Inverting the Bellman Equation: From Q-Values to World Models

Alistair Letcher[α] **Mattie Fellows**[α] **Alexander D. Goldie**[α]

Jonathan Richens[β] **Jakob N. Foerster**[α] **Oliver Richardson**[γ]

[α] FLAIR, University of Oxford [β] Google DeepMind [γ] Mila, University of Montreal

Abstract

Model-based and model-free reinforcement learning are traditionally viewed as separate paradigms: instead of learning a model of the transition kernel P, model-free agents typically estimate value functions tied to a specific policy and reward. In this paper, we challenge this dichotomy by proving that value-based agents trained on a sufficiently rich set of reward functions, e.g. using goal-conditioned RL, implicitly encode a unique and accurate world model. To extract this model in practice, we introduce *P-learning*, an inverse analogue to Q-learning that samples from an agent's Q-values, policies and rewards to decode its internal model of the environment. We then provide sufficient conditions on the type and number of goals for which agents encode the true kernel P, covering both stochastic and deterministic MDPs over finite or continuous state spaces. Even when our assumptions are violated, we empirically demonstrate that agents trained on a handful of reward functions encode accurate dynamics in `Reacher`, `MountainCar` and stochastic variants of `FourRooms`. Surprisingly, we find that policies trained exclusively on a `Reacher` agent's implicit world model are quasi-optimal on out-of-distribution, velocity-based goals despite position-only training – suggesting that agents contain hidden generalisation capabilities and providing a new lens into the connection between model-based, model-free, and goal-conditioned RL.

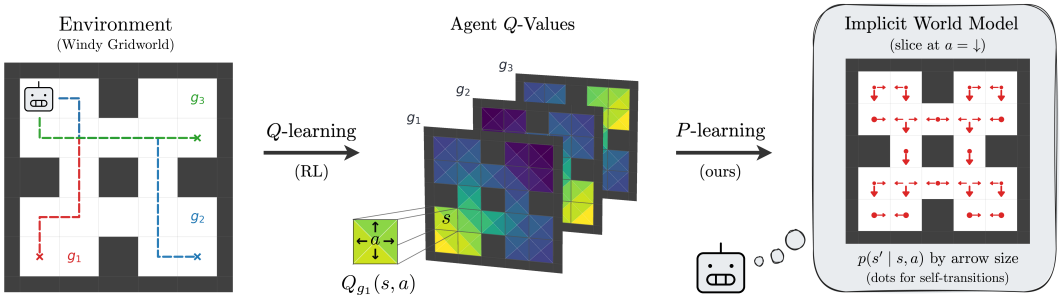


Figure 1: Illustration of P-learning: extracting the world model contained in an agent's Q-values. Code available at `github.com/aletcher/inverting-bellman`.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

The model-free paradigm of reinforcement learning (RL) has proven effective in training highly capable agents across a variety of domains [1, 2, 3], sidestepping the challenges associated with modelling complex world dynamics [4, 5]. However, the policies or value functions learnt by such agents are tied to reward functions, making it difficult to repurpose them for new tasks without additional data, or interpret what they *understand* of the world – as opposed to what they *value*. Accessing an agent's world model would enable us to inspect its beliefs, support transfer learning, and bound its generalisation capabilities in safety-critical settings [6].

Unfortunately, distinct world models can give rise to identical value functions and optimal policies, an obstacle formalised as *value equivalence* [7, 8]. For example, in a deterministic maze with a binary-reward goal, the optimal value of a state is γ^d, where γ is the discount factor and d is the path length to the goal. In particular, Q-values encode whether an action moves the agent closer to the goal, but do not pin down the state that an action leads to, i.e. the transition kernel, if there are equidistant states. This underdeterminacy implies that model-free agents do not uniquely encode the environment they inhabit, suggesting a fundamental divide between model-free and model-based RL.

However, agents are increasingly trained on a *variety* of goals – e.g. via successor features [9, 10], forward-backward learning [11] or goal-conditioned RL (GCRL) [12] – which may provide additional constraints on the underlying dynamics. This raises the core question motivating this paper:

When do model-free agents implicitly encode an accurate model of their environment?

We hypothesise that value-based agents trained on a sufficiently rich set of goals implicitly learn a *unique* model of the world dynamics, even in stochastic and continuous MDPs, and that this model is *accurate* insofar as Q-values are accurate. Taking GCRL as the broader framework for agents trained on goals parameterised by arbitrary reward functions, our work makes three contributions to test this hypothesis bridging model-free, model-based and goal-conditioned RL.

Methodologically, we introduce P-learning (Section 3), an analogue to Q-learning illustrated in Figure 1. Instead of iteratively updating value estimates for a fixed environment, P-learning updates a candidate world model (WM) to be progressively more consistent with fixed value functions, effectively 'inverting' the Bellman equation. We make this precise by proving that tabular P-learning converges to a solution involving the Moore-Penrose pseudo-inverse (Theorem 1). Our algorithm enables the efficient extraction of an implicit WM from a collection $(Q_g, \pi_g, r_g)_{g\in\mathcal{G}}$ of Q-functions, policies and reward functions – noting that policies are usually induced by Q-functions via argmax or softmax, so (Q, π, r) is often equivalent to (Q, r) in practice.

Theoretically, we prove sufficient conditions on the goals and reward functions for which P is uniquely identifiable from Q-values (Section 4), breaking value equivalence. Four regimes emerge: in *deterministic finite* MDPs, a single generic goal suffices (Theorem 2); in *stochastic finite* MDPs, goals must span the state space (Theorem 3), with tight error bounds in both regimes when Q-values are ϵ-approximate. In the *deterministic continuous* case, a finite number of goals is, surprisingly, sometimes sufficient (Theorems 8 and 9). In *stochastic continuous* MDPs, identifiability is driven by *coverage* instead of count: a goal set with non-empty interior for Gaussian rewards (Theorem 6), or covering the state space for indicator rewards (Theorem 7). Taken together, our results show that methods like GCRL can bridge model-free and model-based RL, with the collection $(Q_g, \pi_g, r_g)_{g\in\mathcal{G}}$ becoming informationally equivalent to the kernel P when $\mathcal{G}$ is sufficiently rich.

Empirically, we show that P-learning is highly effective in extracting accurate world models from agents, even with very few goals (Section 5). We validate our theory across three increasingly stochastic variants of `FourRooms`, with $|\mathcal{G}| = 1$, $|\mathcal{G}| = 4$ and $|\mathcal{G}| = 20$ goals respectively sufficient for recovery. In `Reacher`, with only $|\mathcal{G}| = 4$ goals despite a continuous state space, we find that implicit WMs are even accurate over variables that rewards never directly depend on. Surprisingly, this accuracy transfers to near-optimal planning inside the WM for out-of-distribution goals including reaching specific angular velocities, despite position-only training. We study these "implicit generalisation capabilities" in `MountainCar` and show that inverting the setup, by training on velocity-based goals, yields a near-identical WM – revealing that agents with widely different objectives can secretly encode similar models. Finally, we identify a strong correlation (Spearman $\rho = 0.98$) between an agent's performance and the accuracy of its internalised WM, suggesting that GCRL is an implicitly hybrid method linking model-free and model-based RL. Related work is deferred to Section 6.

2 Background

Goal-Conditioned RL (GCRL). We consider goal-augmented MDPs $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{G}, P, \mu, r, \gamma)$ with finite or continuous state space $\mathcal{S}$, arbitrary action space $\mathcal{A}$ and goal set $\mathcal{G}$, Markov transition kernel $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, initial state distribution $\mu \in \Delta(\mathcal{S})$, discount factor $\gamma \in [0, 1)$, and goal-conditioned reward function $r : \mathcal{S} \times \mathcal{G} \to \mathbb{R}$. As in early GCRL literature [12], goals $g \in \mathcal{G}$ are specified by *arbitrary* functions of the state, as opposed to exclusively *indicator* goals defined by $r(s, g) = \delta_{gs}$, also called *goal states* [13, 14]. Denoting the return of a trajectory $\tau = (s^t, a^t)_{t\geq 0}$ with respect to g by $R(\tau, g) \coloneqq \sum_{t\geq 0} \gamma^t r(s^{t+1}, g)$, the objective of GCRL is to learn a goal-conditioned policy $\pi : \mathcal{S} \times \mathcal{G} \to \Delta(\mathcal{A})$ such that each $\pi_g \coloneqq \pi(\cdot, g)$ maximises the objective $J_g(\pi) \coloneqq \mathbb{E}_{\tau\sim\pi_g}[R(\tau, g)]$, where $\tau \sim \pi_g$ is shorthand for independently sampling $s_0 \sim \mu$, $a_t \sim \pi(s_t, g)$, $s_{t+1} \sim P(s_t, a_t)$ for all t. The *value* and *action-value* functions are defined by

$$V^\pi(s, g) \coloneqq \mathbb{E}_{\tau\sim\pi_g}\left[R(\tau, g) \mid s^0 = s\right] \text{ and } Q^\pi(s, a, g) \coloneqq \mathbb{E}_{\tau\sim\pi_g}\left[R(\tau, g) \mid s^0 = s, a^0 = a\right] .$$

Our finite-MDP results extend to goals with arbitrary *termination indicators* [15], while some of our continuous-MDP results include explicit termination conditions; see Appendix D.1 for details.

Q-Iteration. For a fixed goal-augmented MDP, we define $\mathcal{Q} \coloneqq \{Q : \mathcal{S} \times \mathcal{A} \times \mathcal{G} \to \mathbb{R}\}$ as the space of action-value functions, and the goal-conditioned Bellman operators $\mathcal{T}^\pi, \mathcal{T}^\star : \mathcal{Q} \to \mathcal{Q}$ as

$$\mathcal{T}^\pi(Q)(s, a, g) \coloneqq \mathbb{E}_{s'\sim P(s,a), a'\sim\pi(s',g)}\left[r(s', g) + \gamma Q(s', a', g)\right] ,$$
$$\mathcal{T}^\star(Q)(s, a, g) \coloneqq \mathbb{E}_{s'\sim P(s,a)}\left[r(s', g) + \gamma \sup_{a'\in\mathcal{A}} Q(s', a', g)\right] ,$$

with the corresponding Bellman and Bellman optimality equations $\mathcal{T}^\pi(Q^\pi) = Q^\pi$, $\mathcal{T}^\star(Q^\star) = Q^\star$. The aim of value-based methods is to approximate a solution to the Bellman optimality equation. A classical approach, known as Q*-iteration* [1, 16], starts with an initial guess Q_0 and iteratively applies the Bellman operator $Q_{n+1} \coloneqq \mathcal{T}^\star(Q_n)$ to get a sequence $(Q_n)_{n\in\mathbb{N}}$ converging to a fixed-point. In practice, with a parametric family $\theta \mapsto Q_\theta \in \mathcal{Q}$, we fix a reference distribution $d \in \Delta(\mathcal{S} \times \mathcal{A} \times \mathcal{G})$, e.g. induced by an exploration policy, and measure fixed-point violation via the *Bellman residual* $\mathcal{L}(\theta, \theta_n) \coloneqq \|\mathcal{T}^\star(Q_{\theta_n}) - Q_\theta\|_d^2$, where $\|X\|_d^2 \coloneqq \frac{1}{2}\mathbb{E}_d[X^2]$. One can then update θ_n by minimising $\mathcal{L}(\theta, \theta_n)$ with respect to θ. Writing $\delta_{\theta,\theta_n} \coloneqq \mathcal{T}^\star(Q_{\theta_n}) - Q_\theta$ for the *temporal difference* error, the residual is usually minimised using gradient-based methods, with parameter updates driven by

$$\nabla_\theta \mathcal{L}(\theta, \theta_n) = -\mathbb{E}_{(s,a,g)\sim d}\left[\delta_{\theta,\theta_n}(s, a, g)\nabla_\theta Q_\theta(s, a, g)\right] .$$

Notably, this *does not* correspond to gradient descent on a fixed objective; the target $\mathcal{T}^\star(Q_{\theta_n})$ depends on current parameters, so $\mathcal{L}(\theta, \theta_n)$ shifts with each iteration. Estimating this expectation using environment samples yields a family of temporal difference (TD) algorithms [17] including Q-learning [18, 19], DQN [2, 20] and PQN [21], adapted here with goal-conditioning. These differ in the iterative optimization method, choice of sampling distribution d, number of steps performed on $\mathcal{L}(\theta, \theta_n)$ before updating θ_n, parametric family, and resulting convergence guarantees.

3 P-Learning

While Q-iteration treats the kernel P as fixed and searches for a function Q^* that satisfies $\mathcal{T}^*(Q^*) = Q^*$, we consider the inverse problem of extracting an "internal" model $P^\star$ of the environment from a fixed agent with goal-conditioned Q-values, policy π, and known reward function r. Formally, let $\mathcal{P} \coloneqq \{\mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})\}$ be the set of transition kernels, and consider a parametric family $\varphi \mapsto P_\varphi \in \mathcal{P}$. We write $\mathcal{T}_\varphi^\pi = \mathcal{T}_{P_\varphi}^\pi$ for the corresponding Bellman operator, making explicit the dependence on WM parameters φ. Our guiding observation is that for an agent (Q, π), a model P_φ satisfying the Bellman equation $\mathcal{T}_\varphi^\pi(Q) = Q$ is compatible with the agent's behavior, and may thus be viewed as (one possible) subjective belief of the agent about the environment.

A natural objective is therefore to minimise the Bellman residual $\mathcal{L}(\varphi) = \|\mathcal{T}_\varphi^\pi(Q) - Q\|_d^2 =: \|\delta_\varphi\|_d^2$, this time with respect to φ. We call this P**-learning**. If P_φ is differentiable, we can minimise this loss with gradient-based methods, where the direction of parameter updates is driven by

$$\nabla_\varphi \mathcal{L}(\varphi) = \mathbb{E}_{(g,s,a)\sim d}\left[\delta_\varphi(s, a, g)\nabla_\varphi \mathcal{T}_\varphi^\pi(Q)(s, a, g)\right] .$$

Unpacking this, we define the TD estimate $\hat{\delta}_\varphi(s,a,g,s',a') \coloneqq r(s',g) + \gamma Q(s',a',g) - Q(s,a,g)$, and prove in Appendix A that under standard regularity assumptions,

$$\nabla_\varphi \mathcal{L}(\varphi) = \mathbb{E}_{g,s,a,s',a'}\left[\delta_\varphi \hat{\delta}_\varphi \nabla_\varphi \log P_\varphi(s'|s,a)\right],$$

where $(s,a,g) \sim d$ and $s' \sim P_\varphi(s,a), a' \sim \pi(s',g)$. For large or continuous state spaces, the gradient can instead be estimated via the reparameterisation trick [22, 23]. Intuitively, while Q-learning finds a Q-function $Q_{\theta^\star}$ that is most consistent with the MDP (best satisfies the Bellman equation $\mathcal{T}^\star[Q_{\theta^\star}] = Q_{\theta^\star}$), P-learning finds a model $P_{\varphi^\star}$ that is most consistent with a Q-function (best satisfies the Bellman equation $\mathcal{T}^\pi_{\varphi^\star}[Q] = Q$). See Algorithm 1 for pseudocode.

What does P-learning converge to? Unlike Q-learning [24], P-learning solves a supervised learning problem using a *stationary* objective $\mathcal{L}(\varphi)$, inheriting the standard convergence guarantees of SGD algorithms.[1] But what does the solution look like? Fixing the policy π and omitting the superscript for brevity, let us first rewrite the goal-conditioned Bellman operator as

$$\mathcal{T}_{P_\varphi}(Q)(s,a,g) = \mathbb{E}_{s'\sim P_\varphi(s,a)}\left[M(s',g)\right], \quad \text{where} \quad M(s',g) \coloneqq r(s',g) + \gamma V(s',g). \tag{1}$$

When $\mathcal{S}, \mathcal{A}, \mathcal{G}$ are finite, we can assemble a matrix $M \in \mathbb{R}^{|\mathcal{G}|\times|\mathcal{S}|}$ with entries $M_{lk} = M(s'_k, g_l)$, and the Bellman equation becomes a linear system, viewing $Q(s,a) \in \mathbb{R}^{|\mathcal{G}|}, P_\varphi(s,a) \in \mathbb{R}^{|\mathcal{S}|}$ as vectors:

$$Q(s,a,g_l) = \mathcal{T}_\varphi(Q)(s,a,g_l) = \sum\nolimits_k P_\varphi(s,a,s'_k) M_{lk} \quad \Longrightarrow \quad MP_\varphi(s,a) = Q(s,a) \tag{2}$$

for each s, a. Writing $J_\varphi(s,a)$ for the Jacobian of $P_\varphi(s,a)$ with respect to φ, P-learning by gradient descent takes the explicit form $\varphi_{n+1} = \varphi_n - \alpha_n \sum_{s,a} J_\varphi(s,a)^\top M^\top (MP_\varphi(s,a) - Q(s,a))$. In the tabular setting where $\mathcal{S}, \mathcal{A}, \mathcal{G}$ are finite and $P_\varphi(s,a) = \varphi(s,a)$ with $\varphi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^{|\mathcal{S}|}$, each Jacobian is the identity $J_\varphi(s,a) = I$ and the update decouples, using a constant step size α, as

$$\textbf{Tabular P-learning:} \quad P_{n+1}(s,a) = P_n(s,a) - \alpha M^\top (MP_n(s,a) - Q(s,a)). \tag{3}$$

As shown below, tabular P-learning converges to a particularly friendly closed-form solution.

Theorem 1 (Convergence of P-learning)**.** *In the tabular setting, fix a policy π, a Q-function $Q \in \mathcal{Q}$, initial $P_0(s,a) \in \Delta(\mathcal{S})$ for each (s,a), and let $\sigma_{\max}$ denote the largest singular value of M (Equation 1). Then P-learning (Equation 3) with step size $0 < \alpha < 2/\sigma^2_{\max}$ converges to*

$$P_\infty(s,a) = M^+ Q(s,a) + \left(I - M^+ M\right) P_0(s,a),$$

where M^+ is the Moore–Penrose pseudo-inverse of M. Equivalently, P_∞ is the orthogonal projection of P_0 onto the set of global minima $\mathcal{Z} = \arg\min_P \|\mathcal{T}_P(Q) - Q\|_2^2$.

link to proof

Theorem 1 reveals key conditions for P-learning to converge to the true world model. The term $(I - M^+M)P_0$ represents the underdetermination from value equivalence: when M is rank-deficient ($M^+M \neq I$), multiple WMs yield identical Q-values, and P-learning selects among them based on the initial guess P_0. [2] When M has **full rank**, the iteration converges to a *unique* solution M^+Q, in which case M^+ can be viewed as an **inverse Bellman map**, noting that $M^+ = M^{-1}$ when M is square. In this regime, P-learning can informally be viewed as a way to uniquely extract the agent's 'beliefs' about the world (implicit model P) by sampling its 'beliefs' about the value of states and actions (learnt function Q). If Q-values are moreover *exact*, i.e. satisfy $\mathcal{T}_P(Q) = Q$, then P-learning converges to the *true* kernel P. We provide intuition for how goals break value equivalence in Appendix B.1, and formal guarantees in Section 4.

Deterministic MDPs. If dynamics are deterministic, we can parameterise the model as a successor function $P_\varphi : \mathcal{S}\times\mathcal{A} \to \mathcal{S}$. The drawback is that backpropagation flows through $r(P_\varphi(s,a),g)$ and $Q(P_\varphi(s,a),a',g)$ via the chain rule, requiring these to be differentiable almost everywhere. This is usually satisfied when the Q-function is a neural network and r is distance-based (including $r_g(s) = \mathbf{1}[||s-g|| < \sigma]$); otherwise, methods like evolutionary strategies [26, 27] are appropriate. If $\mathcal{S}$ is finite, P-learning can instead be performed via *column-matching* (Section 4.1). The WM is then unique provided *column-injectivity*, a much weaker requirement than full-rankness.

[1] E.g. if the Robbins-Monro conditions are satisfied, stochastic gradient descent (SGD) converges to a stationary point [25].

[2] Note that P-learning can desert the probability simplex if M is rank-deficient; see Appendix B.2 for a projected variant.

4 Theoretical Results

Our core theoretical results are sufficient conditions on the goals and reward functions $(\mathcal{G}, r)$ that resolve the value equivalence problem via Q-values, meaning that a unique world model can be extracted from (Q, π, r). When Q-values are *exact*, we prove that the model coincides with the true transition kernel P. When Q-values are *approximate*, we provide bounds on the accuracy of the extracted world model, with further results for *arbitrary* Q-values in Appendix C.7.

For finite MDPs, our results characterise the number of goals $|\mathcal{G}|$ needed to identify P uniquely, for various families of reward functions r. For MDPs with continuous $\mathcal{S} \subseteq \mathbb{R}^d$, uniqueness is typically driven not by count, but coverage of $\mathcal{S}$ (excluding Theorems 8 and 9). In this regime, our results therefore provide guarantees for goals sampled from a *distribution* with sufficiently broad support, e.g. uniform over a bounded goal region, as is common in GCRL benchmarks like JaxGCRL [28]. However, the next section will demonstrate that even agents trained with a handful of goals, despite not covering $\mathcal{S}$, often encode P with high precision. We summarise our results in Table 1.

Notation. We write Q for arbitrary goal-conditioned Q-values, with M the corresponding function from Equation 1, and Q^π for the exact Q-values of policy π, with $M = M^\pi$. We say Q is ϵ-*approximate* if $\|Q(s,a,\cdot) - Q^\pi(s,a,\cdot)\|_1 \leq \epsilon$ for all s, a. The norm $\|\cdot\|$ without subscript refers to the standard Euclidean norm $\|\cdot\|_2$, and $B_\sigma(0) \coloneqq \{s \in \mathbb{R}^d \mid \|s\| < \sigma\}$ is the open ball of radius σ, with $\mathcal{S} + B_\sigma(0) \coloneqq \{s + x \mid s \in S, x \in B_\sigma(0)\}$. Finally, we write $\exists! P(Q, \pi, r, \gamma)$ or simply $\exists! P(Q)$ to assert that P is the unique world model which satisfies the Bellman equation $\mathcal{T}_P^\pi(Q) = Q$.

Table 1: Sufficient conditions on the goals $(\mathcal{G}, r)$ for which $\exists! P(Q^\pi)$.

State space $\mathcal{S}$	**Deterministic** P		**Stochastic** P		Reward r
Finite $\mathcal{S}$	$\lvert\mathcal{G}\rvert \geq 1$	*Prop 1*	$\lvert\mathcal{G}\rvert \geq \lvert\mathcal{S}\rvert$	*Prop 9*	**Gaussian**
	$\lvert\mathcal{G}\rvert \geq 1$	*Thm 2a*	$\lvert\mathcal{G}\rvert \geq \lvert\mathcal{S}\rvert$	*Thm 3a*	**Generic**
	$\lvert\mathcal{G}\rvert \geq \lvert\mathcal{S}\rvert$	*Thm 3b*	$\lvert\mathcal{G}\rvert \geq \lvert\mathcal{S}\rvert$	*Thm 3b*	**Spanning**†
Continuous $\mathcal{S} \subseteq \mathbb{R}^d$	$\lvert\mathcal{G}\rvert$ sufficiently large	*Thms 8,9*	$\mathcal{G}$ non-empty interior*	*Thm 6*	**Gaussian**
	$\mathcal{G} \supseteq \mathcal{S} + B_\sigma(0)$	*Thm 4*	$\mathcal{G} \supseteq \mathcal{S} + B_\sigma(0)^\star$	*Thm 7*	**Indicator**

* For unconditional policies; see Appendix D.3. † Including indicator goals, i.e. $r(s,g) = \delta_{sg}$.

4.1 Finite Space

Using our formulation of the Bellman equation (1) as $\mathbb{E}_{s' \sim P(s,a)}[M(s', g)] = Q(s, a, g)$, we show that unique recovery of P is guaranteed by different structural conditions on M depending on kernel stochasticity, but not on the choice of parameterisation for Q-values or policies.

Deterministic MDPs. For deterministic kernels $P : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$, the Bellman equation (1) collapses to $M(P(s,a), g) = Q(s,a,g)$. Writing $M(s') = M(s', \cdot)$ and $Q(s,a) = Q(s,a,\cdot)$, extracting P reduces to identifying which column of M matches $Q(s,a)$ via *column-matching*, i.e.

$$P(s,a) \coloneqq \arg\min_{s'} \|M(s') - Q(s,a)\|_1 .$$

In particular, the $\arg\min$ (and therefore the kernel P) is unique provided M is **column-injective**: $M(s) \neq M(s')$ for all $s \neq s'$. To quantify injectivity, we define the *column separation* of M as $\Delta(M) \coloneqq \min_{s \neq s'} \|M(s) - M(s')\|_1$. Theorem 2 shows that even with a single goal, (a) P is uniquely determined by exact Q-values Q^π for generic ("almost all") reward functions and (b) P can also be recovered *exactly* from ϵ-approximate Q-values if ϵ is sufficiently small.

Theorem 2 (Deterministic, Finite MDPs)**.** *Fix any deterministic finite MDP and goal set* $|\mathcal{G}| \geq 1$.

(a) For any policy π*, the set of reward functions* $r : \mathcal{S} \times \mathcal{G} \to \mathbb{R}$ *for which* M^π *is column-injective, and thus* $\exists! P(Q^\pi)$*, has full Lebesgue measure.*

(b) If Q *is* ϵ*-approximate and* M *is column-injective with column separation* Δ*, the estimator* $\hat{P}(s,a) = \operatorname{argmin}_{s'} \|Q(s,a) - M(s')\|_1$ *satisfies* $\hat{P} = P$ *for all* $\epsilon < \Delta/(2(1+\gamma m))$*, and any estimator fails to recover* P *in the worst-case for* $\epsilon \geq \Delta/2$.

link to proof

As a corollary, for any base reward r_0 and any random vector $\xi \in \mathbb{R}^{|\mathcal{S}|}$ whose distribution is absolutely continuous with respect to Lebesgue measure (e.g. Gaussian, or uniform on a bounded open region), the perturbed reward $r = r_0 + \xi$ satisfies $\exists! P(Q^\pi)$ almost surely (proof in Appendix C.1). This gives a recipe for ensuring identifiability by initialising rewards with noise, as demonstrated empirically in `FourRooms` (Appendix F). The same conclusion does not follow for general parametric families like indicator rewards; however, when the state space can be embedded in continuous space (e.g. in a 2D environment), we can guarantee $\exists! P(Q^\pi)$ for Gaussian rewards:

Proposition 1 (Gaussian Rewards). *Fix any deterministic finite MDP, policy π, and variance $\sigma > 0$. For any injective map $\varphi : \mathcal{S} \hookrightarrow \mathbb{R}^d$, the single-goal set $\mathcal{G} = \{g\}$ with Gaussian reward function $r_g(s) = \exp(-\|\varphi(s) - g\|^2 / 2\sigma^2)$ satisfies $\exists! P(Q^\pi)$ for Lebesgue-almost every $g \in \mathbb{R}^d$.* link to proof

Stochastic MDPs. For stochastic kernels, the Bellman equation (2) becomes a set of linear systems $M^\pi P(s,a) = Q^\pi(s,a)$, implying that P is unique if M is **full-rank**. In order to specify a concrete class of rewards for which $\exists! P(Q^\pi)$, we say that goals $(\mathcal{G}, r)$ **span the state space** if $\operatorname{rank}(R) = |\mathcal{S}|$, where $R_{lk} \coloneqq r(s_k, g_l)$. This includes the common indicator goals [13, 14] given by $\mathcal{G} = \mathcal{S}$ and $r(s,g) = \delta_{sg}$, since $\operatorname{rank}(R) = \operatorname{rank}(I) = |\mathcal{S}|$. Using this condition, we state our main result.

Theorem 3 (Finite MDPs). *Fix any finite MDP and goal set $|\mathcal{G}| \geq |\mathcal{S}|$.*

(a) For any policy π, the set of reward functions $r : \mathcal{S} \times \mathcal{G} \to \mathbb{R}$ for which M^π is full-rank, and thus $\exists! P(Q^\pi)$, has full Lebesgue measure.

(b) For any goals $(\mathcal{G}, r)$ that span $\mathcal{S}$ and any discount factor $\gamma \in [0,1)$, the set of policies π for which M^π is full-rank has full Lebesgue measure. Similarly, for any π, the set of $\gamma \in [0,1)$ for which M^π is full-rank has full Lebesgue measure. In particular, there exist $\gamma^-, \gamma^+ \in (0,1)$ such that M^π is full-rank for all $\gamma < \gamma^-$ or $\gamma > \gamma^+$. link to proof

(c) If Q is ϵ-approximate and M has full rank, the estimator $\hat{P}_{ij} = M^+ Q_{ij}$ satisfies

$$\|\hat{P}_{ij} - P_{ij}\|_1 \leq \|M^+\|_1 (1 + \gamma m)\epsilon\,,$$

and any estimator has worst-case error $(1+\gamma)\epsilon/2$.

Theorem 3(a) focuses on *generic* reward functions, in line with Theorem 2(a), while part (b) shows that spanning goals are also sufficient if we shift genericity to policies or discount factors. We provide a stronger version of this result in Appendix C.3 (Proposition 6), including optimal and entropy-regularised policies, and corollaries for perturbed and Gaussian rewards in Appendix C.4.

Noting that RL agents typically have much higher Q-errors off-distribution, we prove in Appendix C.6 that our bounds from Theorem 2(b) and Theorem 3(c) hold verbatim for *on-support* state-action pairs, with ϵ replaced by $\epsilon_o \coloneqq \sup_{s \in \mathcal{S}_o,\, a \in \mathcal{A}_o(s)} \|Q(s,a,\cdot) - Q^\pi(s,a,\cdot)\|_1$ over visited states and actions $\mathcal{S}_o, \mathcal{A}_o(s)$. Unique identification of a world model on $\mathcal{S}_o$ then requires only $|\mathcal{G}| \geq |\mathcal{S}_o|$ goals. Going beyond approximate value functions, Appendix C.7 shows that a unique (though likely inaccurate) world model also exists for arbitrary Q-values, e.g. early in training, under similar conditions.

In some settings, such as single-goal contrastive RL [29], the actor learns a single *unconditional* policy $\pi_g = \pi_{g'}$ for all g, g', but a critic encodes Q-values Q^π_g with respect to all indicator goals $g \in \mathcal{S}$. In this regime, Theorem 3(b) strengthens to M^π *always* being full-rank (cf. Appendix C.5). If we also assume indicator goals, the upper bound improves to $\|\hat{P}_{ij} - P_{ij}\|_1 \leq 4\epsilon/(1-2\epsilon)$ for all $\epsilon < 1/2$ (Proposition 7), and M^π reduces to the *occupancy measure* [30].

Are multiple goals necessary? Could value equivalence instead be broken with multiple policies or discount factors, e.g. Multi-Rainbow [31] or Horde [32]? In Appendix C.8 we prove this is generally impossible: for any $|\mathcal{S}| \geq 4$ and any single reward function $r : \mathcal{S} \to \mathbb{R}$, there is a family of MDPs in which P cannot be identified from the set $\{Q^\pi_{r,\gamma} \mid \pi \in \Pi,\ \gamma \in [0,1)\}$ over *all* π and *all* γ.

Local MDPs. Sitting between deterministic and fully-stochastic settings, we say that an MDP is N-*local* if $\operatorname{supp}(s,a) \coloneqq |\{s' : P(s' \mid s,a) > 0\}| \leq N$ for every (s,a), e.g. a robot who can only move to a neighbourhood of its position in a single timestep. Appendix C.9 shows that P can be identified uniquely in local MDPs for (a) $|\mathcal{G}| \geq N - 1$ generic goals when the support is known, and (b) $|\mathcal{G}| \geq 2N - 1$ otherwise. This generalises the stochastic case (where the support $N = |\mathcal{S}|$ is trivially known) and the deterministic case (where $N = 1$ but the successor state is unknown).

4.2 Continuous Space

For continuous $\mathcal{S} \subseteq \mathbb{R}^d$, the matrix M is replaced by a possibly infinite family $\{m_g^\pi\}_{g \in \mathcal{G}}$, and invertibility or column-injectivity generalises to whether this family is **measure determining** or **point separating** for stochastic or deterministic MDPs; see Appendix D.2. We prove results for two classes of goals $\mathcal{G} \subseteq \mathbb{R}^d$: (sparse) indicator goals given by $r_g(s) = \mathbf{1}[||s - g|| < \sigma]$ with a threshold $\sigma > 0$, commonly used in GCRL, and Gaussian goals $r_g(s) = \exp(-||s - g||^2 / 2\sigma^2)$ with a variance $\sigma > 0$ controlling reward sparsity, which admit stronger guarantees. Theorem 4, stated below, is our most relevant result for MuJoCo environments like `Reacher`, which often use indicator goals and *termination upon arrival*, i.e. a termination criterion $D_g(s) = \mathbf{1}[||s - g|| \geq \sigma]$ that prevents further accumulation of reward once the goal is reached; see formal definition in Appendix D.1.

Theorem 4. *Fix any continuous deterministic MDP, policy π, threshold $\sigma > 0$ and a goal set $\mathcal{G} \supseteq \mathcal{S} + B_\sigma(0)$ with indicator rewards $r_g(s') = \mathbf{1}[\|s' - g\| < \sigma]$ and termination upon arrival. Then $\{m_g^\pi : g \in \mathcal{G}\}$ is point separating on $\mathcal{S}$, hence $\exists! P(Q^\pi)$.*

link to proof

Our full set of results, including those for stochastic kernels, are stated and proven in Appendix D. In the deterministic setting, we show that a sufficiently large but finite number of terminating Gaussian goals is sufficient to identify P (Theorem 8), while only $2d + 1$ non-terminating Gaussian goals are enough when V^π is real-analytic (Theorem 9). For stochastic kernels, Theorems 6 and 7 show that a goal set with non-empty interior suffices for Gaussian goals, and $\mathcal{G} \supseteq \mathcal{S} + B_\sigma(0)$ for indicator goals, but only for unconditional policies. The challenging intersection of stochastic kernels and general (goal-conditioned) policies is beyond the scope of this paper.

5 Experiments

We verify that our findings hold empirically in `MountainCar` [33] and MuJoCo `Reacher` [34], both of which are continuous and deterministic. We defer the finite and stochastic setting to Appendix F, where agents are shown to contain highly accurate world models using only $|\mathcal{G}| = 1, 4, 20$ goals for deterministic, local (windy) and fully stochastic (teleporting) variants of `FourRooms` [35] respectively. We train agents using PQN [21] with Hindsight Experience Replay (HER) [36], and extract each agent's implicit world model (WM) at the end of training using P-learning. Hyperparameters and environment details are in Appendix E. All results are reported over 10 seeds.

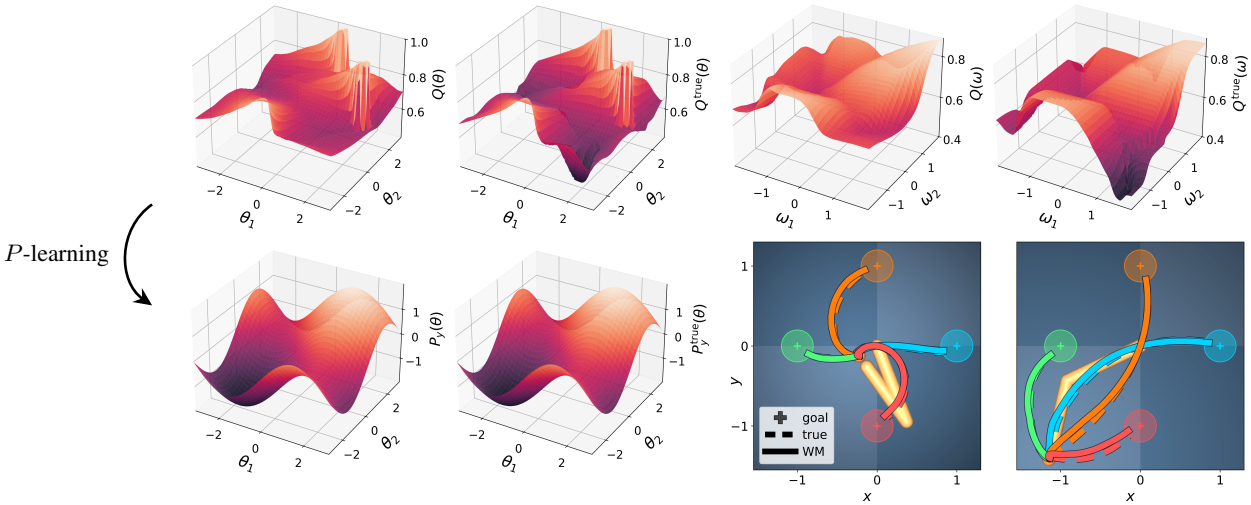


Figure 2: Q-values (top) and implicit WM (bottom) of a `Reacher` agent trained on $|\mathcal{G}| = 4$ goals. Learnt values $Q_g(\theta, \omega, a)$ are coarsely similar ($\texttt{NMSE} = 5.7 \times 10^{-1}$) to ground truth Q^{true}, shown for $g = (0, 1), a = (1, 1)$ and two slices $\omega = (1, -1), \theta = (1, -1)$. Nevertheless, the extracted WM $P(\theta, \omega, a)$ is highly accurate ($\texttt{NMSE} = 1.2 \times 10^{-4}$), shown for prediction of y-position P_y vs true P_y^{true} (same slice). Bottom right: WM/true rollouts of goal-conditioned policy for two start states.

5.1 Reacher

Our theoretical guarantees for continuous $\mathcal{S} \subseteq \mathbb{R}^d$ require an infinite number of indicator goals covering all dimensions of $\mathcal{S}$. However, we find that accurate world model extraction in Reacher is feasible even with a small number $|\mathcal{G}| = 4$ of goals defined only with respect to fingertip position $(x, y) = (\cos\theta_1 + \cos\theta_2, \sin\theta_1 + \sin\theta_2)$. We consider the goals with unit distance in the cardinal directions, $\mathcal{G} = \{(\pm 1, 0), (0, \pm 1)\}$, covering a limited portion of the fingertip space $[-2, 2]$. The indicator (sparse) reward function is $r_g(\theta, \omega, x) = \mathbf{1}\{\|x - g\| < \sigma\}$, with $\sigma = 0.2$ matching ReacherHard in MuJoCo Playground [37], and episodes terminate upon reaching the goal region.

We visualise slices of the learnt Q-values, extracted WM and trajectory rollouts in Figure 2. Despite imperfect Q-values (normalised mean squared error $\texttt{NMSE} = 5.7 \times 10^{-1}$), the WM matches ground truth with extremely high fidelity ($\texttt{NMSE} = 1.2 \times 10^{-4}$). To further validate accuracy, and determine whether P-learning reveals hidden generalisation capabilities, we train policies exclusively inside the extracted WM using value iteration, for three unseen goals that are increasingly far from the training distribution: **(i) far fingertip:** reach fingertip position $(x, y) = (\sqrt{2}, \sqrt{2})$, which is twice as far as any training goal; **(ii) target angle:** reach joint configuration $\theta = (0, \pi)$, noting that distinct angle configurations can induce identical positions; and **(iii) target velocity:** reach velocity $\omega = (2, 2)$, where ω covers a different dimension than the training goals. Despite the agent only training on position-based goals, the WM-learnt policy is quasi-optimal on all three tasks, as shown in Figure 3. Notably, the second and third goals would be impossible to achieve using position-based successor features, since the optimal Q-functions for these goals are not a linear combination of training Q-functions.

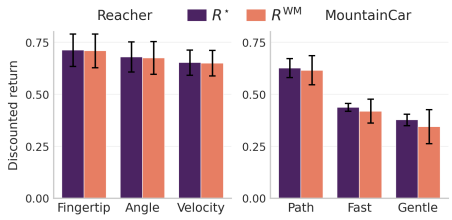


Figure 3: Mean return ± SE (10 training seeds and 512 env. resets), for optimal ($R^\star$) vs WM-trained (R^{WM}) policies on 3 unseen goals.

To further study the relationship between model-free training and implicit world models, we analyse whether more performant agents tend to encode more accurate WMs. We follow recent work suggesting that network capacity improves goal-conditioned performance [38] and plot agent return, MSE of the extracted WM, and return of a WM-trained policy on unseen goals, against architecture capacity (Q-network depth/width) in Figure 4. We run each configuration over ten seeds, and fix all other hyperparameters. The resulting Spearman rank correlation coefficient between agent return and WM error is $\rho = -0.98$ with 95% confidence interval $[-0.99, -0.95]$, and similarly $\rho = +0.95$ $[+0.89, +0.97]$ for agent return and WM-policy return on unseen goals. These strong correlations suggest that performant goal-conditioned agents tend to have better world models and better implicit generalisation capabilities. We expand this analysis in Appendix E.2.

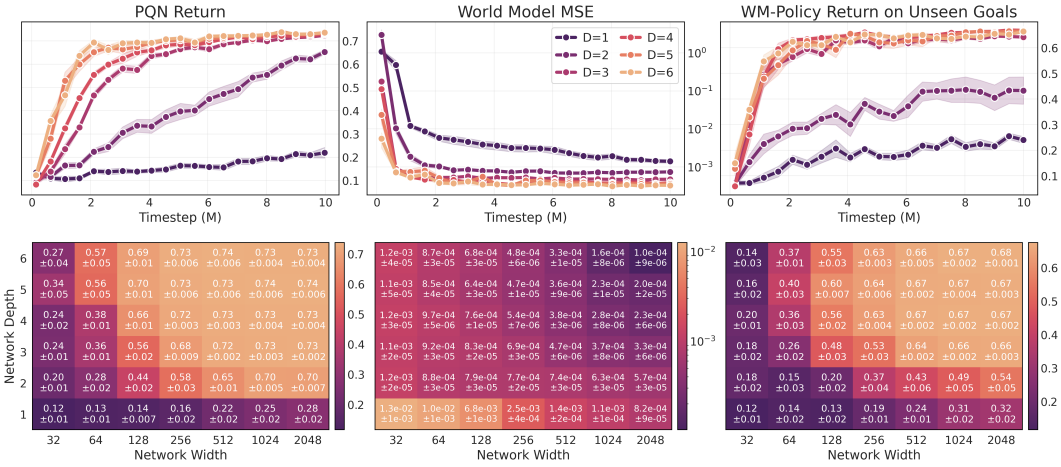


Figure 4: Agent return on training goals (left), implicit world model MSE (middle) and return of WM-extracted policies on unseen goals (right), for a `Reacher` agent trained with just $|\mathcal{G}| = 4$ goals. Top: metrics as a function of training time for fixed width $W = 512$ and increasing depth D. Bottom: Heatmap over all widths/depths at the end of training, with mean ± SE over 10 seeds. The Spearman correlation coefficient between agent return and WM error is $\rho = -0.98\,[-0.99, -0.95]$.

Why is recovery possible with a few sparse goals? For deterministic MDPs, $P(s,a)$ is identified by (Q,π,r) when the Bellman map $m \coloneqq (m^\pi_{g_1},\dots,m^\pi_{g_L}) : \mathcal{S} \to \mathbb{R}^L$ is injective (Definition 6). With terminating sparse goals and *exact* Q-values, m^π_g is a step function of the hitting time of g, so m takes countably many values and is nowhere injective, motivating the infinite goal set assumption of Theorem 4. However, *learnt* value functions can interpolate smoothly between these step functions, and while smoothness alone does not ensure injectivity, a *generic* smooth map m is locally injective almost everywhere for $L \geq d$ (which may be sufficient for recovery under local dynamics), and globally injective for $L \geq 2d+1$ [39, 40]. A rigorous counterpart to this informal argument is proven in Theorem 9 for Gaussian goals, which we can view as a smooth version of sparse goals.

5.2 MountainCar

In `MountainCar`, the environment goal is to reach position $x = 0.5$ under 200 timesteps. Unlike `Reacher`, which has goal-conditioning baked in, we augment the environment to support arbitrary goals. We train a `MountainCar` agent using 4 linearly spaced, position-based goals, with a sparse reward function $r_g(x,v) = \mathbf{1}\{|x-g| < \sigma\}$, where $g \in [-1.2, 0.6]$, with termination upon arrival and threshold $\sigma = 0.1$. The original environment goal corresponds exactly to $g = 0.6$. Again, the dynamics extracted from a trained PQN agent match the true transition kernel with high fidelity ($\texttt{NMSE} = 6.7 \times 10^{-3}$). To further validate accuracy, we consider three out-of-distribution goals: **(i) Fast car:** reach velocity $v = 0.06$; **(ii) Gentle car:** reach the top of the hill $x = 0.6$ without exceeding $|v| = 0.05$ (the unconstrained optimal policy uses $|v| \geq 0.06$); **(iii) Shortest path:** find the shortest path to one of two goals $x = -1.0$ and $x = 0.4$, which, as noted by Touati and Ollivier [11], could not be achieved using successor features based on the indicator-goal family.

Again, planning inside the implicit WM reveals implicit generalisation capabilities (Figure 3). Investigating further, we train a second `MountainCar` agent with four *velocity*-based goals. The resulting implicit WM is near-identical to that of the position-trained agent, namely, they are ~42 times closer to each other ($\texttt{NMSE} = 1.7 \times 10^{-4}$) than to ground truth ($\texttt{NMSE} = 7.2 \times 10^{-3}$), as illustrated in Figure 5. This leads us to an informal "local-global" hypothesis, left for future exploration: *value-based agents, trained to predict returns on a small number of local goals (i.e. with reward functions defined on a subset of variables), tend to encode an accurate world model over all dependent environment variables.*

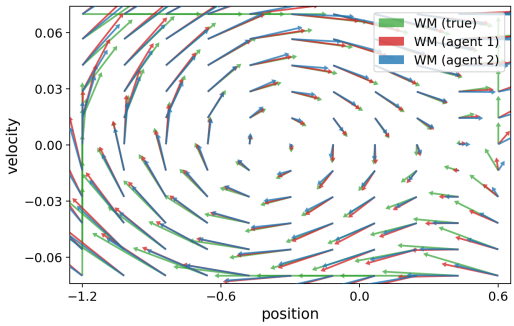


Figure 5: Quiver plots of WMs extracted from two agents trained with position- and velocity-based goals, vs ground truth, for action $a =$ right. Each arrow starts from a state s and points to the predicted or true next-state $s' = P(s,a)$.

6 Related Work

World models from agents. Richens et al. [6] proved that general agents encode approximate dynamics, but their results require (i) goals specified in linear temporal logic, which is rarely the case in RL, (ii) competence on all multi-step goals of depth $N > 1$, a set of size $O(|\mathcal{S}|^N)$, whereas real agents are trained on no more than $|\mathcal{S}|$ single-step goals, (iii) finite state-action spaces, and (iv) deterministic policies. Follow-up work [41, 42, 43] relaxes some assumptions but still use linear temporal logic or causal influence diagrams, while our work deals with standard RL agents in finite and continuous environments. Adamczyk [44] showed that converged Q-values encode dynamics under a "reverse Lipschitz" assumption on V^π, limited to deterministic MDPs and a single goal. Moreover, no function $V : \mathcal{S} \to \mathbb{R}$ can be reverse Lipschitz when $\dim \mathcal{S} \geq 2$ (it would be injective with a Lipschitz left-inverse, so $\mathcal{S}$ would be the Lipschitz image of a subset of $\mathbb{R}$ and have Hausdorff dimension at most one), so their results only apply to one-dimensional state spaces. Reddy et al. [45] infer a demonstrator's internal dynamics from behaviour alone via latent soft Q-values, but address identifiability only heuristically. Turner [46, Theorem F.117] proved that $V^\star$ recovers deterministic MDPs up to isomorphism, but not in stochastic environments. Bush et al. [47] provide mechanistic evidence that model-free agents develop implicit forward models, but do not address uniqueness.

Value equivalence. Grimm et al. [7] introduced the value equivalence (VE) principle, and showed that the equivalence class collapses to a single point for value functions that form a basis of $\mathbb{R}^{|\mathcal{S}|}$ combined with $|\mathcal{A}|$ pointwise-linearly-independent policies, neither of which we assume. Their work focuses on learning a surrogate model to enhance planning using environment samples, while ours addresses the problem of extracting an agent's internal world model with access only to (Q, π, r). Algorithms such as MuZero [48], VAML [49, 4] and unified frameworks [50] also learn value-equivalent models without addressing identifiability. Fellows et al. [51] demonstrate that Bayesian model-free methods can learn Bayes-optimal policies, implying that model-free approaches implicitly encode a posterior over environment dynamics, and hence a space of world models. Relatedly, delayed Q-learning [52] achieves PAC sample complexity comparable to model-based RL despite never representing P explicitly. Conversely, Zhu et al. [53] exhibit MDP families where P admits exponentially smaller circuit representations than $Q^\star$, which motivates extracting P as a compact summary of an agent.

Decoupling dynamics from reward. The successor representation [54] and successor features [9, 10] decompose value functions into transition- and reward-dependent components, but only address families of reward functions that can be decomposed as linear functions of shared features. This excludes e.g. the indicator functions in continuous environments [33], which would require infinitely many features to represent. Forward-backward learning [11] relaxes this constraint significantly, and the successor measure of Lehnert and Littman [55, p. 7] and Blier et al. [30, Prop. 1] was shown to be invertible in the special case of unconditional π, indicator goals and exact Q, but does not address identifiability or extraction. Ollivier [56] derives the optimal base features for successor features, which depend on the dynamics and a reference policy alone, but does not address recovering the dynamics. Reward-free RL [57] and inverse RL [58] solve orthogonal problems to ours: *reward-free* exploration produces a dataset sufficient to plan for any downstream reward, whereas we show the dual phenomenon: diverse *reward-rich* training implicitly encodes P. Inverse RL recovers r from (P, π), while we recover P from (r, π, Q). Indeed, RL, Q-learning, and P-learning can all be thought of as different resolutions to the same inconsistency (the Bellman residual) between a world model P, action-values Q, and rewards r – an instance of a general pattern that also has analogues in decision theory and supervised learning [59].

GCRL. GCRL has emerged as a popular field wherein agents are trained on a number of tasks [60, 36, 30], with the bulk of recent research focused on enhancing the performance of goal-conditioned agents, e.g. by scaling compute usage [61], goal set [62] or network size [38]. Our work takes a different tack by analysing what agents encode during and after training, and showing that their implicit representation of the world improves with the number of goals, network size and performance.

7 Conclusion

In this work, we provided sufficient conditions under which it is possible to recover environment dynamics from the Q-values of a goal-conditioned model-free agent, for a variety of MDP classes (deterministic or stochastic, discrete or continuous). We introduced a novel method, P-learning, which enables the extraction of this implicit world model. We subsequently showed, empirically, the accuracy and effectiveness of P-learning in a number of RL environments, despite goal sets that significantly breach our theoretical guarantees. We demonstrated that this WM leads to implicit generalisation capabilities hidden in the agent's value functions, with zero-shot transfer to reward functions defined on orthogonal variables. Finally, we provided evidence for a relationship between the performance of an agent, and the fidelity of its underlying world model.

Limitations. Our results come with a number of caveats. First, our theoretical results for stochastic MDPs with continuous state spaces only apply to unconditional policies, and while we expect them to generalise, we leave them as open conjectures for future work. Secondly, our work does not encompass partially observable MDPs, or model-free agents that do not learn Q-values. Third, we have not run experiments to test whether P-learning scales to very large, finite environments. Finally, P-learning is not a substitute for model-based RL, but a tool for reading out what an agent believes the dynamics to be, which can interpret its behaviour and diagnose generalisation failures.

Future Work. We hope our results will inspire a deeper understanding of the implicit generalisation capabilities lurking in model-free agents; hybrid methods that leverage them; and interpretability or auditing tools based on internal world models, to ensure that agents can efficiently be repurposed when they are trained using reward functions that are, perhaps inevitably, misspecified.

Acknowledgements

We are grateful to David Abel for his timely suggestions and references to existing work, Michael Matthews for his JAX implementation of PQN augmented with HER, Maria Kostylew for her comments and support, Uljad Berdica for his feedback and world model expertise, Shimon Whiteson for feedback on early drafts, and Matthew Farrugia-Roberts for helpful early conversations.

AL conducted this work as part of the MATS Program (ML Alignment & Theory Scholars), and is generously supported by the Cooperative AI PhD Fellowship. **MF** is funded by the UKRI Engineering and Physical Sciences Research Council EP/Y028481/1. **AG** is funded by the EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines and Systems EP/S024050/1.

Author Contributions

AL led the project, designed the method, proved the theoretical results, ran the experiments and wrote the paper. **MF** and **AG** helped with writing, editing and experiment design. **JR** and **JF** provided guidance and helped frame the paper throughout. **OR** contributed to the writing, editing, mathematical precision, and provided guidance throughout.

References

[1] Richard Sutton and Andrew Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[2] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing Atari with deep reinforcement learning. *arxiv:1312.5602*, 2013.

[3] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arxiv:1707.06347*, 2017.

[4] Nathan Lambert, Brandon Amos, Omry Yadan, and Roberto Calandra. Objective mismatch in model-based reinforcement learning. In *Conference on Learning for Dynamics and Control*, 2020.

[5] Thomas M. Moerland, Joost Broekens, Aske Plaat, and Catholijn M. Jonker. Model-based reinforcement learning: A survey. *Foundations and Trends in Machine Learning*, 2023.

[6] Jonathan Richens, Tom Everitt, and David Abel. General agents contain world models. In *International Conference on Machine Learning*, 2025.

[7] Christopher Grimm, André Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

[8] Christopher Grimm, André Barreto, Gregory Farquhar, David Silver, and Satinder Singh. Proper value equivalence. In *Advances in Neural Information Processing Systems*, 2021.

[9] André Barreto, Will Dabney, Rémi Munos, Jonathan J. Hunt, Tom Schaul, Hado van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. In *Advances in Neural Information Processing Systems*, 2017.

[10] Diana Borsa, André Barreto, John Quan, Daniel Mankowitz, Rémi Munos, Hado van Hasselt, David Silver, and Tom Schaul. Universal successor features approximators. In *International Conference on Learning Representations*, 2019.

[11] Ahmed Touati and Yann Ollivier. Learning one representation to optimize all rewards. In *Advances in Neural Information Processing Systems*, 2021.

[12] Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *International Conference on Machine Learning*, 2015.

[13] Benjamin Eysenbach, Tianjun Zhang, Sergey Levine, and Ruslan Salakhutdinov. Contrastive learning as goal-conditioned reinforcement learning. In *Advances in Neural Information Processing Systems*, 2022.

[14] Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. OGBench: Benchmarking offline goal-conditioned RL. In *International Conference on Learning Representations*, 2025.

[15] Martha White. Unifying task specification in reinforcement learning. In *International Conference on Machine Learning*, 2017.

[16] Martin Riedmiller. Neural fitted Q iteration – first experiences with a data efficient neural reinforcement learning method. In *European Conference on Machine Learning*, 2005.

[17] Richard Sutton. Learning to predict by the method of temporal differences. *Machine Learning*, 1988.

[18] Christopher John Cornish Hellaby Watkins. *Learning from Delayed Rewards*. PhD thesis, King's College, University of Cambridge, 1989.

[19] Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 1992.

[20] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 2015.

[21] Matteo Gallici, Mattie Fellows, Benjamin Ellis, Bartomeu Pou, Ivan Masmitja, Jakob Nicolaus Foerster, and Mario Martin. Simplifying deep temporal difference learning. In *International Conference on Learning Representations*, 2025.

[22] Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

[23] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

[24] Leemon C. Baird. Residual algorithms: Reinforcement learning with function approximation. In *International Conference on Machine Learning*, 1995.

[25] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 1951.

[26] I. Rechenberg. Evolutionsstrategien. In *Simulationsmethoden in der Medizin und Biologie*. Springer Berlin Heidelberg, 1978.

[27] Bidipta Sarkar, Mattie Fellows, Juan Agustin Duque, Alistair Letcher, Antonio León Villares, Anya Sims, Clarisse Wibault, Dmitry Samsonov, Dylan Cope, Jarek Liesen, Kang Li, Lukas Seier, Theo Wolf, Uljad Berdica, Valentin Mohl, Alexander David Goldie, Aaron Courville, Karin Sevegnani, Shimon Whiteson, and Jakob Nicolaus Foerster. Evolution strategies at the hyperscale. In *International Conference on Machine Learning*, 2026.

[28] Michał Bortkiewicz, Władek Pałucki, Vivek Myers, Tadeusz Dziarmaga, Tomasz Arczewski, Łukasz Kuciński, and Benjamin Eysenbach. Accelerating goal-conditioned RL algorithms and research. In *International Conference on Learning Representations*, 2025.

[29] Grace Liu, Michael Tang, and Benjamin Eysenbach. A single goal is all you need: Skills and exploration emerge from contrastive RL without rewards, demonstrations, or subgoals. In *International Conference on Learning Representations*, 2025.

[30] Léonard Blier, Corentin Tallec, and Yann Ollivier. Learning successor states and goal-dependent values: A mathematical viewpoint. *arxiv:2101.07123*, 2021.

[31] William Fedus, Carles Gelada, Yoshua Bengio, Marc G. Bellemare, and Hugo Larochelle. Hyperbolic discounting and learning over multiple horizons. *arxiv:1902.06865*, 2019.

[32] Richard S. Sutton, Joseph Modayil, Michael Delp, Thomas Degris, Patrick M. Pilarski, Adam White, and Doina Precup. Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *International Conference on Autonomous Agents and Multiagent Systems*, 2011.

[33] Mark Towers, Ariel Kwiatkowski, and Jordan Terry. Gymnasium: A standard interface for reinforcement learning environments. https://gymnasium.farama.org/, 2023.

[34] Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012.

[35] Richard S. Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 1999.

[36] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, 2017.

[37] Kevin Zakka, Baruch Tabanpour, Qiayuan Liao, Mustafa Haiderbhai, Samuel Holt, Jing Yuan Luo, Arthur Allshire, Erik Frey, Koushil Sreenath, Lueder A. Kahrs, Carlo Sferrazza, Yuval Tassa, and Pieter Abbeel. MuJoCo Playground: An open-source framework for GPU-accelerated robot learning and sim-to-real transfer. *arxiv:2502.08844*, 2025.

[38] Kevin Wang, Ishaan Javali, Michał Bortkiewicz, Tomasz Trzcinski, and Benjamin Eysenbach. 1000 layer networks for self-supervised RL: Scaling depth can enable new goal-reaching capabilities. In *Advances in Neural Information Processing Systems*, 2025.

[39] Hassler Whitney. Differentiable manifolds. *Annals of Mathematics*, 1936.

[40] John M. Lee. *Introduction to Smooth Manifolds*. Springer, 2nd edition, 2013.

[41] Matteo Ceriscioli and Karthika Mohan. Agents robust to distribution shifts learn causal world models even under mediation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[42] Yikai Lu, Yifei Wu, Xinyu Lu, and Tongxin Li. World models in pieces: Structural certification for general agents. In *Forty-third International Conference on Machine Learning*, 2026.

[43] Santiago Cifuentes. General agents contain world models, even under partial observability and stochasticity. *arxiv:2602.03146*, 2026.

[44] Jacob Adamczyk. Inferring transition dynamics from value functions. *The Eighth Workshop on "Generalization in Planning" at the Association for the Advancement of Artificial Intelligence*, 2025.

[45] Siddharth Reddy, Anca D. Dragan, and Sergey Levine. Where do you think you're going?: Inferring beliefs about dynamics from behavior. In *Advances in Neural Information Processing Systems*, 2018.

[46] Alexander Matt Turner. *On Avoiding Power-Seeking by Artificial Intelligence*. PhD thesis, Oregon State University, 2022.

[47] Thomas Bush, Stephen Chung, Usman Anwar, Adrià Garriga-Alonso, and David Krueger. Interpreting emergent planning in model-free reinforcement learning. In *International Conference on Learning Representations*, 2025.

[48] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 2020.

[49] Amir-massoud Farahmand. Iterative value-aware model learning. In *Advances in Neural Information Processing Systems*, 2018.

[50] Benjamin Freed, Thomas Wei, Roberto Calandra, Jeff Schneider, and Howie Choset. Unifying model-based and model-free reinforcement learning with equivalent policy sets. In *Reinforcement Learning Conference*, 2024.

[51] Mattie Fellows, Brandon Gary Kaplowitz, Christian Schroeder De Witt, and Shimon Whiteson. Bayesian exploration networks. In *International Conference on Machine Learning*, 2024.

[52] Alexander L. Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L. Littman. PAC model-free reinforcement learning. In *International Conference on Machine Learning*, 2006.

[53] Hanlin Zhu, Baihe Huang, and Stuart Russell. On representation complexity of model-based and model-free reinforcement learning. In *International Conference on Learning Representations*, 2024.

[54] Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 1993.

[55] Lucas Lehnert and Michael L. Littman. Successor features combine elements of model-free and model-based reinforcement learning. *Journal of Machine Learning Research*, 2020.

[56] Yann Ollivier. Which features are best for successor features? *arxiv:2502.10790*, 2025.

[57] Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, 2020.

[58] Haoyang Cao and Samuel N. Cohen. Identifiability in inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, 2021.

[59] Oliver E Richardson, Mandana Samiei, Mehran Shakerinava, Joseph D Viviano, Abdessamad El Kabid, Ali Parviz, and Yoshua Bengio. Local inconsistency resolution: The interplay between attention and control in probabilistic models. In *International Conference on Artificial Intelligence and Statistics*, 2026.

[60] Leslie Pack Kaelbling. Learning to achieve goals. In *International Joint Conference on Artificial Intelligence*, 1993.

[61] Raj Ghugare, Michał Bortkiewicz, Alicja Ziarko, and Benjamin Eysenbach. On the role of iterative computation in reinforcement learning. *arxiv:2602.05999*, 2026.

[62] Michael Matthews, Matthew Jackson, Michael Beukman, Thomas Foster, Alistair Letcher, Scott Fujimoto, Cédric Colas, and Jakob Foerster. Goal-conditioned agents that learn everything all at once. In *International Conference on Machine Learning*, 2026.

[63] Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte Carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 2020.

[64] A. J. Wilkie. Model completeness results for expansions of the ordered field of real numbers by restricted Pfaffian functions and the exponential function. *Journal of the American Mathematical Society*, 1996.

[65] Steven G. Krantz and Harold R. Parks. *A Primer of Real Analytic Functions*. Birkhäuser Boston, 2nd edition, 2002.

[66] Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Pure and Applied Mathematics: A Wiley-Interscience Series of Texts, Monographs and Tracts. John Wiley & Sons, 1999.

[67] Salomon Bochner. *Harmonic Analysis and the Theory of Probability*. University of California Press, 1955.

[68] Elias M. Stein and Guido Weiss. *Introduction to Fourier Analysis on Euclidean Spaces*. Princeton Mathematical Series. Princeton University Press, 1971.

[69] Robert Tjarko Lange. gymnax: A JAX-based reinforcement learning environment library, 2022. URL http://github.com/RobertTLange/gymnax.

[70] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: Composable transformations of Python+NumPy programs, 2018. URL http://github.com/jax-ml/jax.

A Proofs for P-Learning

A.1 Derivation of P-Learning

We provide pseudocode for P-learning in Algorithm 1 below, with batch size 1 for simplicity. To derive the update, recall the definition of the Bellman residual

$$\mathcal{L}(\varphi) \coloneqq \|\mathcal{T}_\varphi^\pi(Q) - Q\|_d^2 = \frac{1}{2}\,\mathbb{E}_{(g,s,a)\sim d}\left[\left(\mathcal{T}_\varphi^\pi(Q)(s,a,g) - Q(s,a,g)\right)^2\right] = \frac{1}{2}\,\mathbb{E}_d[\delta_\varphi^2]\,,$$

where $\delta_\varphi(s,a,g) = \mathcal{T}_\varphi^\pi(Q)(s,a,g) - Q(s,a,g)$ is the TD error. By the chain rule, and since the expectation does not depend on φ, we have

$$\nabla_\varphi \mathcal{L}(\varphi) = \mathbb{E}_{(g,s,a)\sim d}\left[\delta_\varphi(s,a,g)\nabla_\varphi \mathcal{T}_\varphi^\pi(Q)(s,a,g)\right]\,.$$

We now assume three standard regularity conditions for the score-function gradient estimator [63], namely: (i) P_φ is continuously differentiable in φ, (ii) P_φ has full support, i.e. $P_\varphi(s' \mid s,a) > 0$ for all (s,a,s'), and (iii) there is an integrable function $g : \mathcal{S} \to \mathbb{R}$ with $\sup_\varphi \|\nabla_\varphi P_\varphi(s' \mid s,a)\|_1 \le g(s')$ for all s, a, s'. Then the gradient can be estimated via the log-derivative trick [63]:

$$\begin{aligned}\nabla_\varphi \mathcal{L}(\varphi) &= \mathbb{E}_{(g,s,a)\sim d}\left[\delta_\varphi(s,a,g)\nabla_\varphi\, \mathbb{E}_{s'\sim P_\varphi(s,a), a'\sim\pi(\cdot\mid s',g)}\left[r(s',g) + \gamma Q(s',a',g)\right]\right] \\ &= \mathbb{E}_{(g,s,a)\sim d, s'\sim P_\varphi(s,a), a'\sim\pi(s',g)}\left[\delta_\varphi(s,a,g)(r(s',g) + \gamma Q(s',a',g))\nabla_\varphi \log P_\varphi(s' \mid s,a)\right]\,.\end{aligned}$$

To reduce variance and make the per-sample direction of this update explicit, we subtract the baseline $Q(s,a,g)$ to get a per-sample TD error

$$\hat{\delta}_\varphi(s,a,g,s',a') \coloneqq r(s',g) + \gamma Q(s',a',g) - Q(s,a,g)\,,$$

and the update, as claimed in the main text, simplifies to

$$\nabla_\varphi \mathcal{L}(\varphi) = \mathbb{E}_{(g,s,a)\sim d, s'\sim P_\varphi(s,a), a'\sim\pi(s',g)}\left[\delta_\varphi(s,a,g)\hat{\delta}_\varphi(s,a,g,s',a')\nabla_\varphi \log P_\varphi(s' \mid s,a)\right]\,.$$

Note that the baseline preserves unbiasedness: $Q(s,a,g)$ is constant under the inner sampling of (s',a'), so it factors out and multiplies the expected score $\mathbb{E}_{s'\sim P_\varphi(s,a)}[\nabla_\varphi \log P_\varphi(s' \mid s,a)]$, which is zero because $\int_{\mathcal{S}} P_\varphi(s' \mid s,a)ds' = 1$. Now the TD error $\delta_\varphi(s,a,g)$ does not depend on the variables (s',a') in the outer expectation, so we can factor it out, yielding:

$$\nabla_\varphi \mathcal{L}(\varphi) = \mathbb{E}_{(g,s,a)\sim d}\left[\delta_\varphi(s,a,g)\,\mathbb{E}_{s'\sim P_\varphi(s,a),\, a'\sim\pi(s',g)}\left[\hat{\delta}_\varphi \nabla_\varphi \log P_\varphi(s' \mid s,a)\right]\right]\,.$$

The gradient is a product of two expectations over (s',a'), so an unbiased estimator requires two independent samples: (s'_1,a'_1) to form an unbiased estimate $\hat{\delta}^1$ of the outer factor δ_φ, and an independent (s'_2,a'_2) to estimate the inner expectation. Independence ensures unbiased estimation:

$$\begin{aligned}\nabla_\varphi \mathcal{L}(\varphi) &= \mathbb{E}_{(g,s,a)\sim d}\left[\delta_\varphi(s,a,g)\,\mathbb{E}_{s'\sim P_\varphi(s,a),a'\sim\pi(s',g)}\left[\hat{\delta}_\varphi \nabla_\varphi \log P_\varphi(s' \mid s,a)\right]\right] \\ &= \mathbb{E}_{(g,s,a)\sim d, s_{1,2}\sim P_\varphi(s,a), a_{1,2}\sim\pi(s_{1,2},g)}\left[\hat{\delta}^1\hat{\delta}^2\nabla_\varphi \log P_\varphi(s_2 \mid s,a)\right]\,.\end{aligned}$$

This derivation leads to the pseudocode for P-learning in Algorithm 1 below, with batch size 1 for simplicity. Crucially, both samples come from the model P_φ rather than the environment, so double sampling here is essentially free, costing only an extra forward pass through P_φ and Q, unlike the double-sampling obstruction in TD learning which requires double-sampling in the environment [24].

Deterministic case. When dynamics are known to be deterministic and the model is parameterised as a successor function $P_\varphi : \mathcal{S}\times\mathcal{A}\to\mathcal{S}$, the inner expectation over s' in $\mathcal{T}_\varphi^\pi(Q)$ collapses to

$$\mathcal{T}_\varphi^\pi(Q)(s,a,g) = r(P_\varphi(s,a),g) + \gamma V_g(P_\varphi(s,a))\,, \qquad V_g(s') = \mathbb{E}_{a'\sim\pi(s',g)}[Q(s',a',g)]\,.$$

Hence $\delta_\varphi(s,a,g)$ is computable from a single forward pass, with no Monte-Carlo estimate over s'; the only remaining randomness is over $a' \sim \pi(P_\varphi(s,a),g)$. The gradient

$$\nabla_\varphi \mathcal{L}(\varphi) = \delta_\varphi(s,a,g)\nabla_\varphi\left[r(P_\varphi(s,a),g) + \gamma\, V_g(P_\varphi(s,a))\right]$$

obtained by chain rule (reparameterisation) through P_φ is then unbiased from a single sample. This requires r and Q to be differentiable almost everywhere in their first argument, as discussed in Section 3; otherwise, zero-shot methods like evolutionary strategies [26, 27] are appropriate.

Algorithm 1 P-Learning.

Require: Parametric family P_φ, Q-function Q, policy π, reward r, discount γ, sampling distribution d over (g, s, a), step sizes $\{\alpha_n\}$, initial φ_0.

for $n = 0, 1, \ldots,$ **do**
  Sample $(g, s, a) \sim d$
  Sample $s'_1, s'_2 \sim P_{\varphi_n}(\cdot \mid s, a)$, $a'_1 \sim \pi(\cdot \mid s'_1, g)$, $a'_2 \sim \pi(\cdot \mid s'_2, g)$
  $\delta^i \leftarrow r(s'_i, g) + \gamma\, Q(s'_i, a'_i, g) - Q(s, a, g)$ for $i = 1, 2$
  $\hat{g}_n \leftarrow \delta^1 \delta^2 \nabla_\varphi \log P_\varphi(s'_2 \mid s, a)\big|_{\varphi=\varphi_n}$
  $\varphi_{n+1} \leftarrow \varphi_n - \alpha_n \hat{g}_n$
end for
return Extracted world model P_{φ_n}.

A.2 Proof of Theorem 1

Theorem 1 (Convergence of P-learning). *In the tabular setting, fix a policy π, a Q-function $Q \in \mathcal{Q}$, initial $P_0(s, a) \in \Delta(\mathcal{S})$ for each (s, a), and let $\sigma_{\max}$ denote the largest singular value of M (Equation 1). Then P-learning (Equation 3) with step size $0 < \alpha < 2/\sigma^2_{\max}$ converges to*

$$P_\infty(s, a) = M^+ Q(s, a) + \big(I - M^+ M\big) P_0(s, a)\,,$$

where M^+ is the Moore–Penrose pseudo-inverse of M. Equivalently, P_∞ is the orthogonal projection of P_0 onto the set of global minima $\mathcal{Z} = \arg\min_P \|\mathcal{T}_P(Q) - Q\|_2^2$.

Proof. Fix s, a and write $M \coloneqq M^\pi$ and $Q \coloneqq Q(s, a)$, $P_t \coloneqq P_t(s, a)$ for brevity. The iteration is

$$P_{t+1} = P_t - \alpha M^\top (M P_t - Q)\,.$$

Now let $r = \operatorname{rank}(M)$ and write the SVD $M = U \Sigma V^\top$ where $U \in \mathbb{R}^{L \times r}$, $V \in \mathbb{R}^{n \times r}$ have orthonormal columns, and $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_r) \succ 0$. We extend V to an orthonormal basis $[V\ V_\perp]$ of $\mathbb{R}^n$ with $V_\perp \in \mathbb{R}^{n \times (n-r)}$, so that $V_\perp^\top V_\perp = I_{n-r}$ and $V^\top V_\perp = 0$. In particular, $[V; V_\perp]$ is an orthogonal $n \times n$ matrix with $V V^\top + V_\perp V_\perp^\top = I_n$, and we can decompose each iterate as

$$P_t = (V V^\top + V_\perp V_\perp^\top) P_t = V a_t + V_\perp b_t\,,$$

where $a_t \coloneqq V^\top P_t \in \mathbb{R}^r$ and $b_t \coloneqq V_\perp^\top P_t \in \mathbb{R}^{n-r}$. Using orthogonality $(V_\perp^\top V) = 0$,

$$b_{t+1} = V_\perp^\top P_{t+1} = b_t - \alpha V_\perp^\top M^\top (M P_t - Q) = b_t - \alpha (V_\perp^\top V) \Sigma U^\top (M P_t - Q) = b_t\,,$$

so $b_t = b_0 = V_\perp^\top P_0$ for all t. We now prove by induction that

$$a_t = (I - \alpha \Sigma^2)^t a_0 + \alpha \sum_{k=0}^{t-1} (I - \alpha \Sigma^2)^k \Sigma U^T Q$$

for all $t \geq 0$. The base case $t = 0$ holds trivially. Assuming the claim holds for fixed $t \geq 0$,

$$\begin{aligned} a_{t+1} &= a_t - \alpha \Sigma U^\top (U \Sigma V^\top P_t - Q) = (I - \alpha \Sigma^2) a_t + \alpha \Sigma U^\top Q \\ &= (I - \alpha \Sigma^2)\big[(I - \alpha \Sigma^2)^t a_0 + \alpha \sum_{k=0}^{t-1} (I - \alpha \Sigma^2)^k \Sigma U^T Q\big] + \alpha \Sigma U^\top Q \\ &= (I - \alpha \Sigma^2)^{t+1} a_0 + \alpha \sum_{k=0}^{t} (I - \alpha \Sigma^2)^k \Sigma U^T Q\,, \end{aligned}$$

so the induction is complete. In particular, since $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_r) \succ 0$, we have $\big|1 - \alpha \sigma_i^2\big| < 1$ for all $0 < \alpha < 2/\sigma^2_{\max}$ and hence $(I - \alpha \Sigma^2)^t \to 0$ as $t \to \infty$, thus

$$\begin{aligned} \lim_{t \to \infty} a_t &= \lim_{t \to \infty} \alpha \sum_{k=0}^{t-1} (I - \alpha \Sigma^2)^k \Sigma U^T Q \\ &= \alpha \big[I - (I - \alpha \Sigma^2)\big]^{-1} \Sigma U^T Q = \Sigma^{-1} U^\top Q\,. \end{aligned}$$

Combining the expressions for a_t and b_t, we conclude that

$$P_\infty \coloneqq \lim_{k\to\infty} P_t = V\Sigma^{-1}U^\top Q + V_\perp V_\perp^\top P_0$$

for all $0 < \alpha < 2/\sigma_{\max}^2$. Now $M^+ = V\Sigma^{-1}U^\top$ is the Moore–Penrose pseudo-inverse of M, and

$$M^+M = V\Sigma^{-1}U^\top U\Sigma V^\top = VV^\top \quad \Longrightarrow \quad I - M^+M = V_\perp V_\perp^\top ,$$

so we obtain $P_\infty = M^+Q + (I - M^+M)P_0$ as claimed.

Equivalent characterisation. The set of global minima $\mathcal{Z} = \arg\min_P \|MP - Q\|^2$ can be defined via the normal equations $M^\top MP = M^\top Q$, and we claim that $\mathcal{Z} = M^+Q + \ker(M)$. Indeed, $M^\top M(M^+Q) = V\Sigma^2V^\top \cdot V\Sigma^{-1}U^\top Q = V\Sigma U^\top Q = M^\top Q$, so $M^+Q \in \mathcal{Z}$. Moreover for any P, $M^\top M(P - M^+Q) = 0$ iff $P - M^+Q \in \ker(M^\top M) = \ker(M)$, so $\mathcal{Z} = M^+Q + \ker(M)$. Now orthogonal projection onto an affine subspace $a + V_0$ satisfies $\mathrm{proj}_{a+V_0}(x) = a + \mathrm{proj}_{V_0}(x-a)$. Applying this equality to $a = M^+Q$ and $V_0 = \ker(M)$, we obtain $\mathrm{proj}_{\ker M}(M^+Q) = 0$ and thus

$$\mathrm{proj}_{\mathcal{Z}}(P_0) = M^+Q + \mathrm{proj}_{\ker M}(P_0) = M^+Q + (I - M^+M)P_0 = P_\infty ,$$

using $I - M^+M = V_\perp V_\perp^\top = \mathrm{proj}_{\ker M}$ from above. □

B Value Equivalence and Pseudo-World Models

B.1 Intuition: Breaking Value Equivalence

We provide intuition for how exactly spanning goals break value equivalence using the notion of "test functions" to determine unknown distributions. Recalling the Bellman operator

$$\mathcal{T}_P(Q)(s,a,g) = \mathbb{E}_{s'\sim P(s,a)}\left[M_g(s')\right] , \quad \text{where} \quad M_g(s') \coloneqq r_g(s') + \gamma V_g(s') ,$$

notice that the Bellman equation reduces to $Q(s,a,g) = \mathbb{E}_{s'\sim P(s,a)}\left[M_g(s')\right]$, which can be viewed as imposing $|\mathcal{G}|$ constraints on the expectations of *known* "test functions" or "Bellman probes" M_g under the *unknown* distribution $P(s,a)$. If the agent is trained on a single goal g, it is natural to expect that there will be many world models $\hat{P}$ that have the same expectation under the single test function, i.e. such that $\mathbb{E}_{s'\sim\hat{P}(s,a)}\left[M_g(s')\right] = Q(s,a,g)$, inducing value equivalence.

However, probability distributions can be determined by expectations over multiple test functions: beginning with the finite setting, a distribution $p \in \Delta^n$ is uniquely determined by the first $n-1$ moments of any injective random variable ('test function') $X : [n] \to \mathbb{R}$. For example, in the 1-dimensional case $p = (p_1, 1-p_1)$, if we know the first moment $\mathbb{E}[X] = p_1X(1) + (1-p_1)X(2)$, we can invert the system to determine $p_1 = (\mathbb{E}[X] - X(2))/(X(1) - X(2))$ uniquely. For general $p \in \Delta^n$, the first n moments induce a linear system of equations known as a *Vandermonde system*, which is invertible if and only if X is injective – meaning that p can be recovered uniquely.

However, the Bellman probes are not simple functions like monomials, and it is not clear that they will similarly induce an invertible system. For continuous MDPs Section 4.2, the picture is even more involved, and invertibility generalises to a set of test functions being *measure determining*, meaning that distinct Borel measures have distinct expectations.

B.2 World Models and Pseudo-World Models

We introduce two spaces of **stochastic** world models:

$$\mathcal{P} \coloneqq \{P : \mathcal{S}\times\mathcal{A} \to \Sigma(\mathcal{S})\} \quad \text{and} \quad \mathcal{P}_\Delta \coloneqq \{P : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})\} \subset \mathcal{P} ,$$

where $\Sigma(\mathcal{S})$ denotes the space of finite *signed* measures over $\mathcal{S}$. Elements of $\mathcal{P}_\Delta$ are valid *world models*, while elements of $\mathcal{P}\setminus\mathcal{P}_\Delta$ are *pseudo-world models*: signed measures that may still satisfy the Bellman equation $\mathcal{T}_P(Q) = Q$ but do not correspond to valid probabilistic dynamics. In particular, the P-learning iteration $P_{t+1}(s,a) = P_t(s,a) - \alpha M^\top(MP_t(s,a) - Q(s,a))$ does not preserve positivity or normalisation, so the limit $P_\infty = M^+Q + (I - M^+M)P_0$ generally lies in $\mathcal{P}$.

However, if the conditions of Theorem 3(a) are satisfied, then there exists a true kernel $P \in \mathcal{P}_\Delta$ with $Q = MP$, and M has full rank, so $P_\infty = M^+MP = P \in \mathcal{P}_\Delta$, so the iteration converges to a true

world model. Moreover, in realistic settings with function approximation, the world model is typically parameterised as a neural network $P_\varphi : \mathcal{S} \times \mathcal{A} \to \mathcal{P}$ with a softmax output layer, so P-learning is constrained to the simplex by construction. Finally, in **deterministic** MDPs, a world model is defined as a function $P : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$, so the distinction collapses, and either column-matching or approximate methods always return a valid world model. If none of these conditions hold, the pseudo-world caveat can be mitigated in full generality by a projected variant of P-learning, given by

$$P_{t+1}(s,a) = \Pi_\Delta\big(P_t(s,a) - \alpha M^\top (M P_t(s,a) - Q(s,a))\big)\,,$$

where Π_Δ is the Euclidean projection onto the simplex. Since projection onto a closed convex set is non-expansive, the iteration still converges, and the limit is $P^\Delta_\infty = \arg\min_{P \in \mathcal{P}_\Delta} \|MP - Q\|^2$. This is not generally equal to post-hoc projection $\Pi_\Delta(P_\infty)$ of the unconstrained limit, which ignores the geometry of the loss landscape and can return a strictly worse fit than P^Δ_∞ when M is poorly conditioned. In the approximate setting of Theorem 3(b), where M is full-rank, both post-hoc projection and projected P-learning inherit the same asymptotic upper bound.

B.3 Relationship between P-learning and Q-learning

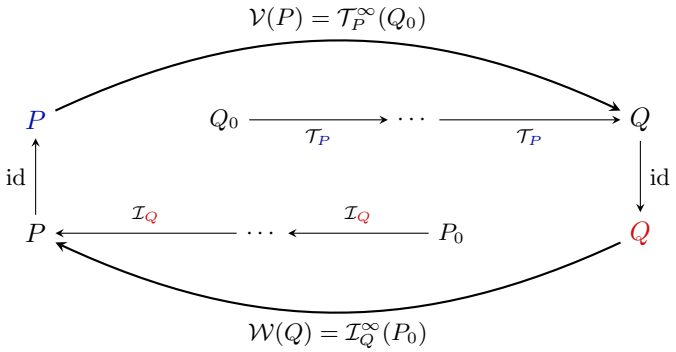


Figure 6: Relationship between Q-learning (represented by $\mathcal{V}$) and P-learning (represented by $\mathcal{W}$). **Top:** Given a transition kernel P, the Bellman operator $\mathcal{T} = \mathcal{T}^\pi$ iteratively maps any initial Q_0 to $Q = Q^\pi$. **Bottom:** Given Q-values, the inverse Bellman operator $\mathcal{I}$ iteratively maps any initial P_0 to some world model P, but the map $\mathcal{W}$ is only independent from P_0 when the conditions of Theorem 3 are satisfied, in which case the diagram commutes (i.e. $\mathcal{V}$ is injective, with inverse $\mathcal{W}$).

In this section, we sketch a sense in which P-learning and Q-learning are in a warm and bijective relationship with each other, illustrated in Figure 6. We work in the tabular setting of Section 3, with finite $\mathcal{S}, \mathcal{A}, \mathcal{G}$ and a fixed policy π, and begin by defining the two maps depicted in the figure.

Recall that $\mathcal{Q} = \{Q : \mathcal{S} \times \mathcal{A} \times \mathcal{G} \to \mathbb{R}\}$ is the space of Q-functions (Section 2), and that $\mathcal{P}_\Delta \subset \mathcal{P}$ are the spaces of world models and signed kernels introduced in Appendix B.2.

The value map $\mathcal{V}$: for any transition kernel $P \in \mathcal{P}_\Delta$, the Bellman operator $\mathcal{T}^\pi_P$ is a γ-contraction on $\mathcal{Q}$, so it admits a unique fixed point Q^π_P, independently from the initial guess $Q_0 \in \mathcal{Q}$. We may therefore define the *value map*, shown as the top edge of Figure 6, by

$$\mathcal{V} : \mathcal{P}_\Delta \to \mathcal{Q}\,, \qquad \mathcal{V}(P) \coloneqq (\mathcal{T}^\pi_P)^\infty(Q_0) = Q^\pi_P\,.$$

The model map $\mathcal{W}$: for any $Q \in \mathcal{Q}$, let $M = M(Q) \in \mathbb{R}^{|\mathcal{G}|\times|\mathcal{S}|}$ be the matrix of Equation 1. Defining the *inverse Bellman operator* $\mathcal{I}_Q$ as a single step of tabular P-learning (Equation 3) with step size $0 < \alpha < 2/\sigma^2_{\max}$, Theorem 1 guarantees that $\mathcal{I}^\infty_Q(P_0) = M^+Q + (I - M^+M)P_0$ for any $P_0 \in \mathcal{P}$. If M has full column rank, we can define the P_0-independent *model map* on the set $\mathcal{Q}_\Delta \coloneqq \{Q \in \mathcal{Q} : M(Q) \text{ has full column rank}\}$, shown as the bottom edge of Figure 6, by

$$\mathcal{W} : \mathcal{Q}_\Delta \to \mathcal{P}\,, \qquad \mathcal{W}(Q)(s,a) \coloneqq \mathcal{I}^\infty_Q(P_0)(s,a) = M(Q)^+ Q(s,a)\,.$$

The unconstrained limit M^+Q may lie outside the probability simplex; see Appendix B.2 above for further discussion. The following proposition shows that, in the full-rank regime of Theorem 3, these maps are mutual inverses, so the diagram commutes in both directions.

Proposition 2 (Bijectivity of the value map)**.** *In the tabular setting, for any policy π and transition kernel $P \in \mathcal{P}_\Delta$, let $Q \coloneqq \mathcal{V}(P)$, and assume the induced matrix $M = M(Q)$ has full column rank. Then $\mathcal{W}(Q)$ is well-defined, independent of P_0, with equations.*

$$\mathcal{W}(\mathcal{V}(P)) = P \qquad \text{and} \qquad \mathcal{V}(\mathcal{W}(Q)) = Q\,.$$

In particular, the Q-learning map $\mathcal{V}$ is injective on $\{P \in \mathcal{P}_\Delta : M(\mathcal{V}(P)) \text{ has full column rank}\}$ with the P-learning map $\mathcal{W}$ as its inverse, so distinct world models in this set induce distinct Q-values and value equivalence is broken. By Theorem 3, this is a generic property for $|\mathcal{G}| \geq |\mathcal{S}|$.

Proof. Fix a state-action pair (s, a) and view $Q(s, a) \in \mathbb{R}^{|\mathcal{G}|}$ and $P(s, a) \in \mathbb{R}^{|\mathcal{S}|}$ as vectors. For any world model $P' \in \mathcal{P}$, Equation 2 gives

$$\mathcal{T}^\pi_{P'}(Q)(s, a) = M\, P'(s, a)\,,$$

where the matrix $M = M(Q)$ depends on (Q, π, r, γ). Since $Q = \mathcal{V}(P)$ is the fixed point of $\mathcal{T}^\pi_P$, we have $\mathcal{T}^\pi_P(Q) = Q$ and so $MP(s, a) = Q(s, a)$. Now M has full column rank, $M^+M = I$, so the consistent system $Mx = Q(s, a)$ has a *unique* solution

$$M^+Q(s, a) = M^+M\, P(s, a) = P(s, a)\,.$$

By Theorem 1, P-learning converges to $M^+Q(s, a) + (I - M^+M)P_0(s, a) = P(s, a)$. Hence $\mathcal{W}(Q)$ is well-defined, independent of P_0, and $\mathcal{W}(\mathcal{V}(P)) = \mathcal{W}(Q) = P$. It follows immediately that $\mathcal{V}(\mathcal{W}(Q)) = \mathcal{V}(P) = Q$. We conclude that $\mathcal{V}$ and $\mathcal{W}$ form a bijection between the set of world models whose value function induces a full-column-rank M and its image under $\mathcal{V}$. The argument applies verbatim to the optimal Bellman operator $\mathcal{T} = \mathcal{T}^\star$ rather than $\mathcal{T}^\pi$, replacing $V^\pi(s', g)$ by $\max_{a'} Q(s', a', g)$ in the matrix M. □

C Proofs for Finite MDPs

In this appendix, we formalise and prove our results for MDPs with finite state spaces: Appendices C.1 and C.2 for deterministic MDPs, Appendix C.3 for stochastic MDPs, and Appendix C.9 for local MDPs. We first state and prove a simple lemma that underpins many of our results.

Lemma 1. *Fix any finite (goal-augmented) MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{G}, P, \mu, r, \gamma)$, and recall the definition*

$$M(s', g) \coloneqq r(s', g) + \gamma V^\pi(s', g)$$

from Equation 1. Then the resolvent identity

$$M_l = R_l\big(I - \gamma(P^{\pi_l})^T\big)^{-1} \tag{4}$$

holds for all $g_l \in \mathcal{G}$, where $M_{lk} \coloneqq M(s_k, g_l)$, $R_{lk} \coloneqq r(s_k, g_l)$ and the policy-induced kernel $P^{\pi_l}(s_k, s_{k'}) \coloneqq \sum_j \pi(a_j \mid s_k, g_l)P(s_{k'} \mid s_k, a_j)$ are viewed as matrices.

Proof. We use $n \coloneqq |\mathcal{S}|$, $m \coloneqq |\mathcal{A}|$, $L \coloneqq |\mathcal{G}|$ throughout, viewing reward functions $R_{lk} \coloneqq r(s_k, g_l) \in \mathbb{R}^{|\mathcal{G}|\times|\mathcal{S}|}$ and similarly M, V as matrices. Rewriting Equation 2 as

$$M^\pi_{lk} = R_{lk} + \gamma V^{\pi_l}_{lk}\,,$$

and dropping the superscript π for simplicity, we define the policy-induced kernel $P^{\pi_l}_{kk'} \coloneqq \sum_j \pi_{lkj}P_{kjk'}$ for each l, and expand the Bellman equation

$$\begin{aligned} M_{lk} &= R_{lk} + \gamma \sum_{j,k'} \pi_{lkj}P_{kjk'}M_{lk'} \\ &= R_{lk} + \gamma \sum_{k'} P^{\pi_l}_{kk'}M_{lk'}\,. \end{aligned} \tag{5}$$

Defining $\bar{P}^{\pi_l} = (P^{\pi_l})^T$, we rewrite this equation as

$$\sum_{k'} M_{lk'}\left(\delta_{k'k} - \gamma\bar{P}^{\pi_l}_{k'k}\right) = R_{lk}\,.$$

The quantity in the brackets is not independent from l, so this is not a matrix product. However, the lth row of M can still be written as a matrix product:

$$M_l = R_l \left(I - \gamma \bar{P}^{\pi_l}\right)^{-1},$$

noting that $\bar{P}^{\pi_l}$ is column-stochastic matrix for each l, so $(I - \gamma \bar{P}^{\pi_l})$ is invertible for each l. □

C.1 Proof of Theorem 2 (Deterministic MDPs)

Theorem 2 (Deterministic, Finite MDPs). *Fix any deterministic finite MDP and goal set $|\mathcal{G}| \geq 1$.*

(a) For any policy π, the set of reward functions $r : \mathcal{S} \times \mathcal{G} \to \mathbb{R}$ for which M^π is column-injective, and thus $\exists! P(Q^\pi)$, has full Lebesgue measure.

(b) If Q is ϵ-approximate and M is column-injective with column separation Δ, the estimator $\hat{P}(s,a) = \operatorname{argmin}_{s'} \|Q(s,a) - M(s')\|_1$ satisfies $\hat{P} = P$ for all $\epsilon < \Delta/(2(1+\gamma m))$, and any estimator fails to recover P in the worst-case for $\epsilon \geq \Delta/2$.

We split this result into three parts: Proposition 3 for part (a), which includes the corollary mentioned below Theorem 2, and Propositions 4 and 5 for part (b).

Proposition 3. *Fix any deterministic finite MDP, goal set $|\mathcal{G}| \geq 1$ and policy π. Then the set of reward functions $r : \mathcal{S} \times \mathcal{G} \to \mathbb{R}$ for which M^π is column-injective, and thus $\exists! P(Q^\pi)$, has full Lebesgue measure. In particular, $|\mathcal{G}| = 1$, fix any $r_0 : \mathcal{S} \to \mathbb{R}$ and any random vector $\xi \in \mathbb{R}^{|\mathcal{S}|}$ whose distribution is absolutely continuous with respect to Lebesgue measure. Then the perturbed reward $r = r_0 + \xi$ satisfies $\exists! P(Q^\pi)$ almost surely, viewing $r \in \mathbb{R}^{|\mathcal{S}|}$ as a vector.*

Proof. First recall from Section 4.1 that for deterministic finite MDPs, the Bellman equation reduces to $M(P(s,a), g) = Q(s,a,g)$ for all $g \in \mathcal{G}$. Writing $M(s') = M(s', \cdot)$, this implies that the true transition kernel satisfies

$$P(s,a) = \arg\min_{s'} \|M(s') - Q(s,a)\|_1 .$$

In particular, a sufficient condition for this argmin to be unique, and thus $\exists! P(Q^\pi)$, is for M to be column-injective, i.e. $M(s_k) \neq M(s_{k'})$ for all $k \neq k'$. Now the resolvent identity (Equation 4)

$$M_l = R_l \left(I - \gamma \bar{P}^{\pi_l}\right)^{-1}$$

implies that the map $F : \mathbb{R}^{L \times n} \to \mathbb{R}^{L \times n}$ given by $F(R) = M$ is linear and invertible. Moreover, for any $k \neq k'$, the set $\{M \in \mathbb{R}^{L \times n} : M_{\cdot,k} = M_{\cdot,k'}\}$ is a proper subspace of $\mathbb{R}^{L \times n}$, so it has Lebesgue measure zero. Taking a finite union over the $\binom{n}{2}$ pairs $k \neq k'$, the set

$$\{R \in \mathbb{R}^{L \times n} \mid \exists k \neq k' \text{ with } F(R)_{\cdot,k} = F(R)_{\cdot,k'}\} = F^{-1}\left(\bigcup_{k \neq k'} \{M \in \mathbb{R}^{L \times n} \mid M_{\cdot,k} = M_{\cdot,k'}\}\right)$$

has Lebesgue measure zero since it is the preimage of a Lebesgue-measure-zero set under an invertible linear map. Therefore, M has pairwise distinct columns for almost every $R \in \mathbb{R}^{L \times n}$, i.e. almost every $r : \mathcal{S} \times \mathcal{G} \to \mathbb{R}$. The first half of the proof is complete. As a corollary, taking $|\mathcal{G}| = 1$ and viewing the single reward function as a vector $r \in \mathbb{R}^{|\mathcal{S}|}$, the set

$$B = \{r \in \mathbb{R}^{|\mathcal{S}|} \mid M^\pi \text{ not column-injective}\}$$

has zero Lebesgue measure. Now fix any base reward $r_0 \in \mathbb{R}^{|\mathcal{S}|}$ and any random vector $\xi \in \mathbb{R}^{|\mathcal{S}|}$ whose distributions μ is absolutely continuous with respect to λ. By translation-invariance of Lebesgue measure, $\lambda(B - r_0) = \lambda(B) = 0$, and absolute continuity ($\mu \ll \lambda$) implies $\mu(B - r_0) = 0$, i.e. $\mathbb{P}_{\xi \sim \mu}(r_0 + \xi \in B) = 0$. By definition of B, we conclude that $M^\pi(r_0 + \xi)$ is almost surely column-injective, hence $\exists! P(Q^\pi, \pi, r_0 + \xi, \gamma)$ almost surely. □

Proposition 4. *Fix any deterministic finite MDP and goal set $|\mathcal{G}| \geq 1$. If Q is ϵ-approximate and M is column-injective with column separation Δ, the estimator $\hat{P}(s,a) = \operatorname{argmin}_{s'} \|Q(s,a) - M(s')\|_1$ satisfies $\hat{P} = P$ for all $\epsilon < \Delta/(2(1+\gamma m))$.*

Proof. Fix a state-action pair (s_i, a_j) and write $k^\star := P(s_i, a_j)$ for the true successor. Throughout, indices $l \in [L], k \in [n], j \in [m]$ correspond to goals, states, and actions respectively, so $\pi_{lkj} = \pi(a_j \mid s_k, g_l)$ and $Q_{lkj} = Q(s_k, a_j, g_l)$; we moreover write $Q_{ij} = Q(s_i, a_j, \cdot) \in \mathbb{R}^L$. It suffices to show that Q_{ij} is closer to the true column $e_{k^\star}$ than to any other column e_k for $k \neq k^\star$, where e_k is the standard basis vector of $\mathbb{R}^n$ corresponding to state $s_k \in \mathcal{S}$. The approximate matrix M satisfies $M_{lk} - M^\pi_{lk} = \gamma \sum_{j'} \pi_{lkj'}(Q_{lkj'} - Q^\pi_{lkj'})$, so for each column k:

$$\|M_{\cdot,k} - M^\pi_{\cdot,k}\|_1 \leq \gamma \sum_l \sum_{j'} \pi_{lkj'} \left|Q_{lkj'} - Q^\pi_{lkj'}\right| \leq \gamma \sum_{j'} \|Q_{kj'} - Q^\pi_{kj'}\|_1 \leq \gamma m \epsilon \,.$$

By the triangle inequality, the distance from Q_{ij} to the true column satisfies

$$\|Q_{ij} - M^\pi_{\cdot,k^\star}\|_1 \leq \|Q_{ij} - Q^\pi_{ij}\|_1 + \|M_{\cdot,k^\star} - M^\pi_{\cdot,k^\star}\|_1 \leq \epsilon + \gamma m \epsilon = \epsilon(1 + \gamma m) \,.$$

For any $k \neq k^\star$, using the Bellman equation $Q^\pi_{ij} = M^\pi_{\cdot,k^\star}$, we apply both standard and reverse triangle inequalities as follows:

$$\begin{aligned}
\|Q_{ij} - M_{\cdot,k}\|_1 &= \|Q_{ij} - Q^\pi_{ij} + M^\pi_{\cdot,k^\star} - M^\pi_{\cdot,k} + M^\pi_{\cdot,k} - M_{\cdot,k}\|_1 \\
&\geq \|M^\pi_{\cdot,k^\star} - M^\pi_{\cdot,k}\|_1 - \left(\|Q_{ij} - Q^\pi_{ij}\|_1 + \|M^\pi_{\cdot,k} - M_{\cdot,k}\|_1\right) \\
&\geq \Delta_1 - \epsilon(1 + \gamma m) \,,
\end{aligned}$$

where $\|M_{\cdot,k^\star} - M_{\cdot,k}\|_1 \geq \Delta_1$ by definition of column separation. Recovery succeeds when the true column is strictly closer than any wrong column, i.e. $\epsilon(1 + \gamma m) < \Delta_1 - \epsilon(1 + \gamma m)$, giving $\epsilon < \Delta_1 / 2(1 + \gamma m)$. Since the choice of (s_i, a_j) was arbitrary, the bound holds uniformly. When π is unconditional, factoring out $\pi_{lkj'} = \pi_{kj'}$ sharpens the column error to $\gamma\epsilon$, giving an improved bound $\epsilon < \Delta_1 / 2(1 + \gamma)$. □

Proposition 5. *For any $n \geq 3$, $m \geq 1$, $L \geq 1$, $\gamma \in [0, 1)$, and $\epsilon > 0$, there exist two distinct deterministic transition kernels P, P', a reward function r, and a policy π such that the induced matrix M^π has column separation $\Delta_1 = \epsilon$, and*

(i) $\|Q^\pi_{ij}(P) - Q^\pi_{ij}(P')\|_1 \leq \epsilon$ for all (i, j);

(ii) $P(s_i, a_j) \neq P'(s_i, a_j)$ for some (i, j).

Consequently, any estimator mapping ϵ-approximate Q-values to a deterministic world model fails to recover the true world for $\epsilon \geq \Delta_1 / 2$.

Proof. Take $n = 3$, $m = 1$, $L = 1$. The construction below extends to any $n \geq 3, m \geq 1, L \geq 1$ by adding identity-transition states $s_4, \ldots, s_n$, self-loop actions $a_2, \ldots, a_m$, and extra goals $g_2, \ldots, g_L$ with $r_{l,k}$ identical at s_2, s_3 for $l \geq 2$ (contributing zero to both column separation and Q-differences). We define two deterministic transition kernels as follows: P and P' agree on the absorbing states $P(s_2, \cdot) = P'(s_2, \cdot) = s_2$ and $P(s_3, \cdot) = P'(s_3, \cdot) = s_3$, but disagree at (s_1, a_1) where $P(s_1, a_1) = s_2$ and $P'(s_1, a_1) = s_3$. Take the deterministic policy $\pi(a_1 \mid s) = 1$, and for any $c \in \mathbb{R}$, set rewards $r(s_2) := (c + \epsilon/2)(1 - \gamma)$, $r(s_3) := (c - \epsilon/2)(1 - \gamma)$, with $r(s_1)$ arbitrary. Since s_2, s_3 self-loop under all actions, the Bellman equation gives $V^\pi(s_k) = r_k / (1 - \gamma)$ at these states for both P, P', so

$$M^\pi_2 = c + \epsilon/2 \quad , \quad M^\pi_3 = c - \epsilon/2$$

and thus $\Delta_1 := \|M^\pi_2 - M^\pi_3\|_1 = \epsilon$, as claimed. Now the Q-values at s_2 or s_3 depend only on M^π_2 and M^π_3 respectively, which are invariant across P, P', so the Q-differences are zero. At (s_1, a_1),

$$Q^\pi_{11}(P) = M^\pi_2 \quad , \quad Q^\pi_{11}(P') = M^\pi_3 \,,$$

so $\|Q^\pi_{11}(P) - Q^\pi_{11}(P')\|_1 = \epsilon$. At (s_1, a_j) for $j \geq 2$, $\pi(a_1 \mid s_1) = 1$ gives $V^\pi_1(P) = Q^\pi_{11}(P) = M^\pi_2$ and similarly $V^\pi_1(P') = M^\pi_3$, so

$$Q^\pi_j(P) = r_1 + \gamma M^\pi_2 \quad , \quad Q^\pi_{1j}(P') = r_1 + \gamma M^\pi_3 \,,$$

giving $\|Q^\pi_j(P) - Q^\pi_j(P')\|_1 = \gamma\epsilon \leq \epsilon$. Hence (i) holds with all Q-differences bounded by ϵ, and (ii) follows from $P(s_1, a_1) = s_2 \neq s_3 = P'(s_1, a_1)$. Now let $\hat{Q} := (Q^\pi(P) + Q^\pi(P'))/2$, which satisfies $\|\hat{Q}_{ij} - Q^\pi_{ij}(P)\|_1 \leq \epsilon/2$ at every (i, j), with equality at (s_1, a_1), and similarly for P'. Hence at $\epsilon = \Delta_1/2$, both $(P, \hat{Q})$ and $(P', \hat{Q})$ satisfy the ϵ-feasibility constraint $\|\hat{Q} - Q^\pi\|_1 \leq \epsilon$. Any estimator $\hat{P}$ returns a world model that necessarily disagrees with either P or P' at (s_1, a_1), since $P(s_1, a_1) \neq P'(s_1, a_1)$, so every estimator fails to recover the true world for $\epsilon \geq \Delta_1 / 2$. □

C.2 Proof of Proposition 1

While Theorem 2 proves that "almost every" single-goal reward function satisfies $\exists! P(Q^\pi)$, this does not necessarily hold for general parametric families. Below we prove that the Gaussian family is particularly well-behaved. Note that $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ is treated as fixed (not dependent on g), but the same conclusion extends to the optimal policy $\pi_g : \mathcal{S} \to \mathcal{A}$ with respect to the varying goal g (cf. proof of the stochastic generalisation, Proposition 9).

Proposition 1 (Gaussian Rewards). *Fix any deterministic finite MDP, policy π, and variance $\sigma > 0$. For any injective map $\varphi : \mathcal{S} \hookrightarrow \mathbb{R}^d$, the single-goal set $\mathcal{G} = \{g\}$ with Gaussian reward function $r_g(s) = \exp(-\|\varphi(s) - g\|^2/2\sigma^2)$ satisfies $\exists! P(Q^\pi)$ for Lebesgue-almost every $g \in \mathbb{R}^d$.*

Proof. Viewing single-goal reward functions $r : \mathcal{S} \to \mathbb{R}$ as vectors $r \in \mathbb{R}^{|\mathcal{S}|}$, the Gaussian family $\{r_g \in \mathbb{R}^{|\mathcal{S}|} \mid g \in \mathbb{R}^d\}$ is a d-dimensional submanifold of $\mathbb{R}^{|\mathcal{S}|}$ which can have zero Lebesgue measure, so we cannot apply Theorem 2(a) directly. We instead exploit the explicit form of the matrix induced by Gaussian rewards. Since $|\mathcal{G}| = 1$, Equation 4 reduces to

$$M^\pi = r_g^\top H\,, \quad \text{where} \quad H := (I - \gamma \bar{P}^\pi)^{-1}\,.$$

Recall that $\exists! P(Q^\pi)$ holds if M^π is column-injective, and define $v_{ss'} := H_{\cdot,s} - H_{\cdot,s'}$. Then column-injectivity of M^π fails for a goal set $\mathcal{G} = \{g\}$ iff $r_g^\top v_{ss'} = 0$ for some $s \neq s'$, which we can rewrite as $h_{ss'}(g) = 0$ for some $s \neq s'$, where

$$h_{ss'}(g) := r_g^\top v_{ss'} = \sum_{k=1}^{n} e^{-\|\varphi(s_k)-g\|^2/2\sigma^2}(H_{ks} - H_{ks'})\,.$$

Since $\|\varphi(s_k) - g\|^2$ is a polynomial in g, each summand is a composition of real-analytic functions, so $h_{ss'}$ is real-analytic in $g \in \mathbb{R}^d$. Now assume for contradiction that $h_{ss'} \equiv 0$. Then

$$\sum_{k=1}^{n} c_k e^{-\|\varphi(s_k)-g\|^2/2\sigma^2} = 0 \quad \forall g \in \mathbb{R}^d\,,$$

where $c_k := H_{ks} - H_{ks'}$. Since φ is injective, the centres $x_k = \varphi(s_k)$ are distinct, and the Gaussians $\{g \mapsto e^{-\|x_k-g\|^2/2\sigma^2}\}_{k=1}^n$ centred at any distinct points $x_1, \ldots, x_n \in \mathbb{R}^d$ are linearly independent as functions on $\mathbb{R}^d$, so $c_k = 0$ for all k, hence $H_{\cdot,s} = H_{\cdot,s'}$. But H is invertible, so it has pairwise distinct columns, which is a contradiction, so we conclude $h_{ss'} \not\equiv 0$. The zero set of a non-zero real-analytic function on $\mathbb{R}^d$ has Lebesgue measure zero, so $\lambda(h_{ss'}^{-1}(0)) = 0$ for each pair $s \neq s'$. Column-injectivity fails only if g lies in $\bigcup_{s\neq s'} h_{ss'}^{-1}(0)$, a finite union of measure-zero sets, hence

$$\lambda\Big(\{g \in \mathbb{R}^d \mid \neg\exists! P(Q^\pi, \pi, r_g)\}\Big) \leq \sum_{s\neq s'} \lambda\big(h_{ss'}^{-1}(0)\big) = 0\,. \qquad \square$$

C.3 Proof of Theorem 3 (Stochastic MDPs)

Theorem 3 (Finite MDPs). *Fix any finite MDP and goal set $|\mathcal{G}| \geq |\mathcal{S}|$.*

(a) For any policy π, the set of reward functions $r : \mathcal{S} \times \mathcal{G} \to \mathbb{R}$ for which M^π is full-rank, and thus $\exists! P(Q^\pi)$, has full Lebesgue measure.

(b) For any goals $(\mathcal{G}, r)$ that span $\mathcal{S}$ and any discount factor $\gamma \in [0, 1)$, the set of policies π for which M^π is full-rank has full Lebesgue measure. Similarly, for any π, the set of $\gamma \in [0, 1)$ for which M^π is full-rank has full Lebesgue measure. In particular, there exist $\gamma^-, \gamma^+ \in (0, 1)$ such that M^π is full-rank for all $\gamma < \gamma^-$ or $\gamma > \gamma^+$.

(c) If Q is ϵ-approximate and M has full rank, the estimator $\hat{P}_{ij} = M^+ Q_{ij}$ satisfies

$$\|\hat{P}_{ij} - P_{ij}\|_1 \leq \|M^+\|_1(1 + \gamma m)\epsilon\,,$$

and any estimator has worst-case error $(1 + \gamma)\epsilon/2$.

We split this result into three parts: Proposition 6 for exact Q-values, and Propositions 7 and 8 for the upper and lower-bounds of the approximate setting. As foreshadowed in the main text, Proposition 6 is stronger than Theorem 3(a–b): informally, it states that fixing any three of the four components in (P, π, r, γ), the set of values of the remaining component for which M^π is full-rank, and thus $\exists! P(Q^\pi, \pi, r, \gamma)$, has full Lebesgue measure. Theorem 3(a) follows from part (b), while Theorem 3(b) follows from parts (c–d), where the failure set for discount factors is in fact *finite*, strictly sharper in the main text. Part (a) of Proposition 6 moreover shows that "almost all" kernels P are identified uniquely for fixed (π, r, γ). Proposition 6 also deals with optimal and optimal entropy-regularised policies, which call for additional care because such policies themselves vary with (P, r, γ).

C.3.1 Proof of Theorem 3(a–b)

To formally state the result, first write $n \coloneqq |\mathcal{S}|$, $m \coloneqq |\mathcal{A}|$, $L \coloneqq |\mathcal{G}|$ and let $\mathcal{P}_\Delta \coloneqq \{P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})\}$ be the space of world models, which is isomorphic to the simplex product $(\Delta^n)^{nm}$ and inherits the Lebesgue measure from $\mathbb{R}^{nm(n-1)}$. We write $\mathcal{R} = \{r : \mathcal{S} \times \mathcal{G} \to \mathbb{R}\}$ for the space of reward functions, $\Pi = \{\pi : \mathcal{S} \times \mathcal{G} \to \Delta(\mathcal{A})\}$ for the set of GC policies, and Π° for its deterministic subset. Finally, denote by $M^\star(P) = M^{\pi_l^\star(P)}(P) \in \mathbb{R}^{L \times n}$ the matrix corresponding to the optimal policy $\pi_l^\star(P)$; similarly for $M^\star(\gamma)$ and $M^\star(r)$. For part (d) we will additionally make use of the soft Bellman operator $\mathcal{T}_\alpha : \mathbb{R}^{n^2 m} \to \mathbb{R}^{n^2 m}$, defined for any $\alpha > 0$ by

$$(\mathcal{T}_\alpha Q)_{lij} = \sum_k P_{ijk} \left[R_{lk} + \gamma \alpha \log \left(\sum_{j'} \exp(Q_{lkj'}/\alpha) \right) \right],$$

the unique fixed point Q^α (cf. Lemma 2), and the optimal entropy-regularised policy $\pi_\alpha^\star$ given by

$$(\pi_\alpha^\star)_{lij} = \frac{\exp(Q^\alpha_{lij}/\alpha)}{\sum_{j'} \exp(Q^\alpha_{lij'}/\alpha)}.$$

Recall that P is identified uniquely by (Q^π, π, r, γ), written $\exists! P(Q^\pi)$, if there is a unique solution P to the Bellman equation $\mathcal{T}_P^\pi(Q^\pi) = Q^\pi$, which reduces to M^π being full-rank (Equation 2).

Proposition 6 (Finite MDPs, Exact)**.** *Fix any finite state space $\mathcal{S}$ and goals $(\mathcal{G}, r)$ that span $\mathcal{S}$.*

- *(a) For any $(\cdot, r, \gamma, \pi)$, the set $\{P \in \mathcal{P}_\Delta \mid \operatorname{rank}(M^\pi(P)) < n\}$ has Lebesgue measure zero. So does the optimal-policy set $\{P \in \mathcal{P}_\Delta \mid \operatorname{rank}(M^\star(P)) < n\}$.*
- *(b) For any $(P, \cdot, \gamma, \pi)$, the set $\{r \in \mathcal{R} \mid \operatorname{rank}(M^\pi(r)) < n\}$ has Lebesgue measure zero. So does the optimal-policy set $\{r \in \mathcal{R} \mid \operatorname{rank}(M^\star(r)) < n\}$.*
- *(c) For any $(P, r, \cdot, \pi)$, the set $\{\gamma \in [0, 1) \mid \operatorname{rank}(M^\pi(\gamma)) < n\}$ is finite. So is the optimal-policy set $\{\gamma \in [0, 1) \mid \operatorname{rank}(M^\star(\gamma)) < n\}$. Moreover, there exist $\gamma^-, \gamma^+ \in (0, 1)$ such that $M^\pi(\gamma)$ and $M^\star(\gamma)$ are full-rank for all $0 \le \gamma < \gamma^-$ or $\gamma > \gamma^+$.*
- *(d) For any $(P, r, \gamma, \cdot)$ the set $\{\pi \in \Pi \mid \operatorname{rank}(M^\pi) < n\}$ has Lebesgue measure zero. Moreover, the set $\{\alpha \in [0, \infty) \mid \operatorname{rank}(M^{\pi_\alpha^\star}) < n\}$ is finite, where $\pi_\alpha^\star$ is the optimal entropy-regularised policy with entropy coefficient α, and there exist $\alpha^-, \alpha^+ \in (0, \infty)$ such that $M^{\pi_\alpha^\star}$ is full-rank for all $\alpha < \alpha^-$ or $\alpha > \alpha^+$.*

Proof. We use $n = |\mathcal{S}|$, $m = |\mathcal{A}|$, $L = |\mathcal{G}|$ throughout, and recall the resolvent identity

$$M_l = R_l \left(I - \gamma \bar{P}^{\pi_l}\right)^{-1}$$

from Lemma 1. Though each row is produced by an invertible matrix (scaled by R_l), the resulting concatenation M may be rank-deficient. However, we can still prove that full-rankness is a generic property over each of the four variables, using the standard characterisation $\operatorname{rank}(M) < n \iff \det(M^T M) = 0$. The order is chosen to facilitate recycling along the way.

Part (c). Fix P, r and first consider **(i)** a fixed GC policy π, taking $M(\cdot)$ as a function of γ, omitting the superscript π for simplicity. Equation 4 gives $M_{lk} = \sum_{k'} R_{lk'} \left(I - \gamma \bar{P}^{\pi_l}\right)^{-1}_{k'k}$, so writing $C^{(l)}_{k'k}$ for the (k', k)th cofactor of $\left(I - \gamma \bar{P}^{\pi_l}\right)$, we use Cramer's rule to obtain

$$M_{lk} = \frac{\sum_{k'} R_{lk'} C^{(l)}_{k'k}}{\det(I - \gamma \bar{P}^{\pi_l})}. \tag{6}$$

Each cofactor $C^{(l)}_{k'k}$ is the signed determinant of an $(n-1)\times(n-1)$ submatrix of $\left(I-\gamma\bar{P}^{\pi_l}\right)$, which is a polynomial of degree no greater than $n-1$ in γ. Since the coefficients $R_{lk'}$ are constants (independent of γ), the numerator of M_{lk} is also a polynomial of degree at most $n-1$. Now for any $I\subset[L]$ with $|I|=n$, writing $M_I\in\mathbb{R}^{n\times n}$ for the submatrix of M with rows indexed by I, the Leibniz formula gives

$$\det(M_I)=\sum_\sigma \operatorname{sgn}(\sigma)\prod_{l\in I}M_{l\sigma(l)}=\frac{\sum_\sigma \operatorname{sgn}(\sigma)\prod_{l\in I}\left(\sum_{k'}R_{lk'}C^{(l)}_{k'\sigma(l)}\right)}{\prod_{l\in I}\det(I-\gamma\bar{P}^{\pi_l})}\,,$$

which is a rational function of γ whose numerator A_I has degree

$$\deg(A_I)\le\max_\sigma\sum_{l\in I}\deg\left(\sum_{k'}R_{lk'}C^{(l)}_{k'\sigma(l)}\right)\le\max_\sigma\sum_{l\in I}(n-1)=n(n-1)\,.$$

By the Cauchy–Binet formula, $\det(M^TM)=\sum_{|I|=n}\det(M_I)^2$, which is rational in γ with numerator of degree at most $2n(n-1)$. Finally note that $\det(M^TM)$ cannot be uniformly zero, since $M=R$ for $\gamma=0$, and spanning goals imply $\operatorname{rank}(R)=n$, hence $\det(R^TR)\neq 0$. In particular, the set of roots of the non-zero rational function $f:=\det(M^\pi(\cdot)^TM^\pi(\cdot)):[0,1)\to\mathbb{R}$ is finite and has cardinality bounded above by the degree of its numerator, so we conclude that $\{\gamma\in[0,1)\mid\operatorname{rank}(M^\pi(\gamma))<n\}$ has cardinality $C\le 2n(n-1)$.

We now turn to **(ii)** optimal policies $\pi^\star$, which depend on γ. The key is noticing that the optimal value function $V^\star(\gamma)$ is componentwise equal to:

$$V^\star(\gamma)=\max_{\pi\in\Pi^\circ}V^\pi(\gamma)\,,$$

where Π° is the set of goal-conditioned *deterministic* policies, which is a finite set bounded of size m^{n^2} (m^n possible deterministic policies for each of the n goals). It follows immediately that

$$M^\star(\gamma)=R+\gamma V^\star(\gamma)=\max_{\pi\in\Pi^\circ}M^\pi(\gamma)\,,$$

hence

$$\{\gamma\in[0,1)\mid\operatorname{rank}(M^\star(\gamma))<n\}\subseteq\bigcup_{\pi\in\Pi^\circ}\{\gamma\in[0,1)\mid\operatorname{rank}(M^\pi(\gamma))<n\}\,.$$

Applying part **(ii)** to each $\pi\in\Pi^\circ$, we obtain

$$\begin{aligned}C^\star&=\left|\{\gamma\in[0,1)\mid\operatorname{rank}(M^\star(\gamma))<n\}\right|\\&\le\left|\bigcup_{\pi\in\Pi^\circ}\{\gamma\in[0,1)\mid\operatorname{rank}(M^\pi(\gamma))<n\}\right|\\&\le\sum_{\pi\in\Pi^\circ}\left|\{\gamma\in[0,1)\mid\operatorname{rank}(M^\pi(\gamma))<n\}\right|\le 2m^{n^2}n(n-1)\,.\end{aligned}$$

In particular, since the zeroes are finite, there exists $\gamma^+\in(0,1)$ such that $M^\pi(\gamma)$ and $M^\star(\gamma)$ are full-rank for all $\gamma>\gamma^+$. Moreover, since M^π is continuous in γ and $M^\pi(0)=M^\star(0)=R$ has full rank by assumption that goals span the state space, there exists γ^- such that $M^\pi(\gamma)$ and $M^\star(\gamma)$ are full-rank for all $0\le\gamma<\gamma^-$ (otherwise the set of zeroes would be infinite).

Part (a). Now fix r,γ and begin with **(i)** a fixed GC policy π, taking $M(\cdot)$ as a function of P. Again, Equation 6 guarantees that each entry $M_{lk}(\cdot)$ is a rational function in the entries of P, so $f:=\det(M(\cdot)^TM(\cdot)):\mathcal{F}_\Delta\to\mathbb{R}$ must also be rational. To prove $f\not\equiv 0$, consider the static world $P_0(s'\mid s,a)=\delta_{ss'}$, which satisfies $\bar{P}_0^{\pi_l}=I$ for any policy π_l. Then using Equation 4,

$$M(P_0)=R\,(I-\gamma I)^{-1}=\frac{R}{1-\gamma}\quad\Longrightarrow\quad f(P_0)=\frac{\det(R^TR)}{(1-\gamma)^{2n}}\neq 0\,,$$

since spanning goals satisfy $\operatorname{rank}(R)=n$. Now consider the set $X=\{x\in\mathbb{R}^{n-1}\mid x_i\ge 0,\sum_i x_i\le 1\}\subset\mathbb{R}^{n-1}$ and the bijective map $g:X\to\Delta^n$ given by $g(x)=(x_1,\dots,x_{n-1},1-\sum_i x_i)$. The

intrinsic Lebesgue measure λ_Δ on $\Delta^n \subset \mathbb{R}^n$ is naturally inherited from the Lebesgue measure on $\mathbb{R}^{n-1}$ by defining

$$\lambda_\Delta(A) \coloneqq \lambda(g^{-1}(A)) \tag{7}$$

for any measurable set A. This extends to the product map $G = g^{mn} : X^{mn} \to (\Delta^n)^{mn} = \mathcal{F}_\Delta$. We now consider the pulled-back function $h : X^{mn} \to \mathbb{R}$ given by $h = f \circ G$. Since f and G are both non-zero rational functions, and G is bijective, h is also a non-zero rational map. The set of roots of non-zero rational functions $\mathbb{R}^d \to \mathbb{R}$ has Lebesgue measure zero, which also holds for non-zero rational maps $U \to \mathbb{R}$ on any subset $U \subset \mathbb{R}^d$ (by monotonicity of the Lebesgue measure), hence

$$\lambda(h^{-1}(0)) = 0\,. \tag{8}$$

We conclude that the bad set $B = \{P \in \mathcal{F}_\Delta \mid \operatorname{rank}(M(P)) < n\}$ satisfies

$$\lambda_\Delta(B) \;= \lambda_\Delta(f^{-1}(0)) \overset{(7)}{=} \lambda(G^{-1}(f^{-1}(0))) \;=\; \lambda\big((f \circ G)^{-1}(0)\big) \;=\; \lambda(h^{-1}(0)) \overset{(8)}{=} 0\,.$$

The argument for **(ii)** optimal policies is similar to part **(a)(ii)**. We have

$$M^\star(P) = R + \gamma V^\star(P) = \max_{\pi \in \Pi^\circ} M^\pi(P)\,,$$

where Π° is the finite set of goal-conditioned deterministic policies, hence

$$\{P \in \mathcal{F}_\Delta \mid \operatorname{rank}(M^\star(P)) < n\} \subseteq \bigcup_{\pi \in \Pi^\circ} \{P \in \mathcal{F}_\Delta \mid \operatorname{rank}(M^\pi(P)) < n\}\,.$$

Applying part **(b)(i)** to each $\pi \in \Pi^\circ$, we obtain

$$\begin{aligned}\lambda_\Delta\big(\{P \in \mathcal{F}_\Delta \mid \operatorname{rank}(M^\star(P)) < n\}\big) &\le \lambda_\Delta\Big(\bigcup_{\pi \in \Pi^\circ} \{P \in \mathcal{F}_\Delta \mid \operatorname{rank}(M^\pi(P)) < n\}\Big) \\ &\le \sum_{\pi \in \Pi^\circ} \lambda_\Delta\big(\{P \in \mathcal{F}_\Delta \mid \operatorname{rank}(M^\pi(P)) < n\}\big) = 0\,.\end{aligned}$$

Part (b). Now fix P, γ and π and take $M^\pi(\cdot)$ as a function of the reward $r \in \mathcal{R}$, written as a matrix $R_{lk} \in \mathbb{R}^{L \times n}$. Since each entry M_{lk} is linear in the entries of R by Equation 4, the function $f \coloneqq \det(M^\pi(\cdot)^T M^\pi(\cdot)) : \mathbb{R}^{L \times n} \to \mathbb{R}$ is a polynomial of degree $2n$ in the entries of R. To prove that $f \not\equiv 0$, consider the reward matrix R^0 with rows $R^0_l = e_l(I - \gamma \bar{P}^{\pi_l})$ for $l \in [n]$ and $R^0_l = 0$ for $l > n$. Then for $l \in [n]$,

$$M_l = R^0_l (I - \gamma \bar{P}^{\pi_l})^{-1} = e_l (I - \gamma \bar{P}^{\pi_l})(I - \gamma \bar{P}^{\pi_l})^{-1} = e_l\,,$$

while $M_l = 0$ for $l > n$, so $M^T M = I$ and $f(R^0) = \det(I) \neq 0$. The zero set of a non-zero polynomial has Lebesgue measure zero, so

$$\lambda\big(\{R \in \mathbb{R}^{L \times n} \mid \operatorname{rank}(M^\pi(R)) < n\}\big) = \lambda(f^{-1}(0)) = 0\,.$$

The result for optimal policies is identical to part **(a)(ii)**.

Part (d). Now fix P, r, γ and take $M^{(\cdot)}$ as a function of π. The proof over policies is identical to parts **(a-c)(i)**: $\det((M^\pi)^T M^\pi)$ is a rational function over the entries of π, and taking any goal-independent policy π_0 (eg the uniform policy) gives

$$M^{\pi_0} = R\left(I - \gamma \bar{P}^{\pi_0}\right)^{-1}$$

which has rank $\operatorname{rank}(R) = n$ by assumption of spanning goals, so $\det((M^{\pi_0})^T M^{\pi_0}) \neq 0$. The pullback argument establishes zero measure.

The second claim for entropy-regularised policies $\pi^\star_\alpha$ is less straightforward, because the matrix $M^{\pi^\star_\alpha}$ is not rational in α. Recall from above that Q^α is the unique fixed point of the soft Bellman operator $\mathcal{T}_\alpha$ (well-defined by Lemma 2), and $\pi^\star_\alpha$ is the softmax of Q^α at temperature α. Now let $f : (0, \infty) \to \mathbb{R}$ be defined by $f(\alpha) = \det((M^{\pi^\star_\alpha})^T M^{\pi^\star_\alpha})$. By Lemma 3, the function $\alpha \mapsto Q^\alpha$ is analytic and definable in $\mathbb{R}_{\exp}$. The matrix $M^{\pi^\star_\alpha}$ is obtained from Q^α and $\pi^\star_\alpha$ using definable operations (polynomials, division, $\exp$, $\log$), and $\det((M^{\pi^\star_\alpha})^T M^{\pi^\star_\alpha})$ is a polynomial in the entries of $M^{\pi^\star_\alpha}$, so f is definable. It is analytic for the same reasons. As $\alpha \to \infty$, the soft-optimal policy

converges to the uniform policy π_0, which is goal-independent, so $\det((M^{\pi_0})^T M^{\pi_0}) \neq 0$. By continuity,

$$\lim_{\alpha\to\infty} f(\alpha) = \det((M^{\pi_0})^T M^{\pi_0}) \neq 0\,,$$

so f is a non-zero, analytic and definable function. Since $\mathbb{R}_{\exp}$ is o-minimal [64, Second Main Theorem], and f is definable, the zero set $\{\alpha \in (0,\infty) \mid f(\alpha) = 0\}$ is a finite union of points and open intervals. But f cannot be zero on an interval by the identity theorem for analytic functions, so the zero set is a finite set of C points. In particular, there exist $\alpha^-, \alpha^+ \in (0,\infty)$ with $\alpha^- < \alpha^+$ such that $f(\alpha) \neq 0$ for all $\alpha \notin (\alpha^-, \alpha^+)$, since otherwise the zero set of f would be infinite, and thus $M^{\pi^\star_\alpha}$ is full-rank for all $\alpha \notin (\alpha^-, \alpha^+)$. Adding the single boundary point $\alpha = 0$ preserves finiteness:

$$\left|\{\alpha \in [0,\infty) \mid \operatorname{rank}(M^{\pi^\star_\alpha}) < n\}\right| \subseteq \left|\{0\} \cup \{\alpha \in (0,\infty) \mid f(\alpha) = 0\}\right| \leq 1 + C < \infty\,. \qquad \square$$

C.3.2 Proof of Theorem 3(c)

We split Theorem 3(c) into two parts: Proposition 7 for the upper bound and Proposition 8 for the worst-case lower bound. Recall that a policy is called *unconditional* if $\pi_g = \pi_{g'}$ for all $g \in \mathcal{G}$.

Proposition 7 (Finite MDPs, Approximate)**.** *Fix a finite MDP and assume $(\mathcal{G}, r)$ span $\mathcal{S}$, Q is ϵ-approximate, and M has full rank. Then the estimator $\hat{P}_{ij} = M^+ Q_{ij}$ satisfies*

$$\|\hat{P}_{ij} - P_{ij}\|_1 \leq \|M^+\|_1 (1+\gamma m)\epsilon\,.$$

Moreover, a sufficient condition for M to be full-rank is for M^π to be full-rank and $\epsilon < 1/\gamma m \|(M^\pi)^+\|_1$. If π is unconditional and goals are states ($\mathcal{G} = \mathcal{S}$), then M^π is always full-rank and the bound improves to

$$\|\hat{P}_{ij} - P_{ij}\|_1 \leq \frac{\epsilon(1+\gamma)^2}{1-\epsilon\gamma(1+\gamma)} \leq \frac{4\epsilon}{1-2\epsilon} \in O(\epsilon)$$

for all $\epsilon < 1/\gamma(1+\gamma)$, which is satisfied whenever $\epsilon < 1/2$.

Proof. We first decompose $M = M^\pi + E$, where $E_{lk} = \gamma \sum_{j'} \pi_{lkj'}(Q_{lkj'} - Q^\pi_{lkj'})$ satisfies

$$\begin{aligned}\|E\|_1 &= \max_k \sum_l \left|E_{lk}\right| \leq \gamma \max_k \sum_l \sum_{j'} \pi_{l,kj'} \left|Q_{lkj'} - Q^\pi_{lkj'}\right| \\ &\leq \gamma \max_k \sum_l \sum_{j'} \left|Q_{lkj'} - Q^\pi_{lkj'}\right| = \gamma \max_k \sum_{j'} \|Q_{kj'} - Q^\pi_{kj'}\|_1 \leq \gamma m \epsilon\,.\end{aligned}$$

If M is full-rank, then $M^+ M^\pi = M^+(M - E) = I - M^+ E$ and

$$\begin{aligned}\hat{P}_{ij} - P_{ij} &= M^+ Q_{ij} - P_{ij} = M^+(Q_{ij} - Q^\pi_{ij}) + (M^+ M^\pi - I) P_{ij} \\ &= M^+(Q_{ij} - Q^\pi_{ij}) - M^+ E P_{ij} = M^+\left(Q_{ij} - Q^\pi_{ij} - E P_{ij}\right).\end{aligned}$$

Taking norms and using $\|P_{ij}\|_1 = 1$,

$$\|\hat{P}_{ij} - P_{ij}\|_1 \leq \|M^+\|_1 (\epsilon + \gamma m \epsilon) = \|M^+\|_1 (1+\gamma m)\epsilon\,.$$

A sufficient condition for M to be full-rank is $\epsilon < 1/\gamma m \|(M^\pi)^+\|_1$, since then $(M^\pi)^+ M = I - (M^\pi)^+ E$ is invertible if $\|(M^\pi)^+\|_1 \|E\|_1 < 1$, and $(M^\pi)^+ M$ being invertible forces M full-rank since $Mx = 0$ implies $x = 0$. When π is unconditional and $\mathcal{G} = \mathcal{S}$, factoring out $\pi_{lkj'} = \pi_{kj'}$ sharpens the bound to $\|E\|_1 \leq \gamma\epsilon$ and Equation 4 gives $M^\pi = (I - \gamma(P^\pi)^T)^{-1}$, implying M^π is always invertible with $\|(M^\pi)^{-1}\|_1 = 1 + \gamma$ by row-stochasticity of P^π. Then

$$\begin{aligned}\|\hat{P}_{ij} - P_{ij}\|_1 &\leq \|(M^\pi)^{-1}\|_1 \|M^\pi \hat{P}_{ij} - M^\pi P_{ij}\|_1 \\ &= (1+\gamma)\|(M^\pi - M)\hat{P}_{ij} + Q_{ij} - Q^\pi_{ij}\|_1 \\ &\leq (1+\gamma)\left(\|E\|_1 \|\hat{P}_{ij}\|_1 + \|Q_{ij} - Q^\pi_{ij}\|_1\right) \\ &\leq (1+\gamma)\left(\gamma\epsilon\Big(\|\hat{P}_{ij} - P_{ij}\|_1 + 1\Big) + \epsilon\right).\end{aligned}$$

Solving for $\|\hat{P}_{ij} - P_{ij}\|_1$, we conclude

$$\|\hat{P}_{ij} - P_{ij}\|_1 \leq \frac{\epsilon(1+\gamma)^2}{1-\epsilon\gamma(1+\gamma)} \leq \frac{4\epsilon}{1-2\epsilon} \in O(\epsilon)$$

for all $\epsilon < 1/\gamma(1+\gamma)$, which is satisfied whenever $\epsilon < 1/2$. $\square$

We now show that the upper bound of Proposition 7 is tight up to a constant factor.

Proposition 8 (Worst-Case Lower Bound)**.** *For any $n \geq 2$, $m \geq 2$, $\gamma \in [0, 1)$, and $\epsilon > 0$ sufficiently small, there exist two MDPs with transition kernels $P, P' \in \mathcal{P}_\Delta$, a goal set $\mathcal{G}$ that spans the state space, and a policy π such that*

(i) $\|Q^\pi_{ij}(P) - Q^\pi_{ij}(P')\|_1 \leq \epsilon$ *for all* (i, j)*;*

(ii) $\|P_{ij} - P'_{ij}\|_1 \geq (1 + \gamma)\epsilon$ *for some* (i, j)*.*

Consequently, any estimator $\hat{P}$ mapping ϵ-approximate Q-values to a world model incurs worst-case error at least $(1 + \gamma)\epsilon/2 \in \Omega(\epsilon)$.

Proof. Take $n = 2$, $m = 2$, $L = 2$ with indicator goals $\mathcal{G} = \mathcal{S}$. The construction below extends to any $n \geq 2, m \geq 2$ by adding identity-transition states $s_3, \ldots, s_n$ and self-loop actions $a_3, \ldots, a_m$. We define two transition kernels as follows: P always swaps states regardless of actions, i.e. $P(s_2 \mid s_1, \cdot) = 1$, $P(s_1 \mid s_2, \cdot) = 1$, and P' agrees with P everywhere except $P'(\cdot s_1, a_1) = (\delta, 1 - \delta)$ for some $\delta \in (0, 1)$. Now take the deterministic policy $\pi(a_2 \mid s_k) = 1$ for $k \in \{1, 2\}$ that always selects the second action a_2. Since a_1 is off-policy, the value function V^π depends only on a_2-transitions, which are identical in P and P', hence $V^\pi(P) = V^\pi(P')$ and $M^\pi(P) = M^\pi(P')$. The policy-induced transition $\bar{P}^\pi = (P^\pi)^T$ is the swap matrix, so

$$M^\pi = (I - \gamma \bar{P}^\pi)^{-1} = \frac{1}{1 - \gamma^2} \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}.$$

Comparing action-value functions, we have $Q_{lij}(P) = \sum_k P_{ijk} M^\pi_{lk} = Q_{lij}(P')$ for all $(i, j) \neq (1, 1)$ since P, P' agree everywhere except at (s_1, a_1). For the latter,

$$Q^\pi_{l11}(P) = \sum_k P(s_k \mid s_1, a_1) M^\pi_{lk} = M^\pi_{l2}, \quad Q^\pi_{l11}(P') = M^\pi_{l2} + \delta(M^\pi_{l1} - M^\pi_{l2}),$$

so

$$\|Q^\pi_{11}(P) - Q^\pi_{11}(P')\|_1 = \delta \sum_l \left| M^\pi_{l1} - M^\pi_{l2} \right| = \frac{2\delta}{1 - \gamma^2} |1 - \gamma| = \frac{2\delta}{1 + \gamma}.$$

Setting $\delta := (1 + \gamma)\epsilon/2$ gives $\|Q^\pi_{ij}(P) - Q^\pi_{ij}(P')\|_1 \leq \epsilon$ for all (i, j) and $\|P_{1,1} - P'_{1,1}\|_1 = 2\delta = (1+\gamma)\epsilon$, as claimed. For $\epsilon < 2/(1+\gamma)$ we have $\delta < 1$, so P' is a valid world model. In particular, for any estimator $\hat{P}$, let $\hat{Q} := Q^\pi(P)$ and notice that $(P, \hat{Q})$ and $(P', \hat{Q})$ satisfy $\|\hat{Q} - Q^\pi(P)\|_1 = 0 \leq \epsilon$ and $\|\hat{Q} - Q^\pi(P')\|_1 \leq \epsilon$. The triangle inequality gives

$$\|\hat{P}(\hat{Q}) - P\|_1 + \|\hat{P}(\hat{Q}) - P'\|_1 \geq \|P - P'\|_1 = (1 + \gamma)\epsilon,$$

so $\max\big(\|\hat{P}(\hat{Q}) - P\|_1, \|\hat{P}(\hat{Q}) - P'\|_1\big) \geq (1 + \gamma)\epsilon/2$. Hence at least one of P, P' is at distance $\geq (1 + \gamma)\epsilon/2$ from $\hat{P}(\hat{Q})$, and since $\hat{P}$ cannot distinguish between them (both are ϵ-feasible given $\hat{Q}$), the worst-case error of $\hat{P}$ on ϵ-approximate inputs is at least $(1 + \gamma)\epsilon/2$. □

C.3.3 Supporting Results

We prove supporting results for Proposition 6, starting with the soft Bellman equation.

Lemma 2 (Soft Bellman)**.** *For any $\alpha > 0$, $\gamma \in [0, 1)$, the soft Bellman operator $\mathcal{T}_\alpha$ given by*

$$(\mathcal{T}_\alpha Q)_{lij} = \sum_k P_{ijk} \left[R_{lk} + \gamma\alpha \log \left(\sum_{j'} \exp(Q_{lkj'}/\alpha) \right) \right]$$

is a γ-contraction in the supremum norm: $\|\mathcal{T}_\alpha Q - \mathcal{T}_\alpha Q'\|_\infty \leq \gamma \|Q - Q'\|_\infty$.

Proof. The log-sum-exp function $\alpha \log(\sum_{j'} \exp(Q_{lkj'}/\alpha))$ is 1-Lipschitz in Q with respect to the supremum norm. Since P_{ijk} sums to 1 over k,

$$\begin{aligned}
|(\mathcal{T}_\alpha Q)_{lij} - (\mathcal{T}_\alpha Q')_{lij}| &= \gamma \left| \sum_k P_{ijk} \alpha \log \left(\frac{\sum_{j'} \exp(Q_{lkj'}/\alpha)}{\sum_{j'} \exp(Q'_{lkj'}/\alpha)} \right) \right| \\
&\leq \gamma \sum_k P_{ijk} \left| \alpha \log \left(\frac{\sum_{j'} \exp(Q_{lkj'}/\alpha)}{\sum_{j'} \exp(Q'_{lkj'}/\alpha)} \right) \right| \\
&\leq \gamma \sum_k P_{ijk} \|Q - Q'\|_\infty = \gamma \|Q - Q'\|_\infty \,. \qquad \square
\end{aligned}$$

We will use the following background on *definable* functions to show that the map $\alpha \mapsto Q^\alpha$ is particularly well-behaved.

Definition 1. A set $S \subseteq \mathbb{R}^n$ is *definable* in $\mathbb{R}_{\exp} = (\mathbb{R}, <, +, \cdot, \exp)$ if it can be described by a first-order formula using the symbols $<, +, \cdot, \exp$ and quantifiers over $\mathbb{R}$. For example, the graph of the logarithm function $G_{\log} = \{(x, y) \mid x > 0 \wedge \exp(y) = x\}$ is definable. A function $f : A \to \mathbb{R}^m$ (where $A \subseteq \mathbb{R}^n$) is *definable* if its graph $\{(x, f(x)) \mid x \in A\} \subseteq \mathbb{R}^{n+m}$ is a definable set. For example, $\log : \mathbb{R}^+ \to \mathbb{R}$ is definable, but $\sin : \mathbb{R} \to \mathbb{R}$ is not.

A foundational result of Wilkie [64, Second Main Theorem] establishes that $\mathbb{R}_{\exp}$ is *o-minimal*: every definable subset of $\mathbb{R}$ is a finite union of points and open intervals. This is a strong tameness property which implies, for instance, that the set of zeroes of a definable function $f : \mathbb{R} \to \mathbb{R}$ is a finite union of points and open intervals. If f is also a non-trivial analytic function, then it cannot be zero on an open interval, so it must have a finite set of zeroes (analogous to polynomials!)

Lemma 3 (Soft Bellman definability)**.** *For each $\alpha > 0$, let Q^α be the unique fixed point of the soft Bellman operator $\mathcal{T}_\alpha$. Then the function $\alpha \mapsto Q^\alpha$ is definable and real analytic in $\mathbb{R}_{\exp}$.*

Proof. The set $R = \{(\alpha, Q) \in (0, \infty) \times \mathbb{R}^{n^2 m} : \mathcal{T}_\alpha Q = Q\}$ is definable in $\mathbb{R}_{\exp}$, since $\mathcal{T}_\alpha$ involves only polynomial operations and the functions exp and log. By Lemma 2, $\mathcal{T}_\alpha$ is a contraction, so there is a unique Q^α such that $\mathcal{T}_\alpha Q^\alpha = Q^\alpha$ for each $\alpha > 0$. In particular, R is the graph of $\alpha \mapsto Q^\alpha$, so the map is definable.

Now define $F : (0, \infty) \times \mathbb{R}^{n^2 m} \to \mathbb{R}^{n^2 m}$ by $F(\alpha, Q) = Q - \mathcal{T}_\alpha Q$, which is analytic because it is the composition of analytic functions (polynomials, exp and log). The Jacobian of F with respect to Q is the $n^2 m \times n^2 m$ matrix $D_Q F = I - D_Q \mathcal{T}_\alpha$. By Lemma 2, $\mathcal{T}_\alpha$ is a γ-contraction in the sup-norm. For any unit vector v with $\|v\|_\infty = 1$,

$$\|D_Q \mathcal{T}_\alpha \cdot v\|_\infty = \lim_{t \to 0} \frac{\|\mathcal{T}_\alpha(Q + tv) - \mathcal{T}_\alpha Q\|_\infty}{|t|} \leq \lim_{t \to 0} \frac{\gamma \|tv\|_\infty}{|t|} = \gamma \|v\|_\infty = \gamma \,,$$

so taking the supremum over unit vectors implies $\|D_Q \mathcal{T}_\alpha\|_\infty \leq \gamma < 1$. It follows that the Neumann series $\sum_{k \geq 0} (D_Q \mathcal{T}_\alpha)^k$ converges in operator norm to $(I - D_Q \mathcal{T}_\alpha)^{-1}$, so $D_Q F = I - D_Q \mathcal{T}_\alpha$ is invertible. In particular, $D_Q F$ and its inverse are continuous in (α, Q), so for any (α_0, Q_0) such that $F(\alpha_0, Q_0) = 0$, the analytic implicit function theorem [65] implies that there is an analytic $G_U : U \to \mathbb{R}^{n^2 m}$ such that $F(\alpha, G(\alpha)) = 0$ for all α in some neighbourhood U of α_0. Uniqueness of solutions Q^α to $F(\alpha, Q^\alpha) = 0$ implies that the local maps G_U agree everywhere: for two different neighbourhoods U, V, the analytic functions G_U and G_V must satisfy $F(\alpha, G_U(\alpha)) = 0 = F(\alpha, G_V(\alpha))$ for all $\alpha \in U \cap V$, but this implies $G_U(\alpha) = Q^\alpha = G_V(\alpha)$, so $G_U = G_V$. In particular, there is a global analytic function $G : (0, \infty) \to \mathbb{R}^{n^2 m}$ such that $F(\alpha, G(\alpha)) = 0$ for all $\alpha > 0$. By uniqueness, $Q^\alpha = G(\alpha)$, so $\alpha \mapsto Q^\alpha$ is real analytic on $(0, \infty)$. $\square$

C.4 Proofs of Corollary 1 (Perturbed Rewards) and Proposition 9 (Gaussian Rewards)

Proposition 3 for deterministic kernels guaranteed that $\exists! P(Q^\pi)$ almost surely under any absolutely continuous perturbation ξ of a single-goal base reward $r_0 \in \mathbb{R}^{|\mathcal{S}|}$. The same translation-invariance argument extends to the stochastic setting of Theorem 3, viewing the reward as a matrix and the perturbation as matrix-valued.

Corollary 1 (Perturbed Rewards, Stochastic)**.** *Fix any finite MDP, policy π goal set $|\mathcal{G}| \geq |\mathcal{S}|$, base reward $R_0 \in \mathbb{R}^{|\mathcal{G}|\times|\mathcal{S}|}$ and random matrix $\Xi \in \mathbb{R}^{|\mathcal{G}|\times|\mathcal{S}|}$ whose distribution μ is absolutely continuous with respect to Lebesgue measure (e.g. entrywise independent Gaussians). Then the perturbed reward $R \coloneqq R_0 + \Xi$ satisfies $\exists! P(Q^\pi)$ almost surely.*

Proof. Viewing reward functions $r : \mathcal{S} \times \mathcal{G} \to \mathbb{R}$ as matrices $R \in \mathbb{R}^{|\mathcal{G}|\times|\mathcal{S}|}$ with $R_{lk} \coloneqq r(s_k, g_l)$, Proposition 6(b) establishes that the set

$$B \coloneqq \{R \in \mathbb{R}^{|\mathcal{G}|\times|\mathcal{S}|} \mid \operatorname{rank}(M^\pi(R)) < |\mathcal{S}|\}$$

has zero Lebesgue measure. Now fix any base reward $R_0 \in \mathbb{R}^{|\mathcal{G}|\times|\mathcal{S}|}$ and any random matrix $\Xi \in \mathbb{R}^{|\mathcal{G}|\times|\mathcal{S}|}$ whose distribution μ is absolutely continuous with respect to Lebesgue measure λ. By translation-invariance of Lebesgue measure, $\lambda(B - R_0) = \lambda(B) = 0$, and absolute continuity ($\mu \ll \lambda$) implies $\mu(B - R_0) = 0$, i.e. $\mathbb{P}_{\Xi\sim\mu}(R_0 + \Xi \in B) = 0$. By definition of B, $M^\pi(R_0 + \Xi)$ is almost surely full-rank, hence $\exists! P(Q^\pi, \pi, R_0 + \Xi, \gamma)$ almost surely. □

We similarly extend Proposition 1, where a single Gaussian reward generically implies $\exists! P(Q^\pi)$ for deterministic MDPs, to the stochastic setting. Extra care is required to define π independently from g, but the result again extends to include optimal policies defined as a function of g.

Proposition 9 (Gaussian Rewards)**.** *Fix any finite MDP, $L \geq |\mathcal{S}|$, policies $\pi_1, \ldots, \pi_L : \mathcal{S} \to \Delta(\mathcal{A})$, and $\sigma_1, \ldots, \sigma_L > 0$. For any injective map $\varphi : \mathcal{S} \hookrightarrow \mathbb{R}^d$, the goal set $\mathcal{G} = \{g_1, \ldots, g_L\}$ with Gaussian rewards $r_{g_l}(s) = \exp(-\|\varphi(s) - g_l\|^2/2\sigma_l^2)$ and policy $\pi(\cdot|s, g_l) \coloneqq \pi_l(\cdot|s)$ satisfies $\exists! P(Q^\pi)$ for Lebesgue-almost every $(g_1, \ldots, g_L) \in (\mathbb{R}^d)^L$. The same conclusion also holds for the optimal (instead of fixed) policy $\pi^\star$ with respect to $(g_1, \ldots, g_L)$.*

Proof. Write $n \coloneqq |\mathcal{S}|$ and recall that $\exists! P(Q^\pi)$ if $\operatorname{rank}(M^\pi) = n$, which for $L \geq n$ is equivalent to $\det((M^\pi)^T M^\pi) \neq 0$. By Equation 4, M^π has rows

$$M_l^\pi = r_{g_l} H_l\,, \quad \text{where} \quad H_l \coloneqq (I - \gamma \bar{P}^{\pi_l})^{-1}\,.$$

Writing $R_{lk} \coloneqq r_{g_l}(s_k) = \exp(-\|\varphi(s_k) - g_l\|^2/2\sigma_l^2)$, the matrix R is real-analytic in the goal tuple $\mathbf{g} \coloneqq (g_1, \ldots, g_L) \in (\mathbb{R}^d)^L$. Since each $M_l^\pi = R_l H_l$ is linear in R_l with goal-independent H_l,

$$f(g_1, \ldots, g_L) \coloneqq \det\left((M^\pi)^T M^\pi\right)$$

is real-analytic on $(\mathbb{R}^d)^L$. Now the zero set of a non-zero real-analytic function on $(\mathbb{R}^d)^L$ has Lebesgue measure zero, so it suffices to exhibit one tuple of goals for which $f \neq 0$. We first show that for any fixed $\sigma > 0$, the family $S_\sigma \coloneqq \{r_{g,\sigma} \in \mathbb{R}^n \mid g \in \mathbb{R}^d\}$ spans $\mathbb{R}^n$. Assume otherwise for contradiction; then there exists $w \neq 0$ such that

$$\sum_{k=1}^{n} w_k e^{-\|\varphi(s_k)-g\|^2/2\sigma^2} = w^T r_{g,\sigma} = 0$$

for all $g \in \mathbb{R}^d$. Since φ is injective, the centres $x_k = \varphi(s_k)$ are distinct, and the Gaussians $\{g \mapsto e^{-\|x_k-g\|^2/2\sigma^2}\}_{k=1}^n$ centred at any distinct points $x_1, \ldots, x_n \in \mathbb{R}^d$ are linearly independent as functions on $\mathbb{R}^d$, so we conclude $w_k = 0$ for all k, a contradiction. Hence $\operatorname{span}(S_\sigma) = \mathbb{R}^n$. We now construct $(g_l)_{l=1}^n$ inductively such that $\{M_l^\pi\}_{l=1}^n = \{r_{g_l} H_l\}_{l=1}^n$ are linearly independent vectors. Proceed by induction on l: at step $l \in \{1, \ldots, n\}$, with g_j already chosen for $j < l$, set $V_0 = \{0\}$ and

$$V_{l-1} \coloneqq \operatorname{span}\{M_j^\pi\}_{j<l}$$

for $l - 1 \geq 1$, of dimension no greater than $l - 1 < n$. Since H_l is invertible,

$$\operatorname{span}(\{r_{g,\sigma_l} H_l : g \in \mathbb{R}^d\}) = \operatorname{span}(\{r_{g,\sigma_l} : g \in \mathbb{R}^d\}) = \operatorname{span}(S_{\sigma_l}) = \mathbb{R}^n\,,$$

so the family $\{r_{g,\sigma_l} H_l : g \in \mathbb{R}^d\}$ cannot be contained in the proper subspace V_{l-1}, hence there exists $g_l \in \mathbb{R}^d$ with $r_{g_l} H_l \notin V_{l-1}$. After n steps, we obtain $V_n = \operatorname{span}\{M_j^\pi\}_{j=1}^n = \mathbb{R}^n$ and thus $\operatorname{rank}(M^\pi) = n$. Choosing the remaining g_l for $l > n$ arbitrarily, we obtain $f(g_1, \ldots, g_L) \neq 0$ and thus $f \not\equiv 0$. We conclude $\lambda(f^{-1}(0)) = 0$, and finally

$$\lambda\Big(\{\mathbf{g} \in (\mathbb{R}^d)^L \mid \neg\exists! P(Q^\pi)\}\Big) \leq \lambda\Big(\{\mathbf{g} \in (\mathbb{R}^d)^L \mid \operatorname{rank}(M^\pi(\mathbf{g})) < n\}\Big) = \lambda(f^{-1}(0)) = 0\,.$$

For the optimal-policy claim, let Π° denote the finite set of goal-conditioned deterministic policies $\pi : \mathcal{S} \times [L] \to \mathcal{A}$, of size $|\mathcal{A}|^{|\mathcal{S}|\cdot L}$. For each goal tuple $\mathbf{g} \in (\mathbb{R}^d)^L$, define $\pi^\star(\mathbf{g}) \in \Pi^\circ$ by $\pi^\star(\mathbf{g})(\cdot, l) \coloneqq \pi^\star(\cdot, g_l)$, the optimal policy for goal g_l. Then

$$\{\mathbf{g} : \operatorname{rank}(M^{\pi^\star(\mathbf{g})}(\mathbf{g})) < n\} \subseteq \bigcup_{\pi \in \Pi^\circ} \{\mathbf{g} : \operatorname{rank}(M^{\pi}(\mathbf{g})) < n\}\,.$$

For each fixed $\pi \in \Pi^\circ$, decomposing into per-slot policies $\pi_l \coloneqq \pi(\cdot, l)$ and applying the first part of the proposition gives measure zero for the corresponding term. By finite subadditivity,

$$\lambda\big(\{\mathbf{g} : \operatorname{rank}(M^{\pi^\star(\mathbf{g})}(\mathbf{g})) < n\}\big) \le \sum_{\pi \in \Pi^\circ} \lambda\big(\{\mathbf{g} : \operatorname{rank}(M^{\pi}(\mathbf{g})) < n\}\big) = 0\,. \qquad \square$$

C.5 Example of Rank Deficiency

Theorem 3 proves that M^π is almost always full-rank, but singularities can occur for specific (P, π, γ) even when goals are states ($\mathcal{G} = \mathcal{S}$). We construct an explicit example for $n = m = 3$. Let $\mathcal{S} = \{0, 1, 2\} = \mathcal{A}$, with deterministic transition kernel $P(s, a) = a$ and deterministic policies $\pi_0(i) \equiv 0$; $\pi_1(0) = 1, \pi_1(1) = 1, \pi_1(2) = 2$; $\pi_2(0) = 2, \pi_2(1) = 1, \pi_2(2) = 2$. Writing $c = 1/(1-\gamma)$, one can analytically compute:

$$M^\pi = \begin{pmatrix} c & c-1 & c-1 \\ c-1 & c & 0 \\ c-1 & 0 & c \end{pmatrix} = I + (c-1)\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}.$$

The rhs matrix has eigenvalues $1+\sqrt{2}$, 1, $1-\sqrt{2}$, so setting $\gamma = 1/\sqrt{2}$ gives $1+(c-1)(1-\sqrt{2}) = 0$ and thus $\det(M^\pi) = 0$. However, note that the matrix is full-rank for all $\gamma \neq 1/\sqrt{2}$, and for any perturbation of the policies, making singularities "exceptionally rare" as guaranteed by Theorem 3.

C.6 Improved Bounds via On-Support Q-Errors

The bounds of Theorem 2(b) and Theorem 3(c) assume a *uniform* error $\|Q(s, a, \cdot) - Q^\pi(s, a, \cdot)\|_1 \le \epsilon$ over all (s, a), whereas the Q-errors of deep RL agents are typically worst on states the agent never visits. However, the proof can easily be extended to only require on-support Q-errors.

Formally, for each goal g let $\operatorname{Reach}(\pi_g) \subseteq \mathcal{S}$ be the set of states with positive occupancy under π_g from μ, and define the *on-support* states, actions and pairs by

$$\mathcal{S}_o \coloneqq \bigcup_{g \in \mathcal{G}} \operatorname{Reach}(\pi_g)\,, \qquad \mathcal{A}_o(s) \coloneqq \{a : \pi(a \mid s, g) > 0 \text{ for some } g \in \mathcal{G}\}\,,$$

$$\mathcal{O} \coloneqq \bigcup_{g \in \mathcal{G}} \{(s, a) : s \in \operatorname{Reach}(\pi_g),\ \pi(a \mid s, g) > 0\}\,.$$

Note that $\mathcal{O} \subseteq \{(s, a) : s \in \mathcal{S}_o, a \in \mathcal{A}_o(s)\}$ and that $\operatorname{supp} P(\cdot \mid s, a) \subseteq \mathcal{S}_o$ for every $(s, a) \in \mathcal{O}$, since the successor of a visited pair is visited. Write $m_o \coloneqq \max_s |\mathcal{A}_o(s)| \le m$, let $M_o \in \mathbb{R}^{L \times |\mathcal{S}_o|}$ denote the columns of M indexed by $\mathcal{S}_o$ (similarly M_o^π), and define the *on-support error*

$$\epsilon_o \coloneqq \sup_{s \in \mathcal{S}_o,\ a \in \mathcal{A}_o(s)} \|Q(s, a, \cdot) - Q^\pi(s, a, \cdot)\|_1\,,$$

which places no constraint on Q outside $\mathcal{S}_o$.

Corollary 2. *Fix any finite MDP, policy π, and Q-values with on-support error ϵ_o.*

- *(a) If the MDP is deterministic and M_o^π is column-injective with column separation Δ_o, the estimator $\hat{P}(s, a) = \arg\min_{s' \in \mathcal{S}_o} \|Q(s, a) - M(s')\|_1$ satisfies $\hat{P}(s, a) = P(s, a)$ for every $(s, a) \in \mathcal{O}$ whenever $\epsilon_o < \Delta_o/(2(1 + \gamma m_o))$.*
- *(b) If M_o has full column rank (which requires $|\mathcal{G}| \ge |\mathcal{S}_o|$), the estimator $\hat{P}(s, a) = M_o^+ Q(s, a) \in \mathbb{R}^{|\mathcal{S}_o|}$ satisfies $\|\hat{P}(s, a) - P(s, a)\|_1 \le \|M_o^+\|_1 (1 + \gamma m_o)\, \epsilon_o$ for every $(s, a) \in \mathcal{O}$.*

Both conditions are weaker than those of Theorems 2 and 3: column-injectivity of M^π implies that of M_o^π with $\Delta_o \ge \Delta$, full column rank of M implies that of M_o, and the genericity statements of Theorem 2(a) and Theorem 3(a) apply to M_o^π verbatim.

Proof. Fix $(s,a) \in \mathcal{O}$. Since $\operatorname{supp} P(\cdot \mid s,a) \subseteq \mathcal{S}_o$, the Bellman equation (2) reads $M_o^\pi P(s,a) = Q^\pi(s,a)$ with $P(s,a)$ viewed as a vector on $\mathcal{S}_o$, and $\|Q(s,a) - Q^\pi(s,a)\|_1 \le \epsilon_o$ because $s \in \mathcal{S}_o$ and $a \in \mathcal{A}_o(s)$. For a column k with $s_k \in \mathcal{S}_o$, the entries $E_{lk} = M_{lk} - M^\pi_{lk} = \gamma \sum_{j'} \pi_{lkj'}(Q_{lkj'} - Q^\pi_{lkj'})$ only involve actions $a_{j'} \in \mathcal{A}_o(s_k)$, since $\pi_{lkj'} = 0$ otherwise, so

$$\|M_{\cdot,k} - M^\pi_{\cdot,k}\|_1 \le \gamma \sum_{j' : a_{j'} \in \mathcal{A}_o(s_k)} \|Q_{kj'} - Q^\pi_{kj'}\|_1 \le \gamma m_o \epsilon_o \,.$$

These two inequalities are the only places where the uniform bound ϵ was used in the proofs of Proposition 4 and Proposition 7. The remainder of each proof goes through unchanged with M, Δ, m, ϵ replaced by $M_o, \Delta_o, m_o, \epsilon_o$. The final claims follow since a subset of the columns of a column-injective (resp. linearly independent) family is column-injective (resp. linearly independent), the minimum defining Δ_o ranges over fewer pairs, and restricting R to the columns in $\mathcal{S}_o$ maps Lebesgue-null sets to Lebesgue-null sets. □

C.7 Arbitrary Q-Values

Theorems 2 and 3 give conditions under which the true kernel P is identified uniquely by exact Q-values, or approximately by ϵ-approximate Q-values. Propositions 10 and 11 below show that analogous results apply to *arbitrary* $Q \in \mathcal{Q}$. This is particularly relevant for deep RL agents, which seldom have converged Q-values, but may nonetheless encode a unique world model that we denote $P^\star$ – as opposed to P, since there is no guarantee the extracted world model will approximate the true kernel. While the proof techniques are very similar (constructing a map from Q to M, instead of R to M, whose preimages preserve measure-zero sets), neither follows from the other. First, recall that we write M for the matrix from Equation 2 induced by (Q, π, r, γ).

Proposition 10 (General Uniqueness, Stochastic)**.** *Fix a finite MDP, a policy π, and a reward function $r : \mathcal{S} \times \mathcal{G} \to \mathbb{R}$. Then the set $\{Q \in \mathcal{Q} \mid \operatorname{rank}(M) < \min(L,n)\}$ has Lebesgue measure zero. Moreover, if $|\mathcal{G}| \ge |\mathcal{S}|$, then for every Q outside this set, $P^\star(s,a) \coloneqq M^+ Q(s,a)$ is the unique minimiser of the Bellman residual $\|\mathcal{T}^\pi_{P^\star}(Q) - Q\|^2$, and if $|\mathcal{G}| = |\mathcal{S}|$, then $\mathcal{T}^\pi_{P^\star}(Q) = Q$.*

Proof. Recall the notation $L = |\mathcal{G}|, n = |\mathcal{S}|$ and define the map $F : \mathcal{Q} \to \mathbb{R}^{L \times n}$ by $F(Q) = R + \gamma V = R + \gamma \pi Q$, i.e. $V_{li} = \sum_j \pi_{lij} Q_{lij}$. Note that F is affine and surjective, since for any $M \in \mathbb{R}^{L \times n}$, defining $Q_{lij} = (M_{li} - R_{li})/\gamma$ gives $V_{li} = M_{li} - R_{li}$ and thus $F(Q) = M$. Then the set of $Q \in \mathcal{Q}$ for which M has deficient-rank is

$$\{Q \in \mathcal{Q} \mid \operatorname{rank}(M) < \min(L,n)\} = F^{-1}(\{M \in \mathbb{R}^{L \times n} \mid \operatorname{rank}(M) < \min(L,n)\})$$

which has Lebesgue measure zero as the preimage of a Lebesgue-measure-zero set under a surjective affine map: the rank-deficient set $\{M \in \mathbb{R}^{L \times n} \mid \operatorname{rank}(M) < \min(L,n)\}$ is the zero set of $\det(M^T M)$ when $L \ge n$, or of $\det(M M^T)$ when $L \le n$, which is a non-zero polynomial in the entries of M. Therefore, M has full-rank for almost every $Q \in \mathcal{Q}$. Now recall that the Bellman equation $\mathcal{T}^\pi_{P^\star}(Q) = Q$ is a linear system $M P^\star_{ij} = Q_{ij}$ for each i, j (see Equation 2). If $L \ge n$, the rank-nullity theorem gives $\dim(\ker(M)) = n - \dim(\operatorname{im}(M)) = 0$, so the system $M P^\star_{ij} = Q_{ij}$ has at most one solution, and the unique least-squares solution is given by $P^\star_{ij} = M^+ Q_{ij}$. If $L = n$, we have $P^\star_{ij} = M^{-1} Q_{ij}$ and thus $M P^\star_{ij} = Q_{ij}$ exactly, i.e. $\mathcal{T}^\pi_{P^\star}(Q) = Q$. □

Proposition 11 (General Uniqueness, Deterministic)**.** *Fix a finite MDP, a policy π, and a reward function $r : \mathcal{S} \times \mathcal{G} \to \mathbb{R}$. Then the set $\{Q \in \mathcal{Q} \mid \exists k \ne k' \text{ with } M_{\cdot,k} = M_{\cdot,k'}\}$ has Lebesgue measure zero. Moreover, for any $|\mathcal{G}| \ge 1$ and any Q outside this set, $P^\star(s,a) \coloneqq \arg\min_{s'} \|Q(s,a) - M_{\cdot,s'}\|_1$ is the unique deterministic world model minimising the Bellman residual $\|\mathcal{T}^\pi_{P^\star}(Q) - Q\|^2$, and at most one deterministic model $P^\star$ satisfies $\mathcal{T}^\pi_{P^\star}(Q) = Q$.*

Proof. Let $F : \mathcal{Q} \to \mathbb{R}^{L \times n}$ be the affine surjection from the proof of Proposition 10. For any $k \ne k'$, the set $\{M \in \mathbb{R}^{L \times n} : M_{\cdot,k} = M_{\cdot,k'}\}$ is a proper subspace of $\mathbb{R}^{L \times n}$, so it has Lebesgue measure zero. Taking a finite union over the $\binom{n}{2}$ pairs $k \ne k'$, the set

$$\{Q \in \mathcal{Q} \mid \exists k \ne k' \text{ with } M_{\cdot,k} = M_{\cdot,k'}\} = F^{-1}\Big(\bigcup_{k \ne k'} \{M \in \mathbb{R}^{L \times n} \mid M_{\cdot,k} = M_{\cdot,k'}\}\Big)$$

has Lebesgue measure zero since it is the preimage of a Lebesgue-measure-zero set under a surjective affine map. Therefore, M has pairwise distinct columns for almost every $Q \in \mathcal{Q}$. Now the Bellman

equation $\mathcal{T}^{\pi}_{P^\star}(Q) = Q$ reduces in deterministic MDPs to $Q_{ij} = M_{\cdot,P^\star(s_i,a_j)}$ for each (i,j). Since columns are pairwise distinct, the argmin $P^\star(s_i,a_j) = \arg\min_k \|Q_{ij} - M_{\cdot,k}\|_1$ is unique and minimises $\|Q_{ij} - M_{\cdot,P^\star(s_i,a_j)}\|^2 = \|\mathcal{T}^{\pi}_{P^\star}(Q) - Q\|^2$. □

We note that all results above refer to world models $P^\star \in \mathcal{P}$ which may lie outside the probability simplex; we refer to Appendix B.2 for discussion.

C.8 Impossibility Result for Multiple Discount Factors and Policies

Below we prove that, unlike multiple goals, multiple discount factors and multiple policies are not usually sufficient to break value equivalence.

Proposition 12. *For any $n \geq 4$, there is a family of finite MDPs with n states in which P cannot be uniquely identified from the set of Q-values $\{Q^{\pi}_{r,\gamma} \mid \pi \in \Pi,\ \gamma \in [0,1)\}$ over all policies π and all discount factors γ, for* any *reward function $r : \mathcal{S} \to \mathbb{R}$. The non-identifiable family has dimension at least 1 when $\{r_2, r_3, r_4\}$ are not all equal, and dimension at least 2 when $r_2 = r_3 = r_4$.*

Proof. Take $n = 4$, $m \geq 1$, and let s_2, s_3, s_4 be absorbing states: $P(s_k \mid s_k, a_j) = 1$ for $k \in \{2,3,4\}$ and every a_j. From s_1, fix any interior escape, e.g.

$$P(\cdot \mid s_1, a_1) = \left(p_1, \tfrac{1-p_1}{3}, \tfrac{1-p_1}{3}, \tfrac{1-p_1}{3}\right)$$

for some $p_1 \in (0,1)$, and $P(\cdot \mid s_1, a_j)$ for $j \geq 2$ arbitrary. For any reward function $r \in \mathbb{R}^4$, we construct a family of transition kernels $P_\epsilon \neq P$ such that Q-values are identical for all P_ϵ, for all π, γ. First, let

$$\mathcal{C}_r \coloneqq \{\Delta \in \mathbb{R}^4 : \Delta_1 = 0,\ \mathbf{1}^\top \Delta = 0,\ r^\top \Delta = 0\}\,.$$

If $\{r_2, r_3, r_4\}$ are not all equal, note that

$$\Delta_r \coloneqq (0,\ r_3 - r_4,\ r_4 - r_2,\ r_2 - r_3)$$

is a non-zero element of $\mathcal{C}_r$ (and in fact spans the one-dimensional vector space $\mathcal{C}_r$). If on the other hand $r_2 = r_3 = r_4$, the constraint $r^\top \Delta = 0$ reduces to $\mathbf{1}^\top \Delta = 0$. In particular the pair-swaps

$$\Delta^1_r \coloneqq (0,1,-1,0)\,, \qquad \Delta^2_r \coloneqq (0,0,1,-1)$$

are non-zero, linearly independent elements of $\mathcal{C}_r$ (and in fact span the two-dimensional vector space $\mathcal{C}_r$). In either case $\mathcal{C}_r \neq \{0\}$, so the family $\{P_\epsilon\}_{\epsilon\Delta \in \mathcal{C}_r}$ constructed below gives at least a one-parameter family of distinct stochastic kernels (two-parameter in the second case). Now pick any $0 \neq \Delta \in \mathcal{C}_r$. Since $P(\cdot \mid s_1, a_1)$ is interior, there exists $\epsilon_0 > 0$ such that for every $\epsilon \in (-\epsilon_0, \epsilon_0)$,

$$P_\epsilon(\cdot \mid s_1, a_1) \coloneqq P(\cdot \mid s_1, a_1) + \epsilon\Delta$$

is a valid transition function (non-negativity by interior, and $\mathbf{1}^\top \Delta = 0$ preserves the sum); define P_ϵ to agree with P everywhere else. Now fix any policy π and $\gamma \in [0,1)$. For $k \in \{2,3,4\}$, the Bellman equation at an absorbing state gives $V^\pi(s_k) = r_k/(1-\gamma)$, depending on r alone. For $k = 1$, writing $\alpha_l(\epsilon) \coloneqq \sum_j \pi(a_j \mid s_1) P_\epsilon(s_l \mid s_1, a_j)$, the Bellman equation reads

$$V^\pi(s_1) = \alpha_1(\epsilon)(r_1 + \gamma V^\pi(s_1)) + \sum_{l=2}^{4} \alpha_l(\epsilon) r_l/(1-\gamma)\,,$$

depending on the kernel only through $\alpha_1(\epsilon)$ and $\sum_{l=2}^4 \alpha_l(\epsilon) r_l$. Now $\Delta_1 = 0$ implies $\alpha_1(\epsilon) = \alpha_1(0)$ for all ϵ, and since $r^\top \Delta = 0$,

$$\sum_{l=2}^{4} (\alpha_l(\epsilon) - \alpha_l(0)) r_l = \epsilon \pi(a_1 \mid s_1) r^\top \Delta = 0\,,$$

so $\sum_{l=2}^4 \alpha_l(\epsilon) r_l$ is also ϵ-independent. Hence $V^\pi_{P_\epsilon}(s_1) = V^\pi_P(s_1)$. Now consider the Q-values

$$Q^\pi_{P_\epsilon}(s,a;\gamma) = \sum_{l=1}^{4} P_\epsilon(s_l \mid s,a)[r_l + \gamma V^\pi(s_l)]\,.$$

The only entry dependent on ϵ is (s_1, a_1), for which

$$Q^\pi_{P_\epsilon}(s_1,a_1) - Q^\pi_P(s_1,a_1) = \sum_{l=2}^{4} \Delta_l r_l/(1-\gamma) = \epsilon r^\top \Delta/(1-\gamma) = \epsilon r^\top \Delta/(1-\gamma) = 0\,,$$

hence $Q^\pi_{P_\epsilon} = Q^\pi_P$ for every $\epsilon \in (-\epsilon_0, \epsilon_0)$ and every (π, γ). □

C.9 Proof of Theorem 5 (Local MDPs)

As a middle-ground between deterministic and stochastic, we present a result referenced at the end of Section 4.1 for N-*local MDPs*, defined by $\operatorname{supp}_{s,a}(P) \coloneqq |\{s' : P(s' \mid s,a) > 0\}| \leq N$ for all (s,a). To restrict the usual hypothesis class $\mathcal{P} = \{P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})\}$, we introduce two classes corresponding to N-local MDPs with *known* or *unknown* support respectively

$$\mathcal{P}_\Gamma \coloneqq \{P \in \mathcal{P} : \operatorname{supp}_{s,a}(P) = \Gamma(s,a) \;\; \forall s,a\}\,,$$
$$\mathcal{P}_N \coloneqq \{P \in \mathcal{P} : |\operatorname{supp}_{s,a}(P)| \leq N \;\; \forall s,a\}\,,$$

where $\Gamma : \mathcal{S} \times \mathcal{A} \to 2^{\mathcal{S}}$ is a fixed support function with $|\Gamma(s,a)| \leq N$ for all (s,a). For example, a deterministic MDP is a 1-local MDP with typically unknown support (the successor state is not known), while an MDP with locally constrained dynamics may have known support if the agent has some prior knowledge about the environment. In the extreme case, a stochastic MDP is simply an n-local (i.e. global) MDP with known support (the entire state space).

Theorem 5 (Local)**.** *Fix any N-local finite MDP, policy π and support Γ.*

(a) Known support: For any $|\mathcal{G}| \geq N-1$, the set of reward functions $r : \mathcal{S} \times \mathcal{G} \to \mathbb{R}$ for which $P \in \mathcal{P}_\Gamma$ is uniquely determined by Q^π has full Lebesgue measure.

(b) Unknown support: For any $|\mathcal{G}| \geq 2N-1$, the set of reward functions $r : \mathcal{S} \times \mathcal{G} \to \mathbb{R}$ for which $P \in \mathcal{P}_N$ is uniquely determined by Q^π has full Lebesgue measure.

Proof. Write $L = |\mathcal{G}|$, and $\operatorname{supp}(i,j) \coloneqq \{k \in [n] : P_{ijk} > 0\}$ for the support of $P(\cdot \mid s_i, a_j)$.

Part (a). When $\operatorname{supp}(i,j) = \Gamma(s_i,a_j)$ is known, the Bellman equation (Equation 2) reduces to a linear system of size $\leq N$ per state-action pair. Taking $|\operatorname{supp}(i,j)| = N$ without loss of generality (the proof follows identically when the support is smaller), we have

$$M^\pi_{ij} \coloneqq M^\pi_{:,\operatorname{supp}(i,j)} \in \mathbb{R}^{L \times N}$$

and augmenting M^π_{ij} and Q_{ij} with ones to obtain $\mathbf{M}^\pi_{ij} \in \mathbb{R}^{(L+1)\times N}$ and $\mathbf{Q}_{ij} \in \mathbb{R}^{L+1}$, the transition function P satisfies

$$\mathbf{M}^\pi_{ij} P_{ij} = \mathbf{Q}_{ij} \tag{9}$$

for all i,j, where $P_{ij} \coloneqq P_{i,j,\operatorname{supp}(i,j)} \in \Delta^N$. In particular, $P \in \mathcal{P}_\Gamma$ is uniquely determined if $\operatorname{rank}(\mathbf{M}^\pi_{ij}) \geq N$. The proof now follows similarly to that of Proposition 3**(b)**. First recall the resolvent identity from Lemma 1:

$$M^\pi_l = R_l (I - \gamma \bar{P}^{\pi_l})^{-1}\,,$$

where $\bar{P}^\pi = (P^\pi)^T$. Since $(I - \gamma \bar{P}^{\pi_l})^{-1}$ is invertible for each l, the map $F : \mathbb{R}^{L\times n} \to \mathbb{R}^{L \times n}$ given by $F(R) = M^\pi$ is linear and invertible. Now $P \in \mathcal{P}_\Gamma$ is uniquely determined by Equation 9 if $\operatorname{rank}(\mathbf{M}^\pi_{ij}) \geq N$ for all i,j. Taking the converse statement,

$$\begin{aligned}
\{R \in \mathbb{R}^{L\times n} \mid \neg\exists! P(Q)\} &\subseteq \{R \in \mathbb{R}^{L\times n} \mid \exists i,j \text{ s.t. } \operatorname{rank}(\mathbf{M}^\pi_{ij}) < N\} \\
&= F^{-1}\left(\{M \in \mathbb{R}^{L\times n} \mid \exists i,j \text{ s.t. } \operatorname{rank}(\mathbf{M}_{ij}) < N\}\right) \\
&= F^{-1}\left(\bigcup_{i,j} \{M \in \mathbb{R}^{L\times n} \mid \operatorname{rank}(\mathbf{M}_{ij}) < N\}\right).
\end{aligned}$$

For generic $M \in \mathbb{R}^{L\times n}$, the augmented matrix $\mathbf{M} \in \mathbb{R}^{(L+1)\times n}$ is also generic, and so is any slice of columns $\mathbf{M}_{ij}$, with $\operatorname{rank}(\mathbf{M}_{ij}) = \min(L+1, N) = N$ almost surely, for all $L \geq N-1$. In particular,

$$\lambda\left(\{M \in \mathbb{R}^{L\times n} \mid \operatorname{rank}(\mathbf{M}_{ij}) < N\}\right) = 0$$

for each i,j, and under measure-zero preservation through finite unions and F^{-1},

$$\lambda\left(\{R \in \mathbb{R}^{L\times n} \mid \neg\exists! P(Q)\}\right) = 0\,.$$

Part (b). $P \in \mathcal{P}_N$ is uniquely determined if for each i,j, there are no $P_{ij} \neq P'_{ij} \in \Delta^n_N$ such that

$$\mathbf{M}^\pi P_{ij} = \mathbf{Q}_{ij} = \mathbf{M}^\pi P'_{ij}\,,$$

Subtracting one side from the other, note that $P_{ij} - P'_{ij} \in \Lambda^n_{2N} \coloneqq \{T \in \mathbb{R}^n \mid \|T\|_0 \leq 2N\}$, so $P \in \mathcal{P}_N$ is uniquely determined if there is no $0 \neq T_{ij} \in \Lambda^n_{2N}$ such that

$$\mathbf{M}^\pi T_{ij} = 0\,,$$

noting that the last-row constraint $\sum_k T_{ijk} = 0$ is implied by $\sum_k P_{ijk} - P'_{ijk} = 0$. In particular, writing $\mathbf{M}_{:,K}$ for the matrix sliced to the columns indexed by $K \subset [n]$ with $|K| = 2N$, this is equivalent to $\mathbf{M}_{:,K} T_{ij} = 0$ having no non-trivial solutions for $T_{ij} \in \mathbb{R}^{2N}$ for all such K, which is implied by $\operatorname{rank}(\mathbf{M}_{:,K}) \geq 2N$. Now the map $F : \mathbb{R}^{L\times n} \to \mathbb{R}^{L\times n}$ given by $F(R) = M^\pi$ is linear and invertible, and $P \in \mathcal{P}_N$ is uniquely determined if $\operatorname{rank}(\mathbf{M}^\pi_{:,K}) \geq 2N$ for all $K \subset [n]$ with $|K| = 2N$. Taking the converse statement,

$$\begin{aligned}
\{R \in \mathbb{R}^{L\times n} \mid \neg\exists! P(Q)\} &\subseteq \{R \in \mathbb{R}^{L\times n} \mid \exists |K| = 2N \text{ s.t. } \operatorname{rank}(\mathbf{M}^\pi_{:,K}) < 2N\} \\
&= F^{-1}\Big(\{M \in \mathbb{R}^{L\times n} \mid \exists |K| = 2N \text{ s.t. } \operatorname{rank}(\mathbf{M}_{:,K}) < 2N\}\Big) \\
&= F^{-1}\left(\bigcup_{|K|=2N} \{M \in \mathbb{R}^{L\times n} \mid \operatorname{rank}(\mathbf{M}_{:,K}) < 2N\}\right).
\end{aligned}$$

For generic $M \in \mathbb{R}^{L\times n}$, the augmented matrix $\mathbf{M} \in \mathbb{R}^{(L+1)\times n}$ is also generic, and so is any slice of columns $\mathbf{M}_{:,K}$, with $\operatorname{rank}(\mathbf{M}_{:,K}) = \min(L+1, 2N) = 2N$ almost surely, for all $L \geq 2N-1$. In particular,

$$\lambda\left(\{M \in \mathbb{R}^{L\times n} \mid \operatorname{rank}(\mathbf{M}_{:,K}) < 2N\}\right) = 0$$

for each K, and under measure-zero preservation through finite unions and F^{-1},

$$\lambda\left(\{R \in \mathbb{R}^{L\times n} \mid \neg\exists! P(Q)\}\right) = 0\,. \qquad \square$$

D Proofs for Continuous MDPs

This section contains all proofs for continuous MDPs; note that Lemmas 4 and 5 hold for arbitrary Borel state spaces $\mathcal{S}$, but in order to specify concrete reward families (indicator and Gaussian) for which $\exists! P(Q^\pi)$ holds, we specialise the goal-augmented MDP definition of Section 2 to Borel *continuous* state spaces $\mathcal{S} \subseteq \mathbb{R}^d$. Formally, the condition on P being a Markov kernel means that it satisfies two measurability conditions: (i) for each (s,a), $B \mapsto P(B \mid s,a)$ is a Borel probability measure on $\mathcal{S}$, and (ii) for each Borel set $B \in \Sigma_{\mathcal{S}}$, $(s,a) \mapsto P(B \mid s,a)$ is measurable, ensuring that Bellman integrals $\int_{\mathcal{S}} V(s')P(ds' \mid s,a)$ are well-defined. No absolute continuity of P with respect to Lebesgue measure is assumed unless stated explicitly.

D.1 Termination and the Bellman Family

In order to match the goal-dependent termination functions commonly used in GCRL (often written in code as `dones`), and thereby to formally state Theorems 4 and 8, we generalise the goal-augmented MDP formalism (Section 2) to include a *termination function* $D : \mathcal{S} \times \mathcal{G} \to \{0,1\}$ that specifies pairs (s,g) beyond which rewards are set to zero. We use a similar formalism as the state-based discount factors $\gamma : \mathcal{S} \to [0,1]$ used by Sutton et al. [32] and White [15], but with a goal-augmented domain. Writing $r_g(\cdot), D_g(\cdot) \coloneqq r(\cdot,g), D(\cdot,g)$, The return of a trajectory $\tau = (s^t, a^t)_{t\geq 0}$ with respect to g is now given by

$$R(\tau, g) \coloneqq \sum_{t\geq 0} \gamma^t r_g(s_{t+1}) \prod_{k=1}^{t} D_g(s_k)\,,$$

where the empty product ($t = 0$) is defined as 1, and which reduces to the usual return when $D \equiv 1$. Note that the multiplicative factor makes returns path-dependent, and could not be folded into the Markovian reward r. Terminating environments could instead be handled by adding aborbing states [15], but this imposes a modification on the state space, while transition-based termination does not, and cannot handle goal-dependence. With this in hand, the value and action-value functions are defined exactly as before: $V^\pi(s,g) \coloneqq \mathbb{E}_{\tau\sim\pi_g}\left[R(\tau,g) \mid s^0 = s\right]$ and $Q^\pi(s,a,g) \coloneqq \mathbb{E}_{\tau\sim\pi_g}\left[R(\tau,g) \mid s^0 = s, a^0 = a\right]$, with $\tau \sim \pi_g$ shorthand for independently

sampling $s_0 \sim \mu, a_t \sim \pi(s_t, g), s_{t+1} \sim P(s_t, a_t)$. In integral form, the Bellman equation becomes

$$Q_g^\pi(s,a) = \int_{\mathcal{S}} \Big(r_g(s') + \gamma D_g(s')\, V_g^\pi(s') \Big) P(\mathrm{d}s' \mid s, a)\,.$$

While the product $\prod_{k=1}^{t} D_g(s_k)$ induces path-dependent returns, the Bellman operators inherit the usual γ-contraction properties, and their unique fixed point admits an optimal *Markovian* policy [15].

Definition 2 (Bellman family)**.** For any fixed (Q^π, π, r, γ) with goal set $\mathcal{G}$, we define the *Bellman family* as the collection $\{m_g^\pi : g \in \mathcal{G}\}$ of bounded Lebesgue-measurable functions

$$m_g^\pi(s') \coloneqq r_g(s') + \gamma D_g(s') V_g^\pi(s')\,, \quad \text{with} \quad V_g^\pi(s') = \int_{\mathcal{A}} Q_g^\pi(s', a') \pi(\mathrm{d}a' \mid s', g)\,.$$

In particular, for any (s, a), the Bellman equation

$$\forall s, a, g. \quad Q^\pi(s, a, g) = \int m_g^\pi(s') P(\mathrm{d}s' \mid s, a)$$

requires that action-values be a collection of expectations of the *known* test functions m_g^π under the *unknown* measure $P(\cdot \mid s, a)$.

D.2 Determining Measures

The two reward function families we consider for continuous MDPs are the following, adjoined with the notion of *termination upon arrival*.

Definition 3. Given $\mathcal{S}, \mathcal{G} \subseteq \mathbb{R}^d$, we define the Gaussian and indicator reward functions $r : \mathcal{S} \times \mathcal{G} \to \mathbb{R}$ as $r(s, g) = \exp(-\|s - g\|^2/(2\sigma^2))$ and $r(s, g) = \mathbf{1}[\|s - g\| < \sigma]$ respectively, where $\sigma > 0$ is a fixed variance / threshold we omit from r for brevity. For indicator rewards, we say that we assume *termination upon arrival* if $D_g(s) = \mathbf{1}[\|s - g\| \geq \sigma] = 1 - r_g(s)$. For Gaussian goals, we say that we assume *ρ-termination* if $D_g(s) = \mathbf{1}[\|s - g\| \geq \rho]$, noting that the Gaussian variance σ is not necessarily equal to the termination threshold ρ.

To prove that P can be identified uniquely from (Q^π, π, r, γ), we need to show that the Bellman family $\{m_g^\pi : g \in \mathcal{G}\}$ is *measure determining*, defined as follows.

Definition 4 (Measure-determining family)**.** A set $\mathcal{F}$ of bounded measurable functions $f : \mathcal{S} \to \mathbb{R}$ is *measure determining* if for any probability measures $\mu, \nu \in \Delta(\mathcal{S})$,

$$\int f \mathrm{d}\mu = \int f \mathrm{d}\nu \quad \forall f \in \mathcal{F} \quad \Longrightarrow \quad \mu = \nu.$$

When policies are unconditional, Lemma 5 shows that Bellman families are measure-determining if the underlying reward family is signed-measure-determining, as defined below.

Definition 5 (Signed-measure-determining family)**.** A set $\mathcal{F}$ of bounded measurable functions $f : \mathcal{S} \to \mathbb{R}$ is *signed-measure-determining* if for any finite signed measures η and $\tilde{\eta}$ on $\mathcal{S}$,

$$\int f \mathrm{d}\eta = \int f \mathrm{d}\tilde{\eta} \quad \forall f \in \mathcal{F} \quad \Longrightarrow \quad \eta = \tilde{\eta}.$$

Definition 5 is stronger than Definition 4, since probability measures are finite signed measures.

D.3 Proofs for Unconditional Policies (Stochastic & Deterministic MDPs)

The results below apply to arbitrary stochastic continuous MDPs, which includes deterministic ones. However, they are restricted to unconditional policies $\pi_g = \pi_{g'}$ for all $g \in \mathcal{G}$. The next section will deal with general (goal-conditioned) policies, but only apply to deterministic MDPs. Results for the challenging intersection of stochastic kernels and goal-conditioned policies is beyond the scope of this paper, and left for future work.

D.3.1 Proof of Theorem 6 (Gaussian Goals)

Theorem 6. *Fix any continuous stochastic MDP, unconditional policy π and termination function D, $\sigma > 0$, and a goal set $\mathcal{G} \subseteq \mathbb{R}^d$ with non-empty interior and Gaussian rewards $r_g(s') = \exp(-\|s' - g\|^2/(2\sigma^2))$. Then the Bellman family $\{m_g^\pi\}_{g\in\mathcal{G}}$ is measure determining on $\mathcal{S}$, and thus $\exists! P(Q^\pi)$. The conclusion also holds when $\mathcal{G}$ is a countable set dense in some non-empty open subset of $\mathbb{R}^d$.*

Proof. We prove that $\{r_g : g \in \mathcal{G}\}$ is determining for finite signed measures on $\mathcal{S}$, which by Lemma 5 implies that $\{m_g^\pi\}$ is measure determining on $\mathcal{S}$. Let η be a finite signed measure on $\mathcal{S}$ with $\int_{\mathcal{S}} r_g \mathrm{d}\eta = 0$ for all $g \in \mathcal{G}$. Writing $\varphi_\sigma(x) \coloneqq \exp(-\|x\|^2/(2\sigma^2))$, note that $r_g(s') = \varphi_\sigma(g - s')$, so

$$F(g) \coloneqq \int_{\mathcal{S}} \varphi_\sigma(g - s')\, \eta(\mathrm{d}s') = (\eta * \varphi_\sigma)(g)$$

vanishes on $\mathcal{G}$ by hypothesis. By Lemma 6, F is real-analytic on $\mathbb{R}^d$. Since $\mathcal{G} \subseteq \mathbb{R}^d$ has non-empty interior in $\mathbb{R}^d$, F vanishes on an open ball $U \subseteq \mathcal{G}$, and the identity theorem for real-analytic functions on $\mathbb{R}^d$ [65] gives $F \equiv 0$ on $\mathbb{R}^d$. The Fourier transform argument derived in Lemma 7 implies $\eta = 0$. If $\mathcal{G}$ is instead countable and dense in some non-empty open $U \subseteq \mathbb{R}^d$, then F vanishes on a dense subset of U, and continuity (implied by real-analyticity) gives $F = 0$ on U, and the identity theorem applies as before. □

D.3.2 Proof of Theorem 7 (Indicator Goals)

For indicator rewards, the goal set is required to contain the open ball of radius σ around the state space. Recall that $B_\sigma(0) \coloneqq \{s \in \mathbb{R}^d \mid \|s\| < \sigma\}$, and $\mathcal{S} + B_\sigma(0) \coloneqq \{s + x \mid s \in S, x \in B_\sigma(0)\}$

Theorem 7. *Fix any continuous stochastic MDP, unconditional policy π and termination function D, $\sigma > 0$, and goal set $\mathcal{G} \supseteq \mathcal{S} + B_\sigma(0)$ with indicator rewards $r_g(s') = \mathbf{1}[\|s' - g\| < \sigma]$. Then the Bellman family $\{m_g^\pi\}_{g\in\mathcal{G}}$ is measure determining on $\mathcal{S}$, and thus $\exists! P(Q^\pi)$.*

Proof. We prove that $\{r_g : g \in \mathcal{G}\}$ is determining for finite signed measures on $\mathcal{S}$, which by Lemma 5 implies that $\{m_g^\pi\}$ is measure determining on $\mathcal{S}$. Let η be a finite signed measure on $\mathcal{S}$ with $\int_{\mathcal{S}} r_g \mathrm{d}\eta = 0$ for all $g \in \mathcal{G}$. Writing $\psi_\sigma(x) \coloneqq \mathbf{1}[\|x\| < \sigma]$, we have $r_g(s') = \psi_\sigma(g - s')$, so

$$F(g) \coloneqq \int_{\mathcal{S}} \psi_\sigma(g - s')\, \eta(\mathrm{d}s') = (\eta * \psi_\sigma)(g)$$

vanishes on $\mathcal{G}$ by hypothesis. For any $g \notin \mathcal{S} + B_\sigma(0)$, every $s' \in \mathcal{S}$ satisfies $\|g - s'\| \geq \sigma$, so $\psi_\sigma(g - s') = 0$ and thus $F(g) = 0$. Combined with $F = 0$ on $\mathcal{G} \supseteq \mathcal{S} + B_\sigma(0)$, we have $F \equiv 0$ on $\mathbb{R}^d$. The Fourier transform argument proven in Lemma 9 allows us to conclude $\eta = 0$. □

D.4 Proofs for Goal-Conditioned Policies (Deterministic MDPs)

We now consider deterministic MDPs, where $P(\cdot \mid s, a) = \delta_{f(s,a)}(\cdot)$ for a successor function $f : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ and the Dirac delta δ. In this setting, measure determination reduces to a simpler condition called *point separation*.

Definition 6 (Point-separating goals)**.** A family $\mathcal{F}$ of real-valued functions on $\mathcal{S}$ is *point separating* if for any $x \neq y$ in $\mathcal{S}$ there exists $f \in \mathcal{F}$ such that $f(x) \neq f(y)$.

In particular, if the MDP is deterministic and the family $\{m_g^\pi : g \in \mathcal{G}\}$ is point separating on $\mathcal{S}$, the successor function f is uniquely identified by (π, Q^π): any f, f' satisfying the Bellman equation

$$m_g^\pi\big(f(s,a)\big) = Q^\pi(s,a,g) = m_g^\pi\big(f'(s,a)\big) \qquad \forall g, s, a \tag{10}$$

must have $f(s,a) = f'(s,a)$ for every (s,a).

D.4.1 Proof of Theorem 4 (Deterministic MDPs, Indicator Goals with Termination)

Theorem 4. *Fix any continuous deterministic MDP, policy π, threshold $\sigma > 0$ and a goal set $\mathcal{G} \supseteq \mathcal{S} + B_\sigma(0)$ with indicator rewards $r_g(s') = \mathbf{1}[\|s' - g\| < \sigma]$ and termination upon arrival. Then $\{m_g^\pi : g \in \mathcal{G}\}$ is point separating on $\mathcal{S}$, hence $\exists! P(Q^\pi)$.*

Proof. **(i) Reward separation.** Fix any $x \neq y \in \mathcal{S}$. We first exhibit $g \in \mathcal{G}$ with $r_g(x) = 1$ and $r_g(y) = 0$. If $\|x - y\| \geq \sigma$, taking $g = x \in \mathcal{SG}$ gives $\|x - g\| = 0 < \sigma$, hence $r_g(x) = 1$, while $\|y - g\| = \|y - x\| \geq \sigma$ implies $r_g(y) = 0$. Otherwise, let $\delta := \frac{1}{2}\|x - y\| \in (0, \sigma/2)$ and define

$$g := x + (\sigma - \delta)\frac{x - y}{\|x - y\|}\,.$$

Then $\|g - x\| = \sigma - \delta < \sigma$, so $g \in B_\sigma(x) = x + B_\sigma(0) \subseteq \mathcal{S} + B_\sigma(0) \subseteq \mathcal{G}$ and $r_g(x) = 1$. Writing $u := (x - y)/\|x - y\|$, we have $g - y = (x - y) + (\sigma - \delta)u = (\|x - y\| + \sigma - \delta)u$, so

$$\|g - y\| = \|x - y\| + \sigma - \delta = \tfrac{1}{2}\|x - y\| + \sigma > \sigma\,,$$

hence $r_g(y) = 0$. In either case, the family $\{r_g : g \in \mathcal{G}\}$ separates x from y.

(ii) Bellman family expansion. Fix g as constructed in step (i). By Lemma 4,

$$m_g^\pi = \sum_{k \geq 0} \gamma^k T_g^k r_g\,, \quad \text{where} \quad (T_g h)(s) = D_g(s)\ \mathbb{E}_{s' \sim P^{\pi_g}(\cdot|s)}[h(s')]\,,$$

is the per-goal Bellman operator. We claim that for all $k \geq 0$,

$$(T_g^k r_g)(s) = \mathbb{E}_\tau \left[\prod_{t<k} D_g(s_t)\, r_g(s_k) \,\middle|\, s_0 = s \right], \tag{11}$$

where $\tau = (s_t)_{t \geq 0}$ is the policy-induced trajectory with $a_t \sim \pi_g(\cdot \mid s_t)$ and $s_{t+1} = f(s_t, a_t)$. The base case $k = 0$ holds by definition of the empty product equalling 1. For the inductive step, we use the definition of T_g and the inductive hypothesis applied to s_1 to get

$$\begin{aligned}
(T_g^{k+1} r_g)(s) &= D_g(s)\, \mathbb{E}\left[(T_g^k r_g)(s_1) \,\middle|\, s_0 = s \right] \\
&= D_g(s)\, \mathbb{E}\left[\mathbb{E}_{\tau'} \left[\prod_{t<k} D_g(s'_t)\, r_g(s'_k) \,\middle|\, s'_0 = s_1 \right] \middle|\, s_0 = s \right] \\
&= \mathbb{E}\left[D_g(s_0) \prod_{t<k} D_g(s_{t+1})\, r_g(s_{k+1}) \,\middle|\, s_0 = s \right] \\
&= \mathbb{E}\left[\prod_{t<k+1} D_g(s_t)\, r_g(s_{k+1}) \,\middle|\, s_0 = s \right],
\end{aligned}$$

where the third equality uses the Markov property (the inner trajectory τ' started at s_1 has the same law as $(s_{t+1})_{t \geq 0}$ under the original chain) to merge the nested expectations.

(iii) Bellman separation. Since $r_g(x) = 1$, the assumption of termination upon arrival (Definition 3) gives $D_g(x) = 1 - r_g(x) = 0$, so the product in Equation 11 contains the factor $D_g(s_0) = 0$ for every $k \geq 1$, conditional on $s_0 = x$. Therefore $(T_g^k r_g)(x) = 0$ for all $k \geq 1$, and only the $k = 0$ term survives:

$$m_g^\pi(x) = r_g(x) = 1\,.$$

For y, we have $r_g(y) = 0$ and hence $D_g(y) = 1$. Conditioning on $s_0 = y$ throughout, we define the hitting time

$$T := \inf\{t \geq 1 : s_t \in B_\sigma(g)\} \in \{1, 2, \dots\} \cup \{\infty\}\,.$$

Note first that the $k = 0$ term contributes $(T_g^0 r_g)(y) = r_g(y) = 0$. For $k \geq 1$, we evaluate $(T_g^k r_g)(y)$ by splitting on whether $T < k$, $T = k$, or $T > k$, recalling that $s_0 = y$ satisfies $\|y - g\| > \sigma$, so $s_0 \notin B_\sigma(g)$ and $D_g(s_0) = 1$. On $\{T < k\}$, the index $T \in \{1, \dots, k-1\}$

falls inside the product $\prod_{t<k} D_g(s_t)$, and $s_T \in B_\sigma(g)$ gives $D_g(s_T) = 1 - r_g(s_T) = 0$, killing the product. On $\{T = k\}$, for every $1 \le t < k$ we have $s_t \notin B_\sigma(g)$ by definition of T, so $D_g(s_t) = 1$; combined with $D_g(s_0) = 1$, the entire product equals 1, while $s_k \in B_\sigma(g)$ gives $r_g(s_k) = 1$, so the integrand equals 1. On $\{T > k\}$, $s_k \notin B_\sigma(g)$, so $r_g(s_k) = 0$ kills the integrand. Combining, the $\{0,1\}$-valued integrand $\prod_{t<k} D_g(s_t)\, r_g(s_k)$ is precisely the indicator of $\{T = k\}$, so $(T_g^k r_g)(y) = \mathbb{E}_\tau[\mathbf{1}\{T = k\} \mid s_0 = y]$ for $k \ge 1$. Swapping sum and expectation since all terms are nonnegative (Fubini [66]), and using $\sum_{k\ge1} \gamma^k \mathbf{1}\{T = k\} = \gamma^T$ (with the convention $\gamma^\infty = 0$),

$$m_g^\pi(y) = \sum_{k\ge1} \gamma^k\, \mathbb{E}_\tau[\mathbf{1}\{T = k\} \mid s_0 = y] = \mathbb{E}_\tau\left[\gamma^T \,\middle|\, s_0 = y\right] \le \gamma < 1 = m_g^\pi(x)\,,$$

where the final bound uses $T \ge 1$. We conclude that $\{m_g^\pi : g \in \mathcal{G}\}$ separates x from y. Since (x, y) was arbitrary, the Bellman family is point separating on $\mathcal{S}$, and (10) implies that the P is uniquely identified by (π, Q^π, r, γ). □

D.4.2 Proof of Theorem 8 (Deterministic MDPs, Gaussian Goals with Termination)

We now turn to Gaussian goals $r_g(s) = \exp(-\|s - g\|^2/(2\sigma^2))$ paired with *ρ-termination*, i.e. a termination function $D_g(s) = \mathbf{1}[\|s - g\| \ge \rho]$ (see Definition 3). Unlike indicator rewards, $D_g \ne 1 - r_g$, with the termination threshold ρ being independent from the Gaussian variance σ.

Theorem 8. *Fix any continuous deterministic MDP with compact $\mathcal{S} \subseteq \mathbb{R}^d$, policy π and $\sigma, \rho > 0$ satisfying the smallness condition*

$$\sigma^2 < \frac{\rho^2}{-4\log(1-\gamma)}\,.$$

Then there is a finite goal set $\mathcal{G} \subseteq \mathcal{S} + B_\rho(0)$ with Gaussian rewards $r_g(s) = \exp(-\|s - g\|^2/(2\sigma^2))$ and ρ-termination $D_g(s) = \mathbf{1}[\|s - g\| \ge \rho]$ such that $\{m_g^\pi : g \in \mathcal{G}\}$ is point separating on $\mathcal{S}$, hence $\exists! P(Q^\pi)$. In particular, $\exists! P(Q^\pi)$ for any $\mathcal{G} \supseteq \mathcal{S} + B_\rho(0)$.

Proof. **(i) Constructing $\mathcal{G}$.** First note that the smallness condition is equivalent to

$$\exp(-\rho^2/(2\sigma^2)) < (1-\gamma)^2\,, \quad \text{or equivalently,} \quad \sigma^2 < \frac{\rho^2}{-4\log(1-\gamma)}\,. \tag{12}$$

Now pick $\eta \in (0, \rho/2)$ small enough that

$$\exp(-2\eta^2/\sigma^2) > \frac{\exp(-\rho^2/(2\sigma^2))}{1-\gamma} + \gamma\,, \tag{13}$$

which exists because the LHS tends to 1 as $\eta \to 0$ while the RHS is smaller than 1 by Equation 12. By compactness, we can choose locii $c_1, \dots, c_N \in \mathcal{S}$ such that $\mathcal{S} \subseteq \bigcup_{i=1}^N B_\eta(c_i)$. Let $e_1, \dots, e_d$ denote the standard basis of $\mathbb{R}^d$ and define, for each $i \in [N]$ and $j \in [d]$, the finite set of goals $\mathcal{G} := \{g_{ij} \mid 1 \le i \le N,\ 0 \le j \le d\}$ with

$$g_{i0} := c_i\,, \quad \text{and} \quad g_{ij} := c_i + \eta e_j \quad \text{for } j \in [d]\,.$$

Since $c_i \in \mathcal{S}$ and $\|g_{ij} - c_i\| = \|\eta e_j\| = \eta < \rho$, we have $g_{ij} \in c_i + B_\rho(0)$ and thus $\mathcal{G} \subseteq \mathcal{S} + B_\rho(0)$.

(ii) Bellman expansion. Fix any $g \in \mathcal{G}$. By Lemma 4, we have $m_g^\pi = \sum_{k\ge0} \gamma^k T_g^k r_g$ with $(T_g h)(s) = D_g(s)\, \mathbb{E}_{s' \sim P^{\pi_g}(\cdot|s)}[h(s')]$. The trajectory formula (Equation 11) in the proof of Theorem 4 holds for any bounded measurable reward, so for all $k \ge 0$,

$$(T_g^k r_g)(s) = \mathbb{E}_\tau\left[\prod_{t<k} D_g(s_t) r_g(s_k) \,\middle|\, s_0 = s\right]\,.$$

(iii) Bellman separation. Fix distinct $x, y \in \mathcal{S}$ and pick i with $x \in B_\eta(c_i)$. For each $j \in \{0, \dots, d\}$,

$$\|x - g_{ij}\| \le \|x - c_i\| + \|c_i - g_{ij}\| < \eta + \eta = 2\eta < \rho\,,$$

by assumption that $\eta \in (0, \rho/2)$. Hence $D_{g_{ij}}(x) = 0$ and the product in the trajectory formula above contains the factor $D_g(s_0) = 0$ for every $k \geq 1$ when $s_0 = x$. Only the $k = 0$ term survives:

$$m^\pi_{g_{ij}}(x) = r_{g_{ij}}(x) = \exp(-\|x - g_{ij}\|^2/(2\sigma^2)) > \exp(-2\eta^2/\sigma^2)\,. \tag{14}$$

For y, we distinguish two cases.

Case A: some j satisfies $\|y - g_{ij}\| \geq \rho$. Fix any such j and write $g := g_{ij}$. Conditioning on $s_0 = y$, define the hitting time

$$T := \inf\{t \geq 1 : s_t \in B_\rho(g)\} \in \{1, 2, \dots\} \cup \{\infty\}\,,$$

which automatically satisfies $T \geq 1$ since $s_0 = y \notin B_\rho(g)$. We bound the pathwise integrand $\sum_{k\geq 0} \gamma^k \prod_{t<k} D_g(s_t)\, r_g(s_k)$ by case analysis on k relative to T. For $0 \leq k < T$, every s_t with $0 \leq t < k$ satisfies $s_t \notin B_\rho(g)$ (true at $t = 0$ since $\|y - g\| \geq \rho$, and at $1 \leq t < k < T$ by definition of T), so $D_g(s_t) = 1$. On the other hand, $\|s_k - g\| \geq \rho$ gives $r_g(s_k) \leq \exp(-\rho^2/(2\sigma^2))$. For $k = T$ (only when $T < \infty$), the product $\prod_{t<T} D_g(s_t) = 1$ as above and $r_g(s_T) \leq 1$, contributing $\gamma^T \leq \gamma$ since $T \geq 1$. For $k > T$, the product contains the factor $D_g(s_T) = 0$. Summing,

$$\sum_{k\geq 0} \gamma^k \prod_{t<k} D_g(s_t)\, r_g(s_k) \leq \sum_{k=0}^{T-1} \gamma^k \exp(-\rho^2/(2\sigma^2)) + \mathbf{1}\{T < \infty\}\gamma^T \leq \frac{\exp(-\rho^2/(2\sigma^2))}{1-\gamma} + \gamma\,.$$

Taking expectations on both sides (Fubini–Tonelli [66], since all terms are nonnegative) and applying the trajectory formula above,

$$m^\pi_g(y) = \sum_{k\geq 0} \gamma^k (T^k_g r_g)(y) \leq \frac{\exp(-\rho^2/(2\sigma^2))}{1-\gamma} + \gamma < \exp(-2\eta^2/\sigma^2) < m^\pi_g(x)\,,$$

where the strict inequalities use Equations 13 and 14 respectively. Hence $m^\pi_g(y) \neq m^\pi_g(x)$.

Case B: $\|y - g_{ij}\| < \rho$ for every $j = 0, \dots, d$. Then $D_{g_{ij}}(y) = 0$ for all j and the same $k = 0$ collapse gives $m^\pi_{g_{ij}}(y) = r_{g_{ij}}(y)$ for all j. Now assume for contradiction that $m^\pi_{g_{ij}}(x) = m^\pi_{g_{ij}}(y)$ for every j; then substituting $m^\pi_{g_{ij}}(x) = r_{g_{ij}}(x)$ by Equation 14, we have $\|x - g_{ij}\|^2 = \|y - g_{ij}\|^2$ for every j. Expanding squared norms and subtracting the $j = 0$ identity from the j-th gives $2\langle g_{ij} - g_{i0}, x - y\rangle = 0$, which reduces to $\eta\langle e_j, x - y\rangle = 0$ for $j = 1, \dots, d$, forcing $x = y$, a contradiction. Hence $m^\pi_{g_{ij}}(x) \neq m^\pi_{g_{ij}}(y)$ for some goal g_{ij}.

Since (x, y) was arbitrary, $\{m^\pi_g : g \in \mathcal{G}\}$ is point separating on $\mathcal{S}$, and (10) implies that P is uniquely identified by (π, Q^π, r, γ). The last claim is immediate: if a goal set $\mathcal{G}'$ satisfies $\mathcal{G}' \supseteq \mathcal{S} + B_\rho(0)$ then $\mathcal{G}' \supseteq \mathcal{G}$, hence $\{m^\pi_g : g \in \mathcal{G}'\} \supseteq \{m^\pi_g : g \in \mathcal{G}\}$, which inherits point separation. □

Remark 1 (Finite, but large!)**.** For $\mathcal{S} \subseteq [-R, R]^d$, a grid with linear spacing $\eta/\sqrt{d}$ along each axis covers $\mathcal{S}$ by η-balls: each grid cell is a hypercube of side $\eta/\sqrt{d}$ and diagonal η, hence fits in an η-ball. This gives an upper bound

$$N \leq \lceil 2R\sqrt{d}/\eta \rceil^d \quad \Longrightarrow \quad |\mathcal{G}| \leq (d+1)\lceil 2R\sqrt{d}/\eta \rceil^d\,,$$

which is $O(d \exp(d \log(d)))$ in the dimension. The construction, albeit finite, is therefore impractical, but serves to illustrate the stronger mathematical guarantees for Gaussian goals (finite goals) as opposed to indicator goals (infinite goals) in this setting.

D.4.3 Proof of Theorem 9 (Deterministic MDPs, Gaussian Goals with Analytic Regularity)

The construction of Theorem 8 relies on the smallness condition $\sigma^2 < \rho^2/(-4\log(1-\gamma))$ and a goal-ball termination function, producing a finite but exponentially-sized goal set (see Remark 1). Trading termination for real-analytic regularity on D and V^π yields a sharper guarantee: $L \geq 2d + 1$ Gaussian goals in *generic position* suffice for point separation, with no constraint on σ.

Theorem 9. *Fix any continuous deterministic MDP with open $\mathcal{S} \subseteq \mathbb{R}^d$, unconditional policy π, $\sigma > 0$, Gaussian rewards $r_g(s) = \exp(-\|s-g\|^2/(2\sigma^2))$, and assume that V^π and D are real-analytic on $\mathcal{S} \times \mathbb{R}^d$ and $\mathcal{S}$ respectively. Then for every $L \geq 2d+1$, the set of goal tuples $\mathbf{g} = (g_1, \ldots, g_L) \in \mathcal{S}^L$ for which $\{m^\pi_{g_l}\}_{l=1}^L$ is point separating on $\mathcal{S}$ has full Lebesgue measure in $\mathcal{S}^L$. In particular, for Lebesgue-almost every L-element goal set $\mathcal{G} = \{g_1, \ldots, g_L\} \subseteq \mathcal{S}$, we have $\exists! P(Q^\pi)$ for all $L \geq 2d+1$.*

Proof. **(i) Resolvent identity.** Following the proof of Lemma 5, notice that the Markov operator $(Th)(s) \coloneqq D(s) \int_{\mathcal{S}} h(s') P^\pi(\mathrm{d}s' \mid s)$ is independent from g, and its adjoint T^* on finite signed measures satisfies $\|T^*\|_{\mathrm{TV}} \leq 1$ and $\int_{\mathcal{S}} Th \mathrm{d}\eta = \int_{\mathcal{S}} h \mathrm{d}(T^*\eta)$. For each $x \in \mathcal{S}$, define

$$\mu_x \coloneqq (I - \gamma T^*)^{-1} \delta_x = \sum_{t \geq 0} \gamma^t (T^*)^t \delta_x \,, \tag{15}$$

a finite non-negative Borel measure on $\mathcal{S}$. By construction, μ_x satisfies the resolvent identity

$$\mu_x - \gamma T^* \mu_x = \delta_x \,. \tag{16}$$

Using $m^\pi_g = (I - \gamma T)^{-1} r_g$ from Lemma 4 and following the proof of Lemma 5 gives

$$\begin{aligned} m^\pi_g(x) &= \int_{\mathcal{S}} m^\pi_g \mathrm{d}\delta_x = \sum_{t \geq 0} \gamma^t \int_{\mathcal{S}} (T^t r_g) \mathrm{d}\delta_x \\ &= \sum_{t \geq 0} \gamma^t \int_{\mathcal{S}} r_g \mathrm{d}((T^*)^t \delta_x) \\ &= \int_{\mathcal{S}} r_g \mathrm{d}\Big(\sum_{t \geq 0} \gamma^t (T^*)^t \delta_x \Big) = \int_{\mathcal{S}} r_g \mathrm{d}\mu_x = (\mu_x * \varphi_\sigma)(g) \,, \end{aligned}$$

where $\varphi_\sigma(u) \coloneqq \exp(-\|u\|^2/(2\sigma^2))$ and $r_g(s') = \varphi_\sigma(g - s')$.

(ii) Non-vanishing slices. For $x, y \in \mathcal{S}$, we define the slice function $h_{x,y} : \mathbb{R}^d \to \mathbb{R}$ by

$$h_{x,y}(g) \coloneqq m^\pi_g(x) - m^\pi_g(y) = ((\mu_x - \mu_y) * \varphi_\sigma)(g) \,,$$

which is real-analytic on $\mathbb{R}^d$ by Lemma 6 applied to the finite signed measure $\mu_x - \mu_y$. We prove that $h_{x,y} \not\equiv 0$ for all $x \neq y$ via the contrapositive: if $h_{x,y} \equiv 0$, then $(\mu_x - \mu_y) * \varphi_\sigma \equiv 0$, so $\mu_x = \mu_y$ by Lemma 7, and applying Equation 16 yields $\delta_x = \delta_y$, which implies $x = y$. Writing $\Lambda \coloneqq \{(x,x) : x \in \mathcal{S}\} \subseteq \mathcal{S}^2$ for the diagonal, we conclude that $h_{x,y} \not\equiv 0$ for all $(x,y) \in \mathcal{S}^2 \setminus \Lambda$.

(iii) The bad set. In order to characterise the set of goal tuples that fail to induce point-separation, we define $\mathcal{U} \coloneqq (\mathcal{S}^2 \setminus \Lambda) \times (\mathbb{R}^d)^L \subseteq \mathbb{R}^{2d+Ld}$ and the map $H : \mathcal{U} \to \mathbb{R}^L$ by

$$H(x, y, \mathbf{g}) \coloneqq \big(h_{x,y}(g_1), \ldots, h_{x,y}(g_L)\big) \,,$$

where $\mathbf{g} = (g_1, \ldots, g_L)$ is a tuple of L goals, and let

$$E \coloneqq H^{-1}(0) = \{(x, y, \mathbf{g}) \in \mathcal{U} : h_{x,y}(g_l) = 0 \; \forall \, l \in [L]\} \,.$$

Now notice that the Bellman family $\{m^\pi_{g_l}\}_{l=1}^L$ fails to be point-separating on $\mathcal{S}$ if and only if there exist $x \neq y \in \mathcal{S}$ with $h_{x,y}(g_l) = 0$ for every l. Defining $\pi_2 : \mathcal{U} \to (\mathbb{R}^d)^L$ as the coordinate projection $(x, y, \mathbf{g}) \mapsto \mathbf{g}$, this leads us to the "bad set"

$$B \coloneqq \pi_2(E) = \{\mathbf{g} \in (\mathbb{R}^d)^L : \exists \, x \neq y \text{ such that } h_{x,y}(g_l) = 0 \; \forall \, l \in [L]\}$$

of points $\mathbf{g} \in B$ that induce a "bad" Bellman family. Our goal is therefore to prove that B has Lebesgue measure zero in $\mathbb{R}^{Ld}$. To do so, we will partition E into a countable union based on the *vanishing order* of partial derivatives.

(iv) Partitioning E. For a multi-index $\beta \in \mathbb{N}^d$, we write ∂^β_g to denote the corresponding partial derivative in the $\mathbb{R}^d$-variable g, as distinct from the bold tuple $\mathbf{g}$, with order $|\beta| \coloneqq \sum_{i=1}^d \beta^i$. For $(x, y, \mathbf{g}) \in E$ and each $l \in [L]$, we define

$$\Omega_l(x, y, \mathbf{g}) \coloneqq \inf\{|\beta| \geq 0 : \partial^\beta_g h_{x,y}(g_l) \neq 0\}$$

as the corresponding *vanishing order*. Now real-analyticity of D, V^π and r implies that $m_g^\pi(x) = r_g(x) + \gamma D(x) V^\pi(x,g)$ is jointly real-analytic in g and x, and in particular, $(x,y,g) \mapsto h_{x,y}(g)$ is real-analytic on the open set $(\mathcal{S}^2 \setminus \Lambda) \times \mathbb{R}^d$. Since $h_{x,y} \not\equiv 0$ for $x \neq y$ by part **(ii)**, the identity theorem [65] implies that the Taylor series of $h_{x,y}$ at g_l has at least one non-zero coefficient, so $\Omega_l(x,y,\mathbf{g}) < \infty$ for all $(x,y,\mathbf{g}) \in E$. Moreover, $\Omega_l(x,y,\mathbf{g}) \geq 1$ since $h_{x,y}(g_l) = 0$ by assumption that $(x,y,\mathbf{g}) \in E$. We thus have $\Omega_l \in \mathbb{N}$ (positive natural numbers), and can partition E as a countable (disjoint) union

$$E = \bigcup_{\mathbf{N} \in \mathbb{N}^L} E_{\mathbf{N}}\,, \qquad E_{\mathbf{N}} \coloneqq \{(x,y,\mathbf{g}) \in E : \Omega_l(x,y,\mathbf{g}) = N_l \;\forall\, l \in [L]\}\,.$$

(v) Local dimension. Fix $\mathbf{N}$ and $p = (\bar{x}, \bar{y}, \bar{\mathbf{g}}) \in E_{\mathbf{N}}$. By definition, for each $l \in [L]$, there is a multi-index $\beta_l \in \mathbb{N}^d$ with $|\beta_l| = N_l$ and $\partial_g^{\beta_l} h_{\bar{x},\bar{y}}(\bar{g}_l) \neq 0$. Since $|\beta_l| \geq 1$, pick a coordinate $c_l \in [d]$ with $\beta_l^{c_l} \geq 1$, and let $\alpha_l \in \mathbb{N}^d$ be β_l with its c_l-th entry decremented by 1, so $|\alpha_l| = N_l - 1$ and

$$\partial_g^{\beta_l} = \partial_{g^{c_l}} \partial_g^{\alpha_l}\,, \tag{17}$$

where g^{c_l} denotes the c_l-th scalar coordinate of $g \in \mathbb{R}^d$. We now define the map

$$S_p : \mathcal{U} \to \mathbb{R}^L\,, \qquad S_p\big(x,y,\mathbf{g}\big)_l \coloneqq \partial_g^{\alpha_l} h_{x,y}(g_l)\,,$$

and its zero set

$$Z_p \coloneqq S_p^{-1}(0) = \{(x,y,\mathbf{g}) \in \mathcal{U} : S_p(x,y,\mathbf{g}) = 0\}\,.$$

The motivation for this construction is that $E_{\mathbf{N}} \subseteq Z_p$: any $(x,y,\mathbf{g}) \in E_{\mathbf{N}}$ satisfies $\Omega_l(x,y,\mathbf{g}) = N_l$ for every l, so for $|\alpha_l| = N_l - 1 < N_l$, we have

$$S_p(x,y,\mathbf{g})_l = \partial_g^{\alpha_l} h_{x,y}(g_l) = 0 \quad \forall\, l \in [L]\,,$$

hence $(x,y,\mathbf{g}) \in Z_p$. We now leverage this inclusion to show that Z_p has the local structure of a real-analytic submanifold of $\mathcal{U}$ in a neighbourhood of p. First, the map S_p is real-analytic on $\mathcal{U}$ since partial derivatives of real-analytic functions are real-analytic. To invoke the implicit function theorem, we split the coordinates of $\mathcal{U} \subseteq \mathbb{R}^{2d+Ld}$ into $w \coloneqq (g_1^{c_1}, \ldots, g_L^{c_L}) \in \mathbb{R}^L$, comprising the L scalar coordinates of $\mathbf{g}$ singled out by the indices $c_1, \ldots, c_L$ chosen above, and the remaining $2d + L(d-1)$ coordinates v. This identifies $\mathcal{U}$ with an open subset of $\mathbb{R}^{2d+L(d-1)} \times \mathbb{R}^L$ via the bijection $(x,y,\mathbf{g}) \leftrightarrow (v,w)$, and we write $p = (v_p, w_p)$ accordingly. The key is now to notice that the partial Jacobian $\mathrm{D}_w S_p(p) \in \mathbb{R}^{L \times L}$ is diagonal:

$$\mathrm{D}_w S_p(p) = \begin{pmatrix} \partial_g^{\beta_1} h_{\bar{x},\bar{y}}(\bar{g}_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \partial_g^{\beta_L} h_{\bar{x},\bar{y}}(\bar{g}_L) \end{pmatrix}\,,$$

where the diagonal entries $\partial_{g^{c_l}} \partial_g^{\alpha_l} h_{\bar{x},\bar{y}}(\bar{g}_l) = \partial_g^{\beta_l} h_{\bar{x},\bar{y}}(\bar{g}_l)$ use Equation 17, and are non-zero by definition of β_l. In particular, $\mathrm{D}_w S_p(p)$ is invertible. Combined with $S_p(p) = 0$, the real-analytic implicit function theorem [65] provides open neighbourhoods $V_p \subseteq \mathbb{R}^{2d+L(d-1)}$ of v_p and $W_p \subseteq \mathbb{R}^L$ of w_p, and a real-analytic function $G_p : V_p \to W_p$ such that $S_p(v, G_p(v)) = 0$ for all $v \in V_p$, and moreover these are the only zeros of S_p in $V_p \times W_p$:

$$Z_p \cap (V_p \times W_p) = \{(v, G_p(v)) : v \in V_p\}\,.$$

In other words, setting $U_p \coloneqq V_p \times W_p$ and $\psi_p(v) \coloneqq (v, G_p(v))$ for brevity, the zero set Z_p is the image of the real-analytic map ψ_p in a neighbourhood of p, namely,

$$Z_p \cap U_p = \psi_p(V_p)\,.$$

Returning to our set $E_{\mathbf{N}} \subseteq Z_p$, we obtain the local inclusion $E_{\mathbf{N}} \cap U_p \subseteq \psi_p(V_p)$.

(vi) Measure-zero conclusion. The open set $\mathcal{U} \subseteq \mathbb{R}^{2d+Ld}$ is second-countable, so the open cover $E_{\mathbf{N}} \subseteq \bigcup_{p \in E_{\mathbf{N}}} U_p$ admits a countable subcover $E_{\mathbf{N}} \subseteq \bigcup_{n \in \mathbb{N}} U_{p_n}$, and the local inclusion $E_{\mathbf{N}} \cap U_p \subseteq \psi_p(V_p)$ from part **(v)** gives

$$E_{\mathbf{N}} = E_{\mathbf{N}} \cap \bigcup_{n \in \mathbb{N}} U_{p_n} = \bigcup_{n \in \mathbb{N}} E_{\mathbf{N}} \cap U_{p_n} \subseteq \bigcup_n \psi_{p_n}(V_{p_n})\,.$$

At last, we invoke the assumption $L \geq 2d+1$ to get

$$2d + L(d-1) = 2d + Ld - L \leq 2d + Ld - (2d+1) < Ld\,,$$

so the parameterising domain $V_{p_n} \subseteq \mathbb{R}^k$ with $k \coloneqq 2d + L(d-1)$ has dimension strictly less than Ld. In particular, since projection is also real-analytic, $\pi_2(\psi_{p_n}(V_{p_n}))$ is the image of the real-analytic map $\pi_2 \circ \psi_{p_n} : V_{p_n} \to \mathbb{R}^{Ld}$, from an open subset of $\mathbb{R}^k$ with $k < Ld$. By Sard's theorem [40], or a direct proof covering V_{p_n} with countably many cubes, it immediately follows that $\lambda_{Ld}(\pi_2(\psi_{p_n}(V_{p_n}))) = 0$ for every n. Moreover, since π_2 is monotonic and preserves unions, $\pi_2(E_{\mathbf{N}}) \subseteq \bigcup_n \pi_2(\psi_{p_n}(V_{p_n}))$. We sidestep a direct proof that $\pi_2(E_{\mathbf{N}})$ is Lebesgue measurable by working with the *Lebesgue outer measure* λ^*_{Ld}, which is defined everywhere, agrees with the Lebesgue measure on measurable sets, and satisfies countable subadditivity [66]. Using this property, have

$$\lambda^*_{Ld}(\pi_2(E_{\mathbf{N}})) \leq \sum_{n\in\mathbb{N}} \lambda^*_{Ld}(\pi_2(\psi_{p_n}(V_{p_n}))) = 0\,.$$

Using union preservation once more, the bad set is given by $B = \pi_2(E) = \bigcup_{\mathbf{N}\in\mathbb{N}^L} \pi_2(E_{\mathbf{N}})$, hence

$$\lambda^*_{Ld}(B) \leq \sum_{\mathbf{N}\in\mathbb{N}^L} \lambda^*_{Ld}(\pi_2(E_{\mathbf{N}})) = 0\,.$$

Completeness of Lebesgue measure [66] guarantees that every set with Lebesgue outer measure zero is Lebesgue-measurable with measure zero, so we conclude $\lambda_{Ld}(B) = 0$. Now recall that a Bellman family $\{m^\pi_{g_l}\}_{l=1}^L$ is point-separating on $\mathcal{S}$ iff $\mathbf{g} \notin B$. Hence the goal tuples in $\mathcal{S}^L$ that fail to induce point-separation are exactly $B \cap \mathcal{S}^L$, of λ_{Ld}-measure zero, and the complement has full Lebesgue measure in $\mathcal{S}^L$. In particular, λ_{Ld}-almost every L-element goal set $\mathcal{G} = \{g_1, \ldots, g_L\} \subseteq \mathcal{S}$ satisfies $\exists! P(Q^\pi)$ since the diagonal $\bigcup_{i\neq j}\{g_i = g_j\} \subseteq \mathcal{S}^L$ has λ_{Ld}-measure zero, so λ_{Ld}-almost every tuple has distinct entries and corresponds to an L-element goal set. □

D.5 Supporting Results

Lemma 4 (Resolvent)**.** *Let $\gamma \in [0,1)$, a measurable policy π, measurable termination function D and bounded measurable reward function r. For any goal g, let T_g be the Markov operator*

$$(T_g h)(s) \coloneqq D_g(s) \int_{\mathcal{S}} h(s') P^{\pi_g}(\mathrm{d}s' \mid s)\,, \quad \textit{where} \quad P^\pi_g(\cdot \mid s) \coloneqq \int_{\mathcal{A}} P(\cdot \mid s, a)\pi_g(\mathrm{d}a \mid s)\,.$$

Then for each $g \in \mathcal{G}$, $m^\pi_g = (I - \gamma T_g)^{-1} r_g = \sum_{t\geq 0} \gamma^t T^t_g r_g$ as bounded measurable functions on $\mathcal{S}$.

Proof. We note that the identity $m^\pi_g = (I - \gamma T^\pi)^{-1} r_g$ is essentially identical to Blier et al. [30, Theorem 2], but with a goal-conditioned policy baked in; we include the proof for completeness. First recall the Bellman equations

$$V^\pi_g(s) = \int_{\mathcal{A}} Q^\pi_g(s,a)\pi(\mathrm{d}a \mid s, g) \qquad \text{and} \qquad Q^\pi_g(s,a) = \int_{\mathcal{S}} m^\pi_g(s') P(\mathrm{d}s' \mid s, a)$$

from Definition 2. Substituting, marginalising over $\pi(\cdot \mid s, g)$ and multiplying by $D_g(s)$:

$$D_g(s) V^\pi_g(s) = D_g(s) \int_{\mathcal{S}} m^\pi_g(s') P^{\pi_g}(\mathrm{d}s' \mid s) = (T_g m^\pi_g)(s)\,.$$

Now $m^\pi_g(s) = r_g(s) + \gamma D_g(s) V^\pi(s,g) = r_g + \gamma T_g m^\pi_g$, hence $(I - \gamma T_g) m^\pi_g = r_g$. Since $0 \leq D \leq 1$ and $P^{\pi_g}(\mathcal{S} \mid s) = 1$, we have $\|T_g h\|_\infty \leq \|D\|_\infty \cdot \|h\|_\infty \leq \|h\|_\infty$, so $\|T_g\|_{\infty\to\infty} \leq 1$. For any $\gamma < 1$, this implies that $I - \gamma T_g$ is invertible on bounded measurable functions, with Neumann series $(I - \gamma T_g)^{-1} = \sum_{t\geq 0} \gamma^t T^t_g$. Boundedness of r_g ensures m^π_g is bounded. □

Lemma 5 (Determining rewards ⇒ determining Bellman family)**.** *Assume $\pi(\cdot \mid s, g) = \pi(\cdot \mid s)$ and $D(s, g) = D(s)$ are unconditional. If the family $\{r_g : g \in \mathcal{G}\}$ is signed-measure-determining (Definition 5), then the Bellman family $\{m^\pi_g : g \in \mathcal{G}\}$ is measure determining (Definition 4).*

Proof. By unconditionality, the Markov operator $(T_g h)(s) \coloneqq D_g(s) \int_{\mathcal{S}} h(s') P^{\pi_g}(\mathrm{d}s' \mid s)$ from Lemma 4 is independent from g, so we write $T \coloneqq T_g$. Now let $\mu, \nu \in \Delta(\mathcal{S})$ satisfy $\int m_g^\pi \mathrm{d}\mu = \int m_g^\pi \mathrm{d}\nu$ for all g, and write $\eta \coloneqq \mu - \nu$. Define the adjoint T^* on finite signed measures by

$$(T^*\eta)(B) \coloneqq \int_{\mathcal{S}} D(s) P^\pi(B \mid s) \eta(\mathrm{d}s)$$

for any Borel set $B \subseteq \mathcal{S}$. By Fubini's theorem [66], $T^*\eta$ is a finite signed measure satisfying, for any bounded measurable function h, the following adjoint relation:

$$\begin{aligned}\int_{\mathcal{S}} (Th)\mathrm{d}\eta &\coloneqq \int_{\mathcal{S}} \int_{\mathcal{S}} D(s) h(s') P^\pi(\mathrm{d}s' \mid s) \eta(\mathrm{d}s) \\ &= \int_{\mathcal{S}} h(s') \int_{\mathcal{S}} D(s) P^\pi(\mathrm{d}s' \mid s) \eta(\mathrm{d}s) = \int_{\mathcal{S}} h(s') T^*\eta(\mathrm{d}s') =: \int_{\mathcal{S}} h \mathrm{d}(T^*\eta)\,.\end{aligned}$$

Moreover, iterating this relation gives $\int (T^t h)\mathrm{d}\eta = \int h \mathrm{d}(T^t \eta)$ for all $t \in \mathbb{N}$. Now $\|T^*\eta\|_{\mathrm{TV}} \leq \|\eta\|_{\mathrm{TV}}$ since $0 \leq D \leq 1$ and $P^\pi(\mathcal{S} \mid s) = 1$, which implies $\gamma \|T^*\|_{\mathrm{TV}} < 1$, so the resolvent

$$\tilde{\eta} \coloneqq (I - \gamma T^*)^{-1} \eta = \sum_{t \geq 0} \gamma^t (T^*)^t \eta$$

converges in total variation. Now, by Lemma 4, $m_g^\pi = \sum_{t \geq 0} \gamma^t T^t r_g$. Interchanging sum and integral (by dominated convergence), applying the adjoint relation and summing the Neumann series gives

$$\begin{aligned}\int_{\mathcal{S}} m_g^\pi \mathrm{d}\eta &= \sum_{t \geq 0} \gamma^t \int_{\mathcal{S}} (T^t r_g)\mathrm{d}\eta \\ &= \sum_{t \geq 0} \gamma^t \int_{\mathcal{S}} r_g \mathrm{d}((T^*)^t \eta) \\ &= \int_{\mathcal{S}} r_g \mathrm{d}\Big(\sum_{t \geq 0} \gamma^t (T^*)^t \eta \Big) =: \int_{\mathcal{S}} r_g \mathrm{d}\tilde{\eta}\,.\end{aligned}$$

Notice the structural reduction from the Bellman family to the reward family: the LHS integrates m_g^π against the *original* measure η, while the RHS integrates r_g against the *resolvent* measure $\tilde{\eta} = (I - \gamma T^*)^{-1}\eta$. To conclude the proof: by hypothesis, $\int_{\mathcal{S}} m_g^\pi \mathrm{d}\eta$ vanishes for all g, so $\int r_g \mathrm{d}\tilde{\eta} = 0$ for all g. Since $\{r_g\}$ is determining for finite signed measures, $\tilde{\eta} = 0$. This immediately implies $\eta = (I - \gamma T^*)\tilde{\eta} = 0$, so we obtain $\mu = \nu$ as required. □

Lemma 6. *Let μ be a finite signed measure on $\mathbb{R}^d$ and $\varphi_\sigma(x) \coloneqq \exp(-\|x\|^2/(2\sigma^2))$ with $\sigma > 0$. Then $g \mapsto (\mu * \varphi_\sigma)(g)$ is real-analytic on $\mathbb{R}^d$.*

Proof. Write $(\mu * \varphi_\sigma)(g) = \int_{\mathbb{R}^d} \varphi_\sigma(g - s')\, \mu(\mathrm{d}s')$. For each $s' \in \mathbb{R}^d$, the map

$$z \mapsto \varphi_\sigma(z - s') = \exp\left(-\frac{1}{2\sigma^2} \sum_{k=1}^d (z_k - s'_k)^2 \right)$$

is the composition of polynomials and the exponential function, which are entire (i.e. holomorphic on $\mathbb{C}^d$), so $z \mapsto \varphi_\sigma(z - s')$ is entire for all $s' \in \mathbb{R}^d$. Now for any $z = x + iy \in \mathbb{C}^d$ with $x, y \in \mathbb{R}^d$ and $\|y\|_\infty \leq \sigma$, we have

$$\mathrm{Re} \sum_{k=1}^d (z_k - s'_k)^2 = \sum_{k=1}^d (x_k - s'_k)^2 - \sum_{k=1}^d y_k^2 \geq -\sum_{k=1}^d y_k^2 \geq -d\sigma^2\,,$$

so $|\varphi_\sigma(z - s')| \leq e^{d/2}$ uniformly in $s' \in \mathbb{R}^d$ and $z \in \{z \mid \|\operatorname{Im} z\|_\infty \leq \sigma\}$. The polydisc of radius σ centred at any $g \in \mathbb{R}^d$, $\{z \mid |z_k - g_k| \leq \sigma \text{ for all } k\}$, lies entirely in the strip $\{z \mid \|\operatorname{Im} z\|_\infty \leq \sigma\}$, so by Cauchy's estimate on polydiscs [65],

$$\sup_{s' \in \mathbb{R}^d} \left| \partial_g^\alpha \varphi_\sigma(g - s') \right| \leq \frac{\alpha!}{\sigma^{|\alpha|}} e^{d/2}$$

for every multi-index $\alpha \in \mathbb{N}^d$. Since this bound is uniform in s' and $|\mu|(\mathbb{R}^d) < \infty$, dominated convergence justifies differentiating under the integral, giving

$$\left|\partial_g^\alpha (\mu * \varphi_\sigma)(g)\right| \le \frac{\alpha!}{\sigma^{|\alpha|}} e^{d/2}\, |\mu|(\mathbb{R}^d)\,.$$

A smooth function whose derivatives satisfy $|\partial^\alpha f| \le AB^{|\alpha|}\alpha!$ uniformly on $\mathbb{R}^d$ has a convergent Taylor series at every point, hence is real-analytic on $\mathbb{R}^d$ [65]. □

Lemma 7. *Let η be a finite signed measure on $\mathbb{R}^d$ and $\varphi_\sigma(x) \coloneqq \exp(-\|x\|^2/(2\sigma^2))$ with $\sigma > 0$. If $(\eta * \varphi_\sigma)(g) = 0$ for all $g \in \mathbb{R}^d$, then $\eta = 0$.*

Proof. Taking Fourier transforms

$$\widehat{\eta}(\xi) \coloneqq \int_{\mathbb{R}^d} e^{-2\pi i \xi\cdot s'} \eta(\mathrm{d}s') \quad \text{and} \quad \widehat{\varphi_\sigma}(\xi) = \int_{\mathbb{R}^d} e^{-2\pi i \xi\cdot x} \varphi_\sigma(x)\mathrm{d}x\,,$$

the convolution theorem [66] gives $\widehat{\eta}\cdot\widehat{\varphi_\sigma} = \widehat{\eta * \varphi_\sigma} = \widehat{0} = 0$ on $\mathbb{R}^d$. Now

$$\widehat{\varphi_\sigma}(\xi) = (2\pi\sigma^2)^{d/2} e^{-2\pi^2\sigma^2\|\xi\|^2} > 0$$

for all ξ, so we conclude $\widehat{\eta} = 0$ everywhere. Using the Jordan decomposition $\eta = \eta^+ - \eta^-$ into finite positive Borel measures $\eta^\pm$ on $\mathbb{R}^d$, linearity of the Fourier transform gives $\widehat{\eta^+} = \widehat{\eta^-}$ on $\mathbb{R}^d$, so Bochner's theorem [67] implies $\eta^+ = \eta^-$, and thus $\eta = \eta^+ - \eta^- = 0$. □

Lemma 8. *For $\sigma > 0$ and $\psi_\sigma(x) \coloneqq \mathbf{1}[\|x\| < \sigma]$ on $\mathbb{R}^d$, the Fourier transform of ψ_σ is*

$$\widehat{\psi_\sigma}(\xi) = \sigma^{d/2}\, \frac{J_{d/2}(2\pi\sigma\|\xi\|)}{\|\xi\|^{d/2}}\,,$$

where J_ν denotes the Bessel function of the first kind, and the zero set $\{\xi \in \mathbb{R}^d : \widehat{\psi_\sigma}(\xi) = 0\}$ has Lebesgue measure zero.

Proof. Since ψ_σ is radial, i.e. depends only on $\|x\|$, the Fourier transform $\widehat{\psi_\sigma}$ depends only on $\|\xi\|$. Writing $r = \|\xi\|$, we use the radial Fourier formula [68]

$$\widehat{\psi_\sigma}(\xi) = 2\pi r^{-(d/2-1)} \int_0^\sigma J_{d/2-1}(2\pi r t) t^{d/2}\mathrm{d}t\,.$$

Substituting $u = 2\pi r t$ gives $\mathrm{d}t = \mathrm{d}u/(2\pi r)$, with $t \in [0,\sigma]$ becoming $u \in [0, 2\pi r\sigma]$, and applying the Bessel identity $\int_0^T u^{\nu+1} J_\nu(u)\,\mathrm{d}u = T^{\nu+1} J_{\nu+1}(T)$ with $\nu = d/2 - 1$ and $T = 2\pi r\sigma$:

$$\begin{aligned}\widehat{\psi_\sigma}(\xi) &= \frac{2\pi}{r^{d/2-1}} \times \frac{1}{(2\pi r)^{d/2+1}} \int_0^{2\pi r\sigma} u^{d/2} J_{d/2-1}(u)\,\mathrm{d}u \\ &= \frac{2\pi}{r^{d/2-1}} \times \frac{(2\pi r\sigma)^{d/2} J_{d/2}(2\pi r\sigma)}{(2\pi r)^{d/2+1}} = \sigma^{d/2}\frac{J_{d/2}(2\pi\sigma r)}{r^{d/2}}\,,\end{aligned}$$

which settles the first claim. First notice that $\xi = 0$ is not a zero of $\widehat{\psi_\sigma}$, since $J_{d/2}(z)/z^{d/2}$ converges to $1/(2^{d/2}\Gamma(d/2+1)) \neq 0$ as $z \to 0$.[a] For $\xi \neq 0$, the Bessel function $J_{d/2}$ has a sequence of strictly increasing positive zeros $\{a_{d/2,k}\}_{k\in\mathbb{N}}$ with no finite accumulation point. In particular, $\widehat{\psi_\sigma}(\xi) = 0$ iff $J_{d/2}(2\pi\sigma r) = 0$, i.e. $r = a_{d/2,k}/(2\pi\sigma)$ for some $k \ge 1$. Hence Z is the countable union

$$\bigcup_{k\ge 1} \{\xi : \|\xi\| = a_{d/2,k}/(2\pi\sigma)\}\,,$$

where each set is a $(d-1)$-dimensional sphere in $\mathbb{R}^d$, which has Lebesgue measure zero. Countable unions of measure-zero sets have zero measure, so we are done. □

[a] This is irrelevant for the conclusion of Lebesgue measure zero, but gives a neater characterisation of the zero set.

Lemma 9. *Let η be a finite signed measure on $\mathbb{R}^d$ and $\psi_\sigma(x) := \mathbf{1}[\|x\| < \sigma]$ with $\sigma > 0$. If $(\eta * \psi_\sigma)(g) = 0$ for all $g \in \mathbb{R}^d$, then $\eta = 0$.*

Proof. Taking Fourier transforms, the convolution theorem [66] gives $\widehat{\eta} \cdot \widehat{\psi_\sigma} = \widehat{\eta * \psi_\sigma} = \widehat{0} = 0$ on $\mathbb{R}^d$. Letting $Z = \{\xi : \widehat{\psi_\sigma}(\xi) = 0\}$, it follows that $\widehat{\eta}(\xi) = 0$ for all $\xi \in \mathbb{R}^d \setminus Z$. We claim that $\widehat{\eta}$ is continuous: if $\xi_n \to \xi$ in $\mathbb{R}^d$, then $e^{-2\pi i \xi_n \cdot s'} \to e^{-2\pi i \xi \cdot s'}$ pointwise, while $|e^{-2\pi i \xi_n \cdot s'}| = 1$ is integrable with respect to the finite measure $|\eta|$, so dominated convergence gives $\widehat{\eta}(\xi_n) \to \widehat{\eta}(\xi)$. By Lemma 8, Z has Lebesgue measure zero, so $\mathbb{R}^d \setminus Z$ is dense in $\mathbb{R}^d$ (every open ball has positive Lebesgue measure, so meets $\mathbb{R}^d \setminus Z$). Continuity of $\widehat{\eta}$ combined with vanishing on the dense set $\mathbb{R}^d \setminus Z$ gives $\widehat{\eta} \equiv 0$ on $\mathbb{R}^d$. As in the proof of Lemma 7, we conclude $\eta = 0$. □

E Reacher and MountainCar Experiments

In this section we provide additional details for the continuous-control experiments of Section 5. We train goal-conditioned Q-functions with PQN [21] and HER [36], then fit a world model P_φ to the converged Q-values via P-learning (Section 3). Throughout, the WM is parameterised as a residual MLP $P_\varphi(s, a) = s + h_\varphi(s, a)$, trained with ℓ_1 Bellman residual loss function on independently sampled (s, a, g) batches. The sampling distribution is given by $d = \mu_\mathcal{S} \otimes \mathrm{Unif}(\mathcal{A}) \otimes \mathrm{Unif}(\mathcal{G})$, with $\mu_\mathcal{S}$ and all hyperparameters specified per-environment in Table 3 and Table 4. To train policies on unseen goals using the WM, we run value iteration on $\hat{P}$ over a discretised grid (resolution 50 for `Reacher`, 200 for `MountainCar`) to obtain a WM-derived policy.

Throughout, *ground-truth* action-values $Q^{\text{true}} := Q^\pi$ are the exact Q-values of the agent's policy $\pi(s, g) = \arg\max_a Q_\theta(s, a, g)$, computed by policy evaluation on the same discretised grid. We report the error of *learnt* Q-values Q_θ relative to Q^π rather than $Q^\star$ because world model recovery depends on Bellman *consistency* and not Bellman *optimality*.

Compute. All experiments were run on a single H100 GPU, leveraging vectorised environments implemented by Gymnax [69] in JAX [70]. PQN training takes < 5min per `Reacher` run and < 2min per `MountainCar` or `FourRooms` run; P-learning takes < 10s in all cases. The largest single experiment is the 42-architecture `Reacher` sweep (Appendix E.2), totalling ∼35 GPU-hours.

E.1 Reacher: Environment Details

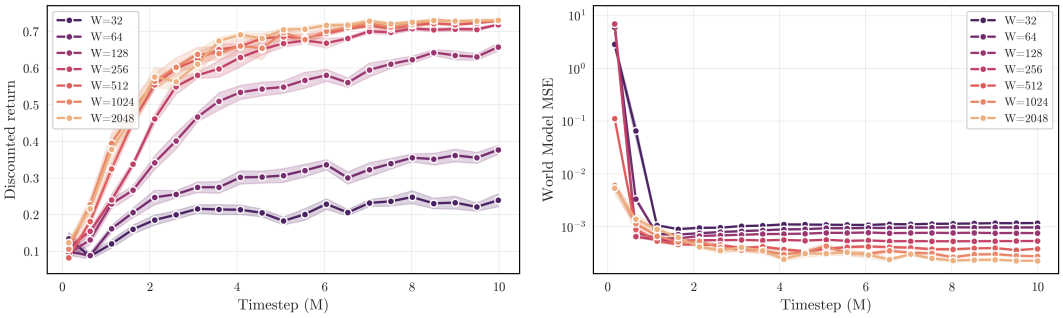
The `Reacher` state $(\theta_1, \theta_2, \omega_1, \omega_2, x, y) \in \mathbb{R}^6$ comprises the two joint angles θ_i, their angular velocities ω_i, and the fingertip position $(x, y) = (\cos\theta_1 + \cos\theta_2, \sin\theta_1 + \sin\theta_2)$. Since the fingertip is a deterministic function of the joints, training goals only specify (x, y) and the world model is restricted to the dynamic dimensions $(\theta_1, \theta_2, \omega_1, \omega_2)$. The state distribution $\mu_\mathcal{S}$ used by P-learning is the env reset distribution, $\theta_i \sim U[-\pi, \pi]$ and $\omega_i \sim U[-1, 1]$. The 2D action space is discretised into 9 actions (3 per joint) in order to train agents with PQN. Full hyperparameters in Table 3.

E.2 Reacher: Correlation Analysis

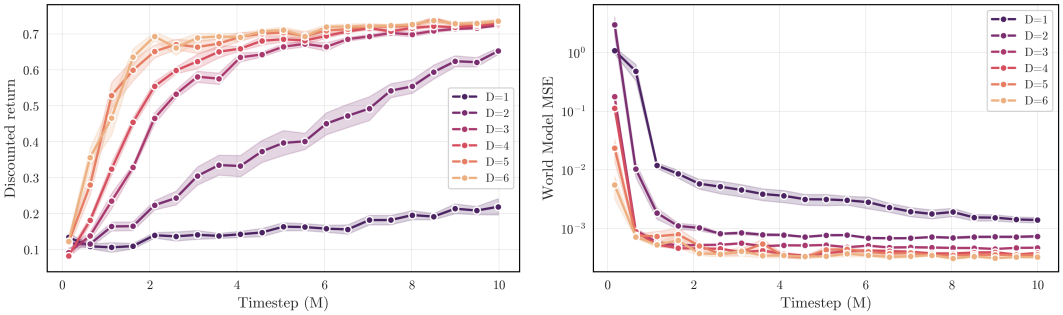
We expand the architecture-capacity analysis behind Figure 4 by sweeping 42 architectures over depths $D \in \{1, 2, 3, 4, 5, 6\}$ and widths $W \in \{32, 64, 128, 256, 512, 1024, 2048\}$, with 10 seeds each, and record agent return and extracted WM error (MSE) over 20 checkpoints (a total of 420 `Reacher` runs and 8400 checkpoints). All hyperparameters match Table 3 apart from the Q-network

Table 2: `Reacher` correlations between PQN agent return, WM dynamics error, and WM-extracted policy return on three unseen goals, across the $n = 42$ network architectures, at the end of training. 95% confidence intervals in brackets (2000 resamples).

Pair	Spearman ρ	Pearson r
PQN return vs. WM error	-0.98 $[-0.99, -0.95]$	-0.82 $[-0.88, -0.76]$
PQN return vs. WM-policy return	$+0.95$ $[+0.89, +0.97]$	$+0.95$ $[+0.92, +0.97]$
WM error vs. WM-policy return	-0.96 $[-0.98, -0.91]$	-0.82 $[-0.88, -0.77]$



(a) PQN return, slice at $D = 4$ (one curve per W). (b) WM dynamics MSE (log-y), slice at $D = 4$.



(c) PQN return, slice at $W = 512$ (one curve per D). (d) WM dynamics MSE (log-y), slice at $W = 512$.

Figure 7: PQN return (left) and WM dynamics MSE (right) vs environment steps for `Reacher`, with varying depth (top) / width (bottom) at fixed $D = 4$ and $W = 512$. Mean $\pm$ SE over 10 seeds.

depth and width, which we vary. We record three quantities: (i) the PQN return averaged over the four training goals, (ii) the WM dynamics MSE on a held-out grid, and (iii) the WM-derived policy return on three unseen goals (*far fingertip*, *target angle*, *target velocity*). Figure 4 in the main text shows the results across all 42 architectures at the end of training. Table 2 below reports the final-checkpoint correlations, confirming that more performant goal-conditioned agents tend to have more accurate implicit world models, and that these in turn yield higher zero-shot returns on unseen goals.

To investigate whether this correlation only holds at the end of training, we plot the PQN return and WM MSE *during* training in Figure 7, sliced at fixed depth $D = 4$ with varying W (top) and at fixed width $W = 512$ with varying D (bottom). The plots visually suggest a strong correlation between agent performance and WM accuracy, not only at the end of training but throughout. We confirm this numerically via the Pearson and Spearman correlations between return and MSE error over time, reported in Figure 8.

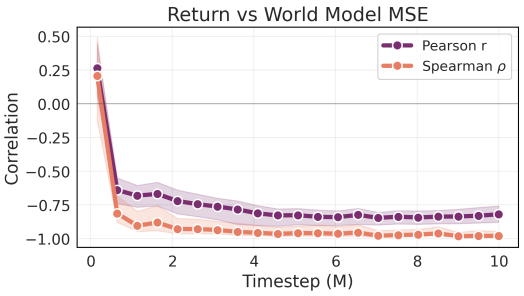


Figure 8: Pearson and Spearman correlations over time between `Reacher` agent return and WM MSE across the $n = 42$ architectures.

This correlation is an empirical regularity rather than a consequence of our theory, which ties WM accuracy to the Bellman *consistency* of Q-values and not to *optimality*; in other words, a poor policy with accurate Q-values would encode an equally accurate WM. However, such a policy (e.g. the uniform random policy) typically reaches a very limited portion of the state space, so its Q-values are rarely updated beyond its reach, and the resulting WM will therefore have high error when measured by the MSE across the entire state space. Moreover, it could have been the case that Q-values only become accurate once the policy converges (because of co-training), so that the WM accuracy would suddenly jump up towards the end of training, but this is not the case (cf. Figure 7).

E.3 MountainCar

`MountainCar` has a 2D position-velocity state space $\mathcal{S} = [-1.2, 0.6] \times [-0.07, 0.07]$ and three acceleration actions $\mathcal{A} = \{\text{left}, \text{none}, \text{right}\}$, with initial position uniform on $[-0.6, -0.4]$ and zero initial velocity. The state distribution $\mu_{\mathcal{S}}$ used by P-learning is uniform over the full state space $\mathcal{S}$. We consider two agents that share the hyperparameters listed in Table 4: a position-trained agent with 4 linearly spaced goals $g_x \in \{-1.2, -0.6, 0, 0.6\}$ and indicator reward with $\sigma = 0.1$, matching the original environment-provided goal for $g_x = 0.6$, and a velocity-trained counterpart (Figure 5) with linearly spaced goals $g_v \in \{-0.07, -0.023, 0.023, 0.07\}$ and threshold $\sigma = 0.01$ to match the approximately 10x smaller velocity state space. All results are provided in the main text.

Table 3: Hyperparameters for `Reacher`.

Group	Hyperparameter	Value
Env	State / action dims	$6 \,/\, 9$
	Number of goals $\lvert\mathcal{G}\rvert$	4 (cardinal (x, y) targets on unit circle)
	Reward / radius / termination	indicator $\mathbf{1}[\lVert s - g\rVert < \sigma]$, $\sigma = 0.2$, terminate on arrival
	Discount γ / Episode length	0.99 / 100 steps
PQN	Network	MLP, 6×2048, layer norm, sigmoid outputs
	Backup	1-step TD ($\lambda = 0$)
	Optimiser	RAdam, global-norm gradient clipping at 10
	Learning rate	10^{-4}, linear decay
	Total env steps	10^7 (256 envs $\times$ 64 rollout steps $\times$ 610 updates)
	Minibatch / epochs per update	256 / 1
	ε-greedy schedule	$1.0 \to 0.1$ over first 50% of training
	HER / resets	$k = 4$, "future" strategy; optimistic resets, ratio 16
WM	Network	MLP, 2×256, tanh, no layer norm
	Output	residual $P_\varphi(s, a) = s + h_\varphi(s, a)$ (init scale 0.01)
	Optimiser	Adam, no gradient clipping, weight decay 0
	Loss / LR	ℓ_1; 10^{-4}, linear decay
	Batch / training steps	$1024 \,/\, 1000$

Table 4: Hyperparameters for `MountainCar`. The velocity-conditioned agent of Figure 5 uses an identical configuration with goals $g_v \in \{-0.07, -0.023, 0.023, 0.07\}$ and $\sigma = 0.01$.

Group	Hyperparameter	Value
Env	State / action dims	$2 \,/\, 3$
	Number of goals $\lvert\mathcal{G}\rvert$	4 (x-positions $\{-1.2, -0.6, 0, 0.6\}$)
	Reward / radius / termination	indicator $\mathbf{1}[\lVert s - g\rVert < \sigma]$, $\sigma = 0.1$, terminate on arrival
	Discount γ / Episode length	0.99 / 200 steps
PQN	Network	MLP, 4×1024, layer norm, sigmoid outputs
	Backup	1-step TD ($\lambda = 0$)
	Optimiser	RAdam, global-norm gradient clipping at 10
	Learning rate	10^{-4}, linear decay
	Total env steps	5×10^6 (256 envs $\times$ 64 rollout steps $\times$ 305 updates)
	Minibatch / epochs per update	256 / 1
	ε-greedy schedule	$1.0 \to 0.1$ over first 50% of training
	HER / resets	$k = 4$, "random" strategy; optimistic resets, ratio 16
WM	Network	MLP, 2×256, tanh, no layer norm
	Output	residual $P_\varphi(s, a) = s + h_\varphi(s, a)$ (init scale 0.01)
	Optimiser	Adam, no gradient clipping, weight decay 0
	Loss / LR	ℓ_1; 10^{-4}, cosine decay
	Batch / training steps	$4096 \,/\, 20{,}000$

F FourRooms Experiments (Deterministic and Stochastic)

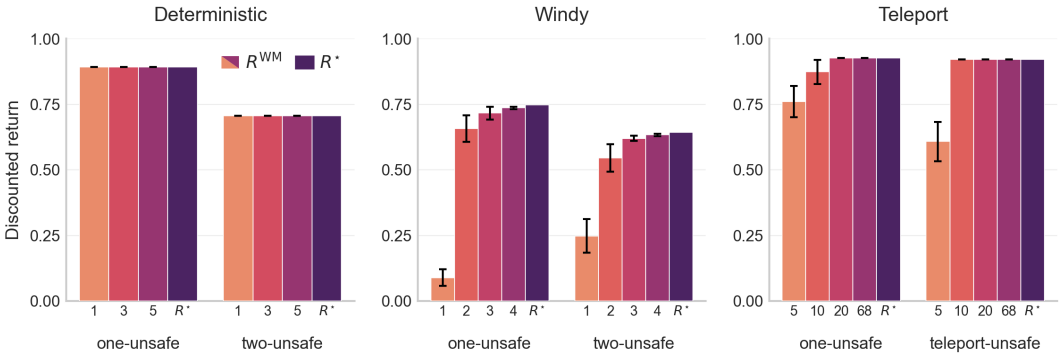


Figure 9: Mean return ± SE (10 training seeds), for optimal ($R^\star$) vs WM-trained (R^{WM}) policies, on two unseen goals, in each `FourRooms` variant, for varying number of training goals $|\mathcal{G}|$ on the x-axis. In the **deterministic** variant, recovery is exact at every $|\mathcal{G}| \geq 1$, in line with Theorem 2. In the **windy** variant, $|\mathcal{G}| = 4$ training goals already drive the WM-derived policy to within $\sim 1\%$ of optimal return. In the **teleporting** variant, $|\mathcal{G}| = 20$ training goals – well short of the $|\mathcal{G}| = |\mathcal{S}| = 68$ worst case of Theorem 3 – also induce quasi-optimal policies on unseen goals.

We complement the continuous-control experiments with a finite-state-space study in `FourRooms` [35], a 11×11 gridworld divided into 4×4 rooms by interior walls and connected by single-cell corridors. After removing wall cells, the open state space has $|\mathcal{S}| = 68$ cells, the action space is $\mathcal{A} = \{\text{up}, \text{right}, \text{down}, \text{left}\}$ with $|\mathcal{A}| = 4$, and the episode horizon is $T_{\max} = 200$. To cover the regimes of our finite-MDP theory (deterministic in Theorem 2; local in Theorem 5; stochastic in Theorem 3), we instantiate three variants of increasing stochasticity. Across all three, the WM-derived policy reaches near-optimal return on unseen goals from a *small fraction* of the $|\mathcal{S}| = 68$ candidate training goals: 1 goal for deterministic, 4 for windy, and roughly 20 for teleporting – consistent with our identifiability theory in the first two cases, and exceeding our expectations in the third. To summarise the environments and results:

1. **Deterministic** (Figures 9 and 10). Each action moves the agent one cell in the chosen cardinal direction; walls keep the agent in place. *A single training goal* ($|\mathcal{G}| = 1$) already drives $\hat{P}$ to zero world-model MSE within 8M env steps and the WM-derived policy is exactly optimal on both unseen goals, in line with Theorem 2.
2. **Windy** (Figures 9 and 11). The chosen action is realised w.p. $\frac{1}{2}$, perturbed $90°$ counter-clockwise or clockwise each w.p. $\frac{1}{4}$. Here, $|\mathcal{G}| = 4$ training goals already drive the WM-derived policy to within $\sim 1\%$ of optimal return on both unseen goals.
3. **Teleporting** (Figures 9 and 12). Four cells $(4,4), (4,6), (6,4), (6,6)$ teleport the agent *uniformly* into the 16 cells of the diagonally opposite 4×4 room. Here, $|\mathcal{G}| = 20$ training goals – well short of the $|\mathcal{G}| = |\mathcal{S}| = 68$ worst case of Theorem 3 – already drive the WM-derived policy to within $\sim 0.1\%$ of optimal return on both unseen goals.

For each variant, the training goals are the first $|\mathcal{G}|$ cells of a fixed seeded permutation of the 68 states, so that the goal subsets nest as $|\mathcal{G}|$ grows; all experiments use 10 seeds.

The reward function differs across variants in line with the regime each one targets: the windy and teleport variants use the indicator reward $r_g(s) = \delta_{gs}$, while the deterministic variant uses the noise-perturbed form $r_g(s) = \delta_{gs} - \xi_{gs}$ with each $\xi_{gs} \sim U[0,1]$ drawn i.i.d. at initialisation, instantiating the noise-injection recipe stated as a corollary of Theorem 2. We take negative perturbation in order to guarantee that the optimal policy still seeks the goal, noting that $r_g(g) \geq 0 \geq r_g(s)$ if $s \neq g$.

We extract $\hat{P}$ from the trained PQN agent by P-learning via linear programming (LP), as defined in Appendix F.1, which we found to be fastest and most effective in this finite, one-hot-encoded setting, compared with function approximation. For the deterministic variant we use column matching, which is exact under the genericity condition of Theorem 2. As for `MountainCar` and `Reacher`, we evaluate $\hat{P}$ directly via the MSE, and on two unseen goals via value-iteration on $\hat{P}$. Final-step returns on unseen goals are summarised in Figure 9, and Table 5 lists the shared hyperparameters.

Our results for each variant are shown via a four-panel figure (Figures 10 to 12): the transition MSE over training steps (one curve per $|\mathcal{G}|$, with mean $\pm$ SE), the discounted return on each unseen goal over training steps, and a $V^\star$ heatmap overlaid with the optimal (dashed red) and WM-derived (solid blue) trajectories at the final checkpoint with $|\mathcal{G}| = |\mathcal{G}|_{\max}$.

F.1 P-learning via linear programming

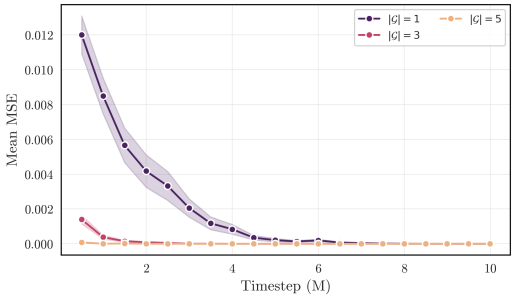
Recall from Equation 2 that for finite MDPs, the P-learning fixed-point equation $\mathcal{T}_P^\pi(Q) = Q$ decouples per (s, a) into the linear system $MP(s, a) = Q(s, a)$, with $M \in \mathbb{R}^{|\mathcal{G}|\times|\mathcal{S}|}$, $M_{lk} = r(s'_k, g_l) + \gamma V(s'_k, g_l)$, and unknown row $P(s, a) \in \Delta^{|\mathcal{S}|}$. For higher efficiency, we directly solve the per-(s, a) system with constrained linear programming (LP) via SciPy, instead of using tabular P-learning. The full sweep over $|\mathcal{S}||\mathcal{A}| = 272$ pairs is computed on a single CPU in < 1 second.

Local variant. The windy variant is an N-local MDP as defined in Section 4.1; we assume this local prior is known in order to illustrate our result for local MDPs, Theorem 5, namely, that $N - 1$ generic goals are sufficient to recover the world model when the support is known. We therefore restrict the LP support to the cardinal-4 neighbours of s plus s itself, reducing each per-(s, a) LP from $|\mathcal{S}| = 68$ to 5 unknowns. This is a geometric prior derived from the gridworld topology, which injects no information about the probability masses in the true kernel.

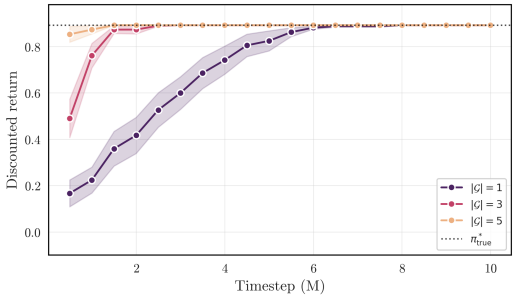
Deterministic variant. For deterministic `FourRooms`, we use *column matching*: $\hat{P}(s, a) = e_{k^\star}$ with $k^\star = \arg\min_k \|M_{\cdot,k} - q\|_1$. This is the LP above with the simplex constraint replaced by a pure argmin; it is exact when M is column-injective (Theorem 2) and requires no LP solver.

In all three cases, the LP solver is a numerical-stability and speed tradeoff over gradient-based P-learning, not a methodological one: both target the same Bellman residual $\|\mathcal{T}_P^\pi(Q) - Q\|^2$ and inherit the same identifiability guarantees of Appendix C.

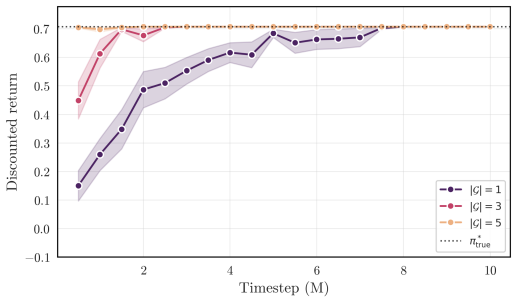
F.2 Deterministic FourRooms



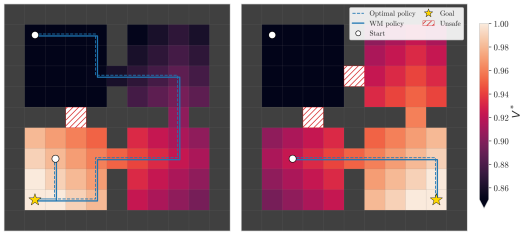
(a) Transition MSE vs. training step.



(b) Discounted return on *one-unsafe*.



(c) Discounted return on *two-unsafe*.

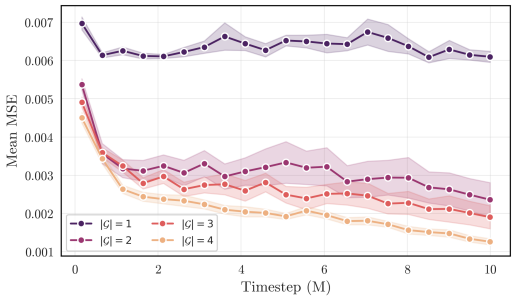


(d) $V^\star$ heatmap with trajectories of optimal (dashed red) and WM-derived (solid blue) policies on both unseen goals, for an agent trained with $|\mathcal{G}| = 1$ goal.

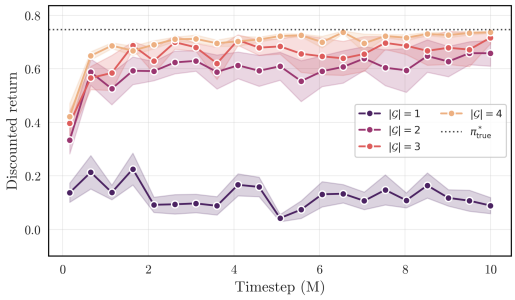
Figure 10: Deterministic `FourRooms`: a single generic goal drives the extracted WM to zero error during training, in line with Theorem 2, and training a policy inside of it produces a policy (solid) that is optimal (dashed) on both unseen goals.

In the deterministic variant, we use the noise-perturbed reward $r_g(s) = \delta_{gs} - \xi_{gs}$ with each $\xi_{gs} \sim U[0, 1]$ drawn i.i.d. at initialisation, which guarantees column-injectivity of M at any $|\mathcal{G}| \geq 1$ via Theorem 2. We train PQN on $|\mathcal{G}| \in \{1, 3, 5\}$ goals over 10 seeds, recover $\hat{P}$ via column matching, and evaluate by training policies in the WM for two unseen goals: *one-unsafe* (target $(9, 1)$ with 'unsafe' state $(5, 3)$ that terminates the episode with zero reward) and *two-unsafe* (target $(9, 9)$ with two 'unsafe' states $(3, 5)$ and $(5, 3)$). Consistent with Theorem 2, a single training goal already drives $\hat{P}$ to *zero* world model MSE within 8M env steps (Figure 10), and the WM-derived policy is optimal on both unseen goals for any number of goals $|\mathcal{G}| \geq 1$ (Figure 9).

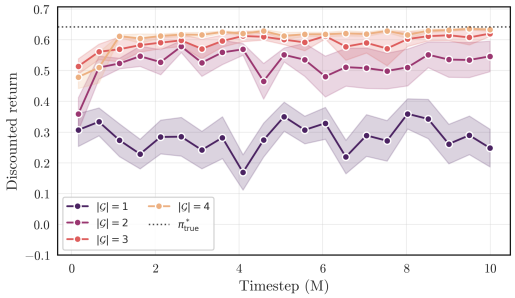
F.3 Windy FourRooms



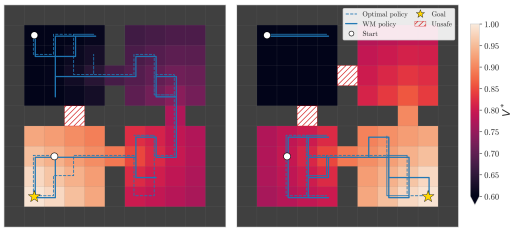
(a) Transition MSE vs. training step.



(b) Discounted return on *one-unsafe*.



(c) Discounted return on *two-unsafe*.



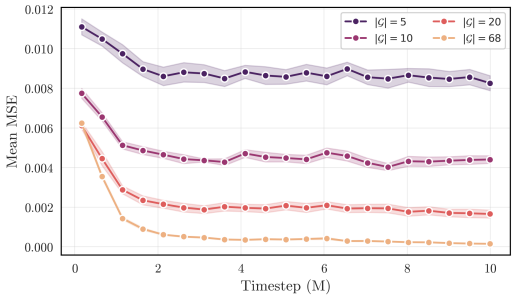
(d) $V^\star$ heatmap with trajectories of optimal (dashed red) and WM-derived (solid blue) policies on both unseen goals, for an agent trained with $|\mathcal{G}| = 4$ goals.

Figure 11: Windy `FourRooms`: the single-goal regime is no longer sufficient for stochastic dynamics, but $|\mathcal{G}| = 4$ drives the WM-derived policy to within a few percent of the oracle on both unseen goals.

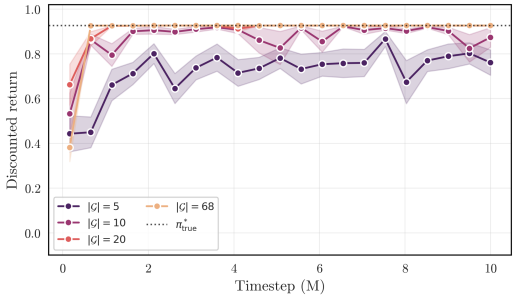
The windy variant introduces local stochasticity: the chosen action is realised faithfully with probability $\frac{1}{2}$, and rotated $90°$ counter-clockwise or clockwise each with probability $\frac{1}{4}$. This places the variant in the local-stochastic regime of Theorem 5. We train PQN on $|\mathcal{G}| \in \{1, 2, 3, 4\}$ training goals over 10 seeds, recover $\hat{P}$ via the local LP, and evaluate on the same two unseen goals as the deterministic variant. As expected for stochastic dynamics, the single-goal regime is under-determined, but $|\mathcal{G}| \geq 3$ already drives the WM-derived policy to within a few percent of the oracle on both goals (Figure 9).

F.4 Teleporting FourRooms

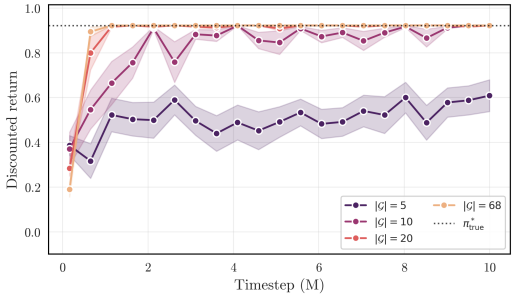
The teleporting variant injects fully-stochastic transitions at four *teleport* cells $(4, 4), (4, 6), (6, 4), (6, 6)$, each of which re-emits the agent uniformly into the 16 cells of the diagonally opposite 4×4 room; all other transitions are deterministic 4-connected moves. We train PQN on $|\mathcal{G}| \in \{5, 10, 20, 68\}$ goals over 10 seeds, and evaluate on *one-unsafe* (as before) and *teleport-unsafe* (all four teleport cells "unsafe" giving zero reward and terminating the environment). Surprisingly, $|\mathcal{G}| \geq 20$ already induces a highly accurate WM, as further validated by a WM-derived policy which matches the optimal policy on both unseen goals to within 0.1% – well below the $|\mathcal{G}| = 68$ goals that Theorem 3 provides guarantees for.



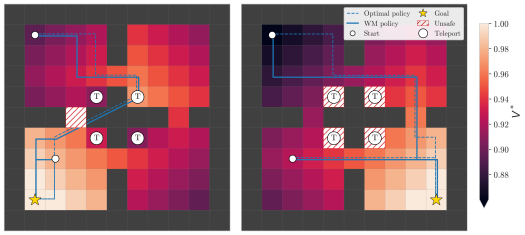
(a) Transition MSE vs. training step.



(b) Discounted return on *one-unsafe*.



(c) Discounted return on *two-unsafe*.



(d) $V^\star$ heatmap with trajectories of optimal (dashed) and WM-derived (solid) policies on both unseen goals, for an agent trained with $|\mathcal{G}| = 20$ goals.

Figure 12: Teleporting `FourRooms`: for $|\mathcal{G}| = 20$, the extracted WM produces quasi-optimal policies on OOD goals, well below our worst-case theoretical bound $|\mathcal{G}| \geq |\mathcal{S}| = 68$.

Table 5: Hyperparameters shared across the three `FourRooms` variants.

Group	**Hyperparameter**	**Value**
Env	State / action dims	$68 \,/\, 4$
	Number of training goals $\lvert\mathcal{G}\rvert$	$\{1, 3, 5\}$ det., $\{1, 2, 3, 4\}$ windy, $\{5, 10, 20, 68\}$ teleport
	Reward	terminate on arrival; indicator $r_g(s) = \delta_{gs}$ for windy/teleport; else perturbed $r_g(s) = \delta_{gs} - \xi$ with $\xi \sim U[0, 1]$
	Discount γ / Episode length	0.99 / 200 steps
PQN	Network	MLP, 4×1024, layer norm, sigmoid outputs
	Backup	1-step TD ($\lambda = 0$)
	Optimiser	RAdam, global-norm gradient clipping at 10
	Learning rate	10^{-4}, linear decay
	Total env steps	10^7 (256 envs $\times$ 64 rollout steps $\times$ 610 updates)
	Minibatch / epochs per update	256 / 1
	ε-greedy schedule	$1.0 \rightarrow 0.1$ over first 50% of training
	HER / resets	$k = 4$, "future" strategy; optimistic resets, ratio 16
WM	Method (det. / windy / tele.)	column matching / local-prior LP / global LP
	LP norm	ℓ_1 (per-(s, a), simplex-constrained)
	Local prior (windy)	cardinal-4 neighbours $\cup \{s\}$
	Sampling	full sweep of the discrete grid $\mathcal{S} \times \mathcal{A}$ (272 pairs)

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Intro (Section 1) and abstract list the three contributions: P-learning (Section 3), identifiability results (Section 4), experiments on `FourRooms`, `MountainCar`, `Reacher` (Section 5). Each is backed by the corresponding section.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See the *Limitations* paragraph in Section 7: stochastic continuous-MDP results restricted to unconditional policies, no POMDPs or non-Q-value agents, no large-scale envs (Atari/Craftax).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Assumptions stated in each theorem (MDP class, policy class, goal-set conditions, reward families); full proofs in Appendices C and D. Supporting lemmas (resolvent, signed-measure-determining, Fourier injectivity) are proved separately and cross-referenced.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Appendices E and F give environment details, architectures, hyperparams (Tables 3 to 5), and seed counts. The released code (`https://github.com/aletcher/inverting-bellman`) regenerates all results and figures in the main text from a single script.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code, configurations and step-by-step instructions to reproduce every figure and reported number are released under the MIT licence at `https://github.com/aletcher/inverting-bellman` (linked in Figure 1).

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Per-env tables (Tables 3 to 5) list architecture, optimiser, LR + schedule, batch sizes, env steps, ε-greedy schedule, HER config, optimistic-resets ratio, WM loss + sampling, and per-variant LP solver.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All FourRooms / Reacher / MountainCar experiments report mean and standard error across 10 seeds (Figures 3, 4 and 9 + the line-plot figures Figure 7). Arch-sweep correlations (Table 2) use 95% bootstrap CIs over 2000 resamples.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See the *Compute* paragraph at the start of Appendix E: GPU type, per-run wall-clock, and total project compute (incl. ablations) are reported there.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Foundational RL identifiability work using standard simulated environments. No human subjects, no scraped data, no high-risk model release.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: This work is focused on foundational mathematical results. Plausible positive impact via interpretability and auditing of RL agents (mentioned in related work / conclusion). We don't foresee specific negative impacts beyond the generic risks associated with RL methodology improvements, e.g. hybrid methods based on P-learning.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The released code trains small task-specific agents and world models for low-dimensional simulated control (`FourRooms`, `MountainCar`, `Reacher`); no pretrained models or datasets are released and there is no misuse risk.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All dependencies are open-source and cited in the paper: JAX [70] (Apache 2.0), Gymnax [69] (Apache 2.0), Gymnasium `MountainCar` [33] (MIT), MuJoCo `Reacher` [34] (Apache 2.0), PQN/PureJaxQL [21] (Apache 2.0).

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We release the code (MIT licence) with a README documenting installation, the full training/extraction pipeline, and the expected results per environment. No datasets or model checkpoints are released.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: No crowdsourcing or human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were used only for writing assistance and helping with JaX debugging and Seaborn plotting. Not part of the core methodology, theory or experimental pipeline.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.