quantum noise limit at the single-photon level," *npj Quantum Information*, vol. 3, 12 2017.

[2] H. Krovi, S. Guha, Z. Dutton, and M. P. da Silva, "Optimal measurements for symmetric quantum states with applications to optical communication," *Physical Review A*, vol. 92, Dec 2015.
[3] P. H. D. Rengaswamy, Narayanan, "Quantum advantage in classical communications via belief-propagation with quantum messages," 2020.
[4] A. Assalini, N. Dalla Pozza, and G. Pierobon, "Revisiting the Dolinar receiver through multiple-copy state discrimination theory," *Phys. Rev. A*, vol. 84, p. 022342, Aug 2011.
[5] C. W. Helstrom, "Quantum detection and estimation theory," *Journal of Statistical Physics*, vol. 1, no. 2, pp. 231–252, 1969.
[6] A. H. Kiilerich and K. Mølmer, "Multistate and multihypothesis discrimination with open quantum systems," *Physical Review A*, vol. 97, May 2018.
[7] R. Koenig, R. Renner, and C. Schaffner, "The operational meaning of min- and max-entropy," *IEEE Transactions on Information Theory*, vol. 55, p. 4337–4347, Sep 2009.
[8] R. Bellman, "The theory of dynamic programming," *Bull. Amer. Math. Soc.*, vol. 60, pp. 503–515, 11 1954.
[9] S. Brandsen, M. Lian, K. D. Stubbs, N. Rengaswamy, and H. D. Pfister, "Adaptive procedures for discrimination between arbitrary tensor-product quantum states," 2019.
[10] T. Fösel, P. Tighineanu, T. Weiss, and F. Marquardt, "Reinforcement learning with neural networks for quantum feedback," *Phys. Rev. X*, vol. 8, p. 031084, Sep 2018.
[11] G. Tesauro, "Practical issues in temporal difference learning," *Mach. Learn.*, vol. 8, p. 257–277, May 1992.
[12] G. J. Gordon, "Stable fitted reinforcement learning," in *Proceedings of the 8th International Conference on Neural Information Processing Systems*, NIPS'95, (Cambridge, MA, USA), p. 1052–1058, MIT Press, 1995.
[13] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with deep reinforcement learning," 2013.
[14] V. Mnih, K. Kavukcuoglu, D. Silver, A. Rusu, J. Veness, M. Bellemare, A. Graves, M. Riedmiller, A. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–33, 02 2015.
[15] A. Acín, E. Bagan, M. Baig, L. Masanes, and R. Muñoz Tapia, "Multiple-copy two-state discrimination with individual measurements," *Phys. Rev. A*, vol. 71, p. 032338, 2005.
[16] P. Hausladen and W. K. Wootters, "A 'pretty good' measurement for distinguishing quantum states," *Journal of Modern Optics*, vol. 41, no. 12, pp. 2385–2390, 1994.
[17] M. Ban, "Optimum measurements for discrimination among symmetric quantum states and parameter estimation," *International Journal of Theoretical Physics*, vol. 36, no. 6, pp. 1269–1288, 1997.
[18] S. M. Barnett, "Minimum-error discrimination between multiply symmetric states," *Phys. Rev. A*, vol. 64, p. 030303, Aug 2001.
[19] Y. C. Eldar, A. Megretski, and G. C. Verghese, "Optimal detection of symmetric mixed quantum states," *IEEE Transactions on Information Theory*, vol. 50, June 2004.
[20] Y. C. Eldar and G. D. Forney, "On quantum detection and the square-root measurement," *IEEE Transactions on Information Theory*, vol. 47, pp. 858–872, March 2001.
[21] Y. Eldar, A. Megretski, and G. Verghese, "Designing optimal quantum detectors via semidefinite programming," *IEEE Transactions on Information Theory*, vol. 49, p. 1007–1012, Apr 2003.
[22] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017.
[23] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, and I. Stoica, "Tune: A research platform for distributed model selection and training," *arXiv preprint arXiv:1807.05118*, 2018.
[24] E. Liang, R. Liaw, P. Moritz, R. Nishihara, R. Fox, K. Goldberg, J. E. Gonzalez, M. I. Jordan, and I. Stoica, "Rllib: Abstractions for distributed reinforcement learning," 2017.
[25] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "OpenAI gym," 2016.
[26] M. Sasaki, K. Kato, M. Izutsu, and O. Hirota, "Quantum channels showing superadditivity in classical capacity," *Phys. Rev. A*, vol. 58, pp. 146–158, Jul 1998.
[27] C. Mochon, "Family of generalized "pretty good" measurements and the minimal-error pure-state discrimination problems for which they are optimal," *Phys. Rev. A*, vol. 73, p. 032328, Mar 2006.
[28] S. Brandsen, K. D. Stubbs, and H. D. Pfister, "Reinforcement learning with neural networks for quantum multiple hypothesis testing," *placeholder for arXiv version to be posted*, 2020.
[29] V. Belavkin, "Optimum distinction of non-orthogonal quantum signals," *Radio Engineering and Electronic Physics*, vol. 20, pp. 39–47, June 1975.
[30] G. Weir, C. Hughes, S. M. Barnett, and S. Croke, "Optimal measurement strategies for the trine states with arbitrary prior probabilities," 2018.