

An unbounded quantum memory advantage for reinforcement learning world models

Josep Lumbreras, Hailan Ma, Jayne Thompson, and Mile Gu

Introduction. Reinforcement learning (RL) studies sequential decision making, where actions have consequences over time. Besides learning values directly from experience, a central route to this problem is to learn a model that can be queried before acting, from classical model-based RL to modern learned *world models* [19, 27, 28]. In this setting, a world model serves as an internal simulator: conditioned on the agent’s realized history and a candidate action, it predicts future observations and rewards for *planning* [13, 14, 24]. The promise is compelling: simulated experience can improve decision-making without relying solely on trial-and-error interaction with the real environment [14, 24].

For control, the relevant object is not only an unconditional sample from the environment. A planner asks conditional questions: after this history, what would happen if we take this action? World models can generate future trajectories, but their utility for control depends on accurate predictions conditioned on both the realised history and the candidate action [11, 22, 26]. This distinction becomes important in rare but critical safety situations, such as a pedestrian suddenly entering the road. Although such histories may have negligible probability under the natural traffic distribution, inaccurate conditional predictions can still lead to catastrophic decisions once they occur. Accordingly, safety evaluation increasingly considers targeted rare-event scenarios alongside average-case performance [18, 23]. We focus on exact conditional simulation for *planning*: given a history relevant to a control decision, the model predicts the outcomes, rewards, and updated memory associated with each candidate action.

In partially observable environments, optimal decisions depend on the action-observation history. A compact world model therefore compresses the past without discarding information relevant for control [16, 17, 20]. Classical finite-memory models encode histories as probability distributions over finitely many distinguishable latent states. Quantum stochastic models suggest a different possibility: predictive information can be stored in nonorthogonal quantum states, reducing the required memory to simulate classical stochastic processes, including rare-event regimes [1, 6, 7, 10, 12, 21]. In this work, we ask *whether this quantum memory advantage in stochastic simulation can extend to sequential decision making*. We show that: finite quantum memory can preserve decision-relevant information that no finite classical memory can retain, enabling exact conditional planning where finite classical world models necessarily misestimate the consequences of actions.

Contributions. The main result is an unbounded separation for conditional planning between classical and quantum world models. We construct a fully classical controlled world with an infinite family of histories $\mathcal{H} = \{h_t\}_{t \geq 0}$ such that every finite-state classical world model eventually loses decision-relevant information across this infinite family of histories. As a result, any planner using such a model persistently misestimates the future return of at least one candidate action, independently of the number of classical memory states. The same world admits an exact finite-dimensional quantum model, which predicts the correct conditional future after every history and action, enabling exact planning and optimal control.

First, we formulate world models as planning interfaces: after a realised history, the model must answer conditional queries of the form “what would happen if action a were taken now?” and support rollouts from the corresponding updated memory. Second, we use the Fox–Rubin–Dharmadhikari–Nadkarni (FRDN) renewal process [4, 5, 9] as a benchmark for a decision-relevant memory effect that finite classical positive models cannot retain [29, 30]. By adding actions that continue, reset, or probe the process, this predictive limitation becomes a persistent error in the

expected returns used for planning. A finite-dimensional quantum model, in contrast, stores the relevant information in nonorthogonal memories and predicts the benchmark exactly. Third, we complement this exact separation with a quantum bisimulation distance, which quantifies how approximate quantum memories affect the values and action-values obtained by planning.

World models and quantum realizations for RL. Consider a realised history $h_t = (a_0, y_1, \dots, a_{t-1}, y_t)$, where each outcome $y_s = (o_s, r_s)$ contains a classical observation o_s and reward r_s . A world model M compresses h_t into a memory state $z = E_M(h_t)$. When queried with a candidate action a , it samples $Y = (O, R) \sim P_M(\cdot \mid z, a)$ and updates its memory to $T_Y^a(z)$, so that subsequent simulated queries represent the extended history h_{t+1} . Repeated queries therefore generate simulated rollouts from the updated memory. The quantity optimized in RL is the return, namely the discounted cumulative reward $G = \sum_{k \geq 0} \gamma^k R_{k+1}$, with $\gamma \in [0, 1)$. A policy $\pi_M(\cdot \mid z)$ maps each model memory state z to a distribution over actions and describes the agent interacting with the world model. Let $\mathbb{E}_{M,\pi}^{z,a}$ denote expectation over model-generated trajectories initialized at memory z , taking action a first and following π_M thereafter; $\mathbb{E}_{M,\pi}^z$ denotes the analogous expectation when π_M is followed from z immediately. The action-value $Q_M^\pi(z, a) = \mathbb{E}_{M,\pi}^{z,a}[G]$ is therefore the model-predicted return of choosing a as the first action. The value function $V_M^\pi(z) = \mathbb{E}_{M,\pi}^z[G]$ is the predicted return when the policy is followed from the start. Optimal planning compares candidate actions through $Q_M^*(z, a) = \sup_\pi Q_M^\pi(z, a)$, with $V_M^*(z) = \max_a Q_M^*(z, a)$. Thus, planning uses the world model to compare candidate actions through their predicted returns; the model is judged by the values its conditional predictions and memory updates induce.

The classical baseline is the usual finite-state POMDP or hidden-state generative model [3, 16], written in controlled positive-realization form [15, 29]. An N -state model stores a belief $z \in \Delta_{N-1}$, and each action–outcome branch (a, y) is a non-negative matrix $D_y^{(a)}$ with $\mathbf{1}^\top \sum_y D_y^{(a)} = \mathbf{1}^\top$ ($\mathbf{1}$ is the all-ones vector). It predicts and updates by $P_C(y \mid z, a) = \langle \mathbf{1}, D_y^{(a)} z \rangle$ and $T_{C,y}^a(z) = D_y^{(a)} z / \langle \mathbf{1}, D_y^{(a)} z \rangle$. We note that the belief simplex is continuous and includes the Bayesian belief-state representation used for planning under partial observability.

A quantum world model realizes the same classical input–output interface with a finite-dimensional quantum memory. The memory is a density operator $\rho \in \mathcal{S}(\mathcal{H}_Q)$, and each action is represented by an instrument $\{\mathcal{E}_y^a\}_y$ of completely positive trace-nonincreasing maps, with $P_Q(y \mid \rho, a) = \text{Tr}[\mathcal{E}_y^a(\rho)]$ and $T_{Q,y}^a(\rho) = \mathcal{E}_y^a(\rho) / \text{Tr}[\mathcal{E}_y^a(\rho)]$. For a history h , products of classical branch matrices are replaced by compositions of instrument branches $\mathcal{E}_h$; the encoded memory is $E_Q(h) = \mathcal{E}_h(\rho_0) / \text{Tr}[\mathcal{E}_h(\rho_0)]$. The agent, actions, observations, rewards, and planner remain classical; only the internal memory used to answer conditional planning queries is quantum. Classical models correspond to the special case where memory states remain diagonal in a fixed basis, so each instrument branch acts only as a classical update of the associated probability distribution.

FRDN benchmark and separation. Our benchmark is based in the construction of a classical world using the FRDN process [4, 5, 9]. After each reset, a hidden lifetime L is drawn and the process emits Ticks until the lifetime ends. For parameters $0 < \lambda \leq 1/2$ and $\alpha/\pi \notin \mathbb{Q}$, the probability of surviving at least t ticks and the conditional one-more-Tick probability are

$$\Phi(t) = \Pr(L \geq t) = \sum_{\ell=t}^{\infty} \lambda^\ell \sin^2(\ell\alpha/2), \quad S(t) = \frac{\Phi(t+1)}{\Phi(t)} = f(t\alpha).$$

Here f is continuous, nonconstant, and periodic, so the irrational phase $t\alpha$ never collapses to a finite cycle. This is the feature that finite positive memory cannot reproduce exactly. We turn this prediction benchmark into a controlled world with three actions. At a history h_t^{tick} of t consecutive tick branches since the last reset, **Wait** produces another Tick with probability $S(t)$ which rewards the agent and otherwise resets after a breakdown and the agent pays a cost; **Maintain** resets immediately with a maintenance cost; **Probe** tests the same Tick event as **Wait** but resets after

either outcome. The reward and cost constants are chosen so that the cumulative rewards of the candidate actions depend nontrivially on $S(t)$.

For any finite classical world model, the all-Tick histories are generated by repeated powers of one non-negative Tick branch matrix D_T : $P_C(t) = \langle \mathbf{1}, D_T^t z_0 \rangle$ and $S_C(t) = P_C(t+1)/P_C(t)$. Perron–Frobenius theory [2, 25] forces such predictions, along finitely many residue classes, to converge to finitely many limiting phases. The benchmark prediction $S(t) = f(t\alpha)$ keeps sampling an irrational orbit. Consequently there is a constant $\varepsilon_0 > 0$, independent of N , such that every N -state classical world model satisfies

$$\liminf_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \max_{a \in \mathcal{A}} \left| Q_\star^*(h_t^{\text{tick}}, a) - Q_{C_N}^*(E_C(h_t^{\text{tick}}), a) \right| \geq \varepsilon_0.$$

The same argument also give an optimal-value gap ε_0 -error. This benchmark certifies an unbounded finite-classical-memory limitation for planning, and states that any classical agent planning in this classical world model will incur an error of at least ε_0 on its estimated (action)-value functions.

The quantum side uses a qutrit realization of the FRDN process as a memory primitive [7]. On $\mathcal{H}_Q \simeq \mathbb{C}^3$, the Wait–Tick branch is $\mathcal{E}_{y_T}^W(\rho) = A_T \rho A_T^\dagger$, with $A_T = \sqrt{\lambda} e^{-r\sigma_x} e^{i\alpha\sigma_z/2} e^{r\sigma_x} \Pi_{01}$, where X and Z act on the $\{|0\rangle, |1\rangle\}$ subspace and the reset state ρ_0 and constant r are fixed by the FRDN parameters. The remaining branches complete the instruments: breakdown and Maintain reset to ρ_0 , while Probe reads the same Tick event and then resets. If $\tilde{\rho}_t = (\mathcal{E}_{y_T}^W)^t(\rho_0)$ and $\rho_t = \tilde{\rho}_t / \text{Tr } \tilde{\rho}_t$, then $\text{Tr } \tilde{\rho}_t = \Phi(t)$, $P_Q(y_T | \rho_t, W) = \text{Tr } \tilde{\rho}_{t+1} / \text{Tr } \tilde{\rho}_t = S(t)$, and $T_{Q, y_T}^W(\rho_t) = \rho_{t+1}$. Therefore the updates are exact at every history and achieves zero model-induced error in V and Q .

Quantum bisimulation distance. Classical bisimulation metrics compare states through the rewards they produce and the future states they lead to, giving a planning-aware notion of state similarity [8]. We extend this to quantum world models to quantify robustness: two quantum memories are close when replacing one by the other has little effect on the values and action-values induced by planning. We use subnormalized quantum states τ , whose trace is the branch probability and whose normalized state is the posterior memory. Define the branch-weighted reward $r_Q(\tau, a) = \sum_y r_y \text{Tr}[\mathcal{E}_y^a(\tau)]$. The quantum bisimulation distance is the fixed point

$$d_Q(\tau, \omega) = \max_{a \in \mathcal{A}} \left[(1 - \gamma) |r_Q(\tau, a) - r_Q(\omega, a)| + \gamma \sum_{y \in \mathcal{Y}} d_Q(\mathcal{E}_y^a(\tau), \mathcal{E}_y^a(\omega)) \right].$$

The recursion compares immediate reward consequences and then compares the future branch memories reached under each possible outcome. This distance has direct planning meaning. For the optimal value and action-value functions, $|\tilde{V}_Q^*(\tau) - \tilde{V}_Q^*(\omega)| \leq d_Q(\tau, \omega)/(1 - \gamma)$ and $|\tilde{Q}_Q^*(\tau, a) - \tilde{Q}_Q^*(\omega, a)| \leq d_Q(\tau, \omega)/(1 - \gamma)$; if rewards are bounded by $R_{\max}$, then $d_Q(\tau, \omega) \leq R_{\max} \|\tau - \omega\|_1$. For the FRDN quantum world model, we analyze different setups where $d_Q(\rho_t, \rho_s) \approx |S(t) - S(s)|$. Thus the metric identifies precisely the distinctions between quantum memories that the available actions can convert into reward or future planning differences.

Outlook. Our results identify finite quantum memory as a resource for world modeling and show that memory itself matters: quantum memory can genuinely improve prediction and planning. This suggests a hybrid role for quantum processors as compact conditional simulators, queried by classical planners when rare histories make errors most costly. A natural next step is to understand when approximate quantum world models yield practical advantages.

[1] C. Aghamohammadi, S. P. Loomis, J. R. Mahoney, and J. P. Crutchfield. “*Extreme quantum memory advantage for rare-event sampling*”. Physical Review X **8**(1):011025 (2018).

- [2] A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. volume 9 of *Classics in Applied Mathematics*, SIAM (1994).
- [3] A. R. Cassandra, L. P. Kaelbling, and M. L. Littman. “Acting optimally in partially observable stochastic domains”. In *Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence*, AAAI’94, page 1023–1028, (1994).
- [4] S. W. Dharmadhikari. “Sufficient conditions for a stationary process to be a function of a finite Markov chain”. *The Annals of Mathematical Statistics* **34**(3):1033–1041, (1963).
- [5] S. W. Dharmadhikari and M. G. Nadkarni. “Some regular and non-regular functions of finite Markov chains”. *The Annals of Mathematical Statistics* **41**(1):207–213, (1970).
- [6] T. J. Elliott, M. Gu, A. J. P. Garner, and J. Thompson. “Quantum Adaptive Agents with Efficient Long-Term Memories”. *Phys. Rev. X* **12**:011007 (2022).
- [7] M. Fanizza, J. Lumbrellas, and A. Winter. “Quantum theory in finite dimension cannot explain every general process with finite memory”. *Communications in Mathematical Physics* **405**(2): 50 (2024).
- [8] N. Ferns, P. Panangaden, and D. Precup. “Metrics for Finite Markov Decision Processes.”. In *Uai*, volume 4, pages 162–169, (2004).
- [9] M. Fox and H. Rubin. “Functions of processes with Markovian states”. *The Annals of Mathematical Statistics* **39**(3): 938–946, (1968).
- [10] F. Ghafari, N. Tischler, J. Thompson, M. Gu, L. K. Shalm, V. B. Verma, S. W. Nam, R. B. Patel, H. M. Wiseman, and G. J. Pryde. “Dimensional quantum memory advantage in the simulation of stochastic processes”. *Physical Review X* **9**(4):041013 (2019).
- [11] C. Grimm, A. Barreto, S. Singh, and D. Silver. “The value equivalence principle for model-based reinforcement learning”. *Advances in neural information processing systems* **33**: 5541–5552, (2020).
- [12] M. Gu, K. Wiesner, E. Rieper, and V. Vedral. “Quantum mechanics can reduce the complexity of classical models”. *Nature communications* **3**(1): 762 (2012).
- [13] D. Ha and J. Schmidhuber. “World models”. *arXiv preprint arXiv:1803.10122* , (2018).
- [14] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi. “Dream to control: Learning behaviors by latent imagination”. In *International Conference on Learning Representations*, (2020).
- [15] H. Jaeger. “Observable Operator Models for Discrete Stochastic Time Series”. *Neural Comput.* **12**(6): 1371–1398 (2000).
- [16] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. “Planning and acting in partially observable stochastic domains”. *Artificial Intelligence* **101**(1): 99–134 (1998).
- [17] M. Littman and R. S. Sutton. “Predictive representations of state”. *Advances in neural information processing systems* **14**, (2001).
- [18] H. X. Liu and S. Feng. “Curse of rarity for autonomous vehicles”. *nature communications* **15**(1): 4808 (2024).
- [19] T. M. Moerland, J. Broekens, A. Plaat, and C. M. Jonker. “Model-based reinforcement learning: A survey”. *Foundations and Trends in Machine Learning* **16**(1): 1–118, (2023).
- [20] N. Meuleau, L. Peshkin, K.-E. Kim, and L. P. Kaelbling. “Learning finite-state controllers for partially observable environments”. *arXiv preprint arXiv:1301.6721* , (2013).
- [21] A. Monràs and A. Winter. “Quantum learning of classical stochastic processes: The completely positive realization problem”. *Journal of Mathematical Physics* **57**(1):015219, (2016).
- [22] J. Oh, S. Singh, and H. Lee. “Value prediction network”. *Advances in neural information processing systems* **30**, (2017).
- [23] M. O’Kelly, A. Sinha, H. Namkoong, R. Tedrake, and J. C. Duchi. “Scalable end-to-end autonomous vehicle testing via rare-event simulation”. *Advances in neural information processing systems* **31**, (2018).
- [24] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, et al. “Mastering atari, go, chess and shogi by planning with a learned model”. *Nature* **588**(7839): 604–609, (2020).
- [25] E. Seneta. *Non-negative Matrices and Markov Chains*. Springer (2006).
- [26] D. Silver, H. Hasselt, M. Hessel, T. Schaul, A. Guez, T. Harley, G. Dulac-Arnold, D. Reichert, N. Rabinowitz, A. Barreto, et al. “The predictron: End-to-end learning and planning”. In *International Conference on Machine Learning*, pages 3191–3199, (2017).
- [27] R. S. Sutton. “Dyna, an integrated architecture for learning, planning, and reacting”. *ACM Sigart Bulletin* **2**(4): 160–163, (1991).
- [28] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press (2018).

- [29] M. Vidyasagar. *“The Complete Realization Problem for Hidden Markov Models: A Survey and Some New Results”*. Mathematics of Control, Signals, and Systems **23**: 1–65 (2011).
- [30] M. Vidyasagar. *Hidden Markov Processes: Theory and Applications to Biology*. Princeton University Press (2014).