OPERATOR-VALUED Q-LEARNING: COMPOSITIONAL BELLMAN RECURSIONS FOR ORDER-SENSITIVE DECISION MAKING

Anonymous authors
Paper under double-blind review

ABSTRACT

In quantum pulse sequences, spacecraft attitude control, and in-hand reorientation, actions act on a hidden physical state as non-commuting operators, so the outcome of a plan depends on the order of its actions. Standard value functions either observe this latent state and learn a nonlinear function of it, or record the action history, which in tabular form costs K^T entries for K actions and horizon T. We instead let the value function carry order: each $Q(s, a)$ is an $N \times N$ operator, and the Bellman backup applies the action's operator to the continuation value, $Q(s, a) = \mathcal{K}_a^\dagger\big(R(s, a) + \gamma\, \mathbb{E}\, V(s')\big)$. Unrolled along a plan, the recursion builds the ordered product of the actions' operators, and the value at any latent state is a linear trace readout. For policies that do not observe the latent, this representation is exact, the backup contracts for unitary and orthogonal dynamics, and an asynchronous operator temporal-difference (TD) scheme learns it with almost-sure convergence. Control needs a maximum over actions, which operators lack: by Kadison's anti-lattice theorem, self-adjoint operators generally have no least upper bound. An operator log-sum-exp instead gives one operator per state–action pair whose readout upper-bounds the optimal value at every latent state, reducing to soft Q-learning when actions commute. We also prove that scalar values on observations that ignore action order must confuse swapped histories, that exact history tables grow exponentially for free action groups but polynomially for nilpotent ones, and a Hankel-rank lower bound on the operator dimension. On random and $SO(3)$-pointing control tasks, operator Q-learning reaches a normalised score of 0.95 within 100 episodes and transfers without retraining to unseen initial states (0.95–0.99), where a recurrent baseline drops to 0.21–0.53. A DQN that observes the latent eventually matches or exceeds it, consistent with the certificate's slack.

1 INTRODUCTION

Value functions in reinforcement learning (RL) are scalars (Sutton & Barto, 2018; Watkins & Dayan, 1992), or, in distributional RL, probability measures (Bellemare et al., 2017); either way, their values can be combined in any order. Yet in quantum control, attitude control, and robotic manipulation, actions compose as non-commuting operators, and the order of a plan determines its outcome. The issue is particularly relevant when each episode is expensive, as in quantum and robotic control. Exact history-based fixes grow as K^T, and generic function approximators learn the order-dependence from data without guarantees. We show that the value function itself can carry this non-commutative structure, with exact representation, convergence and certificate guarantees. Entropy-regularised RL (Haarnoja et al., 2018; Ziebart et al., 2010) supplies the aggregation we use for control.

A simple example of the problem. Consider a rectangular object, and rotate it $90°$ about its vertical axis, then $90°$ about its horizontal axis. Now perform the same two rotations in the opposite order: the object ends up facing a different way. The same signature appears whenever a quantum gate is applied before versus after another: the objects being composed are matrices, and matrix multiplication does not commute. If the agent's observation records *which* actions were taken (a step counter, action counts, a task phase) but not *in which order*, a scalar Q-function on that observation

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

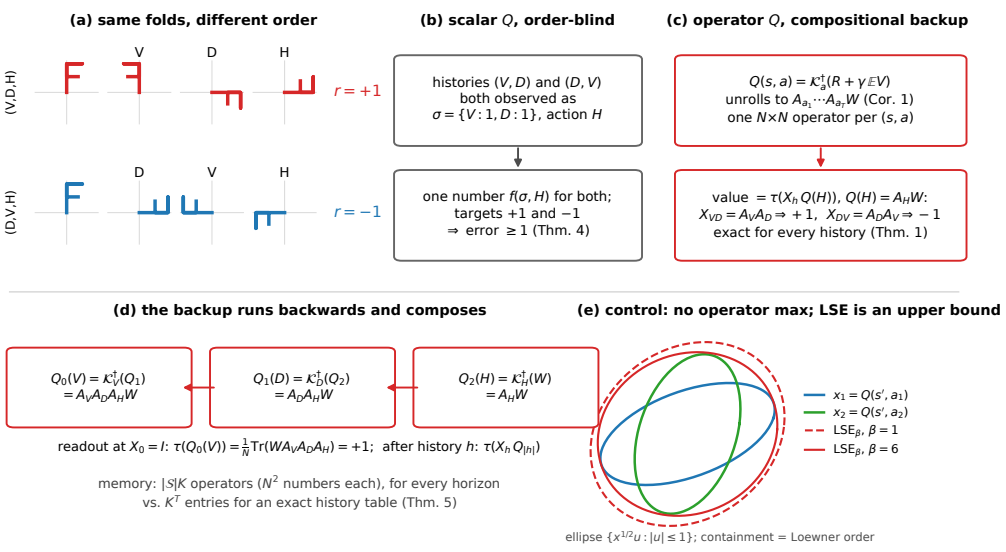


Figure 1: **The idea in one example** (Appendix A). (a) (V, D, H) and (D, V, H) use the same folds but end mirrored, with rewards ± 1. (b) After two steps both histories contain one V and one D, so a scalar value on this order-blind observation errs by at least 1 (Thm. 4). (c) The compositional backup forms the ordered product, and one operator $Q(H) = A_H W$, read against the history operator X_h, returns both values exactly (Thm. 1). (d) Unrolled, the recursion applies the dual maps backwards along the plan, so each stored operator already contains the rest of the plan (Cor. 1). (e) For control there is no operator maximum (Kadison); the operator log-sum-exp is a Löwner upper bound of every $Q(s', a')$ (ellipse containment), which yields the upper-bound certificate of Theorem 3.

assigns the same number to both histories. Section 5 makes this precise: with three steps or more, two histories that differ only by a swap receive values that no scalar function of an order-blind observation can both match (Theorem 4).

In **quantum control**, an agent selects unitary pulses $\{U_k = e^{-iH_k \Delta t}\}$; since $[H_j, H_k] \neq 0$ for generic Hamiltonians, pulse j then k reaches a physically different quantum state than k then j (Niu et al., 2019). In **rigid-body manipulation**, control inputs induce rotations composing the same way, formalised by $SO(3)$ (Schuck et al., 2025) and inherited by in-hand reorientation, spacecraft attitude control, and locomotion on $SE(3)$. In each domain a plan $a_1, \ldots, a_T$ has cumulative effect equal to an *ordered product* $X_T = A_{a_1} \cdots A_{a_T}$, with reward $r = \frac{1}{N} \operatorname{Tr}(W X_T)$ for an observable W. Figure 1 shows the phenomenon and the operator remedy, on a three-fold example worked out in Appendix A.

Observing the latent directly. If the latent X_t is observed, the process on (s_t, X_t) *is* Markov, and a scalar $Q(s, X, a)$ suffices in principle. The difficulty is then representational: X_t ranges over a continuous non-abelian group, so tabular methods do not apply and a generic approximator must learn a nonlinear-looking function on it. If X_t is *not* observed, the problem is partially observed, and the conventional fix—augmenting the state with the action history—costs K^T table entries in the worst case (Theorem 5). We treat both regimes in one framework (Sec. 3). Our representation avoids this cost: the value is *linear* in the latent, so one learned operator per state–action pair is valid at every latent state, including ones never seen in training (Theorem 1). This gives operator Q-learning the ability to evaluate unseen latent states without retraining; in E18, DQN and recurrent baselines degrade in this setting.

The central idea is to place order information in the *codomain* of the value function. If $Q(s, a)$ is an operator and the Bellman backup transports the continuation value through the action's operator, $Q(s, a) = \mathcal{K}_a^\dagger(R + \gamma \mathbb{E}V)$, then unrolling the recursion produces exactly the ordered product $A_{a_1} \cdots A_{a_T}$, and the scalar value at any latent state is a trace readout. For every policy that does not look at the latent state this representation is *exact* (Theorem 1); the history enters only through an N^2-dimensional *history operator* X_h, which the agent can maintain by one matrix product per step.

Operator aggregation for control. Control requires aggregating $\{Q(s', a')\}_{a'}$ into a continuation value. For scalars this is a maximum; for self-adjoint operators the Löwner order is only partial, and by Kadison's anti-lattice theorem two incomparable self-adjoint operators have *no* least upper bound (Kadison, 1951). We use the operator log-sum-exp, which is a Löwner upper bound of every argument, and prove that the resulting soft operator recursion certifies an upper bound on the optimal value at every latent state (Theorem 3); in the commuting case it reduces exactly to soft Q-learning (Prop. 2). We also delimit what does *not* survive: the operator log-sum-exp is not monotone and not non-expansive outside the commuting sector (Props. 4, 5).

Scope. Three settings should be kept apart. The exact representation theorem (Thm 1) concerns *latent-blind policy evaluation*: the policy sees only s_t, and its value is linear in the latent. The certificate (Thm 3) bounds the value of every policy, latent-aware ones included. The control experiments are *latent-aware*: the agent observes ρ_t, or can compute it from the action history because ρ_0 and the $\mathcal{K}_a$ are known, and uses it only to perform the greedy readout $\arg\max_a \tau(\rho_t Q(s_t, a))$. We do not claim to solve a POMDP in which the latent must be inferred from rewards alone, nor the case of unknown action maps, which we leave open (Sec. 7).

Relation to prior work. The closest conceptual connection is to α-vector representations in POMDPs (Smallwood & Sondik, 1973; Kaelbling et al., 1998): in both cases the value is linear in a latent quantity. The main difference is in how the latent dynamics enter the Bellman backup. An α-vector is propagated through a stochastic belief update, whereas here the continuation operator is transported through the dual $\mathcal{K}_a^\dagger$ of a known, generally non-commuting action map, which produces ordered products $A_{a_1} \cdots A_{a_T} W$ (Corollary 1). The optimal POMDP value is the upper envelope of a set of α-vectors that can grow exponentially with the horizon; we keep one operator per state–action pair, which is exact for latent-blind policy evaluation, while the operator log-sum-exp provides an upper certificate for control (Thm 3), and E18 measures the cost of this compression. Because the latent carries a representation of the action algebra, the required operator dimension is related to the Hankel rank of the ordered-product readout (Thm 6), and the nilpotency step of the algebra controls the size of exact history statistics (Thm 5).

In the commuting case Theorem 3 reduces to the familiar "latent revealed" (QMDP-type) upper bound (Hauskrecht, 2000). The ordered-product readout is a weighted automaton, and our dimension results import Hankel-rank theory (Carlyle & Paz, 1971; Fliess, 1974), as used for observable operator models and predictive state representations (Jaeger, 2000; Littman et al., 2002; Balle et al., 2014); our contribution is to place these operators *inside* a Bellman recursion with an order-theoretic aggregation. Distributional RL (Bellemare et al., 2017) enriches the codomain with measures that commute under convolution; Bayesian Bellman operators (Fellows et al., 2021), operator DQN (Liu et al., 2022), quantum RL (Dong et al., 2008; Skolik et al., 2022), geometric RL on $SO(3)$ (Schuck et al., 2025), and the analyses of Lee (2024); Lee et al. (2023); Meyn (2024) retain scalar Bellman equations. Hayase et al. (2025) use free probability for input embeddings, without changing the Bellman operator. Recurrent Q-learning (SeyedSalehi et al., 2023) compresses history without a K^T table, and equivariant Q-learning (Wang et al., 2021) exploits invariance under a group *action* rather than non-commutativity of action *composition*. We compare against recurrent (GRU) value functions in E13–E16 and E18. Höftmann et al. (2026) propose a non-scalar Bellman operator on antisymmetric value differences, unrelated to operator non-commutativity; Zhang et al. (2026) prove value preservation when Lie-group operators commute with the controlled generator, parallel to Proposition 2.

Contributions. (1) A formal problem class, operator-latent decision processes, that separates latent-aware from latent-blind control and makes the state-aliasing structure explicit (Sec. 3). (2) A *compositional* operator Bellman recursion: exact representation of action values (Thm 1), contraction and almost-sure convergence of an asynchronous operator TD scheme under its sampling and step-size assumptions (Thm 2), and policy improvement (Prop. 1). (3) An operator soft-Bellman recursion whose fixed point is a Löwner upper bound on every policy's value (Thm 3), exact reduction to soft Q-learning under commutativity (Prop. 2), and a corrected Lipschitz analysis of the operator log-sum-exp, including counterexamples to β-free constants (Prop. 4). (4) Representation theory of order: an aliasing lower bound (Thm 4), a free-versus-nilpotent dichotomy for exact *tabular* history representation (Thm 5), and a Hankel-rank lower bound on the operator dimension that is tight for generic full-rank instances (Thm 6). (5) Finite-instance numerical checks of

every statement, end-to-end control experiments against scalar tabular, DQN and recurrent baselines, and representation experiments on four non-abelian domains (Sec. 6); all code is released at https://anonymous.4open.science/r/opq-RL-2027-0D23.

2 PRELIMINARIES

Classical RL. For an MDP $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ the optimal Q^* satisfies $Q^*(s,a) = \mathbb{E}[r + \gamma \max_{a'} Q^*(s', a')]$, and the Bellman operator is a γ-contraction in $\|\cdot\|_\infty$ (Watkins & Dayan, 1992; Sutton & Barto, 2018).

Operators. We work in $\mathcal{M}_N = \mathbb{C}^{N\times N}$ with the normalised trace $\tau = \frac{1}{N}\operatorname{Tr}$, a faithful tracial state, so $(\mathcal{M}_N, \tau)$ is the canonical finite-dimensional non-commutative probability space (Nica & Speicher, 2006); the scalar case is $N = 1$. We write $\|\cdot\|$ for the operator norm, $\|x\|_2^2 = \tau(x^*x)$ for the normalised Hilbert–Schmidt norm, $\|x\|_1 = \tau(|x|)$ for the normalised trace norm, and $x \preceq y$ for the Löwner order on self-adjoint operators. For a linear map $\mathcal{K}$ on $\mathcal{M}_N$ its *dual* (Heisenberg picture) $\mathcal{K}^\dagger$ is defined by $\tau\big(Y\,\mathcal{K}(\rho)\big) = \tau\big(\mathcal{K}^\dagger(Y)\,\rho\big)$. For self-adjoint $x_1, \dots, x_K$ the operator log-sum-exp is

$$\mathrm{LSE}_\beta(x_1, \dots, x_K) = \tfrac{1}{\beta}\log\Big(\textstyle\sum_i e^{\beta x_i}\Big), \tag{1}$$

defined by functional calculus (Bhatia, 1997).

Remark 1 (Group and algebra views)**.** *Writing $A_a = \exp(\Omega_a)$ with Ω_a in a Lie algebra $\mathfrak{g} \subset \mathfrak{gl}(N)$, the latent evolves as $X_t = X_{t-1}\exp(\Omega_{a_t})$. For $\mathfrak{g} = \mathfrak{so}(3)$ the Ω_a are real skew-symmetric; for quantum pulses $\Omega_a = -iH_a\Delta t$. The algebra view is used in Sec. 5 to state nilpotency.*

3 OPERATOR-VALUED Q-LEARNING

OPERATOR-LATENT DECISION PROCESSES

Definition 1 (Operator-latent decision process)**.** *An OLDP is a tuple $(\mathcal{S}, \mathcal{A}, P, \{\mathcal{K}_a\}, R, \gamma)$: a finite* observable *state space $\mathcal{S}$ and action set $\mathcal{A}$ ($|\mathcal{A}| = K$) with transition kernel $P(s' \mid s, a)$ that does not depend on the latent; a* latent *$\rho_t \in \mathcal{M}_N$ updated by linear maps, $\rho_{t+1} = \mathcal{K}_{a_t}(\rho_t)$; and rewards linear in the latent, $r_t = \tau\big(R(s_t, a_t)\,\rho_{t+1}\big)$, with $\|R(s,a)\| \le R_{\max}$. Rewards are real: in CP, R and ρ are self-adjoint; in MP we use real matrices, or replace $\tau(\cdot)$ by $\operatorname{Re}\tau(\cdot)$, which is still linear over $\mathbb{R}$. We write $\kappa := \max_a \|\mathcal{K}_a^\dagger\|_{\mathcal{M}_N\to\mathcal{M}_N}$ for the operator-norm bound of the dual maps.*

Two cases cover all our examples. In the *multiplicative picture* (MP), $\mathcal{K}_a(X) = XA_a$ with $A_a \in GL(N)$, so $X_T = X_0 A_{a_1}\cdots A_{a_T}$ and $\mathcal{K}_a^\dagger(Y) = A_a Y$; $\kappa = \max_a\|A_a\|$, which equals 1 for orthogonal or unitary A_a. The terminal reward $\frac{1}{N}\operatorname{Tr}(WX_T)$ of Sec. 1 is the case $R = W$ at the final step and $R = 0$ before (time is recorded in s). In the *conjugation picture* (CP), $\mathcal{K}_a(\rho) = U_a\rho U_a^*$ with U_a unitary and ρ a density operator (quantum control; or the dyad $\rho = nn^\top$ of a marked body axis in rigid-body control), so $\mathcal{K}_a^\dagger(Y) = U_a^* Y U_a$, and $\kappa = 1$. In CP the dual maps are positive and unital, and R is self-adjoint.

A *latent-aware* policy chooses actions from (s_t, ρ_t); the process (s_t, ρ_t) is an MDP for such policies. A *latent-blind* policy $\pi(a \mid s)$ sees only s_t. When s_t does not record the order of past actions (e.g. s_t is the step counter or the action counts), latent-blind control is a POMDP with deterministic, action-driven latent dynamics. If ρ_0 and the $\mathcal{K}_a$ are known, a latent-blind agent can nevertheless reconstruct $\rho_t = \mathcal{K}_{a_t}\circ\cdots\circ\mathcal{K}_{a_1}(\rho_0)$ from its own actions; in MP with $X_0 = I$ this is the *history operator* $X_h = A_{a_1}\cdots A_{a_t}$, an N^2-dimensional sufficient statistic of the history.

THE COMPOSITIONAL BELLMAN RECURSION

For a latent-blind policy π define, for $Q : \mathcal{S}\times\mathcal{A} \to \mathcal{M}_N$,

$$(\mathcal{T}^\pi Q)(s,a) \;=\; \mathcal{K}_a^\dagger\Big(R(s,a) + \gamma \sum_{s'} P(s' \mid s,a) \sum_{a'} \pi(a' \mid s')\, Q(s', a')\Big). \tag{2}$$

The continuation value is transported *through the action's dual map* before it is combined with the current action; this is where composition happens. Unrolling (2) along a deterministic plan in MP

gives $Q = A_{a_1} A_{a_2} \cdots A_{a_T} W$, the ordered product. (Setting $\mathcal{K}_a^\dagger = \mathrm{id}$ removes the composition and gives a permutation-symmetric recursion whose target, in a single-state terminal-reward problem, does not depend on the current action at all; such a recursion cannot represent order, and (2) is the minimal change that can.) For a finite horizon H we write $Q_t^\pi = \mathcal{T}^\pi Q_{t+1}^\pi$, $Q_H^\pi = 0$.

Theorem 1 (Exact operator representation of action values)**.** *Let π be latent-blind, and let $q^\pi(s,\rho,a) = \mathbb{E}^\pi\big[\sum_{t\ge 0}\gamma^t r_t \,\big|\, s_0 = s, \rho_0 = \rho, a_0 = a\big]$ be its action value in the full process. For a finite horizon, the backward recursion $Q_t^\pi = \mathcal{T}^\pi Q_{t+1}^\pi$ satisfies $q_t^\pi(s,\rho,a) = \tau\big(\rho\, Q_t^\pi(s,a)\big)$ for all s, a and all $\rho \in \mathcal{M}_N$. For an infinite horizon with $\gamma\kappa < 1$, $\mathcal{T}^\pi$ has a unique fixed point Q^π and $q^\pi(s,\rho,a) = \tau(\rho\, Q^\pi(s,a))$. (For ρ outside the physical set, q^π is defined by the same expectation and is linear in ρ.)*

Corollary 1 (Ordered products are Bellman fixed points)**.** *In MP with a single observable state per step, terminal reward W and a deterministic plan $(a_1, \dots, a_T)$, the recursion (steps $t = 0, \dots, T-1$, action a_{t+1} at step t) gives $Q_t(a_{t+1}) = A_{a_{t+1}} \cdots A_{a_T} W$, so the return from $X_0 = I$ is $\tau(Q_0(a_1)) = \frac{1}{N}\mathrm{Tr}(W A_{a_1} \cdots A_{a_T})$. After a history $h = (a_1, \dots, a_t)$, the value of the remaining plan is $\tau(X_h\, Q_t(a_{t+1}))$.*

The product readout used in Sec. 5 and in experiments E8, E13–E16 is therefore not a separate model: it is the fixed point of (2) along each plan.

Theorem 2 (Contraction and convergence of operator TD)**.** *(i) $\mathcal{T}^\pi$ is a $\gamma\kappa$-contraction in $\|Q\|_{\infty,\mathrm{op}} := \max_{s,a}\|Q(s,a)\|$; if every $\mathcal{K}_a^\dagger$ is also non-expansive in $\|\cdot\|_2$ (unitary conjugation, orthogonal or unitary multiplication), it is a γ-contraction in $\max_{s,a}\|Q(s,a)\|_2$. (ii) Let (s_k, a_k) be drawn i.i.d. from a distribution μ with full support, $s'_k \sim P(\cdot \mid s_k, a_k)$, $a'_k \sim \pi(\cdot \mid s'_k)$, and update only the visited pair,*

$$Q_{k+1}(s_k,a_k) = Q_k(s_k,a_k) + \alpha_k\Big(\mathcal{K}_{a_k}^\dagger\big(R(s_k,a_k) + \gamma\, Q_k(s'_k,a'_k)\big) - Q_k(s_k,a_k)\Big), \tag{3}$$

with $\sum_k \alpha_k = \infty$, $\sum_k \alpha_k^2 < \infty$. If $\gamma\kappa < 1$ then $Q_k \to Q^\pi$ almost surely, without projection.

Part (ii) is a statement about the specific asynchronous scheme above (i.i.d. state–action sampling with full support, on-policy next actions, square-summable steps); it is not a convergence result for arbitrary behaviour policies.

The update samples the observable transition and the next action only; it uses the dual maps $\mathcal{K}_a^\dagger$ and the reward observable R, which are known in the control problems that motivate us (gate unitaries and a target observable; rotation generators and a target attitude). When the A_a are unknown they can be fitted from returns (E8, Sec. 5); combining such fitting with TD, with guarantees, is left open.

CONTROL

Proposition 1 (Latent-aware policy improvement)**.** *Let π be latent-blind with $\gamma\kappa < 1$ and let $\pi^+(s,\rho) \in \arg\max_a \tau(\rho\, Q^\pi(s,a))$. Then $V^{\pi^+}(s,\rho) \ge V^\pi(s,\rho)$ for all (s,ρ).*

Proposition 1 is the practical controller when the latent is observed or reconstructed: evaluate a latent-blind base policy in operator form and act greedily on the scalar readout. To obtain a single operator per (s,a) that speaks for *all* policies, we need an aggregation over actions in the operator codomain. In CP (Q self-adjoint, $\mathcal{K}_a^\dagger$ positive and unital) define

$$(\mathcal{T}_\beta Q)(s,a) = \mathcal{K}_a^\dagger\Big(R(s,a) + \gamma \sum_{s'} P(s' \mid s,a)\ \mathrm{LSE}_\beta\big(\{Q(s',a')\}_{a'}\big)\Big). \tag{4}$$

Theorem 3 (Löwner certificate)**.** *In CP, let $V_H^\beta = 0$, $V_t^\beta(s) = \mathrm{LSE}_\beta(\{Q_t^\beta(s,a')\}_{a'})$ and $Q_t^\beta(s,a) = \mathcal{K}_a^\dagger\big(R(s,a) + \gamma\sum_{s'} P(s' \mid s,a) V_{t+1}^\beta(s')\big)$ (the finite-horizon form of (4)). For every policy π—latent-aware and history-dependent included, with q_t^π conditioned on any history that reaches (s,ρ) at step t—every a, and every $\rho \succeq 0$,*

$$q_t^\pi(s,\rho,a) \ \le\ \tau\big(\rho\, Q_t^\beta(s,a)\big).$$

In particular $q_t^*(s,\rho,a) \le \tau(\rho\, Q_t^\beta(s,a))$, and $Q_t^\pi(s,a) \preceq Q_t^\beta(s,a)$ for every latent-blind π. For an infinite horizon, no contraction is needed: for every policy and every $n \ge 1$, $q^\pi(s,\rho,a) \le$

$\tau\big(\rho\, Q_0^{\beta,(n)}(s,a)\big) + \gamma^n R_{\max}\tau(\rho)/(1-\gamma)$*, where* $Q^{\beta,(n)}$ *is the* n*-step recursion. If, in addition,* $\mathcal{T}_\beta$ *contracts on the ball* $\{\|Q\|_{\infty,\mathrm{op}} \le M_\beta\}$*,* $M_\beta = (R_{\max} + \gamma\beta^{-1}\log K)/(1-\gamma)$*—with modulus* γ *on the commuting sector, or* $\gamma e^{2\beta M_\beta}$ *in general (Prop. 4)—then its fixed point is itself a certificate.*

The certificate needs only positivity of $\mathcal{K}_a$ and the Löwner bound $\mathrm{LSE}_\beta(x) \succeq x_i$; it does *not* need monotonicity of LSE_β, which fails in general (Prop. 5). Its slack has two sources: softness ($\beta^{-1}\log K$ per step) and non-commutativity—by Kadison's theorem no single operator can be the least upper bound of incomparable $Q(s',a')$. E17 measures both.

Proposition 2 (Exact commutative reduction)**.** *In CP, suppose all* U_a *and* $R(s,a)$ *lie in a common maximal abelian subalgebra* $\mathcal{D}$ *(simultaneously diagonal in a basis* $\{e_j\}$*). Then (i)* $\mathcal{T}_\beta$ *preserves* $\mathcal{D}$ *and acts on each coordinate* j *as the classical soft Bellman operator* $q_j(s,a) = r_j(s,a) + \gamma\sum_{s'} P(s' \mid s,a)\,\mathrm{lse}_\beta(q_j(s',\cdot))$*; (ii) for a pure eigenstate* $\rho = Ne_je_j^*$ *the certificate of Theorem 3 is the soft-optimal value of coordinate* j*, exceeding the optimal value by at most* $\beta^{-1}\log K\sum_{k=1}^{H-t-1}\gamma^k$ *and converging to it as* $\beta\to\infty$*; (iii) for* $N=1$*,* $\mathcal{T}_\beta$ *is soft Q-learning's Bellman operator.*

In the commuting sector only the diagonal of the latent affects values, and it never changes ($U_a^* Y U_a = Y$ for $Y \in \mathcal{D}$). Order is then irrelevant and nothing is lost relative to classical RL; the operator structure is paid for only when a non-commuting action appears.

4 Properties of the Operator Log-Sum-Exp

Throughout, x_i, y_i are self-adjoint with $\|x_i\|, \|y_i\| \le M$.

Proposition 3 (Well-posedness and Löwner properties)**.** LSE_β *is well defined and self-adjoint, and (i)* $\mathrm{LSE}_\beta(x) \succeq x_i$ *for every* i*; (ii)* $\mathrm{LSE}_\beta(x_1+cI,\dots,x_K+cI) = \mathrm{LSE}_\beta(x)+cI$*; (iii)* $\|\mathrm{LSE}_\beta(x)\| \le \max_i\|x_i\| + \beta^{-1}\log K$*; (iv)* $U^*\,\mathrm{LSE}_\beta(x)U = \mathrm{LSE}_\beta(U^*x_1U,\dots,U^*x_KU)$ *for unitary* U*; (v) if all* x_i *commute,* LSE_β *acts as the scalar log-sum-exp on the joint spectrum. By Kadison's theorem (Kadison, 1951) a pair* x_1, x_2 *has a least upper bound in the Löwner order only if* $x_1 \preceq x_2$ *or* $x_2 \preceq x_1$*; property (i) is therefore the natural replacement for "*$\max$*" in the non-commutative case.*

Proposition 4 (Lipschitz behaviour)**.** *Let* $h_i = x_i - y_i$ *and* $\Delta = \mathrm{LSE}_\beta(x) - \mathrm{LSE}_\beta(y)$*. (a)* Commuting sector *(all* x_i, y_j *in one eigenbasis):* $\|\Delta\| \le \max_i\|h_i\|$ *and* $\|\Delta\|_2 \le \big(\sum_i\|h_i\|_2^2\big)^{1/2}$*, for every* β*; the first constant is attained (uniform shifts) and the second is approached as* $\beta\to\infty$*. (b)* General*: for* $p \in \{2,\infty\}$*,* $\|\Delta\|_p \le e^{2\beta M}\,\frac{1}{K}\sum_i\|h_i\|_p$*. (c)* No β-free constant for max-aggregated Hilbert–Schmidt perturbations*: for* $K = N$ *and* $x_i = e_ie_i^*$*, perturbing each* x_i *in its own coordinate gives* $\|\Delta\|_2/\max_i\|h_i\|_2 \to p_\beta\sqrt{N}$ *with* $p_\beta = e^\beta/(e^\beta + N - 1)$*; any bound of the form* $\|\Delta\|_2 \le C\max_i\|h_i\|_2$ *therefore needs* $C \ge p_\beta\sqrt{N}$*.*

Numerical observation (E7, 80-digit arithmetic). Outside the commuting sector LSE_β is not non-expansive in operator norm: local ratios $\|\Delta\|/\max_i\|h_i\|$ reach 0.97, 1.17 and 1.48 at $\beta = 1, 5, 10$. We state this as an observation, not a theorem.

Part (a) gives $\mathcal{T}_\beta$ modulus exactly γ on the commuting sector. Part (c) rules out the β-linear spectral-diameter constant one might hope for under max-aggregation (at $K = N = 25$, $\beta = 5$ the ratio is 4.30 while $1 + \frac{\beta}{4}\,\mathrm{diam}(\mathrm{spec}) = 2.25$); the ℓ_2-aggregated bound in (a) is the correct commuting statement. Closing the gap between (b) and the observed growth is open; none of our results depends on it, since Theorems 1 and 2 involve no aggregation and Theorem 3 is proved by finite-horizon induction.

Proposition 5 (Monotonicity)**.** *On the commuting sector,* $x_i \preceq y_i$ *for all* i *implies* $\mathrm{LSE}_\beta(x) \preceq \mathrm{LSE}_\beta(y)$*, and hence* $\mathcal{T}_\beta$ *is monotone. In general it is not: the matrix exponential is not operator monotone, and E12 exhibits violations at every perturbation scale down to* 10^{-4}*.* $\mathcal{T}^\pi$ *is always monotone in CP, since* $\mathcal{K}_a^\dagger$ *is positive.*

5 What Scalar Values Lose, and What Operators Retain

We now work in MP with $X_0 = I$, terminal reward $r(a_1,\dots,a_T) = \frac{1}{N}\,\mathrm{Tr}(WX_T)$, and a history $h = (a_1,\dots,a_t)$ with history operator $X_h = A_{a_1}\cdots A_{a_t}$. An observation map σ on histories is

order-blind if $\sigma(h) = \sigma(h')$ whenever h' is a permutation of h (the step counter and the vector of action counts are examples). Consider the last-step value $q_T(h, a) = \frac{1}{N}\operatorname{Tr}(WX_hA_a)$ for $|h| = T-1$, which does not depend on any continuation policy.

Theorem 4 (Order-aliasing lower bound)**.** *Let $T \geq 3$ and let σ be order-blind. For every scalar function f of $(\sigma(h), a)$,*

$$\sup_{|h|=T-1,\,a} \big|f(\sigma(h), a) - q_T(h, a)\big| \;\geq\; \tfrac{1}{2N} \max_{a,b,c} \big|\operatorname{Tr}\big(WX_g[A_a, A_b]A_c\big)\big|,$$

where g is any fixed prefix of length $T-3$. If $[A_a, A_b] \neq 0$ for some a, b, the right-hand side is positive for Lebesgue-almost every W. In contrast, $q_T(h, a) = \tau\big(X_h\, Q_T(a)\big)$ exactly, with $Q_T(a) = \mathcal{K}_a^\dagger(W) = A_aW$ (Corollary 1).

The condition $T \geq 3$ is necessary: for $T = 2$ the histories (a) and (b) are different multisets, an order-blind map may separate them, and no aliasing occurs. The theorem is a statement about state aliasing, not about the Markov property of the underlying process: with the latent observed, the process is Markov and the bound does not apply.

The next result concerns *exact tables* indexed by a summary of the history. It is not a lower bound for compact function approximators: a recurrent network needs no K^T table and can approximate these values well (E13–E16, E18). What the theorem isolates is when an exact, finite history statistic exists at all, and how large it must be.

Theorem 5 (Cost of exact tabular history representation: free versus nilpotent)**.** *(a) If the A_a generate a free group of rank K—for $SO(3)$ and $SE(3)$ this holds for Lebesgue-almost every K-tuple, $K \geq 2$ (Epstein, 1971)—then for almost every W any exact scalar table indexed by $(\sigma(h), a)$, for any map σ of histories, must separate all K^{T-1} histories of length $T-1$, i.e. has at least K^T entries; in particular no order-blind σ is exact. (b) If $A_a = e^{\Omega_a}$ and the Lie algebra generated by the Ω_a is nilpotent of step 2, then X_h depends on h only through the counts $n_a(h)$ and the signed pair counts $L_{ab}(h) = \#\{i < j : h_i = a, h_j = b\} - \#\{i < j : h_i = b, h_j = a\}$, $a < b$; an exact table therefore needs at most $K(T+1)^K(2T^2+1)^{K(K-1)/2}$ entries, polynomial in T.*

For step $m > 2$ the analogous statistics are the iterated counts of degree $< m$ (Chen's expansion, Chen, 1957); we state and prove only $m = 2$, which covers the Heisenberg algebra $\mathfrak{n}_3$ and bosonic displacements (E10, E15). For compact groups the scalar deficit itself is bounded by $\|W\|$ uniformly in T; the distinction between the two regimes is the *size of an exact history representation*, not a growing error.

Theorem 6 (Operator sufficiency and minimal dimension)**.** *Let $f(w) = \frac{1}{N}\operatorname{Tr}(WA_w)$ on action words w, with Hankel matrix $\mathbf{H}_f(u, v) = f(uv)$. (i) The K operators A_a and the readout W ($(K+1)N^2$ parameters) represent f exactly for every horizon; by Corollary 1 this is the fixed point of* (2) *along each plan. (ii) Every d-dimensional operator model $g(w) = \frac{1}{d}\operatorname{Tr}(W_dQ_w)$, $Q_a \in \mathcal{M}_d$, has $\operatorname{rank}\mathbf{H}_g \leq d^2$; since $\operatorname{rank}\mathbf{H}_f \leq N^2$, exact representation requires $d \geq (\operatorname{rank}\mathbf{H}_f)^{1/2}$, and $d = N$ suffices; that is, the minimal exact dimension satisfies $(\operatorname{rank}\mathbf{H}_f)^{1/2} \leq d_{\min} \leq N$, with $d_{\min} = N$ whenever $\operatorname{rank}\mathbf{H}_f = N^2$. (iii) For any finite prefix set $\mathcal{U}$ and suffix set $\mathcal{V}$, every d-dimensional model has $\max_{u,v}|g(uv) - f(uv)| \geq \big(\sum_{j>d^2}\sigma_j(\mathbf{H}_f[\mathcal{U}, \mathcal{V}])^2\big)^{1/2}/\sqrt{|\mathcal{U}||\mathcal{V}|}$.*

Part (ii) is the Hankel-rank theory of weighted automata (Carlyle & Paz, 1971; Fliess, 1974) applied to the ordered-product readout: a $d \times d$ operator model is a weighted automaton of dimension at most d^2. The lower bound is not a full characterisation of $d_{\min}$ for arbitrary f; it is tight when the Hankel matrix has full rank N^2. For unitary actions, Proposition 8 refines it using the structure of the algebra the actions generate. For generic orthogonal generators we find $\operatorname{rank}\mathbf{H}_f = N^2$ (E17), so exact recovery occurs exactly at $d = N$, as observed in E8, E14 and E16. The comparison with history augmentation is therefore specific: an *exact tabular* history representation needs K^T entries (Thm. 5a); compact recurrent models do not need a table, and E13/E14 test whether they recover the product structure from data at comparable parameter counts.

Extensions and practical cost. Appendix E proves three extensions, each checked in E17b. *Robust improvement* (Prop. 6): acting greedily on an estimate $\hat{Q}$ with $\max_{s,a}\|\hat{Q} - Q^\pi\| \leq \varepsilon$ loses at most $2\varepsilon\|\rho\|_1/(1-\gamma\kappa)$, so the TD guarantee of Theorem 2 carries over to control. *Noisy and random*

Agent	C1 random OLDP fixed	C1 random OLDP random	C1 random OLDP transfer	C2 $SO(3)$ pointing fixed	C2 $SO(3)$ pointing random	C2 $SO(3)$ pointing transfer	C3 two-qubit prep. fixed	C3 two-qubit prep. random	C3 two-qubit prep. transfer
OpQ-soft (learned R)	0.84	0.95	0.95	0.98	0.99	0.99	0.83	0.68	0.68
OpQ-soft (known R)	0.84	0.95	–	0.98	0.99	–	0.79	0.68	–
OpQ-PI	0.86	0.97	0.97	0.91	0.98	0.98	0.28	0.70	0.70
Scalar Q on (t, m)	0.87	0.51	–	0.69	0.06	–	0.19	0.10	–
Scalar Q, history table	0.94	0.50	–	0.75	0.05	–	0.28	-0.04	–
DQN on ρ_t	1.00	0.99	0.87	0.99	0.99	0.96	1.00	0.87	0.49
DRQN (GRU) on history	1.00	0.95	0.53	0.97	0.94	0.21	0.96	-0.01	0.11

Table 1: E18: normalised return of the greedy policy after 3000 episodes (mean over 5 seeds; standard deviations, scores after 100 episodes and learning curves are in Appendix G). *Transfer*: trained with a fixed ρ_0, evaluated on 8 unseen initial latents.

latent dynamics (Prop. 7): Theorems 1 and 3 hold for any positive maps (e.g. decohering quantum channels) and for randomly perturbed actions, using the dual of the averaged map. *Structure-refined dimension* (Prop. 8): for unitary actions generating $\bigoplus_k \mathcal{M}_{n_k}$, $(\sum_k n_k^2)^{1/2} \le d_{\min} \le \sum_k n_k$ for almost every W. One TD update costs $O(N^3)$, a soft update $O(KN^3)$, and memory is $|\mathcal{S}|KN^2$ numbers, independent of the horizon. Because $\sum_i e^{\beta x_i}$ has condition number up to $e^{2\beta M}$, we use double precision only for $\beta M \lesssim 10$ and extended precision beyond.

6 Experiments

We report three kinds of evidence. **(C) End-to-end control** (E18): agents learn from sampled episodes, and their policies are evaluated exactly against the optimal latent-aware value. **(A) Finite-instance checks of the theory** (E17, E7, E9, E12). **(B) Representation** (E8, E10, E13–E16): the ordered-product readout, which by Corollary 1 is the fixed point of (2) along each plan, is fitted to exact sequence returns. Every number below is produced by the released code (https://anonymous.4open.science/r/opq-RL-2027-0D23; Appendix H).

E18: End-to-end control

Setup. We use three OLDPs in the conjugation picture, all with observable state $s_t = (t, m_t)$: **C1**, a random OLDP (N=3, K=4, two stochastic modes, H=6, noisy rewards); **C2**, $SO(3)$ pointing: six ± 0.6 rad body rotations of a marked axis $\rho = 3nn^\top$, reward $(t_m \cdot n)^2$ for a mode-dependent target, H=8; **C3**, two-qubit state preparation with $\{H_1, H_2, T_1, T_2, \mathrm{CNOT}\}$ and a terminal fidelity reward, H=8. The initial latent is either fixed or drawn at random in every episode and observed; a *transfer* test trains with the fixed ρ_0 and evaluates on 8 unseen ones. **Agents.** *OpQ-soft* learns one $N \times N$ operator per (s, a) with the sample form of (4) (β=5), estimates $R(m, a)$ from scalar rewards by least squares, and acts greedily on $\tau(\rho_t Q(s_t, a))$. *OpQ-PI* evaluates the uniform policy by operator TD (3) and applies Prop. 1. Scalar baselines: Q-learning on (t, m) (order-blind), Q-learning on the full history (the K^T table), a DQN with the latent ρ_t as input, and a GRU-based DRQN on the history. The OpQ agents use the known dual maps $\mathcal{K}_a^\dagger$; DQN is given ρ_t itself. Scores are normalised so that 1 is the exact dynamic-programming optimum and 0 is the uniform policy.

Results (Table 1). *(i) Order-blind and history-table scalar agents fail once the latent varies* (random ρ_0: 0.51, 0.06, 0.10 for Q on (t, m)), as Theorem 4 predicts. OpQ reaches 0.95–0.99 on C1 and C2 while storing 432 and 864 numbers, compared with $K^H = 4096$ and 1.7×10^6 histories. *(ii) OpQ shows a strong sample-efficiency advantage in these environments.* After 100 episodes with a random ρ_0, OpQ-soft scores $0.95/0.95$ on C1/C2, against $0.89/0.73$ for DQN and $0.44/0.28$ for DRQN. *(iii) OpQ transfers zero-shot.* Theorem 1 makes the readout valid at every ρ, so OpQ trained on one initial latent keeps 0.95–0.99 on C1/C2 at unseen latents, where DQN drops to 0.87/0.96 and DRQN to 0.53/0.21. On C3, OpQ's transfer score (0.68–0.70) exceeds both DQN (0.49) and DRQN (0.11). *(iv) With enough episodes, a DQN that observes ρ_t matches or exceeds OpQ* (C1 fixed: 1.00 vs. 0.84; C3: $1.00/0.87$ vs. $0.83/0.68$). With the exact model, the greedy readout of Q^β scores 0.86 (C1 fixed), 0.79 (C3 fixed) and 0.68 (C3 random), and the learned agents reach (or, on C3, slightly exceed) these ceilings (Table 5). The exact-model experiment therefore separates this gap from optimisation error; the gap is consistent with the non-commutative slack of a single-operator certificate (Sec. 3), since one operator per (s, a) cannot represent a value that is a maximum over latent-dependent continuations. Where that slack is small (C2), OpQ is within 0.02 of the optimum.

FINITE-INSTANCE CHECKS OF THE THEORY (E17, E7, E9, E12)

E17 uses a random OLDP with $|\mathcal{S}|$=2, K=N=3, H=4 (Appendix H). *Theorem 1*: the readout matches the exact expected return to 6.7×10^{-16} (CP) and 3.6×10^{-16} (MP). *Theorem 2*: operator TD reduces $\max_{s,a}\|Q_k - Q^\pi\|$ from 1.69 to 0.14 over 4×10^5 steps. *Theorem 3*: for $\beta \in \{1, 5, 20, 50\}$ there are $0/30$ violations, and $\lambda_{\min}(Q^\beta - Q^\pi) > 1.7$. The mean slack (1.2–1.9, against a mean $|q^*|$ of 3.45) does not vanish as β grows, which isolates its non-commutative part. In the commuting sector the slack at pure eigenstates is 4.8×10^{-12} at β=200. *Theorem 4*: a swapped prefix gives $r(a, b, c) = -1.143$ and $r(b, a, c) = -0.807$, so any order-blind scalar errs by at least 0.168. *Theorem 6*: the Hankel rank is $9 = N^2$. **E7**: over 20,000 commuting configurations the largest operator-norm ratio is 1.000000 (Prop. 4a); the construction of Prop. 4(c) at K=N=25, β=5 gives 4.304, exceeding $1 + \beta\,\mathrm{diam}/4 = 2.25$, with ℓ_2-aggregated ratio 0.861; in 80-digit arithmetic, non-commuting ratios reach 0.970, 1.171 and 1.477 at $\beta = 1, 5, 10$. **E9**: with commuting actions, the operator recursion and the sample updates coincide with N decoupled soft Q-learners to 1.6×10^{-14} (Prop. 2). **E12**: monotonicity holds in all 2000 commuting trials but fails in $72/2000$ generic ones, and continues to fail at perturbation scale 10^{-4} (Prop. 5). The proof of the certificate does not require monotonicity.

REPRESENTATION (E8, E10, E13–E16)

Each benchmark fits the ordered-product readout of Theorem 6 to exact returns (3000 training and 1000 test words). The benchmarks are random orthogonal (N=5, E8/E13), $\mathfrak{so}(3)$ (N=3, E14), an $\mathfrak{se}(3)$ hinge chain (N=4, E16) and truncated bosonic displacements (real dimension 12, E15). OpQ(d) becomes exact at the predicted dimension and not below it: test MSE is $\leq 3 \times 10^{-30}$ in the first three benchmarks and 9.7×10^{-15} for displacements, with 63–1008 parameters, against 4^6–6^8 history-table entries. The best other models are recurrent (GRU or LSTM), with test MSE 0.003–0.09 (return variances 0.09–1.28). Attention, the 1D-CNN and the windowed table stay close to the order-blind baseline (Table 6, Fig. 4). **E10** compares $\mathfrak{so}(3)$ with the Heisenberg algebra $\mathfrak{n}_3$ at matched generator norm. The per-depth commutator term is constant (0.221) for $\mathfrak{so}(3)$ and exactly zero beyond depth one for $\mathfrak{n}_3$, the nilpotency step that makes exact history statistics polynomial (Thm. 5). The permutation-averaged deficit stays bounded for both algebras (Fig. 5), as $|r| \leq \|W\|$ requires.

7 CONCLUSION

We introduced operator-valued Q-learning for decision processes in which the effects of actions compose non-commutatively. The compositional Bellman recursion gives an exact trace representation for latent-blind policy evaluation, almost-sure convergence of an asynchronous operator TD scheme, and a Löwner upper certificate for control that reduces to soft Q-learning when actions commute. On the representation side, order-blind scalar values must alias swapped histories, and the operator dimension is bounded below by a Hankel rank that is tight for generic instances. The exact representation and the certificate also hold for noisy, positive latent dynamics such as decohering quantum channels (Appendix E). In our control experiments, operator Q-learning learned within a few hundred episodes and evaluated unseen latent states without retraining. A DQN that observes the latent eventually matched or exceeded it, and exact planning showed that this gap is consistent with the non-commutative slack of the certificate. A useful next step is to determine whether the certificate can be tightened without giving up the single-operator representation, for example by keeping a small set of operators per state as α-vector methods do. Other open problems are a tight Lipschitz theory for the operator log-sum-exp, and learning unknown action operators from rewards, for which spectral methods for weighted automata (Balle et al., 2014) offer one route. These results suggest that representing action composition directly in the value function is useful when order information is costly to recover from data.

Reproducibility statement

Assumptions and proofs are in Appendices B, C, D and E; the full experimental protocols, hyperparameters and seeds are in Appendix H. The anonymized repository at https://anonymous.4open.science/r/opq-RL-2027-0D23 contains the code for every experiment, the raw results, and the scripts that produce every table and figure; the complete suite runs on one CPU core in about three hours.

Ethics statement

This is a methods paper on reinforcement learning with operator-valued value functions, and we see no direct ethical concerns. Applications to quantum or robotic control should follow the safety practice of those domains.

AI use statement

We used an AI assistant for language editing, restructuring the manuscript, literature search, and checking theorem statements, assumptions, and experimental design. All outcomes from the AI assistant were independently reviewed and verified by the authors. The authors take full responsibility for the content, methodology, code, and reported results of the paper.

References

Borja Balle, Xavier Carreras, Franco M. Luque, and Ariadna Quattoni. Spectral learning of weighted automata: A forward-backward perspective. *Machine Learning*, 96(1–2):33–63, 2014.

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

Rajendra Bhatia. *Matrix Analysis*. Springer, 1997.

Rajendra Bhatia. *Positive Definite Matrices*. Princeton University Press, 2007.

Vivek S. Borkar and Sean P. Meyn. The O.D.E. method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000.

Jack W. Carlyle and Azaria Paz. Realizations by stochastic finite automata. *Journal of Computer and System Sciences*, 5(1):26–40, 1971.

Kuo-Tsai Chen. Integration of paths, geometric invariants and a generalized Baker–Hausdorff formula. *Annals of Mathematics*, 65(1):163–178, 1957.

Daoyi Dong, Chunlin Chen, Hanxiong Li, and Tzyh-Jong Tarn. Quantum reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 38(5), 2008.

D. B. A. Epstein. Almost all subgroups of a Lie group are free. *Journal of Algebra*, 19(2):261–262, 1971.

Matthew Fellows, Kristian Hartikainen, and Shimon Whiteson. Bayesian Bellman operators. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Michel Fliess. Matrices de Hankel. *Journal de Mathématiques Pures et Appliquées*, 53:197–222, 1974.

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)*, 2018.

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

Tomohiro Hayase, Benoît Collins, and Yuki Inoue. Free probability for in-context reinforcement learning. *arXiv preprint*, 2025.

Marc Höftmann, Jan Robine, and Stefan Harmeling. Relative value learning. In *International Conference on Learning Representations (ICLR)*, 2026.

Herbert Jaeger. Observable operator models for discrete stochastic time series. *Neural Computation*, 12(6):1371–1398, 2000.

Richard V. Kadison. Order properties of bounded self-adjoint operators. *Proceedings of the American Mathematical Society*, 2(3):505–510, 1951.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.

Donghwan Lee. Final-iteration convergence bounds of Q-learning: Switching system approach. *IEEE Transactions on Automatic Control*, 2024.

Donghwan Lee, Jianghai Hu, and Niao He. Discrete-time switching system analysis of Q-learning. *SIAM Journal on Control and Optimization*, 2023.

Michael L. Littman, Richard S. Sutton, and Satinder Singh. Predictive representations of state. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 14, 2002.

Ziyang Liu et al. Operator deep Q-learning: Zero-shot reward transferring in reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Sean Meyn. Projected Bellman equations in reinforcement learning. *IEEE Transactions on Automatic Control*, 2024.

Alexandru Nica and Roland Speicher. *Lectures on the Combinatorics of Free Probability*. Cambridge University Press, 2006.

Murphy Yuezhen Niu, Sergio Boixo, Vadim N. Smelyanskiy, and Hartmut Neven. Universal quantum control through deep reinforcement learning. *npj Quantum Information*, 5:33, 2019.

T. Schuck et al. Geometric reinforcement learning on SO(3). *arXiv preprint*, 2025.

Erfan SeyedSalehi, Nima Akbarzadeh, Amit Sinha, and Aditya Mahajan. Approximate information state based convergence analysis of recurrent Q-learning. *arXiv preprint arXiv:2306.05991*, 2023.

Andrea Skolik, Sofiene Jerbi, and Vedran Dunjko. Quantum agents in the gym: A variational quantum algorithm for deep Q-learning. *Quantum*, 6, 2022.

Richard D. Smallwood and Edward J. Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973.

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

Dian Wang, Robin Walters, and Robert Platt. Equivariant Q learning in spatial action spaces. In *Conference on Robot Learning (CoRL)*, 2021.

Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3):279–292, 1992.

Zuyuan Zhang, Fei Xu Yu, and Tian Lan. Operator-guided invariance learning for continuous reinforcement learning, 2026. URL https://arxiv.org/abs/2605.06500.

Brian D. Ziebart, J. Andrew Bagnell, and Anind K. Dey. Modeling interaction via the principle of maximum causal entropy. In *International Conference on Machine Learning (ICML)*, 2010.

Appendix

This appendix contains the full proofs of all theoretical results, further theoretical extensions, and the complete experimental results and protocols.

- Appendix A works through a fully explicit low-dimensional example (paper folding).
- Appendices B, C and D prove the results of Sections 3, 4 and 5, respectively.
- Appendix E states and proves the extensions: robust improvement, noisy and random latent dynamics, and a structure-refined operator dimension.
- Appendix F summarises all experiments.
- Appendix G gives the full results.
- Appendix H gives the experimental protocols, hyperparameters and computational requirements.

A A Worked Example: Folding Paper

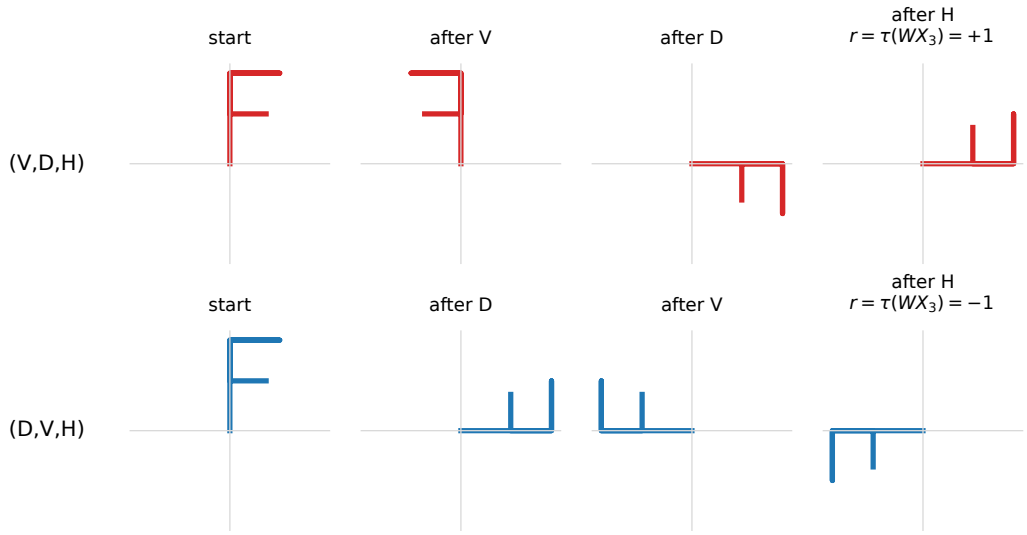


Figure 2: The plans (V, D, H) and (D, V, H) applied to a marked flap (drawn as the letter F). After two folds both histories contain one V and one D, and the third action is the same, yet the readouts are $+1$ and -1. An order-blind scalar Q must err by at least 1 on one of them; the operator recursion gives both values exactly.

A warm-up: folding paper. Track one marked flap of a square sheet, and idealise a fold along a crease through the centre with unit normal n as the reflection it applies to that flap (layer and collision constraints ignored, as in rigid origami):

$$A_n \;=\; I - 2nn^\top \;=\; e^{\,i\pi nn^\top} \in \mathbb{C}^{2\times 2}.$$

Take the vertical, horizontal and diagonal creases,

$$A_V = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \quad A_H = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad A_D = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

A plan $a_1, \ldots, a_T$ moves the flap by the ordered product $X_T = A_{a_1} \cdots A_{a_T}$ (acting on row vectors, $x^\top \mapsto x^\top X_T$). This is the multiplicative picture of Definition 1 with $X_0 = I$.

Where order enters. A_V and A_H are simultaneously diagonal, $[A_V, A_H] = 0$: folding in half twice gives $-I$ in either order. The diagonal crease breaks this:

$$A_V A_D = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = R_{90^\circ}, \qquad A_D A_V = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = R_{90^\circ}^\top, \qquad [A_V, A_D] = \begin{pmatrix} 0 & -2 \\ 2 & 0 \end{pmatrix} \neq 0.$$

Two reflections compose to a rotation by twice the angle between the creases, with sign fixed by the order.

Why a scalar value on an order-blind observation misses it. With two folds, a scalar Q on the observation "one fold made" is *not* aliased: after V the next action is D and after D it is V, so the two plans query different entries $Q(1, D)$ and $Q(1, V)$. Aliasing requires a third step (Theorem 4). Consider the plans (V, D, H) and (D, V, H) with $W = (A_V A_D A_H)^\top$ and $N = 2$, so that $r = \frac{1}{N}\operatorname{Tr}(W X_3)$ is the cosine of the orientation error. Then $r(V, D, H) = \tau(W\, R_{90^\circ} A_H) = 1$ and $r(D, V, H) = \tau\big(A_H R_{90^\circ}^\top R_{90^\circ}^\top A_H\big) = \tau(-I) = -1$. Before the third fold both histories contain one V and one D, so any observation that records which folds were made (fold count, fold counts per crease) is identical, the current action H is identical, and every scalar Q on that observation errs by at least $\frac{1}{2}|1-(-1)| = 1$ on one of them. The unweighted trace is blind as well, $\operatorname{Tr}(A_V A_D A_H) = \operatorname{Tr}(A_D A_V A_H) = 0$; order becomes visible only through the readout.

What the operator recursion does. The last-step operator is $Q_2(H) = \mathcal{K}_H^\dagger(W) = A_H W$, and the recursion (2) along the plan gives $Q_0(a_1) = A_{a_1} A_{a_2} A_{a_3} W$ (Corollary 1). After either history h the value of folding H is $\tau(X_h\, A_H W)$, with the history operator $X_h \in \{R_{90^\circ}, R_{90^\circ}^\top\}$: exactly $+1$ and -1. The model stores $K = 3$ operators and a readout, $(K+1)N^2 = 16$ numbers, for every horizon. In the commuting sector $\{V, H\}$ the operators share an eigenbasis and order is irrelevant.

B Proofs for Section 3

Throughout, $\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|$ are the normalised trace, Hilbert–Schmidt and operator norms, and $|\tau(Y\rho)| \le \|Y\|\,\|\rho\|_1$ (Hölder). Since the trace pairing makes $(\mathcal{M}_N, \|\cdot\|_1)$ and $(\mathcal{M}_N, \|\cdot\|)$ dual, $\|\mathcal{K}_a\|_{1\to 1} = \|\mathcal{K}_a^\dagger\|_{\infty\to\infty} \le \kappa$.

Proof of Theorem 1 and Corollary 1

Proof. Finite horizon. Let q_t^π be the expected discounted reward collected from step t to $H-1$. We show $q_t^\pi(s, \rho, a) = \tau(\rho\, Q_t^\pi(s, a))$ by backward induction; the case $t = H$ is $0 = 0$. Since P does not depend on the latent and $\pi(\cdot \mid s')$ does not depend on it either, the Markov property of (s_t, ρ_t) gives

$$q_t^\pi(s, \rho, a) = \tau\big(R(s, a)\mathcal{K}_a(\rho)\big) + \gamma \sum_{s'} P(s' \mid s, a) \sum_{a'} \pi(a' \mid s')\, q_{t+1}^\pi\big(s', \mathcal{K}_a(\rho), a'\big).$$

By the induction hypothesis the second term is $\gamma\, \tau\big(\mathcal{K}_a(\rho)\, Y\big)$ with $Y = \sum_{s'} P(s' \mid s, a) \sum_{a'} \pi(a' \mid s') Q_{t+1}^\pi(s', a')$. Using the defining identity of the dual map and its linearity, $q_t^\pi(s, \rho, a) = \tau\big(\rho\, \mathcal{K}_a^\dagger(R(s, a) + \gamma Y)\big) = \tau(\rho\, Q_t^\pi(s, a))$.

Infinite horizon. By Theorem 2(i), $\mathcal{T}^\pi$ is a $\gamma\kappa$-contraction, so $Q_n := (\mathcal{T}^\pi)^n 0 \to Q^\pi$, the unique fixed point. Q_n is the time-0 operator of the n-step problem, so by the finite-horizon case the n-step truncated return equals $\tau(\rho Q_n(s, a))$. The tail is controlled by $|r_t| \le R_{\max}\|\rho_{t+1}\|_1 \le R_{\max}\kappa^{t+1}\|\rho\|_1$, which is summable against γ^t when $\gamma\kappa < 1$. Letting $n \to \infty$ gives $q^\pi(s, \rho, a) = \tau(\rho Q^\pi(s, a))$.

Corollary. With $s_t = t$, a deterministic plan, $R = 0$ before the last step and $R = W$ at it, the recursion reads $Q_{T-1}(a_T) = \mathcal{K}_{a_T}^\dagger(W) = A_{a_T} W$ and $Q_t(a_{t+1}) = \mathcal{K}_{a_{t+1}}^\dagger(Q_{t+1}(a_{t+2})) = A_{a_{t+1}} Q_{t+1}(a_{t+2})$, whence $Q_t(a_{t+1}) = A_{a_{t+1}} \cdots A_{a_T} W$ and, from $X_0 = I$, $\tau(Q_0(a_1)) = \frac{1}{N}\operatorname{Tr}(A_{a_1} \cdots A_{a_T} W) = \frac{1}{N}\operatorname{Tr}(W A_{a_1} \cdots A_{a_T})$ by cyclicity. After $h = (a_1, \ldots, a_t)$ the latent is X_h, and the finite-horizon case gives the value $\tau(X_h Q_t(a_{t+1})) = \frac{1}{N}\operatorname{Tr}(W A_{a_1} \cdots A_{a_T})$. □

PROOF OF THEOREM 2

Proof. (i) Rewards cancel and $\|(\mathcal{T}^\pi Q_1 - \mathcal{T}^\pi Q_2)(s,a)\| = \gamma\|\mathcal{K}_a^\dagger(\sum_{s'} P \sum_{a'} \pi\,(Q_1 - Q_2)(s',a'))\| \le \gamma\kappa\|Q_1 - Q_2\|_{\infty,\mathrm{op}}$, since $\sum_{s'} P \sum_{a'} \pi$ is a convex combination. The same computation in $\|\cdot\|_2$ applies when $\|\mathcal{K}_a^\dagger(Y)\|_2 \le \|Y\|_2$, which holds for $Y \mapsto U^*YU$ and $Y \mapsto AY$ with A unitary or orthogonal. A consequence is the stability bound $\|Q - Q^\pi\|_{\infty,\mathrm{op}} \le (1-\gamma\kappa)^{-1}\|\mathcal{T}^\pi Q - Q\|_{\infty,\mathrm{op}}$.

(ii) Identify Q with a vector in $\mathbb{C}^{|\mathcal{S}||\mathcal{A}|N^2}$ and write $e_{(s,a)}$ for the block embedding. The update is $Q_{k+1} = Q_k + \alpha_k\big(F(Q_k) + M_{k+1}\big)$ with $F(Q) = D_\mu(\mathcal{T}^\pi Q - Q)$, where D_μ multiplies block (s,a) by $\mu(s,a) > 0$, and $M_{k+1} = e_{(s_k,a_k)}\big[\mathcal{K}_{a_k}^\dagger(R + \gamma Q_k(s_k',a_k')) - Q_k(s_k,a_k)\big] - F(Q_k)$. Given the past, (s_k, a_k, s_k', a_k') is drawn afresh, so M_{k+1} is a martingale difference with $\mathbb{E}[\|M_{k+1}\|^2 \mid \mathcal{F}_k] \le C(1 + \|Q_k\|^2)$. F is affine, $F(Q) = GQ + g$ with $G = D_\mu(\gamma\mathcal{P}^\pi - I)$ and $(\mathcal{P}^\pi Q)(s,a) = \mathcal{K}_a^\dagger(\sum_{s'} P \sum_{a'} \pi\, Q(s',a'))$. For $0 < h \le 1/\max\mu$ and every block, $\|((I + hG)Q)(s,a)\| \le (1 - h\mu(s,a))\|Q(s,a)\| + h\mu(s,a)\gamma\kappa\|Q\|_{\infty,\mathrm{op}} \le \big(1 - h\mu_{\min}(1-\gamma\kappa)\big)\|Q\|_{\infty,\mathrm{op}}$, so the logarithmic norm of G in $\|\cdot\|_{\infty,\mathrm{op}}$ is at most $-\mu_{\min}(1-\gamma\kappa) < 0$. Every eigenvalue of G therefore has negative real part, and both $\dot{Q} = F(Q)$ and the scaled limit $\dot{Q} = GQ$ have globally asymptotically stable equilibria (Q^π and 0 respectively; D_μ is invertible, so $F(Q) = 0$ iff $\mathcal{T}^\pi Q = Q$). The conditions of the ODE method of Borkar & Meyn (2000) hold, giving $\sup_k\|Q_k\| < \infty$ and $Q_k \to Q^\pi$ almost surely. □

PROOF OF PROPOSITION 1

Proof. A latent-blind policy is a latent-aware policy that ignores ρ, and by Theorem 1 its action-value function in the MDP on (s,ρ) is $\tau(\rho Q^\pi(s,a))$. Hence $V^\pi(s,\rho) = \sum_a \pi(a \mid s)\tau(\rho Q^\pi(s,a)) \le \tau(\rho Q^\pi(s,\pi^+(s,\rho)))$. Unrolling as in the classical policy-improvement theorem (Sutton & Barto, 2018, §4.2), $V^\pi(s_0,\rho_0) \le \mathbb{E}^{\pi^+}[\sum_{t<n}\gamma^t r_t] + \gamma^n\mathbb{E}^{\pi^+}[V^\pi(s_n,\rho_n)]$. Since $|V^\pi(s,\rho)| \le \|Q^\pi\|_{\infty,\mathrm{op}}\|\rho\|_1$ and $\|\rho_n\|_1 \le \kappa^n\|\rho_0\|_1$, the last term vanishes when $\gamma\kappa < 1$, and the first converges to $V^{\pi^+}(s_0,\rho_0)$. □

PROOF OF THEOREM 3

Proof. In CP each $\mathcal{K}_a$ and $\mathcal{K}_a^\dagger$ is positive, so Q_t^β is self-adjoint for all t and $\mathcal{K}_a(\rho) \succeq 0$ whenever $\rho \succeq 0$. For an arbitrary (possibly history-dependent, latent-aware) policy, let q_t^π denote its action value from step t; the continuation of π after one more step is again a policy. We prove by backward induction that $q_t^\pi(s,\rho,a) \le \tau(\rho Q_t^\beta(s,a))$ for all policies, s, a and $\rho \succeq 0$. At $t = H-1$ both sides equal $\tau(R(s,a)\mathcal{K}_a(\rho)) = \tau(\rho\,\mathcal{K}_a^\dagger(R(s,a)))$. For $t < H-1$, writing $\rho' = \mathcal{K}_a(\rho) \succeq 0$ and $\bar{\pi}(\cdot)$ for the (history-dependent) distribution of the next action,

$$\begin{aligned} q_t^\pi(s,\rho,a) &= \tau(R(s,a)\rho') + \gamma\sum_{s'} P(s' \mid s,a)\sum_{a'}\bar{\pi}(a')\,q_{t+1}^{\pi'}(s',\rho',a') \\ &\le \tau(R(s,a)\rho') + \gamma\sum_{s'} P(s' \mid s,a)\sum_{a'}\bar{\pi}(a')\,\tau\big(\rho'\,Q_{t+1}^\beta(s',a')\big) \\ &\le \tau(R(s,a)\rho') + \gamma\sum_{s'} P(s' \mid s,a)\,\tau\big(\rho'\,V_{t+1}^\beta(s')\big) = \tau(\rho\,Q_t^\beta(s,a)). \end{aligned}$$

The first inequality is the induction hypothesis at $\rho' \succeq 0$; the second uses $Q_{t+1}^\beta(s',a') \preceq V_{t+1}^\beta(s')$ (Prop. 3(i)) and that $Y \mapsto \tau(\rho' Y)$ is monotone for $\rho' \succeq 0$; the last equality is the dual identity. Taking the supremum over policies gives the statement for q_t^*. For latent-blind π, Theorem 1 gives $\tau(\rho(Q_t^\beta - Q_t^\pi)) \ge 0$ for every $\rho \succeq 0$, i.e. $Q_t^\pi \preceq Q_t^\beta$.

Infinite horizon. Unital positive maps have operator norm 1 (Russo–Dye; Bhatia, 2007, Ch. 2), so by Prop. 3(iii) $\|\mathcal{T}_\beta Q\|_{\infty,\mathrm{op}} \le R_{\max} + \gamma(\|Q\|_{\infty,\mathrm{op}} + \beta^{-1}\log K)$ and the ball of radius M_β is invariant. On it LSE_β is Lipschitz with constant 1 on the commuting sector (which $\mathcal{T}_\beta$ preserves when all $U_a, R(s,a)$ lie in one maximal abelian subalgebra) and $e^{2\beta M_\beta}$ in general (Prop. 4(a,b)), giving the stated moduli.

Truncation bound (no contraction needed). Fix n. The finite-horizon argument above, applied with horizon n, bounds the n-step truncated return of every policy by $\tau(\rho Q_0^{\beta,(n)}(s,a))$. In CP, $\rho_t \succeq 0$ and $\tau(\rho_t) = \tau(\rho)$, so $|r_t| \le R_{\max}\tau(\rho)$ and the remaining rewards contribute at most $\sum_{t\ge n}\gamma^t R_{\max}\tau(\rho) = \gamma^n R_{\max}\tau(\rho)/(1-\gamma)$. This proves the truncation bound. When $\mathcal{T}_\beta$ contracts, $Q_0^{\beta,(n)} \to Q^\beta$ as $n \to \infty$ and the tail vanishes, so the fixed point is a certificate. □

Proof of Proposition 2

Proof. (i) For $Y \in \mathcal{D}$ and $U_a \in \mathcal{D}$ unitary, $U_a^* Y U_a = U_a^* U_a Y = Y$, so $\mathcal{K}_a^\dagger$ is the identity on $\mathcal{D}$; functional calculus acts entrywise on diagonal operators, so LSE_β maps $\mathcal{D}^K$ to $\mathcal{D}$ coordinatewise, and coordinate j of $\mathcal{T}_\beta$ is the classical soft Bellman operator with reward $r_j(s,a) = R(s,a)_{jj}$. (ii) For $\rho = N e_j e_j^*$, $\mathcal{K}_a(\rho) = U_a \rho U_a^* = \rho$, so the latent is constant and every policy faces the classical MDP with reward r_j; hence $q_t^*(s,\rho,a)$ is its optimal value $q_{j,t}^*(s,a)$, while $\tau(\rho Q_t^\beta(s,a)) = (Q_t^\beta(s,a))_{jj}$ is its soft-optimal value. With $\delta_t = \max_{s,a}(q_{j,t}^\beta - q_{j,t}^*) \ge 0$ we have $\delta_{H-1} = 0$ and, since $\mathrm{lse}_\beta(u) \le \max u + \beta^{-1}\log K$, $\delta_t \le \gamma(\beta^{-1}\log K + \delta_{t+1})$, which gives the bound. (iii) $N = 1$ gives $\mathcal{M}_1 = \mathbb{C}$, $\mathcal{K}_a^\dagger = \mathrm{id}$ and $\mathrm{LSE}_\beta = \mathrm{lse}_\beta$. □

C Proofs for Section 4

Proof of Proposition 3

Proof. Each $e^{\beta x_i} \succeq e^{-\beta M} I$, so $S = \sum_i e^{\beta x_i} \succeq K e^{-\beta M} I \succ 0$ and $\log S$ is defined and self-adjoint. (i) $S \succeq e^{\beta x_i}$ and log is operator monotone (Bhatia, 1997, Ch. V), so $\log S \succeq \beta x_i$. (ii) $e^{\beta(x_i + cI)} = e^{\beta c} e^{\beta x_i}$. (iii) With $m = \max_i \|x_i\|$, $S \preceq K e^{\beta m} I$ and operator monotonicity of log give $\mathrm{LSE}_\beta(x) \preceq (m + \beta^{-1}\log K) I$, while (i) gives $\mathrm{LSE}_\beta(x) \succeq x_1 \succeq -mI$. (iv) $U^* f(A) U = f(U^* A U)$ for any functional calculus. (v) In a joint eigenbasis all operations are entrywise. □

Lemma 1 (Scalar case)**.** *For real $u, v \in \mathbb{R}^K$ and every $\beta > 0$, $|\mathrm{lse}_\beta(u) - \mathrm{lse}_\beta(v)| \le \max_i |u_i - v_i|$, with equality for uniform shifts.*

Proof. With $D = \max_i |u_i - v_i|$, monotonicity and the shift identity give $\mathrm{lse}_\beta(v) - D \le \mathrm{lse}_\beta(u) \le \mathrm{lse}_\beta(v) + D$. □

Proof of Proposition 4

Proof. (a) In the joint eigenbasis, coordinate j of Δ is $\Delta_j = \mathrm{lse}_\beta(x^{(j)}) - \mathrm{lse}_\beta(y^{(j)})$, and by Lemma 1, $|\Delta_j| \le \max_i |h_i^{(j)}|$. Taking $\max_j$ gives the operator-norm bound; averaging squares gives $\|\Delta\|_2^2 = \frac{1}{N}\sum_j \Delta_j^2 \le \frac{1}{N}\sum_j \sum_i |h_i^{(j)}|^2 = \sum_i \|h_i\|_2^2$. A uniform shift $h_i = cI$ attains the first bound. For the second, take $K = 2$, $x_1 = \Delta_0 I$, $x_2 = 0$ and perturb only x_1: the ratio is the softmax weight $(1 + e^{-\beta\Delta_0})^{-1} \to 1$.

(b) Let $X = \sum_i e^{\beta x_i}$, $Y = \sum_i e^{\beta y_i}$, both $\succeq m := K e^{-\beta M}$. From $\log x = \int_0^\infty [(1+s)^{-1} - (x+s)^{-1}] ds$,

$$\log X - \log Y = \int_0^\infty (Y+s)^{-1}(X - Y)(X+s)^{-1}\, ds,$$

so for $p \in \{2, \infty\}$, $\|\log X - \log Y\|_p \le \|X - Y\|_p \int_0^\infty (m+s)^{-2} ds = \|X - Y\|_p / m$. (The one-line formula $\int_0^1 ((1-t)Y + tX)^{-1}(X - Y)\, dt$ is valid only when X and Y commute.) By Duhamel's formula $e^{\beta x} - e^{\beta y} = \beta \int_0^1 e^{u\beta x}(x - y) e^{(1-u)\beta y} du$, so $\|e^{\beta x_i} - e^{\beta y_i}\|_p \le \beta e^{\beta M} \|h_i\|_p$. Combining and dividing by β, $\|\Delta\|_p \le \frac{e^{\beta M}}{K} e^{\beta M} \sum_i \|h_i\|_p$.

(c) In the common basis, coordinate j of the $K = N$ family has values 1 for action j and 0 otherwise. Perturbing x_i by $\epsilon\, e_i e_i^*$ changes coordinate j only through action j, with softmax weight $p_\beta = e^\beta/(e^\beta + N - 1)$, so $\Delta = \epsilon p_\beta I + O(\epsilon^2)$ and $\|\Delta\|_2 = \epsilon p_\beta + O(\epsilon^2)$, whereas $\|h_i\|_2 = \epsilon/\sqrt{N}$

for each i and $(\sum_i \|h_i\|_2^2)^{1/2} = \epsilon$. Hence the max-aggregated ratio tends to $p_\beta\sqrt{N}$ and the ℓ_2-aggregated ratio to $p_\beta \le 1$. At $N = 25$, $\beta = 5$: $p_\beta\sqrt{N} = 4.304$.

□

Proof of Proposition 5

Proof. On the commuting sector LSE_β is the scalar log-sum-exp in each eigen-coordinate, which is non-decreasing in each argument; $\mathcal{K}_a^\dagger$ is positive, so $\mathcal{T}_\beta$ inherits monotonicity. In general the argument fails because $x \mapsto e^{\beta x}$ is not operator monotone (Bhatia, 1997, Ch. V); E12 gives explicit violations (Sec. 6). $\mathcal{T}^\pi$ is affine with positive linear part in CP, hence monotone. □

D Proofs for Section 5

Proof of Theorem 4

Proof. Fix a prefix g of length $T-3$ and actions a, b, c. The histories $h = (g, a, b)$ and $h' = (g, b, a)$ are permutations of each other, so $\sigma(h) = \sigma(h')$ and $f(\sigma(h), c) = f(\sigma(h'), c)$. Therefore $\max\{|f - q_T(h, c)|, |f - q_T(h', c)|\} \ge \frac{1}{2}|q_T(h, c) - q_T(h', c)| = \frac{1}{2N}|\operatorname{Tr}(W X_g(A_a A_b - A_b A_a)A_c)|$. If $[A_a, A_b] \neq 0$ then $X_g[A_a, A_b]A_c \neq 0$ (the outer factors are invertible), so $W \mapsto \operatorname{Tr}(W X_g[A_a, A_b]A_c)$ is a non-zero linear functional whose zero set is a hyperplane, of Lebesgue measure zero. The exact operator form is Corollary 1. □

Proof of Theorem 5

Proof. (a) Distinct positive words are distinct reduced words, so in a free group of rank K the K^{T-1} histories of length $T-1$ give pairwise distinct X_h. For each of the finitely many pairs $h \neq h'$ and each a, $\operatorname{Tr}(W(X_h - X_{h'})A_a) = 0$ defines a hyperplane in W; outside their union all values $q_T(h, a)$ with distinct h differ for every a. An exact table indexed by $(\sigma(h), a)$ must then have $\sigma(h) \neq \sigma(h')$ for all $h \neq h'$, i.e. at least K^{T-1} observation values and K^T entries. For $SO(3)$ and $SE(3)$ (connected, non-solvable), almost every K-tuple generates a free group of rank K (Epstein, 1971).

(b) Step-2 nilpotency means every bracket is central. For central $[X, Y]$, $e^X e^Y = e^{X+Y+\frac{1}{2}[X,Y]}$, and by induction $e^{\Omega_{h_1}} \cdots e^{\Omega_{h_t}} = \exp\big(\sum_i \Omega_{h_i} + \frac{1}{2}\sum_{i<j}[\Omega_{h_i}, \Omega_{h_j}]\big)$ (the new bracket at each step is central, so it commutes past the accumulated exponent). The first sum equals $\sum_a n_a(h)\Omega_a$ and the second $\sum_{a<b} L_{ab}(h)[\Omega_a, \Omega_b]$. Hence X_h, and $q_T(h, \cdot)$, are functions of $(n(h), L(h))$, which takes at most $(T+1)^K(2T^2+1)^{K(K-1)/2}$ values.

Bounded deficit for compact groups. For orthogonal X, $\|X\|_1 = 1$ and $|q_T| \le \|W\|$, so the constant $f = 0$ already has error at most $\|W\|$ for every T. □

Proof of Theorem 6

Proof. (i) is Corollary 1. (ii) For any d-dimensional model, $g(uv) = \frac{1}{d}\operatorname{Tr}(W_d Q_u Q_v) = \frac{1}{d}\sum_{i,j}(W_d Q_u)_{ji}(Q_v)_{ij}$ is an inner product between a vector in $\mathbb{C}^{d^2}$ depending only on u and one depending only on v, so every Hankel block has rank at most d^2; with $d = N$, $W_d = W$, $Q_a = A_a$ this also shows $\operatorname{rank}\mathbf{H}_f \le N^2$. Exactness forces $\mathbf{H}_g = \mathbf{H}_f$. (iii) On the block $B = \mathbf{H}_f[\mathcal{U}, \mathcal{V}]$ the model's block G has rank at most d^2, so $\|B - G\|_F \ge (\sum_{j>d^2}\sigma_j(B)^2)^{1/2}$ by the Eckart–Young theorem, and the largest entry of $B - G$ is at least $\|B - G\|_F/\sqrt{|\mathcal{U}||\mathcal{V}|}$. □

Additional Implementation Notes

Spectral clipping. For the soft recursion with function approximation one may clip eigenvalues to $[-M_\beta, M_\beta]$; by Theorem 3 the exact recursion never leaves this ball, so clipping is a numerical safeguard only.

Numerical evaluation of LSE_β**.** $\sum_i e^{\beta x_i}$ has condition number up to $e^{2\beta M}$; for $\beta M \gtrsim 10$ we evaluate LSE_β in extended precision (E7, E17).

E EXTENSIONS

This appendix extends the results of Sections 3 and 5 in three directions: approximate operator values, noisy or random latent dynamics, and reducible action representations. Experiment E17b (`experiments/e17b_extension_checks.py` in https://anonymous.4open.science/r/opq-RL-2027-0D23) checks each statement on finite instances.

Proposition 6 (Robust latent-aware improvement)**.** *Let π be latent-blind with $\gamma\kappa < 1$, and let $\hat{Q}$ satisfy $\max_{s,a}\|\hat{Q}(s,a) - Q^\pi(s,a)\| \le \varepsilon$. Let $\hat{\pi}(s,\rho) \in \arg\max_a \tau(\rho\hat{Q}(s,a))$. Then for all (s,ρ),*

$$V^{\hat{\pi}}(s,\rho) \;\ge\; V^\pi(s,\rho) - \frac{2\varepsilon\,\|\rho\|_1}{1-\gamma\kappa}.$$

Proof. By Hölder, $|\tau(\rho(\hat{Q} - Q^\pi)(s,a))| \le \varepsilon\|\rho\|_1$. Hence $q^\pi(s,\rho,\hat{\pi}(s,\rho)) \ge \tau(\rho\hat{Q}(s,\hat{\pi})) - \varepsilon\|\rho\|_1 \ge \tau(\rho\hat{Q}(s,a)) - \varepsilon\|\rho\|_1 \ge q^\pi(s,\rho,a) - 2\varepsilon\|\rho\|_1$ for every a. Averaging over $a \sim \pi(\cdot \mid s)$ gives $q^\pi(s,\rho,\hat{\pi}(s,\rho)) \ge V^\pi(s,\rho) - 2\varepsilon\|\rho\|_1$. Since $q^\pi(s,\rho,a) = \mathbb{E}[r_0 + \gamma V^\pi(s_1,\rho_1)]$, unrolling n steps under $\hat{\pi}$ gives $V^\pi(s_0,\rho_0) \le \mathbb{E}^{\hat{\pi}}\big[\sum_{t<n}\gamma^t r_t\big] + 2\varepsilon\sum_{t<n}\gamma^t\,\mathbb{E}\|\rho_t\|_1 + \gamma^n\mathbb{E}\,V^\pi(s_n,\rho_n)$. With $\|\rho_t\|_1 \le \kappa^t\|\rho_0\|_1$ and $|V^\pi(s,\rho)| \le \|Q^\pi\|_{\infty,\mathrm{op}}\|\rho\|_1$, letting $n \to \infty$ proves the claim. □

Combined with Theorem 2, this turns an operator TD estimate with error ε into a latent-aware controller whose loss relative to π^+ of Proposition 1 is $O(\varepsilon)$.

Proposition 7 (Noisy and random latent dynamics)**.** *Suppose that at each step, given the action a, the latent is updated by a linear map $\mathcal{K}_a^\omega$, where $\omega \sim \nu_a$ is drawn independently of the past and of s', and $\sup_\omega\|(\mathcal{K}_a^\omega)^\dagger\| \le \kappa$. Let $\bar{\mathcal{K}}_a = \mathbb{E}_\omega\mathcal{K}_a^\omega$. (i) Theorem 1 and Corollary 1 hold with $\mathcal{K}_a^\dagger$ replaced by $\bar{\mathcal{K}}_a^\dagger$. (ii) If every $\mathcal{K}_a^\omega$ is positive—for example, a completely positive trace-preserving quantum channel with decoherence or amplitude damping—then Theorem 3 holds with $\bar{\mathcal{K}}_a^\dagger$, for every policy, including policies that observe the realised latent. In the infinite-horizon truncation bound, $\tau(\rho)$ is replaced by $\sup_t \mathbb{E}\,\tau(\rho_t)$, which equals $\tau(\rho)$ for trace-preserving maps.*

Proof. (i) In the backward induction of Theorem 1, the one-step identity becomes $q_t^\pi(s,\rho,a) = \mathbb{E}_\omega\big[\tau\big(\mathcal{K}_a^\omega(\rho)\,(R(s,a)+\gamma Y)\big)\big]$ with Y as before, using the independence of ω and s'. By linearity this is $\tau(\bar{\mathcal{K}}_a(\rho)(R+\gamma Y)) = \tau(\rho\,\bar{\mathcal{K}}_a^\dagger(R+\gamma Y))$. The infinite-horizon argument is unchanged, since $\|\bar{\mathcal{K}}_a^\dagger\| \le \kappa$. (ii) Apply the induction of Theorem 3 conditionally on ω, with $\rho' = \mathcal{K}_a^\omega(\rho) \succeq 0$. The two inequalities hold for each ω, and taking $\mathbb{E}_\omega$ of the final expression gives $\tau(\rho\,\bar{\mathcal{K}}_a^\dagger(R+\gamma\sum_{s'} PV_{t+1}^\beta))$. Positive maps preserve self-adjointness, so Q_t^β remains self-adjoint. □

Proposition 8 (Structure-refined operator dimension)**.** *Let the A_a be unitary and let $\mathcal{A}$ be the unital algebra they generate. $\mathcal{A}$ is a finite-dimensional C*-algebra, so $\mathcal{A} \cong \bigoplus_{k=1}^r \mathcal{M}_{n_k}$, acting on $\mathbb{C}^N$ with multiplicities m_k ($\sum_k m_k n_k = N$). For Lebesgue-almost every W, $\operatorname{rank}\mathbf{H}_f = \dim\mathcal{A} = \sum_k n_k^2$, and the minimal exact complex operator dimension satisfies*

$$\Big(\sum_k n_k^2\Big)^{1/2} \;\le\; d_{\min} \;\le\; \sum_k n_k.$$

In particular $d_{\min} = N$ when the action representation is irreducible ($\mathcal{A} = \mathcal{M}_N$, the generic case), and $d_{\min} < N$ whenever an irreducible component occurs with multiplicity $m_k > 1$.

Proof. The span of all words A_w, including the empty word, is closed under multiplication and contains I, so it equals $\mathcal{A}$. By Cayley–Hamilton each $A_a^{-1} = A_a^*$ is a polynomial in A_a, so $\mathcal{A}$ is $*$-closed, hence a C*-algebra, and the structure theorem applies. Since words span $\mathcal{A}$ on both sides, $\operatorname{rank}\mathbf{H}_f$ equals the rank of the bilinear form $B_W(X,Y) = \operatorname{Tr}(WXY)$ on $\mathcal{A}\times\mathcal{A}$. At $W = I$ this form is non-degenerate: if $\operatorname{Tr}(XY) = 0$ for all $Y \in \mathcal{A}$, take $Y = X^*$ to get $X = 0$. Its Gram determinant in a basis of $\mathcal{A}$ is a polynomial in W that does not vanish at I, so it is non-zero for almost every W, and then $\operatorname{rank}\mathbf{H}_f = \dim\mathcal{A}$. The lower bound follows from Theorem 6(ii). For the upper

bound, choose a unitary V with $V^* A_w V = \bigoplus_k (I_{m_k} \otimes B_w^{(k)})$. Then $\mathrm{Tr}(W A_w) = \sum_k \mathrm{Tr}(\widetilde{W}_k B_w^{(k)})$, where $\widetilde{W}_k$ is the partial trace over the multiplicity factor of the k-th diagonal block of $V^* W V$. The model $Q_a = \bigoplus_k B_a^{(k)}$ with readout proportional to $\bigoplus_k \widetilde{W}_k$ has dimension $\sum_k n_k$ and reproduces f exactly. Real orthogonal actions are covered by complexification. □

E17b checks these statements on finite instances.

- Proposition 7, with random choices between two amplitude-damping channels: the exact representation holds to 3.9×10^{-16}, and the certificate has $0/20$ violations at $\beta \in \{2, 10\}$.
- Proposition 6, with $\varepsilon = 0.05$: the bound holds with margin 1.12.
- Proposition 8, for a reducible $N{=}4$ representation with blocks $n = (2, 1)$ and multiplicities $m = (1, 2)$: the Hankel rank is $5 = \dim \mathcal{A}$, and an explicit $d{=}3$ model reproduces all 243 words of length 5 to 1.8×10^{-15}, whereas a generic $N{=}3$ representation has Hankel rank 9.

F Summary of All Experiments

Table 2 maps each theoretical claim to the experiment that tests it; Table 3 summarises the experiments.

Table 2: Theory–experiment map.

Claim	Exp	Prediction / outcome
T1 exact representation	E17, E18	readout = exact expected return ($\leq 7 \times 10^{-16}$); zero-shot transfer across ρ_0
T2 operator TD	E17, E18	$\max\|Q_k - Q^\pi\| \to 0$; OpQ-PI
P1 improvement	E18	OpQ-PI improves on the uniform policy
T3 certificate	E17, E18	0 violations; $Q^\beta - Q^\pi \succeq 0$; slack consistent with OpQ-soft's ceiling
P2 reduction	E9, E17	exact collapse to soft Q-learning; slack $\to 0$ at pure eigenstates
P4 Lipschitz	E7	commuting ratio ≤ 1; 4.30 at $K{=}N{=}25$; non-commuting ratio > 1, growing in β
P5 monotonicity	E12	commuting only
T4 aliasing	E17, E18	positive swap gap; order-blind scalar agents fail when the latent varies
T5 free vs. nilpotent	E10	depth profile truncates for $\mathfrak{n}_3$, not for $\mathfrak{so}(3)$
T6 dimension	E8, E14–E17	exact at $d{=}N$ (real form: $2N_F$); Hankel rank N^2
T6 scope	E13–E16	sequence models vs. operator model

Exp	Tests	Setup	Headline result
E7	Prop. 4	Commuting/generic pairs; $K{=}N{=}25$ construction; 80-digit checks	Commuting ratio ≤ 1.000000; 4.304 vs. 2.25; non-commuting $0.97/1.17/1.48$ at $\beta{=}1/5/10$
E8, E13	Thm. 6	Random orthogonal, $N{=}5$, $K{=}4$, $T{=}6$	OpQ at $d{=}5$: 2.6×10^{-30} (125 params); best baseline LSTM 0.033 (17729 params)
E9	Prop. 2	Diagonal unitaries, $N \in \{1, 4\}$	Coincides with soft Q-learning to 1.6×10^{-14}
E10	Thm. 5	$\mathfrak{so}(3)$ vs. $\mathfrak{n}_3$, matched norm, $T \leq 10$	$\mathfrak{so}(3)$: constant per-depth term 0.221; $\mathfrak{n}_3$: zero beyond depth one
E12	Prop. 5	$K{=}4$, $N{=}5$, $\beta{=}2$	$0/2000$ commuting vs. $72/2000$ generic violations; persists at scale 10^{-4}
E14	Thm. 6	$\mathfrak{so}(3)$, $K{=}6$, $T{=}8$	OpQ at $d{=}3$: 6.0×10^{-32} (63 params); best baseline LSTM 0.003 (18241 params)
E15	Thm. 6	Displacements, $N_F{=}6$	OpQ at $d{=}12$: 9.7×10^{-15} (1008 params); best baseline GRU 0.042 (1121 params)
E16	Thm. 6	$\mathfrak{se}(3)$ hinge chain	OpQ at $d{=}4$: 2.2×10^{-31} (112 params); best baseline GRU 0.092 (13697 params)
E17	Thms. 1–3, 4, 6	Random OLDP, $\|\mathcal{S}\|{=}2$, $K{=}N{=}3$, $H{=}4$	See Sec. 6
E18	End-to-end control	C1–C3, 7 agents, 5 seeds	Table 1, Fig. 3

Table 3: All experiments at a glance.

G Additional Results

E18: full control results

Agent	C1 random OLDP fixed 100	C1 random OLDP fixed 3000	C1 random OLDP random 100	C1 random OLDP random 3000	C2 $SO(3)$ pointing fixed 100	C2 $SO(3)$ pointing fixed 3000	C2 $SO(3)$ pointing random 100	C2 $SO(3)$ pointing random 3000	C3 two-qubit prep. fixed 100	C3 two-qubit prep. fixed 3000	C3 two-qubit prep. random 100	C3 two-qubit prep. random 3000
OpQ-soft (learned R)	0.77	0.84±0.01	0.95	0.95±0.00	0.94	0.98±0.00	0.95	0.99±0.00	0.38	0.83±0.08	0.58	0.68±0.00
OpQ-soft (known R)	0.78	0.84±0.01	0.95	0.95±0.00	0.95	0.98±0.00	0.96	0.99±0.00	0.67	0.79±0.00	0.63	0.68±0.00
OpQ-PI	0.84	0.86±0.00	0.96	0.97±0.01	0.95	0.91±0.04	0.96	0.98±0.00	0.28	0.28±0.00	0.59	0.70±0.00
Scalar Q on (t, m)	0.68	0.87±0.06	0.42	0.51±0.03	0.37	0.69±0.01	0.05	0.06±0.07	0.22	0.19±0.42	0.02	0.10±0.12
Scalar Q, history table	0.33	0.94±0.03	0.11	0.50±0.05	0.31	0.75±0.02	-0.01	0.05±0.04	0.28	0.28±0.00	-0.04	-0.04±0.00
DQN on ρ_t	0.96	1.00±0.00	0.89	0.99±0.00	0.85	0.99±0.00	0.73	0.99±0.00	-0.08	1.00±0.00	0.04	0.87±0.03
DRQN (GRU) on history	0.76	1.00±0.00	0.44	0.95±0.01	0.69	0.97±0.01	0.28	0.94±0.01	0.14	0.96±0.08	-0.01	-0.01±0.05

Table 4: E18: normalised return after 100 and 3000 training episodes (mean; ± one standard deviation over 5 seeds at 3000).

Environment / regime	Greedy readout of exact Q^β (known model) β=1	β=2	β=5	β=10	Exact OpQ-PI (Prop. 1)
random OLDP, fixed	0.76 (2.30)	0.76 (1.94)	0.86 (1.88)	0.87 (1.94)	0.86
random OLDP, random	0.97 (1.73)	0.95 (1.48)	0.95 (1.43)	0.96 (1.48)	0.97
$SO(3)$ pointing, fixed	0.88 (2.44)	0.99 (1.72)	0.99 (1.32)	0.98 (1.19)	0.88
$SO(3)$ pointing, random	0.98 (2.43)	0.99 (1.69)	0.99 (1.30)	0.99 (1.18)	0.98
two-qubit preparation, fixed	0.28 (11.69)	0.79 (6.17)	0.79 (3.02)	0.79 (2.01)	0.28
two-qubit preparation, random	0.70 (12.60)	0.69 (6.63)	0.68 (3.23)	0.68 (2.16)	0.70

Table 5: E18: exact-planning references with the known model. These are the ceilings that the learned OpQ agents approach (compare Table 1). Numbers in parentheses give the certificate $\mathbb{E}\,\max_a \tau(\rho_0 Q_0^\beta(s_0, a))$ relative to V^*; it is always at least 1, as Theorem 3 requires.

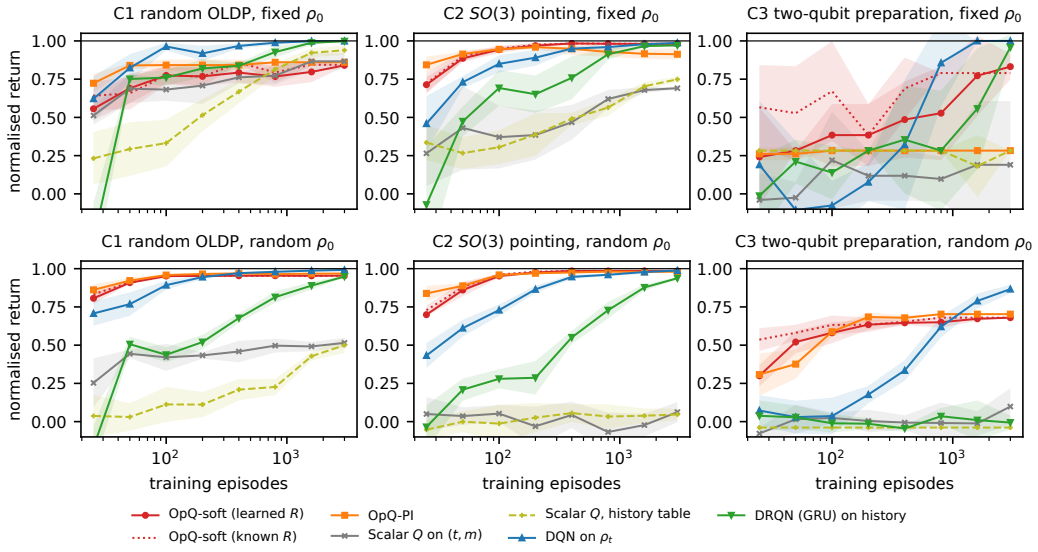


Figure 3: **E18.** Learning curves: normalised return of the greedy policy, evaluated exactly, against the number of training episodes (mean ± one standard deviation over 5 seeds).

Representation benchmark: full results

H Experimental Details

All code is in the anonymized repository at https://anonymous.4open.science/r/opq-RL-2027-0D23; `run_e18.sh` and `run_rest.sh` reproduce every number, and `make_figures.py` every figure

Model	Orthogonal (E8/E13) params	Orthogonal (E8/E13) test MSE	$\mathfrak{so}(3)$ (E14) params	$\mathfrak{so}(3)$ (E14) test MSE	$\mathfrak{se}(3)$ (E16) params	$\mathfrak{se}(3)$ (E16) test MSE	Displacements (E15) params	Displacements (E15) test MSE
Scalar order-blind (t, a_t)	19	0.095	41	0.303	41	0.861	31	0.116
Windowed table (t, a_{t-1}, a_t)	64	0.094	216	0.277	216	0.730	156	0.115
GRU, hidden 4	113	0.077	137	0.062	137	0.363	137	0.058
GRU, hidden 16	1025	0.041	1121	0.010	1121	0.106	1121	0.042
GRU, hidden 64	13313	0.040	13697	0.005	13697	0.092	13697	0.064
LSTM, hidden 64	17729	0.033	18241	0.003	18241	0.107	18241	0.051
Self-attention, width 32	7713	0.064	7841	0.096	7841	0.366	7777	0.059
1D-CNN on action embeddings	973	0.083	979	0.202	979	0.578	979	0.122
OpQ, d=1	5	0.091	7	0.361	7	1.108	–	–
OpQ, d=2	20	0.078	28	0.165	28	0.773	28	0.080
OpQ, d=3	45	0.069	63	6.0×10^{-32}	63	0.361	–	–
OpQ, d=4	80	0.052	–	–	112	2.2×10^{-31}	112	0.045
OpQ, d=5	125	2.6×10^{-30}	–	–	–	–	–	–
OpQ, d=6	–	–	–	–	–	–	252	0.012
OpQ, d=8	–	–	–	–	–	–	448	0.005
OpQ, d=10	–	–	–	–	–	–	700	0.002
OpQ, d=12	–	–	–	–	–	–	1008	9.7×10^{-15}
Planted solution		1.1×10^{-32}		2.1×10^{-32}		1.1×10^{-31}		2.1×10^{-32}
Var(test returns)		0.094		0.412		1.282		0.117
History table	4096	0	1679616	0	1679616	0	46656	0

Table 6: Representation benchmark (E8, E13–E16): held-out test MSE. OpQ(d) is exact at $d = N$ (real form $2N_F = 12$ for displacements, to 10^{-14}), as Theorem 6 predicts.

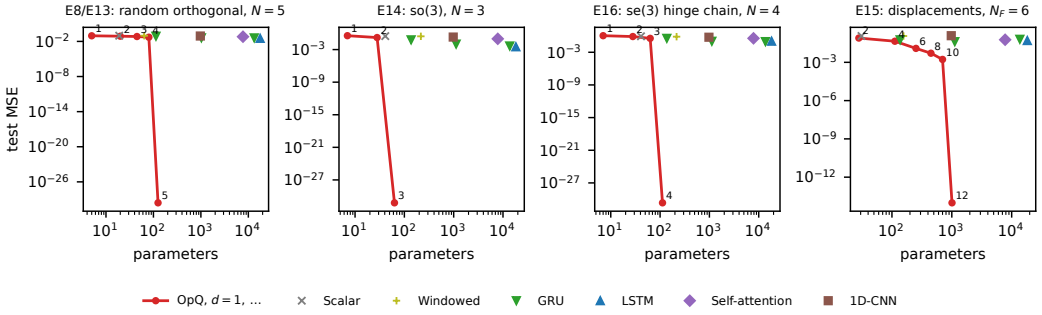


Figure 4: **E8, E13–E16.** Test MSE against the number of parameters. OpQ(d) (red, labelled by d) against sequence-model and table baselines.

and table. Everything runs on one CPU core (the complete suite takes about three hours). Each script fixes its random seeds.

E18: END-TO-END CONTROL

Environments. All three are OLDPs in the conjugation picture with observable state $s_t = (t, m_t)$: a step counter and an observable mode whose transitions do not depend on the latent. *C1 (random OLDP):* N=3, K=4, two modes, H=6, γ=0.95; $U_a = \exp(0.8(G_a - G_a^\top))$ with G_a standard Gaussian; $R(m, a)$ random symmetric; $P(\cdot \mid m, a)$ flat Dirichlet; Gaussian reward noise of standard deviation 0.1. *C2 ($SO(3)$ pointing):* the latent is the dyad $\rho = 3nn^\top$ of a body axis n ($n_0 = e_3$); K=6 actions rotate by ± 0.6 rad about the three body axes; two modes with targets t_1, t_2 (random unit vectors), reward $r_t = (t_{m_t} \cdot n_{t+1})^2 = \tau(t_m t_m^\top \rho_{t+1})$, mode transitions $P(m' \mid 0) = (0.8, 0.2)$ and $P(m' \mid 1) = (0.3, 0.7)$ for every action; H=8, γ=0.95. *C3 (two-qubit preparation):* N=4, gates $\{H \otimes I, I \otimes H, T \otimes I, I \otimes T, \mathrm{CNOT}\}$, $\rho_0 = 4|00\rangle\langle 00|$, H=8, γ=1; the target $|\psi\rangle$ is the output of a random 8-gate word, redrawn until it is entangled (reduced purity < 0.9) and far from $|00\rangle$ ($|\langle 00|\psi\rangle|^2 < 0.5$), so the optimum is 1; the reward is the fidelity $\tau(|\psi\rangle\langle\psi|\rho_H)$ at the last step and 0 before.

Regimes. *Fixed*: one initial latent (C1: a random density operator; C2: $n_0 = e_3$; C3: $|00\rangle$). *Random*: at every episode $\rho_0 = Nvv^*$ with v uniform on the real (C1, C2) or complex (C3) unit sphere; evaluation averages over 8 held-out draws. *Transfer* (Table 1, last column per environment): agents are trained in the fixed regime and evaluated, without retraining, on the same 8 held-out initial latents.

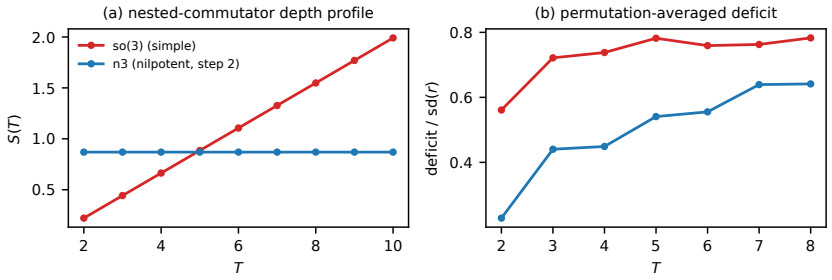


Figure 5: **E10.** (a) The nested-commutator depth profile has a constant per-depth contribution for $\mathfrak{so}(3)$ and is exactly zero beyond depth one for $\mathfrak{n}_3$. (b) The permutation-averaged deficit, normalised by the spread of returns, is bounded for both algebras, as $|r| \le \|W\|$ requires. Nilpotency controls the *size* of an exact history statistic (Thm. 5), not whether order matters.

Agents. All agents see the same information: s_t, the rewards, and either ρ_t (latent-aware agents) or the action history (from which ρ_t is determined, because ρ_0 is observed and the $\mathcal{K}_a$ are fixed). *OpQ-soft*: one $N \times N$ operator per (t, m, a), updated by the sample form of (4), $Q(s,a) \leftarrow Q(s,a) + \alpha\big(\mathcal{K}_a^\dagger(\hat{R}(m,a) + \gamma\,\mathrm{LSE}_\beta(Q(s',\cdot))) - Q(s,a)\big)$, with β=5, and acting greedily on $\tau(\rho_t Q(s_t,a))$. The update is unbiased for (4) because (4) is linear in $P(\cdot \mid s,a)$. The reward observable $\hat{R}(m,a)$ is estimated online by least squares from the scalar rewards (ridge 10^{-6}, orthonormal Hermitian basis), since $r = \tau(R\rho')$ is linear in ρ'; the *known-R* variant uses the true R. *OpQ-PI*: operator TD (3) evaluating the uniform policy (behaviour is uniform throughout), followed by the latent-aware greedy improvement of Prop. 1. *Scalar Q on* (t,m) and *scalar Q, history table* (keyed by $(t, m, a_1 m_1 \ldots a_t m_t)$): tabular Q-learning. The step size of all tabular and operator agents is $\alpha = \max(0.02, n^{-0.6})$, where n counts visits to the entry. *DQN on* ρ_t: MLP [one-hot t, one-hot m, coordinates of ρ_t] $\to 64 \to 64 \to K$ (ReLU), Adam (10^{-3}), batch 64, replay 5×10^4, target network every 250 steps, one update per step. *DRQN*: GRU with 64 hidden units on [previous action, previous reward, mode, one-hot t, ρ_0 at step 0], Adam (10^{-3}), 32 episodes per batch, 4 updates per episode, target network every 50 episodes, replay of 5000 episodes. Exploration is ϵ-greedy, annealed linearly from 1 to 0.05 over the first 60% of 3000 episodes, for every agent except OpQ-PI.

Evaluation. The greedy policy of each agent is evaluated *exactly*: all mode branches are enumerated and mean rewards are used, so the reported curves contain no Monte-Carlo noise. V^* is computed by backward dynamic programming over all K^H action histories and modes; V_{rand} comes from Theorem 1 applied to the uniform policy. Scores are $(V - V_{\mathrm{rand}})/(V^* - V_{\mathrm{rand}})$, averaged over initial latents; we report the mean and standard deviation over 5 seeds. Code: `opq/envs.py`, `opq/agents.py`, `opq/deep_agents.py`, `experiments/e18_control.py`, `experiments/e18_transfer.py`.

E8, E13–E16: REPRESENTATION BENCHMARK

The target is $f(w) = \frac{1}{N}\,\mathrm{Re}\,\mathrm{Tr}(W A_{a_1} \cdots A_{a_T})$ for uniformly random words (3000 training, 1000 test). *Orthogonal* (E8/E13): N=5, K=4, T=6, $A_a = \exp(G_a - G_a^\top)$ with G_a having i.i.d. $\mathcal{N}(0, 0.6^2)$ entries, W random symmetric. $\mathfrak{so}(3)$ (E14): K=6 rotations by ± 0.6 rad about the coordinate axes, T=8, W Gaussian. $\mathfrak{se}(3)$ *hinge chain* (E16): N=4 (homogeneous coordinates); each of K=6 creases is a revolute joint with a random unit axis n_a, a Gaussian hinge point p_a and an angle $\theta_a \sim U[0.5, 1.5]$, with twist $\Omega_a = \left(\begin{smallmatrix} \theta_a[n_a]_\times & -\theta_a[n_a]_\times p_a \\ 0 & 0 \end{smallmatrix}\right)$; T=8, W Gaussian. *Displacements* (E15): Fock space truncated at N_F=6, $D(\alpha_a) = \exp(\alpha_a a^\dagger - \bar{\alpha}_a a)$ for K=6 random α_a ($|\alpha_a| \approx 0.5$), T=6, W random Hermitian; complex matrices are encoded in real form as $X \mapsto \left(\begin{smallmatrix} \mathrm{Re}\,X & -\mathrm{Im}\,X \\ \mathrm{Im}\,X & \mathrm{Re}\,X \end{smallmatrix}\right)$, so the real operator dimension that suffices is $2N_F$=12. In the untruncated algebra $\mathrm{ad}^2_{\Omega_a} = 0$, but truncation breaks nilpotency. The depth profile $S(10)$ is 3.0×10^4, 1.1×10^6 and 2.4×10^7 at $N_F = 6, 12, 24$, so it grows rather than vanishes. The order-dependent phase of one swapped pair, $\arg\langle 0|D(\alpha)D(\beta)|0\rangle / \langle 0|D(\beta)D(\alpha)|0\rangle = -0.3596$ rad, matches the untruncated value $2\,\mathrm{Im}(\alpha\bar{\beta})$ to four digits. In the $\mathfrak{se}(3)$ chain the same two folds applied in opposite orders move the marked point $(1,1,0)$ by up to 2.00.

Before any fitting, a planted-solution check evaluates the operator model with the *true* generators; its test MSE is at floating-point level in every environment (Table 6). ***OpQ***(d): K learned $d \times d$ operators and a readout, $g(w) = \mathrm{Tr}(W_d Q_{a_1} \cdots Q_{a_T})$, initialised at $Q_a = \exp$(random skew-symmetric, scale 0.3); full-batch Adam (0.02, cosine decay, 1500 epochs, gradient clipping at 1), best of 4 restarts by training loss, followed by at most 2000 L-BFGS iterations. ***Baselines***: ordinary least squares on one-hot (t, a_t) (order-blind) and on one-hot (t, a_{t-1}, a_t) (windowed table); GRUs with $4, 16, 64$ hidden units and an LSTM with 64; a single-head self-attention block (width 32, learned positional embeddings, feed-forward block, mean pooling); and a 1D-CNN (learned 3-dimensional action embedding, two convolutions with kernel 3 and 16 channels, mean pooling). Neural baselines are trained with full-batch Adam (0.01; 0.005 for attention; cosine decay, 1000 epochs). For linear models the parameter count is the rank of the design matrix. Code: `opq/rep_models.py`, `experiments/rep_benchmark.py`.

E10: NILPOTENT VERSUS NON-NILPOTENT ALGEBRAS

$\mathfrak{so}(3)$ (generators L_x, L_y, L_z) and the Heisenberg algebra $\mathfrak{n}_3$ (strictly upper-triangular X, Y, Z with $[X, Y] = Z$), both in dimension 3 with every generator rescaled to Frobenius norm $\sqrt{2}$, so that the nilpotency step is the only difference; W is a general Gaussian matrix (a symmetric W would annihilate every $\mathfrak{so}(3)$ bracket term). (a) $S(T) = \sum_{t=2}^{T} \max_{a \neq b} |\mathrm{Tr}(W \,\mathrm{ad}_{\Omega_a}^{t-1} \Omega_b)|/N$ by exact iterated bracketing. (b) For $T = 2, \ldots, 8$ and 400 random words (100 for $T \geq 7$), the return is averaged over all $T!$ reorderings of the word; this average is the best predictor available from the action counts. We record the mean absolute deviation of the true return from it, normalised by the standard deviation of returns at that T. Code: `experiments/e10_depth_profile.py`.

E7, E9, E12: THE OPERATOR LOG-SUM-EXP AND THE COMMUTATIVE REDUCTION

E7.

(i) 20,000 commuting configurations ($N \in \{2, \ldots, 5\}$, $K \in \{2, 3, 4\}$, β cycling through $\{0.05, 0.1, 0.25, 0.5, 1, 2, 5, 10\}$, shared random orthonormal eigenbasis, eigenvalues in $[-1, 1]$); half use local perturbations and half independent draws.

(ii) The construction of Prop. 4(c) at $K{=}N{=}25$, $\beta{=}5$, $\epsilon{=}10^{-6}$.

(iii) 80 non-commuting configurations per $\beta \in \{1, 5, 10\}$ ($N, K \in \{2, 3\}$, perturbations at scale 10^{-12}), evaluated in 80-digit arithmetic with `mpmath`. In double precision, $\sum_i e^{\beta x_i}$ loses its smallest eigenvalues once $e^{2\beta M}$ approaches 10^{16}, which produces spurious ratios in the thousands.

E9. Three observable states, $K{=}3$, $\gamma{=}0.9$, $\beta{=}3$, diagonal unitaries $U_a = \mathrm{diag}(e^{i\phi})$ and diagonal $R(s, a)$, for $N \in \{1, 4\}$. We compare (a) 200 iterations of the exact operator recursion against N independent scalar soft Bellman iterations, and (b) 20,000 sample-based updates fed to both the operator learner and N scalar soft Q-learners with the same samples and step sizes. *E12.* We apply the aggregation $x \mapsto R + \gamma\, \mathrm{LSE}_\beta(x)$ to $K{=}4$ self-adjoint 5×5 inputs ($\gamma{=}0.95$, $\beta{=}2$). We draw $y_i = x_i + P_i$ with $P_i \succeq 0$ at scale 0.5 (2000 trials each for commuting and generic inputs), then sweep the scale over $\{0.5, 0.1, 0.01, 10^{-3}, 10^{-4}\}$ with 1000 generic trials each. A trial counts as a violation when $\lambda_{\min}$ of the output difference is below -10^{-12}. Code: `experiments/e07_lse_lipschitz.py`, `e09_commutative_reduction.py`, `e12_monotonicity.py`.

E17: EXACT CHECKS OF THE COMPOSITIONAL RECURSION

A random OLDP with $|\mathcal{S}|{=}2$, $K{=}3$, $N{=}3$, $H{=}4$, $\gamma{=}0.9$, P flat Dirichlet, $U_a = \exp(0.8(G - G^\top))$, R random symmetric. *Theorem 1*: Q_0^π from the backward recursion is compared with the exact expected return, obtained by enumerating every branch, at 20 random (s, a, ρ), for ρ a density operator (CP) and a rotation (MP). *Theorem 2*: the infinite-horizon Q^π is computed by 400 Bellman iterations; operator TD uses i.i.d. uniform (s, a) and $\alpha_k = (1 + k/200)^{-0.7}$. *Theorem 3*: Q_0^β for $\beta \in \{1, 5, 20, 50\}$, with LSE_β evaluated in 120-digit arithmetic, is compared with q^* from exhaustive latent-aware dynamic programming at 30 random (s, a, ρ), and $\lambda_{\min}(Q_0^\beta - Q_0^\pi)$ is taken over

5 random latent-blind policies. The commuting check uses diagonal sign matrices and diagonal R at β=200 and pure eigenstates. *Theorem 4*: three generators at scale 0.6, Gaussian W, T=3. *Theorem 6*: a Hankel block over the 40 words of length $\leq$ 3, with numerical rank taken at relative threshold 10^{-9}. Code: `experiments/e17_exact_checks.py`.

COMPUTATIONAL REQUIREMENTS

Python 3.12 with NumPy, SciPy, mpmath, Matplotlib, JAX (CPU) and Optax. On one CPU core: E18 learning curves $\approx$ 70 minutes; E18 transfer $\approx$ 25 minutes; E17 $\approx$ 1 minute; E7 $\approx$ 10 minutes (dominated by 80-digit arithmetic); E9, E10, E12 under 2 minutes each; representation benchmark $\approx$ 15 minutes per environment.