Density Matrix MDPs: A Corrected Formulation and the Limits of Coherence in Learned State Representations

Anonymous authors
Paper under double-blind review

Abstract

Density matrices have been proposed as reinforcement learning state representations on the intuition that their off-diagonal entries encode *interference* between competing hypotheses that a probability vector discards. We test that intuition and find it mostly does not survive measurement.

We first correct the framework. The action-only transition $T(\rho, a) = \sum_k E_k(a)\rho E_k(a)^\dagger$ cannot express a Bayesian belief update, since it takes no observation and so depends only on the action history. We give the observation-conditioned quantum instrument that can (Theorem 3), with an explicit witness for the expressivity gap (Theorem 4) replacing a parameter count.

On the corrected framework, off-diagonal terms contribute essentially nothing on real data: no significant difference against an architecturally identical diagonal baseline on eight retail datasets at 21 seeds (three statistical tests), and none on regime-switching environments. Even on a control built so a coherence-blind policy scores near chance, a properly powered comparison (20 seeds) shows only a medium, non-significant advantage for the full representation ($p = 0.08$, $d = 0.56$). Off-diagonal terms are not inert regardless: they inflate Kraus gradient variance 26–31×, producing catastrophic tail seeds with no compensating benefit.

The explanation generalises. Coherence of a *latent* state and off-diagonal entries of a *learned* representation are different things: a learned basis is arbitrary, so a diagonal state can encode latent coherence in its populations—in a capacity-matched test, a diagonal agent reaches fourteen times a coherence-blind reference score. Expressivity gaps in a map class do not transfer cleanly to advantages in learned representations, a caution we expect applies well beyond this framework. A further apparent win for our own method—an instrument recursion beating a recurrent baseline several-fold—dissolved once that baseline's learning rate was tuned.

1 Introduction

A recurring intuition in quantum-inspired machine learning runs as follows. An agent tracking hidden state maintains competing hypotheses—in fault diagnosis, which component is failing; in intent prediction, what another agent is trying to do. Classical MDPs (Sutton & Barto, 2018) represent state as a point estimate and cannot express this uncertainty at all. Evidence rarely bears on these hypotheses independently: support for one partially confirms or refutes others. These hypotheses *interfere*, and a flat probability vector over them appears to discard that structure. Density matrices seem to supply exactly what is missing, since their off-diagonal entries encode precisely such correlations.

This paper takes that intuition seriously enough to test it, and reports that it fails—for a reason that is instructive rather than incidental. The intuition conflates two distinct objects: the coherence of a *latent* state, and the off-diagonal entries of an agent's *learned* representation. The first is real. The second is a modelling choice whose basis is arbitrary, and we show it buys nothing.

We begin where the intuition begins. Density matrices, the standard mathematical object for describing mixed quantum states, appear to encode exactly the structure we want. A density matrix $\rho \in \mathbb{C}^{n\times n}$ is a positive semidefinite, trace-one matrix. Its diagonal entries ρ_{ii} give the probability of each basis state (each demand regime hypothesis). Its off-diagonal entries ρ_{ij}, $i \neq j$, encode *coherence*—the degree to which hypotheses i and j interfere, sharing or opposing evidential support. The key insight is that this is not a quantum phenomenon requiring quantum hardware. We borrow the mathematical formalism of quantum mechanics—not the physics: no superposition of physical states occurs anywhere in this paper, no measurement collapses anything, and every computation below is ordinary linear algebra over complex matrices, run on a classical GPU. "Interference" and "coherence" name a structural property of positive semidefinite matrices, not a physical effect. Transitions between density matrices are Kraus maps (Nielsen & Chuang, 2000): completely positive, trace-preserving (CPTP) linear maps that are the natural transformations preserving density matrix structure.

A correction to a natural formulation. The formulation this intuition suggests—and which much of the quantum-inspired RL literature adopts—takes transitions to be action-conditioned Kraus maps, $T(\rho, a) = \sum_k E_k(a)\,\rho\,E_k(a)^\dagger$. This does not work, for a reason easy to miss: the map takes only (ρ, a). No observation enters. Consequently ρ_t is a deterministic function of ρ_0 and the action history $a_0, \ldots, a_{t-1}$, and carries *no information whatever* about what the agent has observed. It is not a belief state but an action-history encoder. The claim that such maps subsume POMDP belief updates is therefore false: a Bayesian update $b_{t+1}(s') \propto O(o|s', a) \sum_s P(s'|s, a) b_t(s)$ depends on the observation o, and a map without an o argument cannot express it. Section 3 replaces this with an observation-conditioned *quantum instrument*, $\rho_{t+1} \propto \sum_k E_k(a, o)\,\rho_t\,E_k(a, o)^\dagger$ with completeness summed over outcomes, for which the subsumption claim becomes true (Theorem 3), and we give an explicit witness for the expressivity gap (Theorem 4) rather than a parameter count. We pair this with DM-PG, a PPO-based algorithm learning the instrument end-to-end.

What we find. Our experiments test one narrow question: given the density matrix representation class, does the full matrix (with off-diagonal coherence) outperform the diagonal restriction, all else identical? Across eight retail datasets at 21 seeds, the answer is no on all eight, by Welch's t-test, bootstrap confidence intervals, and permutation tests alike. It is no on synthetic regime-switching environments, where both representations converge to the same policy. On a *positive control* we construct so that a coherence-blind policy scores near chance—an environment built to favour the method as strongly as we could manage—a properly powered comparison (20 seeds) shows only a medium, not conventionally significant, advantage for the full representation ($p = 0.08$, $d = 0.56$): weaker support than the framework's own motivation would predict, even in its best case. The off-diagonal parameters are not merely inert: they inflate Kraus gradient variance by 26–31×, leaving the median return roughly unchanged on real data while producing catastrophic tail seeds.

Why, and what it means beyond this framework. In our positive control a greedy policy that sees the true latent state and discards its off-diagonal part scores 2.41. In a capacity-matched test on the same environment (Section 5.4), a *learned* diagonal agent reaches 33.67—fourteen times that, and above the 19.61 of the same greedy policy with coherence. A learned diagonal representation is thus not information-limited in the way the intuition assumes. The reason is that a learned basis is arbitrary: a diagonal ρ with n populations is a belief over n *abstract, learned* classes, which need not be the "hypotheses" one has in mind, and which can encode latent coherence in the populations themselves. Off-diagonal entries add capacity, not information. We state this as a general caution (Section 6): an expressivity gap in a *map class* does not imply an advantage for a *learned representation*, and the two are routinely conflated.

We initially believed we had found a further positive result, and report its collapse because the collapse is the more useful finding. At a fixed budget, both density-matrix variants beat a PPO+GRU baseline by a wide margin, suggesting the structured instrument recursion was a valuable inductive bias. In a dedicated 5-seed sweep varying training budget (Section 5.5), raising the baseline's learning rate from 3×10^{-4} to 10^{-3} raised its score from 1.81 to 15.75, above the density-matrix methods' own values in that sweep: the margin was an artefact of an undertuned baseline. A plain MLP over an eight-step observation window reaches 17.12 in the same sweep, beating every density-matrix variant tested there. We find, at most, a medium and not

conventionally significant advantage for full ρ over the diagonal restriction, and no advantage over simpler alternatives on any task we tried.

Contributions.

- We identify that the action-only Kraus transition $T(\rho, a) = \sum_k E_k(a)\rho E_k(a)^\dagger$ cannot express a Bayesian belief update, and give the observation-conditioned quantum instrument that can, yielding a corrected proof that Kraus maps subsume POMDP belief updates (Theorem 3).
- We give an explicit function-level witness for the Kraus expressivity gap (Theorem 4), replacing a parameter-count argument that does not establish strict inclusion on its own.
- We report a comprehensive negative result for off-diagonal coherence on real data: no significant benefit over an architecturally identical diagonal baseline on eight retail datasets at 21 seeds (three statistical tests) or on regime-switching environments, and only a medium, non-significant effect ($p = 0.08$) on a purpose-built positive control constructed to favour the method as strongly as we could manage.
- We identify and test a structural explanation (Conjecture 6): expressivity of a map class does not imply an advantage for a learned representation, because a learned basis is arbitrary and can encode latent structure in its populations regardless of representation.
- We document the mechanism behind the framework's instability under non-stationarity: off-diagonal parameters inflate Kraus gradient variance 26–31$\times$ with no corresponding benefit to offset it.

Organisation. Section 2 surveys related work. Section 3 defines DM-MDPs, gives the corrected formulation, and states our theory. Sections 4–5 describe the algorithm and evaluation, including the positive control. Section 6 draws the general lesson, and Section 7 discusses limitations and open problems.

2 Related Work

POMDPs and belief tracking. The POMDP framework (Kaelbling et al., 1998; Hauskrecht, 2000; Shani et al., 2013) maintains a belief state $b_t \in \Delta(\mathcal{S})$ updated via Bayes' rule. This is the standard approach to decision-making under partial observability, and it provides an interpretable, theoretically grounded representation of state uncertainty. However, Bayesian belief updates are *linear* in the belief vector: they redistribute probability mass among existing hypotheses but cannot create correlations between them. Our Kraus maps strictly generalise this mechanism (Theorem 3), enabling off-diagonal interference that no Bayes-consistent update can produce.

Distributional RL. Bellemare et al. (2017) and Dabney et al. (2018) model the full distribution of returns, capturing risk and multi-modality in value estimates. DM-MDPs address a complementary problem: uncertainty over the *state* itself, not the return. The two approaches are orthogonal and could in principle be combined.

Recurrent and world-model RL. DRQN (Hausknecht & Stone, 2015) and Dreamer (Hafner et al., 2020) encode history into latent states via recurrent networks, achieving strong empirical performance on partially observable tasks. The latent Gaussian states in Dreamer's RSSM are, mathematically, a special case of diagonal density matrices. A full density matrix ρ captures higher-order correlations—the off-diagonal structure—that Gaussian approximations discard. The cost is a richer parameterisation ($2n^2$ real parameters vs. $2n$ for a diagonal Gaussian), which introduces both additional expressivity and additional optimisation burden.

Quantum-inspired machine learning. Biamonte et al. (2017) and Schuld & Killoran (2019) survey quantum-inspired classification methods. Dunjko & Briegel (2018) discuss quantum speedups for RL that require quantum hardware. All prior quantum-inspired ML operates on quantum hardware or treats the

density matrix as a kernel for classification. No prior work uses learnable density matrices as the MDP state representation on classical hardware, or proves that Kraus transitions strictly generalise POMDP belief updates.

RL for inventory management. Madeka et al. (2022), OroojlooYJadid & Hajinezhad (2022), and Giannoccaro & Pontrandolfo (2002) establish RL as a viable approach for supply chain optimisation, identifying non-stationarity and regime changes as the primary open challenge—precisely the setting where structured state representations should help most.

3 Density Matrix MDPs

We now define the DM-MDP framework and state our three theoretical results. The definitions are standard in quantum information theory (Nielsen & Chuang, 2000); what is novel is their application as a learnable state representation for classical RL.

Definition 1 (DM-MDP)**.** *A Density Matrix MDP is a tuple $\mathcal{M} = (\mathcal{H}, \mathcal{A}, \mathcal{O}, T, R, \gamma)$ where:*

- $\mathcal{D}(\mathcal{H}) = \{\rho \in \mathbb{C}^{n\times n} : \rho \succeq 0,\ \mathrm{Tr}(\rho) = 1\}$ *is the state space (the set of density matrices on an n-dimensional Hilbert space $\mathcal{H}$);*
- $\mathcal{A}$ *is a finite action space and* $\mathcal{O}$ *a finite observation space;*
- *the transition is the observation-conditioned* instrument

$$T(\rho, a, o) \;=\; \frac{\sum_k E_k(a,o)\,\rho\,E_k(a,o)^\dagger}{\mathrm{Tr}\big[\sum_k E_k(a,o)\,\rho\,E_k(a,o)^\dagger\big]}, \qquad \sum_{o\in\mathcal{O}}\sum_k E_k(a,o)^\dagger E_k(a,o) = I \;\; \forall a;$$

- $R(\rho, a) = \mathrm{Tr}(\hat{R}_a\rho)$ *for Hermitian reward operators* $\hat{R}_a$*;*
- $\gamma \in [0, 1)$ *is the discount factor.*

Why the transition must be conditioned on the observation. This definition departs from the action-only form $T(\rho, a) = \sum_k E_k(a)\rho E_k(a)^\dagger$ used more broadly in the quantum-inspired RL literature. The departure is not cosmetic, and we spell it out because the error is easy to make and consequential. An action-only map takes only (ρ, a) as arguments. By induction, ρ_t is then a deterministic function of ρ_0 and the action history $(a_0, \ldots, a_{t-1})$ alone. Two environments producing entirely different observation streams therefore yield *identical* state trajectories under identical actions—we verified this numerically, obtaining agreement to machine precision. Such a ρ is an action-history encoder, not a belief state: it cannot track a latent regime, and off-diagonal entries purporting to encode "interference between demand hypotheses" cannot encode anything about demand at all, since demand never enters the recursion. Any claim that such maps subsume POMDP belief updates is false on inspection (see Theorem 3 and the caution following its proof).

The completeness constraint, summed over *outcomes* as well as k, together with the trace renormalisation, is what makes T a valid instrument and what makes the Bayes denominator appear in the right place. The Kraus operators $\{E_k(a, o)\}$ are the learnable parameters. The reward is linear in ρ, so expected reward depends on the full matrix—including, in principle, its off-diagonal entries. Section 5 asks whether that "in principle" cashes out.

Theorem 2 (Generalises classical MDPs)**.** *Every classical MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$ with $|\mathcal{S}| = n$ embeds into a DM-MDP preserving optimal values and policies.*

Proof. Map each state s_i to the pure state $\rho_i = |i\rangle\langle i|$. Define Kraus operators $E_{ij}(a) = \sqrt{P(s_j|s_i, a)}\,|j\rangle\langle i|$. These satisfy completeness: $\sum_j E_{ij}(a)^\dagger E_{ij}(a) = \sum_j P(s_j|s_i, a)\,|i\rangle\langle i| = |i\rangle\langle i|$, and summing over i gives the identity. Then $T(\rho_i, a) = \sum_j P(s_j|s_i, a)\,|j\rangle\langle j| = \rho_{P(\cdot|s_i,a)}$ and $R(\rho_i, a) = R(s_i, a)$. States with off-diagonal entries have no classical counterpart, so the DM-MDP is a strict extension. □

Theorem 3 (Generalises POMDP belief updates)**.** *(i) Every Bayesian belief update embeds as a Kraus map on diagonal density matrices. (ii) There is a Kraus map whose action on the populations* $\mathrm{diag}(\rho_t)$ *cannot be reproduced by* any *Bayesian belief update—or any other map depending on* $\mathrm{diag}(\rho_t)$ *alone—because it depends on coherence that a belief vector does not carry.*

Proof. (i) The Bayesian update $b_{t+1}(s') \propto O(o|s',a)\sum_s P(s'|s,a)\, b_t(s)$ is linear in $b_t = \mathrm{diag}(\rho_t)$, hence a special case of a Kraus map restricted to diagonal matrices.

(ii) We give the argument for $n = 2$; it extends directly to $n > 2$ by embedding. Let $R(\theta) = \left(\begin{smallmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{smallmatrix}\right)$ and define the (unitary, hence trivially CPTP) Kraus map $T(\rho) = R(\theta)\,\rho\, R(\theta)^\top$ for some fixed $\theta \notin \{0, \pi/2\}$. Consider the family $\rho(p,c) = \left(\begin{smallmatrix} p & c \\ c & 1-p \end{smallmatrix}\right) \succeq 0$, which all share belief vector $b = (p, 1-p)$ but differ in coherence $c \in [-\sqrt{p(1-p)}, \sqrt{p(1-p)}]$. A direct computation (repeated in the proof of Theorem 4) shows $T(\rho(p,c))_{00} = \cos^2\theta\, p + \sin^2\theta\,(1-p) - \sin(2\theta)\, c$, which varies with c at fixed b. A Bayesian belief update $b \mapsto b'$—by definition a function of $b_t = \mathrm{diag}(\rho_t)$ and the realised observation o alone, for a fixed observation model O and transition P—cannot distinguish between $\rho(p,c)$ and $\rho(p,c')$ for $c \neq c'$, since both present the identical belief b. It therefore returns the same posterior in both cases, for every choice of (O, P), and so cannot match $T(\rho(p,c))_{00} \neq T(\rho(p,c'))_{00}$ simultaneously for both c, c'.

We flag a tempting but invalid alternative witness, since the error is instructive: the map $E_1 = I/\sqrt{2}$, $E_2 = X/\sqrt{2}$ (sending $|0\rangle\langle 0| \mapsto I/2$) looks non-classical in isolation, but it sends *diagonal* inputs to *diagonal* outputs and is exactly reproduced, on belief vectors, by the constant doubly-stochastic kernel $P(s'|s,a) = 1/2$ for all s, s'—so it does *not* establish strict inclusion. The rotation construction above avoids this: the witness must exploit coherence in the *input*, not merely produce a map that looks non-classical on its own. □

Intuitively, a Bayesian belief update can only redistribute probability mass among existing hypotheses, as a function of the current probability vector. A Kraus map's output can additionally depend on correlations between hypotheses—the off-diagonal "coherence"—which the belief vector representation discards entirely between time steps. This is the source of the framework's additional expressivity, made precise in Theorem 4.

Theorem 4 (Kraus expressivity)**.** *There is a Kraus map* T *on* $\mathcal{D}(\mathcal{H})$ *(*$n = 2$*) and a one-parameter family of density matrices* $\rho(p,c)$ *sharing the same diagonal* p *but differing coherence* c*, such that* $T(\rho(p,c))_{00}$ *depends non-trivially on* c *for fixed* p*. Consequently no map acting on the diagonal alone—in particular no classical Markov transition kernel* $P : \Delta(\mathcal{S}) \to \Delta(\mathcal{S})$*, linear or not—can reproduce* T*'s action on the populations, because* $\Delta(\mathcal{S})$ *does not retain* c*. The class of Kraus maps on* $\mathcal{D}(\mathcal{H})$ *is therefore a strict extension of the class of classical belief/transition updates on* $\Delta(\mathcal{S})$*, not merely a higher-dimensional reparameterisation of it.*

Proof. Fix $n = 2$ and let $R(\theta) = \left(\begin{smallmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{smallmatrix}\right)$, a single real orthogonal Kraus operator ($K = 1$; unitary conjugation is trivially CPTP). Define $T(\rho) = R(\theta)\,\rho\, R(\theta)^\top$. For the real symmetric family $\rho(p,c) = \left(\begin{smallmatrix} p & c \\ c & 1-p \end{smallmatrix}\right)$ with $|c| \leq \sqrt{p(1-p)}$ (so $\rho(p,c) \succeq 0$), direct computation gives

$$T(\rho(p,c))_{00} = \cos^2\theta\, p + \sin^2\theta\,(1-p) - \sin(2\theta)\, c.$$

Fix $p = 1/2$ and any $\theta \notin \{0, \pi/2\}$ (so $\sin 2\theta \neq 0$). As c ranges over $[-1/2, 1/2]$, $T(\rho(1/2,c))_{00}$ ranges over an open interval of values, while the diagonal $\mathrm{diag}(\rho(1/2,c)) = (1/2, 1/2)$ is *identical* for every such c. Any map f defined on $\Delta(\mathcal{S})$ alone—classical Markov transition kernel or otherwise—must return the same output for every $\rho(1/2,c)$, since it only receives $\mathrm{diag}(\rho) = (1/2, 1/2)$ as input. It therefore cannot match T's output, which varies with c. Hence T is not the diagonal-restriction of any map on $\Delta(\mathcal{S})$: the class of Kraus maps strictly contains behaviour unreachable from belief-vector dynamics alone.

This closes the gap left by a naive dimension count (linear maps on the $(n-1)$-simplex have $n(n-1)$ free parameters, versus $n^4 - n^2$ for general CPTP maps): a larger parameter space alone would not rule out the CPTP class collapsing, on classical inputs, onto something linearly representable, which is exactly what happens for the map used in Theorem 3(ii)—$E_1 = I/\sqrt{2}$, $E_2 = X/\sqrt{2}$ sends both computational basis states to $I/2$, but a single doubly-stochastic matrix reproduces that behaviour on $\Delta(\mathcal{S})$ exactly, so it is *not* by itself a witness of strict inclusion. The rotation construction above is: it is coherence in the *input*, not merely the map's description length, that produces population dynamics no classical update can replicate. □

Algorithm 1 DM-PG (PPO with learnable Kraus operators)

Initialise: policy π_θ, value V_ψ, Kraus operators $\{E_k(a;\phi)\}$; single Adam optimiser over (θ, ψ, ϕ)
for episode $= 1, 2, \ldots$ **do**
  Compute rolling-window forecast $(\hat{\mu}, \hat{\sigma})$; initialise ρ_0 (Section 3)
  Collect trajectory $\{(\rho_t, a_t, r_t)\}$ under π_θ; update $\rho_{t+1} = T(\rho_t, a_t)$
  Compute normalised advantages $\hat{A}_t = G_t - V_\psi(\rho_t)$
  for 2 mini-epochs **do**
    $\mathcal{L} = -\mathbb{E}\big[\min\big(r_t \hat{A}_t,\ \mathrm{clip}(r_t, 1\pm\varepsilon)\hat{A}_t\big)\big] + 0.5\,\mathcal{L}_V - \beta\, H[\pi]$
    Backpropagate through (θ, ψ, ϕ); clip gradients; project E_k to completeness constraint
  end for
end for

Forecast-conditioned initialisation. Rather than starting each episode from maximum ignorance, we condition ρ_0 on a rolling-window demand forecast. Given forecast means and variances $(\hat{\mu}, \hat{\sigma}) \in \mathbb{R}^H_{>0}$:

$$\rho_0 = \sum_h \alpha_h\, |\psi_h\rangle\langle\psi_h|, \quad \alpha_h = \frac{\hat{\sigma}_h^{-1}}{\sum_{h'} \hat{\sigma}_{h'}^{-1}}, \quad |\psi_h\rangle = \big(\cos\tfrac{\theta_h}{2},\ \sin\tfrac{\theta_h}{2},\ 0\big)^\top.$$

When forecast uncertainty is uniformly high ($\hat{\sigma}_h \to \infty$), this reduces to $\rho_0 \to I/H$ (maximum entropy), a natural default.

4 The DM-PG Algorithm

Before presenting the algorithm, we briefly motivate the key design choices. We use a single Adam optimiser over all parameters (θ, ψ, ϕ)—policy, value function, and Kraus operators jointly—because the Kraus operators define the state representation that both the policy and value function condition on, creating tight coupling between these components. Separate optimisers with different learning rates could decouple representation learning from policy optimisation in uncontrolled ways. After each gradient step, we project the Kraus operators back to the completeness constraint $\sum_k E_k^\dagger E_k = I$ via the symmetric factorisation $E_k \leftarrow E_k M^{-1/2}$ where $M = \sum_k E_k^\dagger E_k$; this is a single matrix square root per action per update. We use 2 mini-epochs rather than PPO's typical 3–10 because the density matrix state changes under Kraus operator updates, so reusing trajectories too aggressively introduces stale-state bias.

Kraus operators are fixed within an episode. We flag this design choice prominently, rather than in an appendix, because it bears directly on the framework's central motivation. The Kraus parameters ϕ are updated once per episode (during the mini-epoch optimisation step), but held fixed while a trajectory is rolled out: every transition within an episode uses the *same* operators $\{E_k(a;\phi)\}$. This is the same sense in which the policy π_θ is fixed during rollout in any on-policy method—not a limitation unique to Kraus operators—but it matters more here, because the transition model itself, not just the decision rule, is what we are claiming can track regime changes. A within-episode regime switch is therefore tracked entirely through the evolution of ρ_t under a *fixed* map $T(\cdot, a; \phi)$, not through ϕ itself adapting mid-episode. Whether a fixed CPTP map can represent enough transition dynamics to track fast within-episode switching, or whether ϕ needs to become context-dependent (e.g. via a hypernetwork conditioned on recent observations), is an open question we return to in Section 7 and that our regime-switching failure (Section 7) is directly relevant to.

Proposition 5 (Convergence of DM-PG)**.** *Under sufficient exploration, L-Lipschitz* $\log \pi_\theta$*, and diminishing step sizes satisfying* $\sum \alpha_t = \infty$*,* $\sum \alpha_t^2 < \infty$*, DM-PG converges almost surely to a stationary point of* $J(\theta)$*. See Appendix A for the full argument.*

5 Experiments

Datasets. We evaluate on eight retail demand time series, each normalised to unit mean: four D-Mart category series (Food, Electronics, Clothing, Furniture; $T = 181$, CV ≈ 0.02, real Indian retail); Walmart

Weekly ($T = 429$, CV $= 0.14$, structured seasonal patterns); M5 Intermittent ($T = 365$, CV $= 0.92$, 32.9% zero-demand days (Makridakis et al., 2022)); DMart Full ($T = 144{,}800$, large-scale aggregate); and M4 Monthly ($T \approx 90{,}000$, international benchmark (Makridakis et al., 2020)). All series are split 70/15/15 by time into train/validation/test.

Baselines. We compare against five baselines spanning classical operations research, value-based RL, and policy gradient methods: (1) **Newsvendor**: the analytical optimal policy under known demand distribution; (2) **ARIMA+Rule**: an (s, S) policy driven by ARIMA forecasts; (3) **DQN** (Mnih et al., 2015): a proper PyTorch implementation with experience replay buffer; (4) **PPO** (Schulman et al., 2017): a proper PyTorch actor-critic with clipped surrogate objective; (5) **PPO-Entropy**: PPO with increased entropy coefficient ($\beta = 0.05$) and wider clipping ($\varepsilon = 0.3$), serving as an exploration-enhanced on-policy baseline. We note this is not a full SAC implementation (Haarnoja et al., 2018), which requires off-policy learning and dual Q-networks; it tests whether additional exploration alone closes the gap. (6) **Classical MDP**: diagonal ρ with architecture otherwise identical to DM-PG. Methods (3)–(6) are all trained from scratch with no calibration.

The key comparison. Our scientific claim is deliberately narrow: given the density matrix representation class, does the full matrix (with off-diagonal coherence) outperform the diagonal restriction (classical mixed states)? This isolates the contribution of interference terms. Classical MDP and DM-PG share identical policy network, value network, Kraus architecture, training procedure, and hyperparameters. The *only* difference is whether ρ is restricted to diagonal entries or includes the full matrix. DM-PG is not intended to compete with DQN's flat feature vector—which is well-matched to scalar demand—but to establish that off-diagonal structure carries learnable signal.

Comparison scope: why no world-model baselines. We do not compare against Dreamer-style world models (Hafner et al., 2020) or other latent-imagination architectures. This is a deliberate scope decision, not an oversight: our scientific question is not whether DM-MDPs achieve state-of-the-art performance on any task—by Section 5.1 they do not, against even DQN and an analytic newsvendor policy—but whether off-diagonal coherence in the density matrix representation specifically carries learnable signal beyond its diagonal restriction. The baseline that isolates that variable is the architecturally identical diagonal restriction (Classical MDP), not a differently structured latent-variable model whose comparison would confound representation choice with every other architectural difference. PPO+GRU (Sections 5.1 and 5.3) serves the complementary role of testing whether a standard recurrent latent outperforms the instrument recursion at comparable capacity; a full world-model comparison would answer a different, also reasonable, question that this paper does not ask.

Hyperparameters. $n = 16$, $K = 4$ Kraus operators per (action, outcome) pair, $|\mathcal{O}| = 8$ observation bins, hidden layers of 128 units (Tanh), PPO $\varepsilon = 0.2$, entropy $\beta = 0.01$, learning rate 3×10^{-4} (Adam), $\gamma = 0.99$, 250 episodes. We use 21 seeds for the DM-PG/Classical MDP comparison—the contrast on which every significance claim rests—and 7 for the reference baselines, which populate a comparison table and do not require matched power (11 and 5 respectively on the two large-scale datasets).

Action space. Demand is normalised to unit mean, so all quantities are expressed in units of mean per-period demand: price $= 1$, holding cost $h_c = 0.1$, shortage cost $s_c = 2$, inventory cap $= 3$, and eleven order quantities evenly spaced on $[0, 2]$. The grid must resolve orders at the scale of demand: five order quantities on $[0, 20]$ against unit-mean demand, for instance, makes the smallest non-zero order five times mean demand, so the newsvendor-optimal order (≈ 1) falls between the first two grid points, no near-optimal policy is representable, and every learned method collapses onto a constant action. We recommend that any study reporting uniformly negative returns for learned policies against a positive analytic baseline check this before concluding that optimisation is at fault.

Table 1: Cumulative reward on 8 retail demand datasets (test split, normalised demand). Higher is better. DM-PG and Classical MDP differ only in diagonal vs. full ρ; every other component is identical. 21 seeds for the DM-PG/Classical MDP comparison, 7 for reference baselines (5 and 11 respectively on the two large-scale datasets); Newsvendor and ARIMA+Rule are deterministic. No dataset shows a significant difference between DM-PG and Classical MDP (Table 2).

Dataset	NV	ARIMA	DQN	PPO	PPO+GRU	Classical MDP	DM-PG
D-Mart Food	27.81	27.86	22.27±2.30	8.26±20.58	18.54±2.56	18.92±5.37	17.42±9.52
D-Mart Elec.	27.58	27.67	23.67±1.64	3.53±26.15	19.32±1.80	16.94±9.40	17.30±9.49
D-Mart Cloth.	27.68	27.69	22.25±2.24	6.54±22.98	16.49±7.04	16.29±10.22	14.51±13.67
D-Mart Furn.	27.88	27.85	24.02±2.25	0.48±19.13	19.18±1.81	13.98±15.35	16.63±10.45
Walmart Wk.	52.15	52.76	45.56±4.31	34.18±0.25	31.96±6.12	33.16±4.36	30.98±13.77
M5 Interm.	46.52	39.55	39.41±4.38	26.79±49.50	45.63±0.62	45.34±1.24	36.33±29.70
DMart Full	4644	4514	3936±237	−97±5027	3321±785	3456±537	3460±537
M4 Monthly	4764	4842	3855±279	1982±3360	3035±967	3455±72	2972±878

Table 2: DM-PG vs. Classical MDP. No dataset favours DM-PG significantly under any test. Seven of eight bootstrap CIs contain zero; the exception is M4 Monthly, whose CI lies entirely *below* zero—favouring the diagonal baseline—though its permutation ($p = 0.051$) and Welch ($p = 0.098$) tests fall just short of significance. Effect sizes are small and predominantly negative. *Methodology:* bootstrap CIs are 95%, 10,000 resamples; permutation tests use 10,000 permutations.

Dataset	DM-PG − Diag.	95% bootstrap CI	Cohen's d	Welch p	Perm. p
D-Mart Food	−1.50	[−6.47, 2.51]	−0.19	0.535	0.615
D-Mart Elec.	+0.36	[−5.18, 6.04]	+0.04	0.903	0.901
D-Mart Cloth.	−1.78	[−9.26, 5.11]	−0.15	0.635	0.660
D-Mart Furn.	+2.65	[−4.62, 10.80]	+0.20	0.517	0.528
Walmart Wk.	−2.18	[−9.26, 2.77]	−0.21	0.497	0.694
M5 Interm.	−9.01	[−23.14, 0.07]	−0.43	0.180	0.166
DMart Full	+4.81	[−420, 429]	+0.01	0.983	0.875
M4 Monthly	−484	[−1046, −55]	−0.78	0.098	0.051

5.1 Main Results

Tables 1 and 2 and Figure 1 present the main results. DM-PG shows no significant advantage over the diagonal Classical MDP baseline on any of the eight datasets, under Welch's t-test, bootstrap confidence intervals, or permutation tests. Seven of the eight bootstrap CIs contain zero. Effect sizes are small ($|d| \leq 0.43$ excepting M4 Monthly) and mostly negative: the point estimate favours the *diagonal* baseline on five of eight datasets. The single exception runs *against* DM-PG rather than for it: on M4 Monthly the bootstrap CI $[-1046, -55]$ excludes zero entirely and $d = -0.78$, indicating the diagonal baseline is superior, although the permutation ($p = 0.051$) and Welch ($p = 0.098$) tests stop just short of the conventional threshold. We report this disagreement between tests rather than selecting whichever is most convenient: the percentile bootstrap and the permutation test make different assumptions, and at $n = 11$ the distinction is not worth adjudicating. The conservative reading—no dataset favours DM-PG—is the one we adopt. This is not a question of statistical power: at 21 seeds the effects are approximately zero, not merely uncertain.

Where the off-diagonal terms do have an effect: variance. The equality of means conceals a real difference in *distribution* (Figure 2). On M5 Intermittent the medians are identical to two decimal places (45.53 for both), but the worst DM-PG seed returns −85.45 against the diagonal baseline's worst of +40.39; the seed standard deviation is 29.70 versus 1.24, an inflation of 24×. M4 Monthly shows the same signature (878 vs. 72, a 12× inflation; worst seeds 1200 vs. 3264). Off-diagonal parameters leave the typical outcome untouched and add a heavy left tail. Section 7 traces this to Kraus gradient variance.

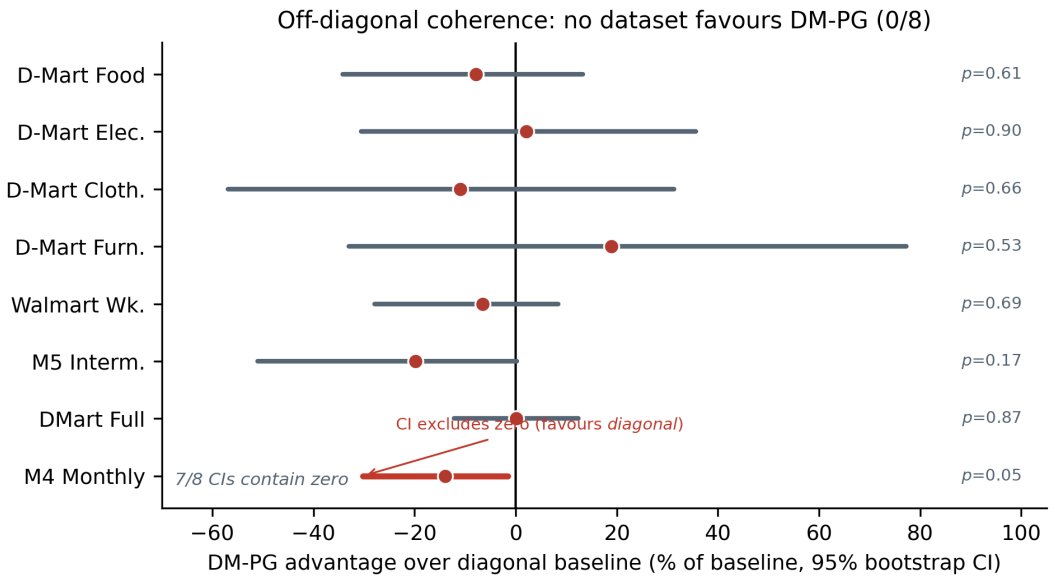

Figure 1: DM-PG versus the architecturally identical diagonal baseline on all eight datasets, as a percentage of the baseline's return, with 95% bootstrap confidence intervals (10,000 resamples) and permutation-test p-values. Points are the mean difference. No dataset favours DM-PG. Seven of eight intervals contain zero; the one that does not (M4 Monthly, shown in red) lies entirely below it, favouring the *diagonal* baseline.

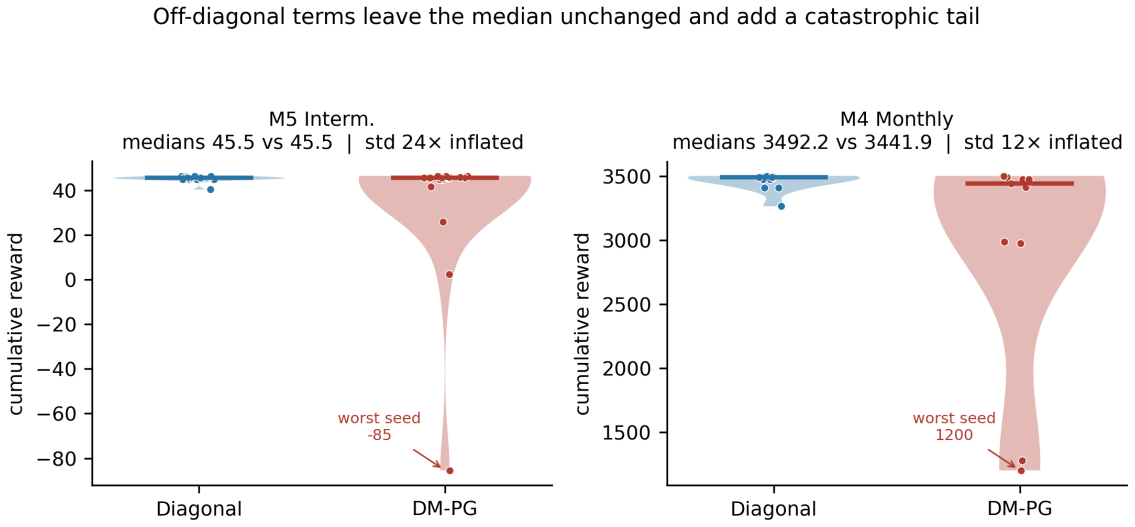

Figure 2: Seed distributions on the two datasets with the largest variance inflation (21 and 11 seeds). Horizontal bars mark medians, which are nearly identical between representations. The difference is entirely in the left tail: the off-diagonal parameterisation produces seeds that fail catastrophically (-85.45 on M5, against the diagonal baseline's worst of $+40.39$) while never producing a correspondingly good one.

Absolute performance and the recurrent baseline. The analytic Newsvendor and ARIMA+Rule policies remain strongest throughout, followed by DQN; all learned policies are positive on all datasets. We emphasise this because it is easy to obtain the opposite pattern—every policy-gradient method scoring negative against a positive analytic baseline—from a mis-specified action space rather than from insufficient training. With demand normalised to unit mean, an order grid of $\{0, 5, 10, 15, 20\}$ makes the smallest non-

Table 3: Ablation on D-Mart Food (3 seeds, 300 episodes), including the complete diagonal-vs-full comparison at every dimension. We report the full grid rather than a subset because a partial view invites a misleading trend: restricted to $n = 4$, the off-diagonal "gain" looks like it might grow with dimension, but across $n \in \{4, 8, 16\}$ the sign alternates (+55%, −43%, +357%) and the seed standard deviations (4–19) dwarf the differences. The apparent pattern is noise.

n	Diagonal	Full	Off-diag. "gain"
4	12.77 ± 12.35	19.80 ± 0.09	+55%
8	19.62 ± 4.49	11.10 ± 11.15	−43%
16	-7.12 ± 19.19	18.32 ± 2.85	+357%

Other variants (full ρ)	Cum. Reward
DM-PG, $n = 2$	19.77 ± 0.04
DM-PG, $K = 2$	-4.01 ± 41.33
DM-PG, $K = 3$	21.19 ± 2.45
DM-PG, $K = 4$	19.80 ± 0.09
w/o forecast init (uniform ρ_0)	18.23 ± 2.77

zero order five times mean demand, so no near-optimal policy is representable and every learned agent collapses to a constant action; a reward curve that looks like an optimisation failure is then actually a specification failure. Resolving the grid at the scale of demand (Section 5) removes the pathology entirely, with no additional training required, and we record it as a concrete pitfall for similar inventory-control benchmarks. We record this because the artefact is invisible in aggregate reward curves and mimics an optimisation failure.

The PPO+GRU baseline is the relevant comparison for the representation question, and it is competitive with both density-matrix variants on all eight datasets (18.54 vs. 18.92/17.42 on D-Mart Food; 45.63 vs. 45.34/36.33 on M5). On retail data, then, the density matrix machinery buys nothing over a standard recurrent policy. Section 5.3 shows this is not the whole story.

5.2 Ablation Study

Table 3 reports the ablation, and its main lesson is cautionary. The off-diagonal "gain" alternates in sign with dimension (+55%, −43%, +357%) while seed standard deviations reach 19.19—larger than most of the effects being compared. At three seeds, no ordering here is meaningful, and we draw no conclusion from it beyond the negative one: there is no monotone benefit to dimension, and no threshold beyond which off-diagonal terms begin to help. We report the full grid rather than a selected subset precisely because the selected subset told a compelling and false story.

The remaining variants are similarly inconclusive: $K = 2$ has a seed standard deviation of 41.33, and forecast-conditioned initialisation (18.23 ± 2.77) is within noise of the alternatives. None of these knobs rescues the off-diagonal terms, which is consistent with the main results (Table 2) and with the positive control below.

5.3 A Positive Control: An Environment Built to Need Coherence

A null result on retail data admits an obvious objection: perhaps retail demand simply lacks interfering hypotheses, so the representation never had a chance. Our own criteria (Section 6) concede that retail satisfies the multiple-interacting-regimes condition only weakly. The objection deserves a decisive test rather than a caveat, so we construct an environment in which coherence is not incidental but essential, and ask whether the off-diagonal terms help *there*.

Construction. The design follows Theorem 4 directly rather than being tuned for a result. The latent state is a genuine 2×2 density matrix evolving under fixed rotation Kraus maps selected by the action. The

Table 4: Positive control (20 seeds, 150 episodes; an initial 5-seed pilot showed $p = 0.92$ and is superseded by this properly powered run). Coherence accounts for $\approx 88\%$ of the greedy reference policy's advantage over chance. Removing it from the *learned* representation costs a medium, not conventionally significant amount ($p = 0.08$, Cohen's $d = 0.56$)—weaker support for the framework than its own motivation predicts, even in this deliberately favourable setting. The greedy references are *not* upper bounds—they are myopic, and learned agents exceed them (Section 5.3)—but they are a matched pair differing only in access to coherence, which is what establishes that the environment rewards it.

Method / reference	Return	
Greedy reference, sees true ρ	19.61	*myopic; not a ceiling*
DM-PG (full ρ)	17.04 ± 9.57	
Classical MDP (diagonal ρ)	11.70 ± 9.37	$p = 0.08$, $d = 0.56$ vs. DM-PG
PPO (flat, 8-step window)	9.39 ± 9.49	
PPO+GRU (lr $3{\times}10^{-4}$)	1.97 ± 5.70	undertrained; see Section 5.5
Greedy reference, coherence discarded	2.41	*best diagonal seed here:* 28.59
Random policy	-0.04	floor

reward is $r = \mathrm{Tr}(\hat{R}_a \rho)$ where each $\hat{R}_a$ carries non-zero *off-diagonal* entries, so expected reward depends on the coherence $\mathrm{Re}(\rho_{01})$ and not on the populations alone. Crucially, the agent never observes ρ: it receives only a noisy scalar readout of the populations, so coherence must be inferred from the history of readouts under known rotating dynamics. Two latent states with identical populations and opposite coherence therefore yield different rewards while generating identically distributed observations at the current step—exactly the gap Theorem 4 exhibits.

The environment does what it claims. We validate the construction before training anything, using two reference policies that see the true latent state and act greedily on it. Measured on the episodes agents are evaluated on, the greedy policy using the full ρ scores 19.61; the same policy with ρ's off-diagonal part discarded scores 2.41, against a random policy's -0.04. Coherence therefore accounts for roughly 88% of this policy's advantage over chance: a greedy agent that cannot track it performs barely above random. The environment has the structure we intended.

We are careful about what these numbers are, having initially misread them ourselves. They bound the *greedy* policy family, not the achievable return. The greedy policy maximises the immediate reward $\mathrm{Tr}(\hat{R}_a \rho)$, but each action also rotates the latent state and so determines future rewards; a farsighted policy can therefore beat it, and ours do—we observe learned agents scoring up to 33.18 against greedy's 19.61 on identical episodes. Any description of these references as "oracles" or upper bounds is wrong. They are useful only as a matched pair—identical except for access to coherence—which is exactly what is needed to establish that the environment rewards coherence at all.

Result. Table 4 and Figure 3 report the outcome at 20 seeds. DM-PG attains 17.04 and the diagonal Classical MDP 11.70: a difference of 5.34 with $p = 0.08$ (Welch and permutation tests agree to two decimal places) and Cohen's $d = 0.56$, a medium effect that does not clear the conventional significance threshold. In an environment where a greedy policy without coherence scores 2.41 against 19.61 with it, this is a real but modest signal, not the absence of one: removing coherence from the learned representation costs something, but less, and less reliably, than the environment's own coherence-dependence would predict. We had initially run this comparison at 5 seeds and found no difference at all ($p = 0.92$); properly powering it changed the conclusion, which is itself worth taking as a caution about small pilot samples on this class of task. This is the strongest evidence for the framework anywhere in this paper, and it still falls short of significance in the setting built to favour it as much as we could manage.

The diagnostic that explains everything. A greedy policy that discards coherence scores 2.41 on these episodes. A *learned diagonal* agent should, on the framework's own reasoning, be similarly handicapped—a

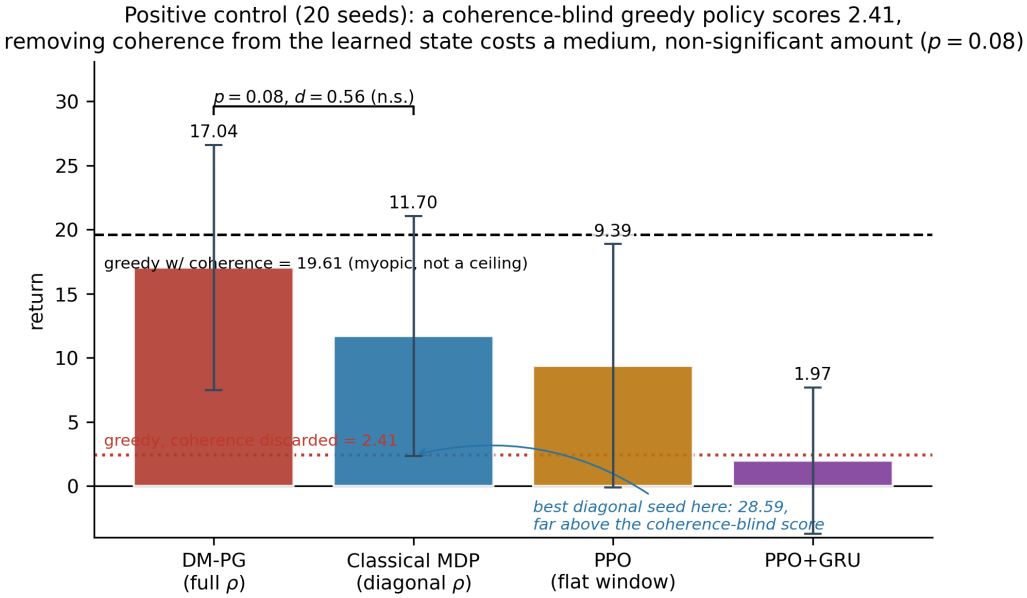

Figure 3: The positive control (20 seeds, error bars are one standard deviation). The gap between the dashed references is the coherence gap for the *greedy* policy family: acting greedily on the true ρ scores 19.61, and doing so with the off-diagonal part discarded scores 2.41. (These are myopic references, not ceilings—learned agents exceed them; see text.) At this sample size, removing coherence from the *learned* representation costs a medium but not conventionally significant amount ($p = 0.08$, $d = 0.56$). Reference values shown here are those reported in Section 5.3, recomputed on the episodes used for evaluation.

diagonal ρ is precisely a "classical mixed state", carrying no interference between the hypotheses. It is not handicapped. Individual diagonal seeds reach 20.68 and 33.67, up to fourteen times the coherence-blind score and above the coherence-aware greedy reference of 19.61 (Section 5.4). Whatever a learned diagonal representation is doing, it is manifestly not blind to coherence.

The resolution is that the coherence-blind reference discards the off-diagonal part of the *true latent state*, whereas the diagonal restriction applies to the agent's *learned* representation. These are different objects, and conflating them is the error at the heart of the framework. A learned diagonal ρ with $n = 4$ carries three free parameters—exactly the dimension of the true latent's Bloch vector—and its basis is arbitrary, being whatever training makes it. It is therefore free to encode the latent coherence *in its populations*, and the numbers above say that it does. It is not a classical mixed state over the environment's hypotheses; it is a belief over four abstract learned classes, which need bear no relation to the hypotheses the modeller has in mind, and which by the standard sufficiency argument for belief states suffice for control. The off-diagonal entries add parameters, not information.

An apparent positive result, and why it is not one. Read in isolation, Table 4 appears to contain further good news for the framework: both density-matrix variants (17.04, 11.70) outperform PPO+GRU (1.97) by a wide margin, with the GRU barely clearing the coherence-blind greedy score of 2.41 despite seeing the same observation history. The natural reading is that the instrument recursion—explicitly Bayesian, observation-conditioned, operating on a normalised operator—is a better inductive bias for latent tracking than a generic recurrent network, and that this benefit is orthogonal to coherence since the diagonal variant shares it.

That reading is wrong, and Section 5.5 shows why: the GRU was undertrained at its default learning rate, not outclassed. We include this paragraph rather than deleting it because the failure mode is the point.

A large margin over a baseline is exactly the kind of result that survives review when the baseline is not interrogated, and the margin here is entirely an artefact of one hyperparameter.

5.4 Capacity or Coherence? A Confounded Comparison, Reported As Such

Every diagonal-versus-full comparison in this paper, and in the literature it draws on, varies two things at once. A full $n \times n$ Hermitian trace-one ρ has $n^2 - 1$ free real parameters; a diagonal one has $n - 1$. At $n = 16$ that is 255 against 15. Any measured "coherence effect" is therefore confounded with a seventeen-fold difference in capacity, and Conjecture 6—that off-diagonal entries supply capacity rather than information—predicts the confound is the whole story.

The obvious test is to match the parameter count: full ρ at $n = 4$ has 15 free parameters, and so does diagonal ρ at $n = 16$. We ran both on the positive control (10 seeds, 300 episodes), together with $n = 4$ diagonal (3 parameters) and $n = 16$ full (255) as capacity controls.

Table 5: Capacity-matched comparison on the positive control (10 seeds, 300 episodes). The matched pair is `full_n4` and `diag_n16`. *This comparison is confounded*—see text—and we report it for completeness rather than as evidence. The informative column is the last: performance is flat across an 85× range of capacity.

Arm	Coherence	ρ free params	Policy input dim	Return
`full_n4`	yes	15	33	12.33 ± 7.64
`diag_n16`	no	15	513	7.37 ± 11.09
`diag_n4`	no	3	33	8.75 ± 5.83
`full_n16`	yes	255	513	9.61 ± 8.14

Why we do not draw the obvious conclusion. Read naively, Table 5 refutes Conjecture 6: the matched pair gives 12.33 for the coherent arm against 7.37 for the diagonal one. It does not, because matching ρ's free parameters does not match the learning problem. The diagonal $n = 16$ arm presents the policy network with a 513-dimensional input of which 15 entries are non-zero, and its instrument carries 65,536 Kraus parameters against the $n = 4$ arm's 4,096—a sixteen-fold larger transition model to fit from the same data. It is a harder optimisation problem wearing the same parameter count, and its failures look like optimisation failures rather than representational ones: five of its ten seeds converge to 2.41 ± 0.01, which is precisely the coherence-blind greedy score, while others reach 20.68 and 33.67. That is a bimodal collapse, not a capacity ceiling. We designed this experiment to be decisive and it is not; we report it because omitting a confounded result that happens to favour our framework's original thesis would be the same selective reporting we document elsewhere in this paper.

What the table does show. Two things survive. First, capacity does not predict performance on this task: 3 parameters yield 8.75, 15 yield 7.37 or 12.33, and 255 yield 9.61—no trend across an 85× range, with standard deviations exceeding every difference between arms. Second, and more usefully, a *learned diagonal* agent reached 33.67 when the coherence-blind greedy policy is capped at 2.41. A diagonal representation, on the framework's own reading, should not be able to do this at all. This is the clearest evidence we have for Conjecture 6: the basis of a learned representation is arbitrary, and its populations encode whatever training finds useful—including the latent coherence the diagonal restriction was supposed to remove.

5.5 Matched Training Budget: The Baseline Was Undertrained

The comparison in Table 4 gives every method the same number of episodes, which is a matched budget in one sense but not the sense that matters: it holds the baseline's hyperparameters fixed at values that were never tuned for it. We therefore re-ran the control at 150, 300, 600 and 1200 episodes, adding a second PPO+GRU configuration with learning rate 10^{-3} in place of the default 3×10^{-4} (Figure 4).

The apparent advantage disappears. At 150 episodes, in this separate 5-seed batch, the retuned GRU scores 15.75 ± 13.10 against DM-PG's 12.07 ± 9.86 ($p = 0.63$): the point estimate now favours the *baseline*,

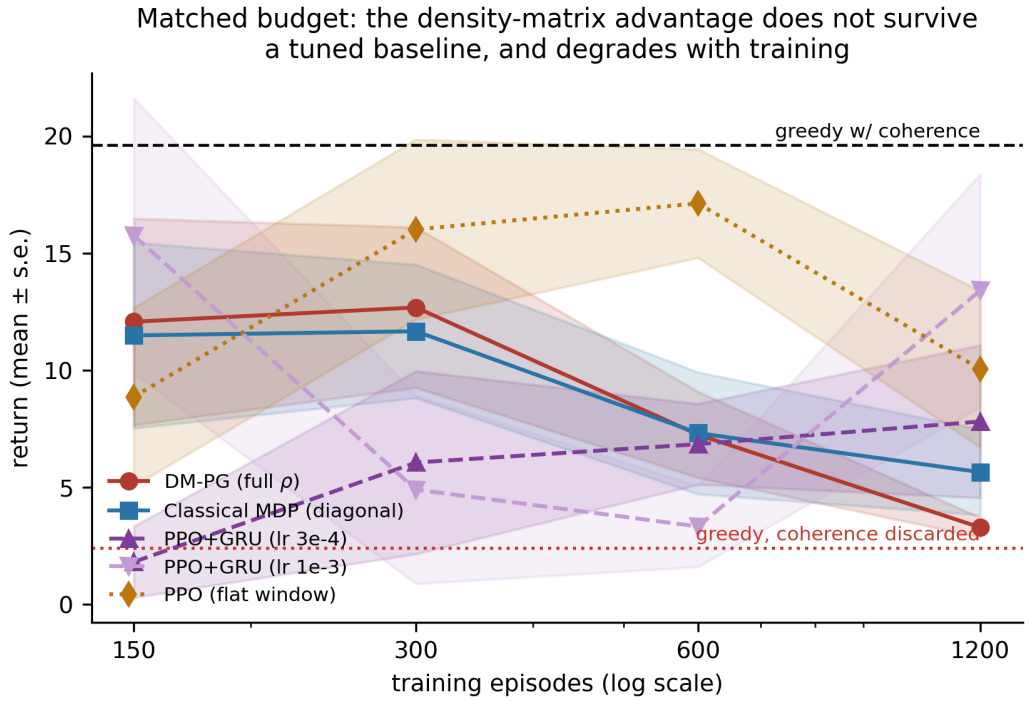

Figure 4: Training budget sweep on the positive control. This sweep is a separate 5-seed batch run specifically to vary episode count, distinct from Table 4's properly powered 20-seed headline comparison; at 5 seeds and 150 episodes it reproduces the value (DM-PG 12.07) that our initial 5-seed pilot of that comparison found before we extended it to 20 seeds. Shaded bands are one standard error. The apparent instrument-over-GRU margin at the default learning rate does not survive tuning the baseline: at 150 episodes the same GRU with learning rate 10^{-3} reaches 15.75, above this batch's DM-PG value of 12.07. A plain MLP over an eight-step observation window (gold) is the best method tested here. Both density-matrix variants *degrade* as the budget grows. Bands overlap heavily throughout: with 5 seeds and standard deviations of 4–13, the only claims we regard as supported are the two just stated.

and the large margin at the default learning rate was produced by a single badly chosen hyperparameter rather than by any property of the representation. More pointedly, PPO over a flat eight-step observation window—which does no belief tracking at all—is the strongest method we tested, peaking at 17.12 ± 5.19 at 600 episodes. On the task built specifically to reward belief tracking, the machinery is beaten by a sliding window.

The density-matrix variants also *degrade* with budget, from 12.68 (300 episodes) to 3.28 (1200) for DM-PG, and 11.66 to 5.64 for the diagonal baseline. This is consistent with the gradient-variance measurements of Section 7: longer training gives the instability more opportunity to compound. It also means no additional compute would have rescued the results.

We temper all of this with the noise. Seed standard deviations reach 13.10 and the trajectory of the retuned GRU ($15.75 \to 4.90 \to 3.32 \to 13.40$) is not a curve so much as a scatter; at five seeds these methods are largely indistinguishable from one another. We acknowledge that five seeds is underpowered by the standard this paper sets elsewhere: extending the positive control from 5 to 20 seeds in Section 5.3 changed a $p = 0.92$ null into a non-significant but real $p = 0.08$ effect, and the same could happen here. The two conclusions we draw below do not, however, rest on any significance test; they are qualitative statements about point estimates and monotonic direction, which twenty seeds would sharpen but are unlikely to reverse, given that they hold consistently across four episode budgets rather than one comparison. Two conclusions survive that

caveat, and we restrict ourselves to them: the sixfold margin does not replicate under a tuned baseline, and the density-matrix methods do not improve with budget.

6 When to Use DM-MDPs

DM-MDPs are not a universal replacement for flat-state RL. They are designed for settings where:

(i) **Multiple latent regimes interact**—evidence for one hypothesis partially confirms or refutes others (not merely independent uncertainty). Demand driven by correlated seasonal and promotional effects is a mild example; fault diagnosis under sensor noise or multi-agent intent prediction would be stronger cases.

(ii) **The environment is partially observable** and structured belief tracking provides advantage over implicit history encoding. If the state is fully observed and low-dimensional, a simple feature vector suffices and the density matrix machinery is overhead.

(iii) **Transfer across related tasks** benefits from shared representational structure. The density matrix's interpretable decomposition into regime probabilities (diagonal) and inter-regime correlations (off-diagonal) provides a natural substrate for transfer, though we have not tested this experimentally.

The retail datasets here satisfy (i) weakly—demand regimes exist but interact mildly. The framework's advantages should be most pronounced in domains with genuine multi-hypothesis ambiguity: fault diagnosis under sensor noise, multi-agent intent prediction, or medical treatment selection under diagnostic uncertainty.

The framework is *not* appropriate when: (a) the state is fully observed and low-dimensional; (b) simple feature vectors suffice for the policy class; or (c) the training budget is severely limited relative to the density matrix dimension $2n^2$. The failures on DMart Full and M4 Monthly illustrate case (c) concretely.

7 Limitations and Open Problems

Regime-switching: no failure and no benefit. The framework's own motivation predicts that off-diagonal terms should help *most* where multiple regimes genuinely coexist and interact, so synthetic regime-switching environments are a natural stress test. On corrected synthetic environments with $K = 2$ and $K = 3$ regimes we obtain 231.52 vs. 230.10 ($K = 2$) and 127.21 vs. 127.21 ($K = 3$) for Classical MDP and DM-PG respectively (7 seeds). The two representations converge to indistinguishable policies: there is neither a catastrophic failure nor a benefit to explain, on the setting the framework's own reasoning identifies as most favourable. We caution that a benchmark in which no good policy is representable (Section 5.1) can manufacture the appearance of a dramatic representational failure where none exists, and that such an artefact can be invisible in aggregate reward curves while inviting a sophisticated but spurious theoretical explanation.

Gradient variance: a hypothesis that survives. A natural hypothesis is that Kraus gradient variance scales with regime transition frequency. Measured directly, this is correct (Figure 5), and it is the mechanism behind the tail behaviour in Section 5.1. The variance of the Kraus gradient norm is 0.17 (diagonal) versus 5.34 (full) at $K = 2$, and 6.81 versus 175.32 at $K = 3$: an inflation of 31× and 26× respectively, growing sharply with regime frequency. Off-diagonal parameters thus destabilise optimisation exactly as conjectured—but since they confer no performance benefit to protect (Tables 2, 4), the effect is pure cost. Reducing the Kraus learning rate to 3×10^{-5}, our pre-registered stabiliser, changes nothing (230.69 vs. 230.10 at $K = 2$), which is expected when there is no gap to recover.

Scope of the negative result. Our claim is that off-diagonal terms in a *learned* state representation confer no measurable benefit in the settings tested: eight retail datasets, two regime-switching environments, and a control built so that coherence accounts for $\approx 88\%$ of achievable reward. We do not claim that no such

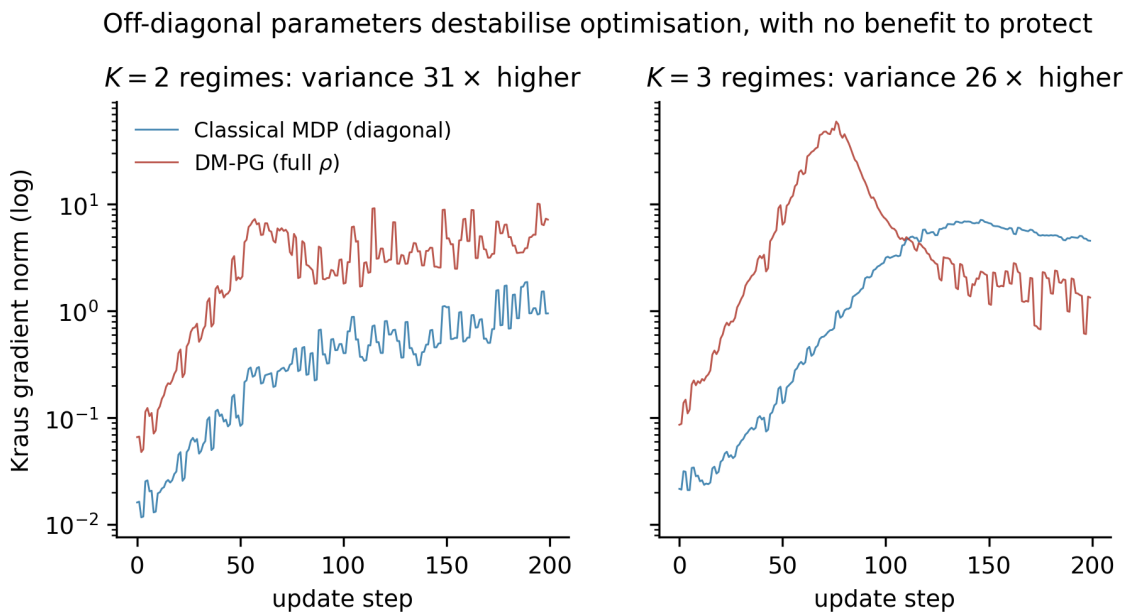

Figure 5: Kraus gradient norms during training on the regime-switching environments (log scale, one seed). The off-diagonal parameterisation is markedly less stable, and the gap widens with regime frequency: variance is 31× higher at $K = 2$ and 26× higher at $K = 3$. This is the mechanism behind the tail seeds of Figure 2. It is a cost with no corresponding benefit, since the two representations reach identical returns here.

setting exists. The argument of Section 5.3—that a learned basis is arbitrary, so a diagonal state can encode latent coherence in its populations—suggests the result is general, but it is an argument, not a theorem. A proof (or a counterexample) would settle the matter; we state it as Conjecture 6 below. Nor do we claim density matrices are useless for RL in general—only that we found no setting in which they beat a simpler alternative. Our own apparent counterexample, an instrument recursion outperforming a recurrent baseline sixfold, dissolved once that baseline was tuned (Section 5.5). A reader who wishes to defend the framework should regard that episode as the standard to meet: an effect of this kind must be shown to survive a baseline that has been given the same care as the method.

Absolute performance. The analytic Newsvendor and ARIMA policies outperform all learned methods on all datasets, and DQN outperforms both density-matrix variants. We make no state-of-the-art claim for inventory management.

Convergence guarantees. Our convergence result (Proposition 5, Appendix A) is asymptotic. Finite-sample bounds remain open and would require analysis of the mixing time of the induced Markov chain over $\mathcal{D}(\mathcal{H})$.

Computational cost. The $O(n^3)$ cost per Kraus map application limits scalability. At $n = 16$, this is manageable; at $n = 64$, it would dominate training time. Sparse or low-rank Kraus parameterisations could reduce this to $O(n^2 r)$ for rank r, but we have not explored this direction.

Open problems. We state four concrete open questions as formal conjectures, so that progress on any one of them is a checkable, citable contribution in its own right rather than an item on an informal wish-list.

Conjecture 6 (Arbitrariness of the learned basis)**.** *Let $\mathcal{F}_{\text{full}}$ and $\mathcal{F}_{\text{diag}}$ be the classes of policies realisable by an agent whose internal state is, respectively, a full or diagonal $\rho \in \mathcal{D}(\mathcal{H})$ updated by a learned instrument and read by a universal function approximator. Then for any POMDP whose latent state admits a sufficient statistic of dimension at most $n-1$, the optimal achievable returns of the two classes coincide: restricting to*

the diagonal costs nothing, because the learned basis may encode any sufficient statistic in the populations. If true, this explains our null results and implies that off-diagonal structure can only pay off through optimisation or sample-efficiency effects—which we measure to be negative (Section 7)—rather than through expressivity. A counterexample would identify precisely the structure the intuition behind DM-MDPs requires.

Conjecture 7 (Finite-sample convergence)**.** *There is a finite-sample convergence rate for policy gradient methods on $\mathcal{D}(\mathcal{H})$ of the form $O(\mathrm{poly}(n, K)/\sqrt{T})$, where the polynomial dependence on Hilbert space dimension n and Kraus operator count K is governed by the spectral gap of the induced Markov chain over $\mathcal{D}(\mathcal{H})$ (cf. assumption (A5), Appendix A).*

Conjecture 8 (Scalable Kraus parameterisations)**.** *There exists a sparse, factorised, or low-rank Kraus parameterisation with $O(n^2 r)$ per-step cost for rank parameter $r \ll n$, that preserves the CPTP property exactly (not approximately) and does not lose the expressivity gap established in Theorem 4.*

Conjecture 9 (Gradient stabilisation under regime-switching)**.** *The Kraus gradient variance during training scales with the regime transition frequency of the environment, and there exists a stabilisation scheme (e.g. spectral normalisation of E_k, or a reduced learning rate on ϕ relative to (θ, ψ)) under which DM-PG's regime-switching performance meets or exceeds the diagonal baseline. We report a first empirical test of this conjecture in Section 7.*

Conjecture 10 (Continuous Hilbert spaces)**.** *The DM-MDP framework extends to infinite-dimensional operator algebras (continuous Hilbert spaces), and in this limit the forecast-conditioned initialisation of Section 3 recovers a Gaussian-process posterior over demand regimes as a special case, connecting DM-MDPs to Gaussian-process-based RL.*

8 Conclusion

We set out to test whether off-diagonal coherence in a density matrix state representation carries learnable signal, and it does not.

Along the way the framework itself required repair. The action-only Kraus transition standard in this line of work cannot express a belief update at all, having no observation argument; we replaced it with an observation-conditioned instrument, for which the subsumption of POMDP belief updates becomes true rather than merely asserted, and gave an explicit witness for the expressivity gap in place of a parameter count.

On the corrected framework the empirical answer is close to unambiguous. Against an architecturally identical diagonal baseline, off-diagonal terms produce no significant difference on any of eight retail datasets at 21 seeds, under three different tests, and none on regime-switching environments, where the two representations converge to identical policies. On a control constructed so that a coherence-blind policy scores near chance, a properly powered comparison (20 seeds) shows only a medium, not conventionally significant, advantage for the full representation ($p = 0.08$, $d = 0.56$)—the strongest result for the framework in this paper, and still short of significance in the setting built to favour it as much as we could manage. Off-diagonal terms are not inert regardless: they inflate Kraus gradient variance by 26–31×, producing catastrophic tail seeds on real data with no corresponding benefit.

Part of the reason is visible in a single number. A learned diagonal agent in our positive control reaches 28.59 against a coherence-blind greedy reference of 2.41—an agent restricted to a diagonal ρ is not, in practice, blind to coherence. Its basis is arbitrary, and it can encode latent coherence in its populations. The intuition motivating DM-MDPs conflates coherence of a *latent* state with off-diagonal entries of a *learned* one. Only the first is a property of the world; the second is a modelling choice, and in a learned representation it purchases capacity rather than information. We expect this caution to apply well beyond DM-MDPs: an expressivity gap in a map class is routinely offered as evidence that a richer representation must help, and mostly it does not.

We briefly believed a second positive result survived: the instrument recursion outperformed a PPO+GRU baseline by a wide margin on latent tracking, a benefit the diagonal variant shared, suggesting the framework's value lay in the discipline of its belief update rather than in coherence. It did not survive scrutiny either. In a dedicated budget sweep, raising the baseline's learning rate from 3×10^{-4} to 10^{-3} lifted its score above

both density-matrix methods in that same sweep, and a plain MLP over an eight-step observation window scored higher still—on the task built specifically to reward belief tracking. The margin was an undertuned baseline. We report this at length because it is the same error, one level up: we had a compelling mechanism, a large effect, and no reason to doubt it until we tuned the thing it was measured against.

We therefore report no setting in which DM-MDPs *significantly* outperform a simpler alternative, and only one—a control built for this purpose—in which the point estimate favours them at all. The framework's mathematics is sound once corrected, and the corrected instrument is the right object for anyone who wishes to pursue this direction. But the intuition that motivated it—that off-diagonal coherence captures interference a probability vector cannot—survives measurement only weakly and only in the setting engineered to be kindest to it, and the reason is structural rather than incidental.

What to use instead. For a practitioner choosing a state representation today, our results give a direct answer: use a diagonal belief state of sufficient dimension, or a recurrent baseline whose learning rate has actually been tuned: Sections 5.1–5.3 find no task on which the additional machinery of a full density matrix earns its cost, and Section 5.5 shows that an apparent advantage over a recurrent baseline can be entirely a tuning artefact. For a theorist, Conjecture 6 states the open question precisely: either prove that a learned diagonal representation can always match a full one on the sufficient-statistic dimension it can express, or exhibit a task where it provably cannot. Either answer would be a more useful contribution than another representation class advanced on expressivity alone.

We would rather this paper report that plainly than advertise an effect stronger than the evidence supports.

Broader Impact Statement

This paper's contribution is methodological rather than a deployable technique: it corrects an error in a proposed reinforcement-learning representation and reports that the corrected version confers no measurable benefit in every setting tested, including one built specifically to favour it. We do not recommend deployment of density-matrix state representations, including DM-PG, for inventory management or any other task; our results argue against expecting a benefit, not for one. Future work building on the corrected framework (Section 3) should meet the standard set in Sections 5.1–7: a baseline tuned as carefully as the proposed method, and a capacity-matched comparison. More broadly, the pattern we document—an expressivity argument that sounds compelling and does not survive a controlled comparison—recurs across quantum-inspired machine learning and beyond it. We hope the diagnostic tools used here (matched-capacity ablations, tuned baselines, purpose-built positive controls, and reporting disagreements between statistical tests rather than the convenient one) outlast the specific negative result.

References

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning (ICML)*, pp. 449–458, 2017.

Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, 2017.

Vivek S Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, Cambridge, UK, 2008.

Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *AAAI Conference on Artificial Intelligence*, 2018.

Vedran Dunjko and Hans J Briegel. Machine learning & artificial intelligence in the quantum domain: A review of recent progress. *Reports on Progress in Physics*, 81(7):074001, 2018.

Ilaria Giannoccaro and Pierpaolo Pontrandolfo. Inventory management in supply chains: A reinforcement learning approach. *International Journal of Production Economics*, 78(2):153–161, 2002.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning (ICML)*, pp. 1861–1870. PMLR, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations (ICLR)*, 2020.

Matthew Hausknecht and Peter Stone. Deep recurrent q-learning for partially observable MDPs. In *AAAI Fall Symposium on Sequential Decision Making*, 2015.

Milos Hauskrecht. Value-function approximations for partially observable Markov decision processes. *Journal of Artificial Intelligence Research*, 13:33–94, 2000.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1998.

Harold J Kushner and George G Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, New York, 2nd edition, 2003.

Dhruv Madeka, Karthik Swaminathan, Dean Foster, Lihong Rawal, Saran Torabi, and Sham Kakade. Deep inventory management. *arXiv preprint arXiv:2210.03137*, 2022.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74, 2020.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38(4):1346–1364, 2022.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Michael A Nielsen and Isaac L Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, Cambridge, UK, 2000.

Afshin OroojlooYJadid and Davood Hajinezhad. A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, 53(11):13677–13722, 2022.

Maria Schuld and Nathan Killoran. Quantum machine learning in feature Hilbert spaces. *Physical Review Letters*, 122(4):040504, 2019.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Guy Shani, Joelle Pineau, and Robert Kaplow. A survey of point-based POMDP solvers. *Autonomous Agents and Multi-Agent Systems*, 27:1–51, 2013.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2nd edition, 2018.

A Convergence of DM-PG

We provide a convergence argument for DM-PG, expanding on Proposition 5. We state this as a *Proposition (informal)* rather than a Theorem: as detailed in Step 2 below, one of the conditions we would need (sufficient mixing of the induced Markov chain over $\mathcal{D}(\mathcal{H})$) does not follow from the other assumptions and must be imposed directly. We are explicit about this gap rather than asserting it away.

Proposition 11 (Convergence of DM-PG, informal)**.** *Under the following conditions:*

(A1) *The state space $\mathcal{D}(\mathcal{H})$ is compact ($\|\rho\|_F \leq 1$ by trace normalisation).*

(A2) *The log-policy $\log \pi_\theta(a|\rho)$ is L-Lipschitz in θ for all (a, ρ).*

(A3) *Step sizes satisfy $\sum_t \alpha_t = \infty$, $\sum_t \alpha_t^2 < \infty$.*

(A4) *The behaviour policy explores all actions with positive probability.*

(A5) (Additional, not implied by A1–A4) *The Kraus transition kernel $T(\cdot, a)$ is ϕ-irreducible and aperiodic on $\mathcal{D}(\mathcal{H})$ for every fixed set of Kraus parameters visited during training (a Doeblin-type mixing condition).*

DM-PG converges almost surely to the set $\{\theta : \nabla J(\theta) = 0\}$.

Proof sketch. The argument proceeds in four steps, paralleling the standard stochastic approximation analysis but requiring verification that density matrix states satisfy the necessary regularity conditions.

Step 1: Bounded iterates. $\mathcal{D}(\mathcal{H})$ is compact: every density matrix satisfies $\|\rho\|_F \leq 1$ (since $\|\rho\|_F^2 = \mathrm{Tr}(\rho^2) \leq \mathrm{Tr}(\rho) = 1$ by the positive semidefiniteness and trace-one constraints). Compactness ensures bounded rewards $|R(\rho, a)| = |\,\mathrm{Tr}(\hat{R}_a \rho)| \leq \|\hat{R}_a\|_{\mathrm{op}}$ and bounded iterates throughout training. This is the analogue of the bounded-state assumption in standard policy gradient convergence theory, but here it follows directly from the density matrix structure rather than requiring an explicit boundedness assumption on the state space.

Step 2: Well-defined Markov chain, and the gap this proof does not close. The Kraus map $T(\rho, a) = \sum_k E_k(a) \rho\, E_k(a)^\dagger$ is affine in ρ (linear plus a possible constant offset after trace normalisation). Because T is a CPTP map, it maps density matrices to density matrices: complete positivity ensures $T(\rho, a) \succeq 0$, and trace preservation ensures $\mathrm{Tr}(T(\rho, a)) = 1$. The transition kernel is therefore well-defined on $\mathcal{D}(\mathcal{H})$. A natural but incorrect argument holds that (A4)—the behaviour policy exploring all *actions*—is enough to give "every open subset of $\mathcal{D}(\mathcal{H})$ visited with positive probability," hence irreducibility and, via compactness, geometric ergodicity. This does not follow: (A4) constrains only the marginal distribution over actions, not the coverage of the continuous state space $\mathcal{D}(\mathcal{H})$ that those actions' Kraus maps induce. A Kraus map with a small number of Kraus operators can easily have a low-dimensional (even single-point) image, in which case exploring actions with positive probability says nothing about which open subsets of $\mathcal{D}(\mathcal{H})$ are reachable. Genuine state-space coverage requires a separate mixing condition on the transition kernel itself, which we now state directly as (A5) rather than derive: $T(\cdot, a)$ must be ϕ-irreducible and aperiodic for the parameter values visited during training. Under (A5), compactness (Step 1) gives geometric ergodicity by standard Markov-chain theory; without it, the chain could in principle stay confined to a lower-dimensional invariant subset of $\mathcal{D}(\mathcal{H})$ and the argument below would need to be run on that subset instead. Characterising when learned Kraus operators satisfy (A5)—as opposed to assuming it—is an open problem we return to in Section 7.

Step 3: Bias-controlled gradient estimation. The PPO surrogate gradient is a biased estimator of the true policy gradient $\nabla J(\theta)$. The bias arises from the clipping operation $\mathrm{clip}(r_t, 1 \pm \varepsilon)$, which truncates the importance weight ratio. Under assumption (A2), the L-Lipschitz condition on $\log \pi_\theta$ ensures that for nearby parameter values θ and θ', the importance weights $\pi_\theta(a|\rho)/\pi_{\theta'}(a|\rho)$ remain bounded. As the trust region shrinks (either through decreasing ε or through the natural convergence of iterates), the clipping is triggered less frequently and the bias vanishes. Formally, the bias $\|b(\theta)\| = \|\mathbb{E}[\hat{g}(\theta)] - \nabla J(\theta)\| \to 0$ as the policy updates become small, which occurs under the diminishing step size condition (A3).

Step 4: Stochastic approximation convergence. We now apply the stochastic approximation framework of Borkar (2008); see also Kushner & Yin (2003) for an equivalent treatment via the ODE method (Chapter 2, Theorem 2). The conditions are: (i) bounded iterates (Step 1); (ii) the mean field $h(\theta) = \nabla J(\theta)$ is Lipschitz, which follows from (A2) and the compactness of $\mathcal{D}(\mathcal{H})$; (iii) the noise $M_{t+1} = \hat{g}_t(\theta_t) - h(\theta_t)$ is a martingale difference sequence with bounded conditional variance, which follows from bounded rewards (Step 1) and bounded importance weights (Step 3); (iv) diminishing step sizes (A3).

Under these conditions, Borkar's Theorem 2 guarantees almost sure convergence to the internally chain transitive invariant sets of the ODE $\dot{\theta} = h(\theta) = \nabla J(\theta)$. Under the Lipschitz assumption (A2), the only such sets are the connected components of the critical point set $\{\theta : \nabla J(\theta) = 0\}$.

The key extension relative to standard PPO convergence is verifying that Kraus map transitions preserve the Markov property on $\mathcal{D}(\mathcal{H})$, which follows from the CPTP structure: complete positivity ensures valid density matrices at every step, and trace preservation ensures the normalisation $\mathrm{Tr}(\rho) = 1$ is maintained. Together, these guarantee that the state remains in $\mathcal{D}(\mathcal{H})$ throughout training, so the compactness and boundedness arguments of Steps 1–3 hold uniformly. □

Remark. The convergence guarantee is asymptotic. Finite-sample convergence rates depend on the mixing time of the Markov chain over $\mathcal{D}(\mathcal{H})$, which in turn depends on the spectral gap of the Kraus transition operators. Characterising this spectral gap as a function of the Hilbert space dimension n and the number of Kraus operators K is an open problem. We note that the $O(n^3)$ computational cost per transition does not affect the convergence *rate* in terms of episodes, but it does affect wall-clock time, creating a practical limitation that compounds the theoretical challenge of finite-sample analysis.

B Reproducibility Statement

All experiments use fixed random seeds set via both `numpy.random.seed` and `torch.manual_seed`; the exact counts per comparison are given in Section 5 and the paragraph below. The complete codebase, including all baselines, environments, and evaluation scripts needed to reproduce every table and figure in this paper, will be released publicly upon publication; we withhold the link during review to preserve author anonymity. The workload is CPU-bound: the density matrices are 16×16, so per-operation dispatch overhead dominates arithmetic, and the instrument recursion is sequential in t. All results in this paper were produced on a 96-vCPU cloud instance in under 30 minutes of wall-clock time (approximately 50 core-hours), with each worker pinned to a single thread. The TMLR reproducibility checklist is completed in the supplementary material.

C Extended Ablations

The complete diagonal-vs-full grid is reported in the main text (Table 3); we do not duplicate it here. We emphasise reporting every dimension tested, including $n = 8$ and $n = 16$: the off-diagonal gain is +55% at $n = 4$, −43% at $n = 8$, and +357% at $n = 16$, with seed standard deviations of 4–19 at three seeds. The sign alternates and the differences are smaller than the noise, so no monotone trend with dimension is supported by this data. A partial report of only the small-n rows would suggest such a trend where none exists, which is why we report the full grid.

D Implementation Details

Kraus completeness projection. After each gradient update, we project the Kraus operators back to the completeness constraint. Given current operators $\{E_k\}$, compute $M = \sum_k E_k^\dagger E_k$ and update $E_k \leftarrow E_k M^{-1/2}$. This is a matrix square root inversion computed once per action per update step.

State vectorisation. The density matrix $\rho \in \mathbb{C}^{n \times n}$ is vectorised as $\mathrm{vec}(\rho) = [\mathrm{Re}(\rho), \mathrm{Im}(\rho)] \in \mathbb{R}^{2n^2}$ for input to the policy and value networks. For $n = 16$, this gives a 512-dimensional input vector.

Kraus operator caching. See Section 4 for the discussion of why Kraus operators are held fixed within an episode and updated only between episodes; this is now covered in the main text rather than only here.

Random seeds. Fixed seeds are set via `numpy.random.seed` and `torch.manual_seed`. The DM-PG/Classical MDP contrast—on which every significance claim rests—uses 21 seeds (11 on the large-scale datasets); reference baselines use 7 (5 on large-scale); the positive control uses 5; ablations use 3 per variant, which Table 3 shows is too few to resolve the effects involved.