

Reinforcement learning for quantum environments with memory

Josep Lumbreras, Ruo Cheng Huang, Yanglin Hu, and Marco Fanizza

Introduction. We ask how a classical agent can efficiently resolve the fundamental exploration-exploitation trade-off over an uncharacterized quantum stochastic process. In reinforcement learning, an agent interacts sequentially with an unknown environment and chooses actions in order to maximize a cumulative reward [14]. The difficulty is already nontrivial in classical partially observed systems. In robotics, navigation, adaptive control, and resource allocation [4, 7, 11, 12], the agent often does not observe the true state of the environment directly. Instead, it receives partial observations whose distribution depends on a hidden memory. The agent must therefore use the history of past actions and observations to infer the latent dynamics, while at the same time choosing actions that give high reward. This is the setting captured by partially observable Markov decision processes (POMDPs), for which recent work has obtained sharp regret guarantees under suitable observability assumptions [3, 8].

The quantum setting adds a further physical constraint. In emerging autonomous quantum experiments, the system to be learned and controlled may be a quantum memory evolving through unknown and possibly noisy quantum channels. Examples include calibration, adaptive quantum control, error correction, and resource extraction, where the experimenter repeatedly applies pulses or measurements and receives only classical data. In these tasks, however, an intervention is not a passive observation: each pulse or measurement both probes the system and changes the post-intervention memory state that will generate future data. Thus, the data collection mechanism is itself part of the dynamics to be controlled. Classical POMDP frameworks do not directly apply when the hidden state, its transition dynamics, and the agent’s access instruments are quantum. This raises the question: *Can a classical agent efficiently learn to control an unknown quantum process with memory, using only the outcomes of its past interventions?*

Contributions. We introduce, to our knowledge, the first reinforcement-learning framework for quantum processes with hidden memory, modeled as input–output *quantum hidden Markov model* (QHMMs) environments, in which a classical agent adaptively probes an unknown quantum memory through quantum instruments and learns only from classical outcomes and rewards. We develop an *optimistic maximum-likelihood learning* (OMLE) algorithm that achieves $\tilde{O}(\sqrt{K})$ regret over K episodes under a stable observability condition, with near-matching lower bounds. The analysis applies not only to discrete action sets, but also to continuous families of quantum measurements, by reducing the action dependence to a finite-dimensional linear span. In the thermodynamic application to state-agnostic work extraction from non-i.i.d. quantum sources, this regret is exactly the cumulative dissipation caused by incomplete knowledge of the source.

Input–output QHMM environment. We model the environment as an L -step input–output QHMM with an S -dimensional hidden quantum memory, initialized in an unknown state ρ and evolving through *unknown quantum channels* $\mathbb{E}_l$ at each step $l \in [L]$.

The agent interacts with the hidden memory and observes classical outcomes in $[O]$ each time it evolves. At each step $l \in [L]$, the agent chooses an action a_l that specifies a quantum instrument $\mathcal{P}^{(a_l)}$. This intervention yields a classical outcome $o_l \in [O]$, a reward $r_l(a_l, o_l)$, and an updated memory state with the CPTNI map $\Phi_o^{(a)}$ that governs future observations. The agent selects actions adaptively from the past trajectory $\tau_{l-1} = (a_1, o_1, \dots, a_{l-1}, o_{l-1})$.

We use ω to denote an input–output QHMM environment specifying the unknown initial memory state, channels, and instruments. Over K independent episodes, the goal is to minimize

cumulative regret,

$$\text{Regret}(K) = \sum_{k=1}^K (V^* - V^{\pi^k}),$$

where V^{π^k} is the expected cumulative reward of the executed policy π^k at episode $k \in [K]$ and $V^* = \sup_{\pi} V^{\pi}$ is the optimal value.

Stable observability condition. A fundamental challenge in learning hidden-memory models is identifiability. Recovering the exact latent CPTP channels from classical data is not the right target: distinct physical realizations may generate the same observable statistics. Our main technical idea is therefore to work in the *Observable Operator Model* (OOM) representation [2]. In this representation, a predictive state x_l^{ω} encodes trajectory probabilities $P_{\omega}(\tau_l)$ and is updated linearly by OOM transition maps $\mathbf{A}_l^{\omega}(o_l, a_l, a_{l+1})$ conditioned on QHMM environment ω , chosen actions a_l, a_{l+1} and observed outcomes o_l , $x_l^{\omega}(o_l) = P_{\omega}(\tau_l)$, $x_{l+1}^{\omega} = \mathbf{A}_l^{\omega}(o_l, a_l, a_{l+1})x_l^{\omega}$. Rather than identifying a unique latent quantum model, the agent learns and plans directly from experimentally accessible trajectory probabilities.

The stable observability condition enabling this representation is the quantum analogue of the undercomplete assumption in classical POMDPs [1, 3, 8, 9]. For every action a , the classical marginal produced by the instrument must contain enough information to stably reconstruct the corresponding post-instrument memory state through a linear recovery map $R^{(a)}$ satisfying for a finite constant κ_{uc} the following constraint: $R^{(a)} \circ \text{Tr}_S \circ \mathcal{P}^{(a)} = \mathcal{P}^{(a)}$, $\|R^{(a)}\|_{1 \rightarrow 1} \leq \kappa_{\text{uc}}$. This observability condition is a condition on the measurement devices available to the agent rather than a hidden constraint on the unknown dynamics, which can be checked by linear algebra once the instruments are fixed. It is satisfied by broad and experimentally natural classes of measurements, including rank-one projective measurements, rank-one measure-and-prepare instruments, and informationally complete POVMs. In the classical limit, it reduces to the standard full-rank/weakly revealing observability condition [1, 3, 8, 9].

Regret guarantees for the OMLE algorithm. On top of this OOM representation, we adapt the OMLE algorithm [8, 9] to learn the trajectory law induced by the input-output QHMM environment ω . For a trajectory τ_L , its observation likelihood can be written compactly as $P_{\omega}(\tau_L)$. Thus the OMLE confidence region is built from the log-likelihood of the observed data τ^i over k rounds, $\ell_k(\omega) = \sum_{i=1}^k \log P_{\omega}(\tau^i)$, up to policy terms that do not depend on ω . The agent then selects the most rewarding plausible model and policy, executes the policy in the true environment, and updates the dataset with the new trajectory.

For finite action sets with $A \in \mathbb{N}$ actions corresponding to general quantum instruments, we prove that OMLE achieves

$$\text{Regret}(K) \leq \tilde{\mathcal{O}}(\text{poly}(A, O, S, L) \kappa_{\text{uc}}^2 \sqrt{K}),$$

where A is the number of actions, O the number of outcomes, S the memory dimension, and L the horizon. The proof combines the OMLE machinery with a quantum-specific error-propagation step. A central part of our proof is the quantum error-propagation argument. While the regret analysis also requires estimation and concentration bounds, this is the step where the input-output QHMM structure is used most directly. We telescope the true and optimistic OOM products to isolate local one-step prediction errors at step l given history τ_{l-1} and current action a_l . Then we propagate this error through the remaining instruments and channels with the contractive property of OOM transition maps $\sum_{\text{future}} \|\mathbf{A}_{L-1:l+1} x\|_1 \leq \sqrt{2} S^2 \kappa_{\text{uc}} \|x\|_1$, for any OOM transition map sequence $\mathbf{A}_{L-1:l+1}$ and any vector x . This contractive property is derived from the stable observability condition.

Linear span technique for continuous actions. A discrete action set is unnatural for a fully quantum process, where an action may be a POVM or instrument chosen from a continuous family.

Existing regret guarantees for partially observable models typically work with discrete action set, or with discretized continuous action set [3, 8, 9]. A second technical highlight of our work is that the regret analysis exploits the finite-dimensional linear structure of quantum measurements to handle continuous quantum action spaces efficiently. The obstruction is that the discrete regret proof counts the contribution of each action separately; over a continuum, this action-by-action accounting would make the bound vacuous. In the input-output QHMM environment, this difficulty can be resolved by exploiting the linearity of the OOM transition maps on a chosen basis. Concretely, after fixing a Hilbert–Schmidt basis for the linear span of the measurement family, each measurement effect can be expressed as a finite linear combination of basis elements. The dependence on the continuous action is then carried entirely by known coefficients in this expansion, while the unknown model-estimation error is controlled on finitely many basis directions. The infinite action class is therefore controlled by the intrinsic linear dimension $d_{\mathcal{A}} \leq (O-1)S^2$ of the measurement family, rather than by any discretization. This yields $\text{Regret}(K) \leq \tilde{O}(\text{poly}(d_{\mathcal{A}}, S, L)\kappa_{\text{uc}}^2\sqrt{K})$, independently of the cardinality of the action set.

Matching lower bound. Finally, we show that the $\tilde{O}(\sqrt{K})$ scaling is statistically optimal up to polylogarithmic factors. Rather than reducing only to a classical diagonal subcase, we use SIC-POVMs and depolarizing channels to embed the multi-armed quantum bandit problem [10] while preserving the observability condition. This shows that the hardness is intrinsic to quantum state discrimination. Classical combinatorial-lock instances [3, 6] further show that the dependence on κ_{uc} cannot be removed.

Thermodynamic interpretation. Beyond abstract learning, the framework has a direct physical implication in quantum thermodynamics. We apply it to state-agnostic work extraction, where an agent seeks to harvest non-equilibrium free energy from a sequence of non-i.i.d. quantum states generated by a hidden memory. If the agent does not know the underlying emission source, it cannot perfectly align its extraction protocol with the incoming states. By the laws of thermodynamics, this informational deficit inevitably results in irreversible dissipation [5, 13].

We formalize this thermodynamic process as an instance of the input-output QHMM. At each time step, the agent chooses an extraction basis and a target purity parameter. The corresponding physical interaction is modeled as a quantum instrument. In this thermodynamic setting, the outcomes include the energy transferred to a work battery, and we set the reward to the realized extracted work, $r_l(a_l, o_l) = w_l(o_l)$. The value $V^\pi = \mathbb{E}_\pi[\sum_{l=1}^L w_l(o_l)]$ is therefore the expected work extracted by policy π W_{ext}^π , while V^* is the maximum expected work extracted by an agent that knows the source. Hence the heat dissipated by incomplete knowledge is exactly the lost extractable work, $W_{\text{diss}}(K) := \sum_{k=1}^K (W_{\text{ext}}^* - W_{\text{ext}}^{\pi_k}) = \text{Regret}(K)$. Thus, regret minimization is equivalent to an adaptive Maxwell’s demon learning to suppress energy waste from past battery outcomes. Our $\tilde{O}(\sqrt{K})$ regret bound implies sublinear total dissipation, and hence vanishing average dissipation per episode.

Outlook. Our results connect partially observable reinforcement learning, quantum information, hidden-memory models, and quantum thermodynamics. They provide a learning-theoretic route to adaptive quantum experiments where learning and control occur simultaneously. Natural next steps include moving beyond the undercomplete regime, treating other resources such as coherence and entanglement, extending the model to quantum combs and process tensors, and developing adversarial or continuous-time variants.

[1] D. Hsu, S. M. Kakade, and T. Zhang. “A spectral algorithm for learning Hidden Markov Models”. *Journal of Computer and System Sciences* **78**(5): 1460–1480 (2012).

- [2] H. Jaeger. “Observable Operator Models for Discrete Stochastic Time Series”. *Neural Comput.* **12**(6):1371–1398 (2000).
- [3] C. Jin, S. M. Kakade, A. Krishnamurthy, and Q. Liu. “Sample-efficient reinforcement learning of undercomplete POMDPs”. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS ’20, Red Hook, NY, USA(2020).
- [4] J. Kober, J. A. Bagnell, and J. Peters. “Reinforcement learning in robotics: A survey”. *Int. J. Rob. Res.* **32**(11):1238–1274 (2013).
- [5] A. Kolchinsky and D. H. Wolpert. “Dependence of dissipation on the initial distribution over states”. *Journal of Statistical Mechanics: Theory and Experiment* **2017**(8):083202 (2017).
- [6] A. Krishnamurthy, A. Agarwal, and J. Langford. “PAC Reinforcement Learning with Rich Observations”. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, (2016).
- [7] S. Levine, C. Finn, T. Darrell, and P. Abbeel. “End-to-end training of deep visuomotor policies”. *J. Mach. Learn. Res.* **17**(1):1334–1373, (2016).
- [8] Q. Liu, A. Chung, C. Szepesvari, and C. Jin. “When Is Partially Observable Reinforcement Learning Not Scary?”. In P.-L. Loh and M. Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 5175–5220, (2022).
- [9] Q. Liu, P. Netrapalli, C. Szepesvari, and C. Jin. “Optimistic MLE: A Generic Model-Based Algorithm for Partially Observable Sequential Decision Making”. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, STOC 2023, page 363–376, New York, NY, USA(2023).
- [10] J. Lumberras, E. Haapasalo, and M. Tomamichel. “Multi-armed quantum bandits: Exploration versus exploitation when learning properties of quantum states”. *Quantum* **6**:749 (2022).
- [11] H. Mao, M. Alizadeh, I. Menache, and S. Kandula. “Resource Management with Deep Reinforcement Learning”. In *Proceedings of the 15th ACM Workshop on Hot Topics in Networks*, HotNets ’16, page 50–56, New York, NY, USA(2016).
- [12] A. Mirhoseini, A. Goldie, M. Yazgan, J. W. Jiang, E. Songhori, S. Wang, Y.-J. Lee, E. Johnson, O. Pathak, A. Nova, et al. “A graph placement methodology for fast chip design”. *Nature* **594**(7862):207–212 (2021).
- [13] P. M. Riechers and M. Gu. “Initial-state dependence of thermodynamic dissipation for any quantum process”. *Phys. Rev. E* **103**:042145 (2021).
- [14] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. MIT Press (1998).
- [15] **Technical version:** J. Lumberras, R. C. Huang, Y. Hu, M. Fanizza, and M. Gu. “Reinforcement learning for quantum processes with memory”. arXiv preprint arXiv:2603.25138 , (2026).