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

Alistair Letcher[α,β] **Oliver Richardson**[γ] **Jakob Foerster**[α]

[α] FLAIR, University of Oxford [β] MATS Fellow [γ] Mila, University of Montreal

Abstract

Model-free agents are trained to achieve goals using value functions, compressing the environment dynamics P and the reward R into a single quantity. This entanglement makes it challenging to repurpose the agent for new goals and safety constraints without retraining from scratch. Orthogonal to inverse RL, we characterise the conditions under which it is possible to reverse-engineer P from Q-values associated to one or many known reward functions, and introduce practical methods to do so. For general MDPs, we prove that P can generically be recovered if the set of goals is as large as the state space $\mathcal{S}$, without any assumptions on policy optimality. For deterministic MDPs, this result strengthens to requiring a *single* goal for finite state spaces, implying that generic *dense* rewards force single-goal Q-learners to encode world models, while sparse rewards may not. This generalises to requiring $2d + 1$ goals for continuous state spaces $\mathcal{S} \subseteq \mathbb{R}^d$. We empirically demonstrate this in (a) stochastic gridworld environments, enabling agents to plan for unseen goals without retraining, and (b) continuous control tasks, where P can be recovered with high fidelity even when Q-values are inaccurate. We note that our method is not an efficient substitute for learning the world model from scratch, but provides a new lens on the implicit environment information encoded in model-free agents, and enables zero-shot transfer without additional environment interaction.

1 Introduction

Motivation. Model-free RL agents learn policies and value functions without explicitly modelling the environment dynamics, from DQN and SAC to PPO [cite]. These methods have proven powerful across a variety of tasks and environments [cite], suggesting the model-free paradigm is sufficient to achieve generally capable agents while side-stepping the challenges inherent in learning a world model. However, model-free agents are opaque to inspection (we cannot query the agent's beliefs about the consequences of its actions), and difficult to repurpose under post-hoc changes to the reward function or environment, e.g. in the case of goal misgeneralisation [cite]. Conversely, model-based agents come with a number of benefits including the application of formal planning methods, formally verifying the safety of plans [1], reducing sample complexity [2] and transfer learning [3].

However, there is increasing evidence that model-free agents may learn implicit world models and exhibit emergent planning, both in the context of RL [4] and LLMs [5, 6]. Recently, Richens et al. [7] proved that the policies of general agents contain world models, provided they are trained on sequential goals defined as Linear Temporal Logic expressions, but RL agents are typically trained on non-sequential goals, for which policies alone are provably *not* sufficient to recover world models [7, Theorem 2]. Conversely, model-free agents typically learn not only policies but value functions, encoding more information than policies alone. We bridge this gap between theory and practice by addressing the following questions:

1. *Do value functions implicitly encode world models?*
2. *Can we extract them in practice?*

Preprint.

This paper establishes conditions under which these can be answered in the affirmative. The key obstacle to the first question is *value equivalence*, whereby different world dynamics can yield the same Q-values. We show that this can be broken when an agent has been trained to reach *various* goals, a setting known as goal-conditioned RL [cite]. Specifically, we prove that the augmented Bellman equation (one for each goal) can be inverted assuming a sufficiently diverse set of goals, and strengthen this result to various hypotheses classes including deterministic and sparse MDPs.

Contributions.

1. *Finite MDPs.* When the state space $\mathcal{S}$ is finite, we prove that P is almost surely[1] determined by exact Q-values provided the set of goals satisfies $|\mathcal{G}| \geq |\mathcal{S}|$ (Theorem 1). In the deterministic setting, recovery is unique provided a *single* $|\mathcal{G}| = 1$ generic goal (Theorem 2), implying that noisy rewards force Q-learners to encode the world model. For MDPs whose transition function is N-sparse, sitting between the general and deterministic settings, $|\mathcal{G}| \geq 2N - 1$ goals suffice (Theorem 3).
2. *Approximate recovery.* In practice, Q-values are only accurate up to some error ϵ. We prove that the resulting world model has error $O(\epsilon)$ when the policy is unconditional,[2] matching an information-theoretic lower bound $\Omega(\epsilon)$. We provide a counterexample in the goal-conditioned case, but demonstrate empirically that such failure-cases are rarely encountered in practical RL environments, where the error is significantly below $O(\epsilon)$.
3. *Continuous MDPs.* We partially extend our results to continuous state spaces $\mathcal{S} \subseteq \mathbb{R}^d$, where we consider three reward function families: Dirac deltas, ball indicators and Gaussian kernels. In the deterministic case, most relevant to robotics and goal-conditioned RL [cite ogbench, purejaxgcrl, gym etc], our key result is that P can be recovered exactly using a *finite* number $2d + 1$ of goals when the policy is unconditional (Theorem 4), despite the state space being infinite. [And a sufficient condition under which this extends to the goal-conditioned setting.]
4. *Experimental validation.* We introduce practical methods to extract P from trained agents in Section 4, and empirically demonstrate them in (a) stochastic gridworld environments, enabling agents to plan for unseen goals and new safety constraints without retraining (Section 5.1), and (b) in continuous control tasks, where P can be recovered with high fidelity from PQN agents even when their value functions are far from convergence (Section 5.2).
5. *Duality.* We establish a connection between our method and TD learning via a novel *inverse Bellman operator*. [And broader duality between world models and value functions via a higher-level mapping between model-free to model-based algorithms.]

Limitations. We make no claims that our method is an efficient substitute for model-based RL, nor that extracting P is always possible or accurate. We focus on the conditions under which the bridge holds, taking a first step towards (i) a complete picture of the relationship between value functions and world models, and (ii) hybrid learning techniques that leverage our method to achieve the best of both worlds. We provide baselines to compare our methods to model-based RL, showing that while model quality and sample efficiency are comparable if the aim is to learn both model and value function, our approach is more computationally intensive. If one only needs a world model, and has access to environment interactions, training from scratch is more efficient. If one wants to repurpose an existing agent without extensive retraining, our method offers a novel path forward.

2 Problem Setting

In this section we introduce the problem setting, focusing on finite MDPs for brevity, with the general case described in Section 3.2. We assume throughout that the reward functions, the discount factor, and the policy are known, as well as (exact or approximate) Q-values. We write $n = |\mathcal{S}|$, $m = |\mathcal{A}|$ and $L = |\mathcal{G}|$, and use indices i, j, k, l for current state, action, next state and goal.

Goal-Conditioned RL. We consider controlled MDPs $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \gamma)$ with state space $\mathcal{S}$, action space $\mathcal{A}$, a transition function $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ describing the environment dynamics, and a discount factor $\gamma \in [0, 1)$. In goal-conditioned RL (GCRL), one additionally defines a goal space $\mathcal{G}$

[1] Meaning this holds for a set of discount factors, reward functions, policies or dynamics with full Lebesgue measure.
[2] Meaning $\pi(\cdot \mid s)$ is independent from g, see Section 2.

with corresponding reward functions $r_g : \mathcal{S} \to \mathbb{R}$.[3] GCRL typically focuses on reward functions that only depend on the landing state, as defined here, though our results generalise to reward functions that also depend on the current state and action. The objective of GCRL is to learn a goal-conditioned policy $\pi_g(\cdot \mid s) : \mathcal{S} \to \Delta(\mathcal{A})$ that maximises the objective $J_g(\pi) \coloneqq \mathbb{E}_{\tau\sim\pi_g}\left[\sum_{t\geq 0} r_g(s^t)\right]$ for each g, where trajectories $\tau = (s^t, a^t)_{t\geq 0}$ are sampled under $\pi(\cdot \mid \cdot, g)$ and the transition P. Q-values are defined by $Q^\pi_g(s,a) \coloneqq \mathbb{E}_{\tau\sim\pi_g}\left[\sum_{t\geq 0}\gamma^t r_g(s^{t+1}) \mid s^0 = s, a^0 = a\right]$, and satisfy the the Bellman equation

$$Q^\pi_g(s,a) = \sum_{s'\in\mathcal{S}} P(s' \mid s,a)\big(r_g(s') + \gamma V^\pi_g(s')\big)\,. \tag{1}$$

Throughout, we assume that the reward functions r, the discount factor γ, and the policy π are known, as well as (exact or approximate) Q-values Q^π. The central question of this paper is: *can the transition function P be uniquely recovered from (r_g, π_g, Q^π_g)?*

Value Equivalence. To extract a world model from Q-values, we rewrite the Bellman equation as

$$Q^\pi_g(s,a) = \int_{\mathcal{S}} m^\pi_g(s')P(s' \mid s,a)\mathrm{d}s'\,, \quad \text{where} \quad m^\pi_g(s') \coloneqq r_g(s') + \gamma V^\pi(s', g)\,, \tag{2}$$

and notice that different $\hat{P} \neq P$ may satisfy this equation for fixed (r_g, π_g, Q^π_g), a phenomenon known as *value equivalence* [ref]. The functions m^π_g can be viewed as Bellman *probes*: for each state-action pair (s,a), the Bellman equation provides expectations of *known* probes m^π_g under the *unknown* distribution $P(\cdot \mid s,a)$. In the finite setting, assembling the probes into a matrix $M^\pi \in \mathbb{R}^{L\times n}$ with entries $M^\pi_{lk} = m^\pi_{g_l}(s_k)$ turns each Bellman equation into a linear system. If M^π is invertible, the **inverse Bellman equation** falls out:

$$P(\cdot \mid s,a) = (M^\pi)^{-1} Q^\pi_\cdot(s,a)\,, \tag{3}$$

recovering P uniquely. If Q-values are only approximate, the hope is that the same procedure yields a world model $\hat{P} \approx P$. To establish conditions under which invertibility holds (Theorem 1), we introduce a notion of goals that are sufficiently *diverse* below. In a deterministic MDP, where $P(\cdot \mid s,a) = \mathbf{1}[s' = f(s,a)]$ for a transition function $f : \mathcal{S}\times\mathcal{A} \to \mathcal{S}$, the Bellman equation collapses to

$$Q^\pi_g(s,a) = M^\pi_g(f(s,a))\,, \tag{4}$$

so the vector $Q_\cdot(s,a) \in \mathbb{R}^L$ is simply the $f(s,a)$-th *column* of M^π. Recovery of f reduces to identifying which column of M^π matches $Q_\cdot(s,a)$, implying that f can be identified uniquely provided M^π is *column-injective*, i.e. $M^\pi_\cdot(s) \neq M^\pi_\cdot(s')$ for all $s \neq s'$. This is a much weaker requirement than invertibility, requiring only a single generic goal (Theorem 2).

State goals and diverse goals. A goal g is a *state goal* if $g \in \mathcal{S}$ and $R_g(s) = \delta_{gs}$, which is the most common type of goal in GCRL environments. A set of goals $\mathcal{G}$ is *diverse* if $\mathrm{rank}(R) = n$, where $R_{lk} \coloneqq r_{g_l}(s_k)$. Note that this implies $|\mathcal{G}| \geq n$. Intuitively, this means that goals *span the state space*, which automatically holds if $\mathcal{G} = \mathcal{S}$ (all state goals) since $R = I$.

Goal-conditioned policies. This paper studies two scenarios, illustrated in Figure 1. (a) The agent has a single unconditional policy π, optimised to pursue one *true* goal given by the environment, but the agent learns Q-values Q^π_g associated to other goals g (eg in single-goal contrastive RL [10], where the critic encodes log Q-values across all state goals). (b) The agent has a goal-conditioned (GC) policy π_g, with corresponding Q-values $Q^{\pi_g}_g$.

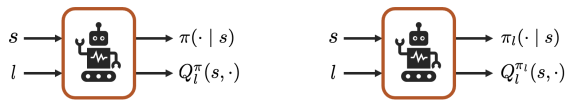


Figure 1: Unconditional (left) vs goal-conditioned policies (right).

[3]Early literature consider goals to be specified by arbitrary reward functions [8], as we do here, while recent GCRL benchmarks only consider goals to be what we call *state* goals [9], with reward 1 upon reaching the state and zero elsewhere.

3 Main Results

Notation. We use the shorthand $P \Leftrightarrow Q$ to mean that P is uniquely determined by Q-values, under the assumption that rewards, discount factor and policy are known. Since Q-values are rarely learnt exactly, we say Q-values are ϵ-approximate if $\|\hat{Q}^\pi_{ij} - Q^\pi_{ij}\|_1 \leq \epsilon$ for all i, j.

3.1 Finite Space

We summarise our results in Table 1.

Table 1: Number of goals to guarantee recovery for finite MDPs, for generic reward functions.

	Stochastic	N**-sparse (known)**	N**-sparse (unknown)**	**Deterministic**
Goals needed	$L \geq n$	$L \geq N - 1$	$L \geq 2N - 1$	$L \geq 1$
Condition on M^π	Invertible	Affine independence	General position	Column-injective
Method	Matrix inversion	Small matrix inversion	Search + inversion	Column matching

3.1.1 Stochastic Setting

We state our main result for $|\mathcal{G}| = n$, noting that the proof extends to $|\mathcal{G}| > n$ by selecting any n-goal subset for which M^π is invertible, which is guaranteed to exist by part (a).

Theorem 1 (Finite)**.** *Consider the class of finite MDPs, and assume $|\mathcal{G}| = n$ is diverse.*

*(a) If Q-values are **exact**, the set of MDPs for which $P \Leftrightarrow Q$ has full Lebesgue measure. More precisely, M^π is invertible for almost every (i) discount factor $\gamma \in [0, 1)$, (ii) transition function P, (iii) set of reward functions R, and (iv) policy π. Moreover, M^π is invertible for all $\gamma < 1$ sufficiently small / large, and for any optimal entropy-regularised policy where the entropy coefficient $\alpha > 0$ is sufficiently small / large.*

*(b) If Q-values are ϵ-**approximate** and M^π is invertible, the estimator $\hat{P}_{ij} = (\hat{M}^\pi)^{-1}\hat{Q}_{ij}$ obtained from the approximate matrix $\hat{M}^\pi = R + \gamma\hat{V}^\pi$ satisfies*

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

for all $\epsilon < 1/\gamma m\|(M^\pi)^{-1}\|_1$.

When the policy is **unconditional** and goals are states, M^π is *always* invertible, $\|(M^\pi)^{-1}\|_1 = (1 + \gamma)$, the factor of m is absorbed, and the upper bound reduces 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 $\epsilon < 1/\gamma(1 + \gamma) < 1/2$. Moreover, any procedure mapping ϵ-approximate Q-values to a world model incurs worst-case error $(1 + \gamma)\epsilon$, implying the minimax error rate is $\Theta(\epsilon)$. When Q-values are exact, and *only* if the policy is unconditional and goals are states, note that M^π reduces to the *occupancy measure* $M^\pi_g(s) = \sum_t \gamma^t P_\pi(s^t = g \mid s^0 = s)$. See Section 6 for a discussion of this connection and existing results.

When the policy is **goal-conditioned**, M^π is not always invertible, even when $\mathcal{G} = \mathcal{S}$. We construct an explicit MDP and policy in [appendix], where $\det(M^\pi) = 0$ if and only if $\gamma = 1/\sqrt{2}$. In practice, none of the Gymnasium [ref] environments we tested [list] exhibit singularities when goals are states, policies are uniform or optimal or entropy-regularised, and $\gamma \in [0, 1)$ [actually run this].

3.1.2 Deterministic Setting

In a deterministic MDP, $P(\cdot \mid s, a) = \mathbf{1}[s' = f(s, a)]$ for a transition function $f : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$. The Bellman equation collapses to

$$Q^\pi_g(s, a) = M^\pi_g(f(s, a))\,, \tag{5}$$

so the vector of Q-values $Q_\cdot(s,a) \in \mathbb{R}^L$ is simply the $f(s,a)$-th *column* of M^π. Recovery of f reduces to identifying which column of M^π matches $Q_\cdot(s,a)$, which we call *column-matching*. In particular, f can be identified uniquely provided M^π is *column-injective*, i.e. $M^\pi_\cdot(s) \neq M^\pi_\cdot(s')$ for all $s \neq s'$. To quantify this, we define the *column separation* of M^π as

$$\Delta_1(M^\pi) \coloneqq \min_{s \neq s'} \|M^\pi_\cdot(s) - M^\pi_\cdot(s')\|_1 .$$

Our main result below shows that for deterministic environments, a *single* generic goal suffices for recovery. Though this may fail for sparse rewards, we provide two families of dense rewards for which recovery is guaranteed, showing that density (almost surely) forces agents to encode the world model, disentangling P from R in Q. [When this fails, e.g. because of sparsity, it is nonetheless still possible to recover a *coarse-grained* world model (up to value-equivalent states), as shown in experiments.]

Theorem 2 (Deterministic)**.** *Consider the class of deterministic finite MDPs, and assume* $|\mathcal{G}| \geq 1$.

(a) If Q-values are ***exact****, the set of reward functions* $R \in \mathbb{R}^{|\mathcal{G}| \times n}$ *for which* $P \Leftrightarrow Q$ *via column-matching has full Lebesgue measure, for any policy and discount factor.*

(b) If Q-values are ϵ*-****approximate*** *and* M^π *is column-injective, the estimator* $\hat{f}(s,a) = \operatorname{argmin}_{s'} \|\hat{Q}(s,a) - \hat{M}^\pi(s')\|_1$ *satisfies* $\hat{f} = f$ *for all* $\epsilon < \Delta_1/2(1+\gamma m)$*.[Add tightness.]*

Proof in Appendix A.3.

Corollary 1. *In a deterministic MDP with* $|\mathcal{G}| = 1$*, the following reward families guarantee unique recovery of the world dynamics via column-matching.*

(a) ***Dense / noisy rewards.*** *For any reward drawn from a density (including a standard state-goal* $r_g(s) = \delta_{sg} + \epsilon_s$ *initialised with any noise* ϵ_s*),* $P \Leftrightarrow Q$ *almost surely.*

(b) ***Gaussian rewards.*** *For any embedding* $\varphi : \mathcal{S} \hookrightarrow \mathbb{R}^d$ *(eg a gridworld), the reward* $r_g(s) = \exp(-\|\varphi(s) - g\|^2/2\sigma^2)$ *gives* $P \Leftrightarrow Q$ *for almost every* $(g, \sigma) \in \mathbb{R}^d \times (0, \infty)$*.*

Proof in Appendix A.4.

When Q-values are approximate, recovery has a qualitatively different behaviour than the stochastic setting – a *threshold* below which recovery is exact, with no guarantees above it, as opposed to the graceful degradation of Theorem 1. [Note that perturbation of γ or π is not sufficient unlike Theorem 1.]

3.1.3 Sparse Setting

An MDP is N*-sparse* if $\operatorname{supp}(s,a) \coloneqq |\{s' : P(s' \mid s,a) > 0\}| \leq N$ for every (s,a), interpolating between deterministic ($N = 1$) and fully stochastic ($N = n$). Theorem 3 show that under this assumption, we can recover the exact world model using only (a) $L \geq N - 1$ generic goals when the support is known, and (b) $L \geq 2N - 1$ otherwise. This is a strict generalisation of the stochastic case (where the support of size $N = n$ is trivially known), and the deterministic case (where $N = 1$ but the support i.e. successor state is unknown, so that $L \geq 2N - 1 = 1$ is required). [We leave the approximate versions to the appendix.]

Theorem 3 (Sparse)**.** *Consider the class of* N*-sparse finite MDPs, and assume* $|\mathcal{G}| \geq N - 1$ *or* $|\mathcal{G}| \geq \min(2N - 1, n - 1)$.

(a) Known support: For any $|\mathcal{G}| \geq N - 1$*, the set of rewards* $R \in \mathbb{R}^{|\mathcal{G}| \times n}$ *for which* $P \Leftrightarrow Q$ *has full Lebesgue measure, for any policy and discount factor.*

(b) Unknown support: For any $|\mathcal{G}| \geq \min(2N - 1, n - 1)$*, the set of rewards* $R \in \mathbb{R}^{|\mathcal{G}| \times n}$ *for which* $P \Leftrightarrow Q$ *has full Lebesgue measure, for any policy and discount factor.*

Proof in Appendix A.5.

3.2 Continuous Space

In the finite case, identifiability of P reduces to invertibility or column-injectivity of the matrix M^π. For continuous state spaces $\mathcal{S} \subseteq \mathbb{R}^d$, the matrix M^π is replaced by the family of Bellman probes $\{m^\pi_g\}_{g \in \mathcal{G}}$, and invertibility generalises to whether this family is *measure determining*. Our proof

techniques apply to arbitrary [Borel] state spaces (not necessarily $\mathbb{R}^d$), incorporating settings where there are both continuous and discrete variables, but we specialise to $\mathbb{R}^d$ in order to specify concrete families of reward functions for which recovery is guaranteed.

We state the result most relevant to practical experiments (Theorem 4), where a *finite* number of Gaussian rewards suffice for recovery in deterministic environments, despite the state space being infinite.. Table 2 summarises other reward families for which recovery is guaranteed in both stochastic and deterministic environments, assuming a sufficiently large number of goals.

Theorem 4. *Assume the MDP is deterministic and consider the family of Gaussian rewards* $r_g(s') = \exp(-\|s' - g\|^2/(2\sigma^2))$ *with fixed* $\sigma > 0$. *For any* $\mathcal{S} \subseteq \mathbb{R}^d$, $\gamma \in [0, 1)$, *unconditional* π *and any* ***finite*** *number of goals* $|\mathcal{G}| \geq 2d + 1$, *the set of tuples* $(g_1, \ldots, g_{|\mathcal{G}|}) \in \mathcal{S}^{|\mathcal{G}|}$ *for which the Bellman family* $\{m^\pi_{g_i}\}$ *is measure determining on* $\mathcal{S}$ *has full Lebesgue measure in* $\mathcal{S}^{|\mathcal{G}|}$. *In particular, the transition function* f *is almost surely identified from* (π, Q^π). *[Add GC conjecture.] Proof in Appendix A.6.*

Table 2: Reward families for which $P \Leftrightarrow Q$ for continuous MDPs, if π is unconditional and $\mathcal{S} \subseteq \mathbb{R}^d$.

	Gaussian	**Ball indicators**	**Dirac**
Stochastic	$\mathcal{G} \subseteq \mathcal{S}$, non-empty interior	$\mathcal{G} \supseteq \mathcal{S} + B_a(0)$	$\mathcal{G} \subseteq \mathcal{S}$, full measure†
Deterministic	$\|\mathcal{G}\| \geq 2d + 1$ (Theorem 4)	$\mathcal{G} \supseteq \mathcal{S}$	$\mathcal{G} = \mathcal{S}$

† Requires $P(\cdot \mid s, a) \ll \lambda^d$ (absolutely continuous kernel).

4 Method

We first describe the naive, exact method that extracts P by inverting the full matrix M^π when the state space is small, and then introduce practical methods addressing large or continuous state-action spaces where the reward function, Q-values and policies can only be *sampled* on batches (g, s, a), whereby M^π cannot be constructed in full.

Warmup. If the state space is finite and moderately small ($n \leq 10^5$), we can probe the goal-conditioned value function/policy on all state-action pairs, and explicitly construct $M^\pi_{lk} = r_{g_l}(s_k) + \gamma V^\pi_{g_l}(s_k)$. The pseudo-inverse $(M^\pi)^+$ can then be computed exactly in $O(Ln^2)$ time, thereby recovering the world model

$$P(\cdot \mid s, a) = (M^\pi)^+ Q^\pi_{\cdot}(s, a)$$

in $O(Ln)$ time per state-action pair, with a total complexity of $O(Ln^2 m) = O(n^3 m)$ for a diverse set of goals. For deterministic MDPs, the column-matching technique described in Section 3.1.2 costs $O(nL)$ per state-action pair, so the resulting total complexity is the same $O(Ln^2 m)$; however, the number of generic goals required to extract P is 1, reducing the complexity to $O(n^2 m)$.

Scaling up. Inverting the matrix is equivalent to minimising the loss function obtained by taking the difference between both sides of the Bellman equation:

$$\mathcal{L}(P) = \|M^\pi P(\cdot \mid s, a) - Q^\pi_{\cdot}(s, a)\|^2 ,$$

which would reduce to the standard TD loss if it were minimised with respect to Q rather than P. For large or continuous state-action spaces, we parameterise the world model as $P_\varphi : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ and minimise

$$\mathcal{L}(\varphi) = \mathbb{E}_{g,s,a}\left[\left(\int_{\mathcal{S}} M^\pi_g(s') P_\varphi(s' \mid s, a)\, \mathrm{d}s' - Q^\pi_g(s, a)\right)^2\right] .$$

The outer expectation over (g, s, a) is estimated by minibatches, uniformly over the goal/state/action spaces. To estimate the inner integral, we draw samples $s'_1, \ldots, s'_K \sim P_\varphi(\cdot \mid s, a)$ to obtain

$$\mathcal{L}(\varphi) = \mathbb{E}_{g,s,a;s'_k \sim P_\varphi(\cdot|s,a)}\left[\left(\frac{1}{K}\sum_{k=1}^{K} M^\pi_g(s'_k) - Q^\pi_g(s, a)\right)^2\right] . \tag{6}$$

Since $M_g^\pi(s') = r_g(s') + \gamma V_g^\pi(s')$, evaluating this loss only requires querying the reward and value functions at sampled next-states. To backpropagate through sampling from P_φ, one can use a standard score function estimator, or the reparameterisation trick if P_φ is e.g. a mixture of Gaussians or normalising flow [ref]. However, in many RL environments of interest – particularly robotics and continuous control – we can take advantage of dynamics being deterministic. In this case, P_φ simplifies to a successor predictor $f_\varphi : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ and the loss reduces to

$$\mathcal{L}(\varphi) = \mathbb{E}_{g,s,a}\left[\left(M_g^\pi(f_\varphi(s,a)) - Q_g^\pi(s,a)\right)^2\right], \tag{7}$$

which requires no integral over $\mathcal{S}$, scales to arbitrary dimensionality, and reduces to column-matching when the state space is small. Unlike the stochastic case, note that backpropagation here passes through $M_g^\pi(f_\varphi)$ via the chain rule, requiring r_g and V_g^π to be (piecewise) differentiable with respect to s'. This is satisfied when V_g^π is a neural network and r_g is distance-based (e.g. L2 distance $-||s-g||^2$, ball indicators $\mathbf{1}[||s-g|| \leq \sigma]$ or Gaussian $\exp(-||s-g||^2/2\sigma^2)$), which holds in many RL environments (e.g. MuJoCo, [...]). This is the setting we focus on for experiments with continuous state spaces (Section 5).

Inverse Bellman operator. Minimising the loss function above by gradient descent is equivalent to iteration of a novel "inverse Bellman operator" that we flesh out in Appendix **??**, providing a cousin to standard **value iteration** that we name **world iteration**. The analogy goes further, whereby standard TD-learning (which corresponds to a *semi-gradient*) also has a cousin in world learning, but only for unconditional policies and a goal set exactly the size of the state space. This is not central to the method, but provides another lens on the equivalence between values and worlds – lifting from the object level (P, Q) to the algorithmic level (value iteration, world iteration).

5 Experiments

5.1 Finite MDP Experiments

See Figure 2 for results on a deterministic gridworld [change to stochastic or poker or other]. We train PQN agents to convergence and extract the world model from their Q-values based on $L = 30$, $L = 35$ and $L = 40$ goals. The recovered world model matches the ground truth perfectly for $L \geq 40$ [not shown in figure], and the agent is able to plan for novel safety constraints, reaching the goal in the optimal number of steps without traversing the forbidden goal. Its world model is not sufficient for $L < 40$, and using it for planning results in dangerous $(L = 30)$ or suboptimal $(L = 35)$ behaviour.

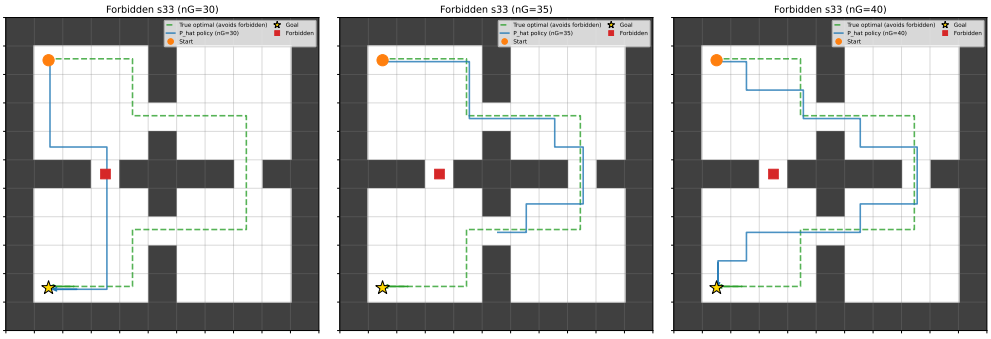
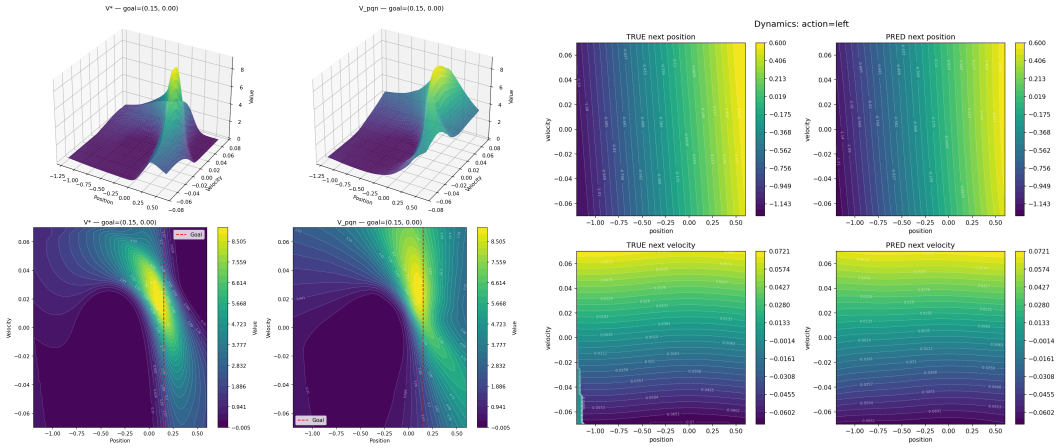


Figure 2: Gridworld – trained agent planning for novel safety constraints using the world model extracted from its Q-values based on $L = 30$ (left), $L = 35$ (middle) and $L = 40$ (right). The agent plans optimally using the extracted world model for $L = 40$, reaching the goal in the optimal number of steps without traversing the forbidden goal.

5.2 Continuous MDP Experiments

See Figures 3 and 4 for results on the 2d environment MountainCar, using only $2d + 1 = 5$ goals with distance-based (Gaussian) rewards at equally spaced positions along the x-axis. We train PQN agents for 10M steps and extract the world model from their Q-values. Despite imperfect Q-values, MSE $= 1.79$, the recovered dynamics match ground truth with high fidelity, MSE $= 0.00013$.



(a) Value functions (b) World models

Figure 3: MountainCar results at the end of training (10M steps). (a) Value functions of PQN agent (right) vs ground truth (left) for the goal of reaching position $p = 0.15$ (one of the $2d+1 = 5$ linearly spaced points between the boundaries $p = -1.2$ and $p = 0.6$ of the environment). The shape is approximately correct but the mean squared error is relatively large, MSE $= 2.99$. The MSE across all 5 goals is MSE $= 1.79$. (b) Recovered world model (right) vs true dynamics (left) for one of the three actions (left). Despite imperfect value functions, the recovered dynamics match ground truth with high fidelity, MSE $= 0.00013$ across actions.

5.3 Further Safety Applications

The experiment in Section 5.1 demonstrates that the world model can be extracted from the Q-values of a trained agent, and used to plan for novel safety constraints. We consider further safety implications of our results with the following experiments: (1) Interpretability via value equivalence: can we spot "blindness" when the goal is not sufficiently informative (e.g. delta function), and how is this helpful? (2) Planning under new safety constraints in continuous environments? (3) Counterfactual reasoning? (4) Bounds on probability of harm based on TD error + our results (at least for unconditional policy, or by computing norm of M)? Compare the bound against the actual harm probability in a gridworld with danger zones.

6 Related Work

World models from agents. Richens et al. [7] proved that the policies of general agents contain world models. Their key result shows that if an agent's policy is ϵ-optimal for a sufficiently diverse set of *multi-step* goal-directed tasks of depth d, the transition dynamics can be recovered with accuracy $O(1/\sqrt{(1-\epsilon)(d-1)})$. This approach has two limitations. First, it requires competence on a multi-step goal set of size at least dn^2m. In realistic settings, agents are only conditioned on *single-step* goals, a set of size n, for which the world model provably *cannot* be extracted from the policy alone. Second, even for exact optimality ($\epsilon = 0$) the error only vanishes as $d \to \infty$, while our method recovers dynamics from n exact Q-values irrespective of whether the policy is optimal. [Even for a small environment with 100 states, 10 actions and horizon $d = 100$, Richens et al. [7] require competence on 10 million goals, while our results require only 100.]

Adamczyk [11] showed that converged Q-values encode *deterministic* dynamics, recovering the next-state function $f(s, a)$. Their approach requires V^π to be δ-separable, without characterising policies or reward functions for which this holds, and is limited to deterministic MDPs with a single reward function. Our work handles arbitrarily stochastic and sparse MDPs, goal-conditioned policies, general goal sets, and characterises the reward function families for which recovery is guaranteed.

Turner [12, Theorem F.117] proved that *optimal* value functions V^* — without access to Q-values — recover deterministic MDPs (up to isomorphism) from n state-indicator rewards, but demonstrated that this fails for stochastic environments. This highlights the role of Q-values: V loses the per-

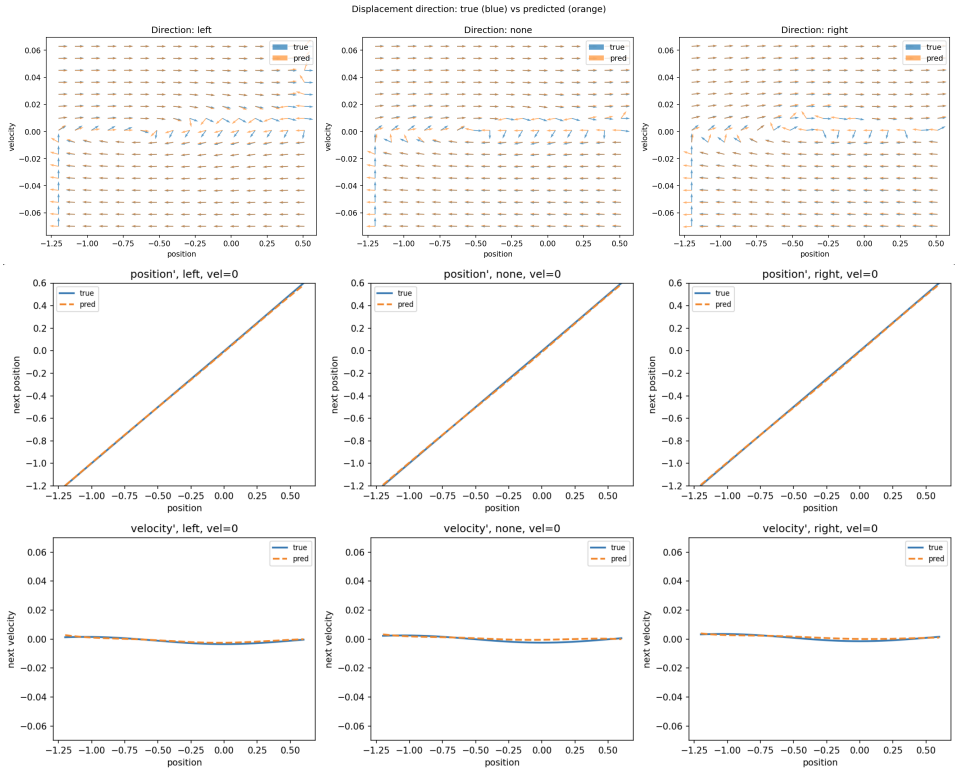


Figure 4: MountainCar results at the end of training (10M steps). (Top) Quiver plot showing the predicted vs true *direction* of the next state (ignoring magnitude) given current state (position on x-axis, velocity on y-axis). (Middle + Bottom) Comparison of predicted vs true position (middle) / velocity (bottom) of the next state for sliced velocity = 0, where the recovered model is v accurate.

action information needed to identify stochastic transitions; our results apply moreover to arbitrary (non-optimal) policies and general reward functions.

Implicit world models. Bush et al. [4] provide empirical evidence via mechanistic interpretability that model-free agents develop implicit forward models during training, complementing our theoretical results. [Also cite Keyon Vafa.]

Successor features and goal-conditioned RL. The successor representation [13] and its extension to successor features [14, 15] decompose the value function into transition-dependent and reward-dependent components, enabling transfer across tasks. Lehnert and Littman [16, p. 7] and Blier et al. [17, Prop. 1] observed that the occupancy measure matrix is invertible, which corresponds to **??** in the special case where (a) the policy is unconditional, (b) goals are state goals ($\mathcal{G} = \mathcal{S}$) and (c) Q-values are exact. However, neither work established a connection to world model recovery from Q-values, nor handles general goal sets, goal-conditioned policies (Theorem 1), or strengthened results for deterministic or sparse MDPs (Theorems 2 and 3).

Value equivalence and model-based RL. The value equivalence (VE) principle [18, 19] shows that multiple environment models can yield identical Bellman updates for a given set of policies and value functions. Algorithms such as MuZero [20] implicitly learn value-equivalent models without state reconstruction, while Farahmand [21] and Lambert et al. [22] motivate value-aware model learning by showing that one-step prediction accuracy need not correlate with control performance. Freed et al. [23] further unify model-based and model-free RL via equivalent policy sets. Our results provide a *converse* to the VE principle: under a sufficiently diverse goal set, all value-equivalent models must agree on the true transition dynamics: goal diversity causes the VE equivalence class to collapse to a single point.

Inverse RL and broader context. Our problem is orthogonal to inverse RL [24], which recovers R from known (or sampled) P and π; we recover P from known (or sampled) R and π – as well as Q, without which recovery is impossible for non-sequential goals [7, Theorem 2]. More broadly, our results connect to the thesis that reward maximisation suffices for intelligence [25], formalising a precise sense in which "goals are enough" to implicitly learn dynamics, namely: a sufficiently diverse set of goal-conditioned Q-values is informationally equivalent to the transition kernel. [Our framework also provides a theoretical lens on hindsight experience replay [26]: information richness of goal relabelling?]

AI Safety. World models come with a number of benefits including the application of formal planning methods, formally verifying the safety of plans [1], reducing sample complexity [2] and transfer learning [3]. More broadly, they are a necessary component of counterfactual reasoning [27, 28], safe exploration [29], predicting human responses [30], auditing agent incentives [31], detecting deception [32], and verifying intent [33].

7 Conclusion

Limitations.

Future directions.

References

[1] David "davidad" Dalrymple, Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, Alessandro Abate, Joe Halpern, Clark Barrett, Ding Zhao, Tan Zhi-Xuan, Jeannette Wing, and Joshua Tenenbaum. Towards guaranteed safe ai: A framework for ensuring robust and reliable ai systems, 2024.

[2] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels, 2019.

[3] Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models, 2018.

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

[5] Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*, 2023.

[6] Yifan Hou, Jiaoda Li, Yu Fei, Alessandro Stolfo, Wangchunshu Zhou, Guangtao Zeng, Antoine Bosselut, and Mrinmaya Sachan. Towards a mechanistic interpretation of multi-step reasoning capabilities of language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4902–4919. Association for Computational Linguistics, 2023.

[7] Jonathan Richens, Tom Everitt, and David Abel. General agents need world models. In *Forty-Second International Conference on Machine Learning*, 2025.

[8] Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal Value Function Approximators. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1312–1320. PMLR, 2015.

[9] Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. OGBench: Benchmarking Offline Goal-Conditioned RL. In *The Thirteenth International Conference on Learning Representations*, 2024.

[10] 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 *The Thirteenth International Conference on Learning Representations*, 2025.

[11] Jacob Adamczyk. Inferring transition dynamics from value functions. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

[13] Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624, 1993.

[14] 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*, volume 30, pages 4058–4068, 2017.

[15] Diana Borsa, André Barreto, John Quan, Daniel Mankowitz, Rémi Munos, Hado van Hasselt, David Silver, and Tom Schaul. Universal Successor Features Approximators, 2018.

[16] Lucas Lehnert and Michael L. Littman. Successor features combine elements of model-free and model-based reinforcement learning. *Journal of Machine Learning Research*, 21(1), 2020.

[17] Léonard Blier, Corentin Tallec, and Yann Ollivier. Learning Successor States and Goal-Dependent Values: A Mathematical Viewpoint, 2021.

[18] 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*, volume 33, pages 5541–5552, 2020.

[19] Christopher Grimm, André Barreto, Gregory Farquhar, David Silver, and Satinder Singh. Proper value equivalence. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[20] 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*, 588(7839):604–609, 2020.

[21] Amir-massoud Farahmand. Iterative value-aware model learning. In *Advances in Neural Information Processing Systems*, volume 31, pages 9072–9083, 2018.

[22] Nathan Lambert, Brandon Amos, Omry Yadan, and Roberto Calandra. Objective mismatch in model-based reinforcement learning. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pages 761–770. PMLR, 2020.

[23] 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.

[24] Haoyang Cao and Samuel N. Cohen. Identifiability in inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[25] David Silver, Satinder Singh, Doina Precup, and Richard S. Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021.

[26] 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*, volume 30, 2017.

[27] Jonathan G. Richens, Rory Beard, and Daniel H. Thompson. Counterfactual harm, 2022.

[28] Yoshua Bengio, Michael K. Cohen, Nikolay Malkin, Matt MacDermott, Damiano Fornasiere, Pietro Greiner, and Younesse Kaddar. Can a bayesian oracle prevent harm from an agent?, 2025.

[29] Lukas Brunke, Melissa Greeff, Adam W. Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P. Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning, 2021.

[30] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction, 2018.

[31] Sebastian Farquhar, Ryan Carey, and Tom Everitt. Path-specific objectives for safer agent incentives, 2022.

[32] Francis Rhys Ward, Francesco Belardinelli, Francesca Toni, and Tom Everitt. Honesty is the best policy: Defining and mitigating ai deception, 2023.

[33] Francis Rhys Ward, Matt MacDermott, Francesco Belardinelli, Francesca Toni, and Tom Everitt. The reasons that agents act: Intention and instrumental goals, 2024.

[34] 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*, 9(4), 1996.

A Proofs

A.1 Proof of Theorem 1(a)

Theorem 1 (Finite)**.** *Consider the class of finite MDPs, and assume $|\mathcal{G}| = n$ is diverse.*

*(a) If Q-values are **exact**, the set of MDPs for which $P \Leftrightarrow Q$ has full Lebesgue measure. More precisely, M^π is invertible for almost every (i) discount factor $\gamma \in [0, 1)$, (ii) transition function P, (iii) set of reward functions R, and (iv) policy π. Moreover, M^π is invertible for all $\gamma < 1$ sufficiently small / large, and for any optimal entropy-regularised policy where the entropy coefficient $\alpha > 0$ is sufficiently small / large.*

*(b) If Q-values are ϵ-**approximate** and M^π is invertible, the estimator $\hat{P}_{ij} = (\hat{M}^\pi)^{-1}\hat{Q}_{ij}$ obtained from the approximate matrix $\hat{M}^\pi = R + \gamma\hat{V}^\pi$ satisfies*

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

for all $\epsilon < 1/\gamma m\|(M^\pi)^{-1}\|_1$.

We split Theorem 1 into two parts, Proposition 1 and Proposition 2. To state the result formally, we let $\mathcal{P}_\Delta := \{P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})\}$ 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 $\mathcal{R}^n = \{R : \mathcal{S} \times \mathcal{G} \to \mathbb{R} \mid |\mathcal{G}| = n\}$ for the space of n-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 $M^*(P) = M^{\pi_l^*(P)}(P)$ the matrix corresponding to the optimal policy $\pi_l^*(P)$; similarly for $M^*(\gamma)$ and $M^*(R)$.

Proposition 1 (Finite, Exact)**.** *Assume $|\mathcal{G}| = n$ is diverse.*

(a) For any $(\cdot, R, \gamma, \pi)$, the set $\{P \in \mathcal{P}_\Delta \mid \det(M^\pi(P)) = 0\}$ has Lebesgue measure zero. So does the optimal-policy set $\{P \in \mathcal{P}_\Delta \mid \det(M^*(P)) = 0\}$.

(b) For any $(P, \cdot, \gamma, \pi)$, the set $\{R \in \mathcal{R}^n \mid \det(M^\pi(R)) = 0\}$ has Lebesgue measure zero. So does the optimal-policy set $\{R \in \mathcal{R}^n \mid \det(M^*(R)) = 0\}$.

(c) For any $(P, R, \cdot, \pi)$, the set $\{\gamma \in [0, 1) \mid \det(M^\pi(\gamma)) = 0\}$ is finite. So is the optimal-policy set $\{\gamma \in [0, 1) \mid \det(M^*(\gamma)) = 0\}$. Moreover, $M^\pi(\gamma)$ and $M^*(\gamma)$ are invertible for all γ sufficiently small / large.

(d) For any $(P, R, \gamma, \cdot)$ the set $\{\pi \in \Pi \mid \det(M^\pi) = 0\}$ has Lebesgue measure zero. Moreover, the set $\{\alpha \in [0, \infty) \mid \det(M^{\pi_\alpha^*}) = 0\}$ is finite, where π_α^* is the optimal entropy-regularised policy with entropy coefficient α. Moreover, $M^{\pi_\alpha^*}$ is invertible for all $\alpha > 0$ sufficiently small / large.

Proof. We first prove a resolvent identity for each row of M^π. Using Equation (2), the matrix

$$M_{lk} = R_{lk} + \gamma V_{lk}^{\pi_l}$$

is independent from i, j. Defining the matrix $P^{\pi_l}_{kk'} \coloneqq \sum_j \pi_{lkj} P_{kjk'}$ for each l, we expand $V = \pi Q = \pi M P$ to obtain

$$\begin{aligned} M_{lk} &= R_{lk} + \gamma \sum_{j,k'} \pi_{lkj} P_{kjk'} M_{lk'} \\ &= R_{lk} + \gamma \sum_{k'} P^{\pi_l}_{kk'} M_{lk'} \,. \end{aligned} \tag{8}$$

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 no longer a simple matrix product. However, the lth row of M can still be written as a matrix product:

$$M_l = R_l \left(I - \gamma \bar{P}^{\pi_l}\right)^{-1} \,, \tag{9}$$

noting that $\bar{P}^{\pi_l}$ is a column-stochastic matrix for each l. Though each row is extracted from an invertible matrix (scaled by R_l), the resulting concatenation M may be singular. We now prove almost-everywhere invertibility over each of the four variables. 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 (9) 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

$$M_{lk} = \frac{\sum_{k'} R_{lk'} C^{(l)}_{k'k}}{\det(I - \gamma \bar{P}^{\pi_l})} \,. \tag{10}$$

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 the determinant of M is given by

$$\det(M) = \sum_{\sigma} \operatorname{sgn}(\sigma) \prod_{l=1}^{n} M_{l\sigma(l)} = \frac{\sum_{\sigma} \operatorname{sgn}(\sigma) \prod_{l=1}^{n} \left(\sum_{k'} R_{lk'} C^{(l)}_{k'\sigma(l)}\right)}{\prod_{l=1}^{n} \det(I - \gamma \bar{P}^{\pi_l})} \,,$$

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

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

Finally note that $\det(M)$ cannot be uniformly zero, since $M = R$ for $\gamma = 0$, and goal diversity implies $\operatorname{rank}(R) = n$, hence $\det(R) \neq 0$. In particular, the set of roots of the non-zero rational function $f \coloneqq \det(M^\pi(\cdot)) : [0, 1) \to \mathbb{R}$ is finite and has cardinality bounded above by the degree of its numerator, so we conclude that $\{\gamma \in [0, 1) \mid \det(M^\pi(\gamma)) = 0\}$ has cardinality $C \leq n(n-1)$.

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

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

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

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

hence

$$\{\gamma \in [0, 1) \mid \det(M^*(\gamma)) = 0\} \subseteq \bigcup_{\pi \in \Pi^\circ} \{\gamma \in [0, 1) \mid \det(M^\pi(\gamma)) = 0\} \,.$$

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

$$\begin{aligned} C^* &= \Big|\{\gamma \in [0,1) \mid \det(M^*(\gamma)) = 0\}\Big| \\ &\leq \Big|\bigcup_{\pi\in\Pi^\circ} \{\gamma \in [0,1) \mid \det(M^\pi(\gamma)) = 0\}\Big| \\ &\leq \sum_{\pi\in\Pi^\circ} \Big|\{\gamma \in [0,1) \mid \det(M^\pi(\gamma)) = 0\}\Big| \leq m^{n^2} n(n-1)\,. \end{aligned}$$

In particular, since the zeroes are finite, $M^\pi(\gamma)$ and $M^*(\gamma)$ are invertible for all $\gamma < 1$ sufficiently large. Moreover, since M^π is continuous in γ and $M^\pi(0) = M^*(0) = R$ with non-zero determinant, $M^\pi(\gamma)$ and $M^*(\gamma)$ are invertible for all $\gamma \geq 0$ sufficiently small.

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 (10) guarantees that each entry $M_{lk}(\cdot)$ is a rational function in the entries of P, so $f \coloneqq \det(M(\cdot)) : \mathcal{F}_\Delta \to \mathbb{R}$ 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 (9),

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

Now consider the set $X = \{x \in \mathbb{R}^{n-1} \mid x_i \geq 0, \sum_i x_i \leq 1\} \subset \mathbb{R}^{n-1}$ and the bijective map $g : X \to \Delta^n$ given by $g(x) = (x_1, \ldots, x_{n-1}, 1 - \sum_i x_i)$. 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

$$\lambda_\Delta(A) \coloneqq \lambda(g^{-1}(A)) \tag{11}$$

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

$$\lambda(h^{-1}(0)) = 0\,. \tag{12}$$

We conclude that the bad set $B = \{P \in \mathcal{F}_\Delta \mid \det(M(P)) = 0\}$ satisfies

$$\lambda_\Delta(B) \;= \lambda_\Delta(f^{-1}(0)) \overset{(11)}{=} \lambda(G^{-1}(f^{-1}(0))) \;=\; \lambda\big((f\circ G)^{-1}(0)\big) \;=\; \lambda(h^{-1}(0)) \overset{(12)}{=} 0\,.$$

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

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

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

$$\{P \in \mathcal{F}_\Delta \mid \det(M^*(P)) = 0\} \subseteq \bigcup_{\pi\in\Pi^\circ} \{P \in \mathcal{F}_\Delta \mid \det(M^\pi(P)) = 0\}\,.$$

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

$$\begin{aligned} \lambda_\Delta\big(\{P \in \mathcal{F}_\Delta \mid \det(M^*(P)) = 0\}\big) &\leq \lambda_\Delta\Big(\bigcup_{\pi\in\Pi^\circ} \{P \in \mathcal{F}_\Delta \mid \det(M^\pi(P)) = 0\}\Big) \\ &\leq \sum_{\pi\in\Pi^\circ} \lambda_\Delta\big(\{P \in \mathcal{F}_\Delta \mid \det(M^\pi(P)) = 0\}\big) = 0\,. \end{aligned}$$

Part (b). Now fix P, γ and π and take $M^\pi(\cdot)$ as a function of the reward $R \in \mathcal{R}^n$, written as a matrix $R_{lk} \in \mathbb{R}^{n\times n}$. Since each entry M_{lk} is linear in the entries of R by Equation (9), the function $f \coloneqq \det(M^\pi(\cdot)) : \mathbb{R}^{n\times n} \to \mathbb{R}$ is a polynomial of degree n in the entries of R. To prove that $f \not\equiv 0$, consider the reward matrix R^0 with rows $R^0_l = e_l(I - \gamma\bar{P}^{\pi_l})$. Then

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

so $f(R^0) = \det(I) \neq 0$. The zero set of a non-zero polynomial has Lebesgue measure zero, so

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

The result for **(ii)** 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 all policies is identical to parts **(a-c)(i)**: $\det(M^\pi)$ is a rational function over the entries of π, and taking any goal-independent policy π_0 (eg the uniform policy) gives

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

by assumption of goal diversity, so $\det(M^{\pi_0}) \neq 0$. The pullback argument establishes zero measure.

The second claim for entropy-regularised policies π^*_α is less straightforward, because the matrix $M^{\pi^*_\alpha}$ is not rational in the entries of α. For each $\alpha > 0$, the soft Q-values $Q^\alpha \in \mathbb{R}^{n^2 m}$ are the unique fixed point of the soft Bellman operator $\mathcal{T}_\alpha$ from Lemma 1. Now let $f : (0, \infty) \to \mathbb{R}$ be defined by by $f(\alpha) = \det(M^{\pi^*_\alpha})$. By Lemma 2, the function $\alpha \mapsto Q^\alpha$ is analytic and definable in $\mathbb{R}_{\exp}$. The matrix $M^{\pi^*_\alpha}$ is obtained from Q^α and the soft-optimal policy

$$(\pi^*_\alpha)_{lij} = \frac{\exp(Q^\alpha_{lij}/\alpha)}{\sum_{j'} \exp(Q^\alpha_{lij'}/\alpha)}$$

using definable operations (polynomials, division, exp, log), and the determinant is a polynomial in the entries of $M^{\pi^*_\alpha}$, 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 $\det(M^{\pi_0}) \neq 0$. By continuity, $\lim_{\alpha\to\infty} f(\alpha) = \det(M^{\pi_0}) \neq 0$, so f is a non-zero, analytic and definable function. Since $\mathbb{R}_{\exp}$ is o-minimal [34, 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. Adding the single boundary point $\alpha = 0$ preserves finiteness:

$$\left|\{\alpha \in [0, \infty) \mid \det(M^{\pi^*_\alpha}) = 0\}\right| \subseteq \left|\{0\} \cup \{\alpha \in (0, \infty) \mid f(\alpha) = 0\}\right| \leq 1 + C < \infty\,.$$

In particular, since the zeroes are finite and isolated, $M^{\pi^*_\alpha}$ is invertible for $\alpha > 0$ sufficiently small and sufficiently large. □

A.1.1 Supporting Results

We prove supporting results for Proposition 1, starting with the soft Bellman equation.

Lemma 1 (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,

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

□

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 [34, 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 2 (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 1, $\mathcal{T}_\alpha$ 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 1, $\mathcal{T}_\alpha$ 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 $\|D_Q \mathcal{T}_\alpha\|_\infty \leq \gamma < 1$. Hence $D_Q F = I - D_Q \mathcal{T}_\alpha$ has eigenvalues $\lambda = 1 - \lambda(D_Q(\mathcal{T}_\alpha)) \geq 1 - \|D_Q \mathcal{T}_\alpha\|_\infty \geq 1 - \gamma > 0$, so it is invertible everywhere. In particular, for any (α_0, Q_0) such that $F(\alpha_0, Q_0) = 0$, the analytic implicit function theorem states that there is an analytic function $G_U : U \to \mathbb{R}^{n^2 m}$ such that $F(\alpha, G(\alpha)) = 0$ for all α in some neighbourhood U of α_0. Now 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 $(0,\infty)$. □

A.2 Proof of Theorem 1(b)

Proposition 2 (Finite, Approximate)**.** *Assume $|\mathcal{G}| = n$ diverse, ϵ-approximate Q-values and invertible M^π. Then for all $\epsilon < 1/\gamma m \|(M^\pi)^{-1}\|_1$, the approximate matrix $\hat{M} = R + \gamma \hat{V}$ is invertible, and the estimator $\hat{P}_{ij} = \hat{M}^{-1} \hat{Q}_{ij}$ satisfies*

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

When π is unconditional, the factor of m vanishes and $\|(M^\pi)^{-1}\|_1 \leq (1+\gamma)\|R^{-1}\|_1$. If goals are states ($\mathcal{G} = \mathcal{S}$), we have $\|R^{-1}\|_1 = 1$ and the bound improves to

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

Proof. Omitting π for convenience, the approximate matrix $\hat{M}$ can be decomposed as $\hat{M} = M + E$, where $E_{lk} = \gamma \sum_{j'} \pi_{kj'} (\hat{Q}_{lkj'} - Q_{lkj'})$. We first bound $\|E\|_1$:

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

Since M is invertible, $\hat{M} = M(I + M^{-1}E)$ is invertible if $\|M^{-1}E\|_1 \leq \|M^{-1}\|_1\|E\|_1 < 1$, which holds for $\epsilon < 1/(\gamma m\|M^{-1}\|_1)$. Now using $MP_{ij} = Q_{ij}$ and $\hat{M}\hat{P}_{ij} = \hat{Q}_{ij}$,

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

where the last line uses $\|P_{ij}\|_1 = 1$. Solving for $\|\hat{P}_{ij} - P_{ij}\|_1$:

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

When π is unconditional, factoring out $\pi_{lkj'} = \pi_{kj'}$ sharpens the bound to $\|E\|_1 \leq \gamma\epsilon$. Equation (9) gives $M = R(I - \gamma(P^\pi)^T)^{-1}$, implying $M^{-1} = (I - \gamma(P^\pi)^T)R^{-1}$ under assumption of goal diversity and thus $\|M^{-1}\|_1 = (1+\gamma)\|R^{-1}\|_1$, by row-stochasticity of P^π. When additionally $\mathcal{G} = \mathcal{S}$ (ie $R = I$), we obtain $\|M^{-1}\|_1 = 1 + \gamma$, implying the improved bound. □

A.3 Proof of Theorem 2

Theorem 2 (Deterministic)**.** *Consider the class of deterministic finite MDPs, and assume* $|\mathcal{G}| \geq 1$.

(a) If Q-values are ***exact****, the set of reward functions* $R \in \mathbb{R}^{|\mathcal{G}|\times n}$ *for which* $P \Leftrightarrow Q$ *via column-matching has full Lebesgue measure, for any policy and discount factor.*

(b) If Q-values are ϵ*-****approximate*** *and* M^π *is column-injective, the estimator* $\hat{f}(s,a) = \operatorname{argmin}_{s'} \|\hat{Q}(s,a) - \hat{M}^\pi(s')\|_1$ *satisfies* $\hat{f} = f$ *for all* $\epsilon < \Delta_1/2(1+\gamma m)$*.* [Add tightness.]

Proof in Appendix A.3.

Proof. Part (a) is implied by Theorem 3 below for $N = 1$. We prove part (b). Fix a state-action pair (s_i, a_j) and write $k^* := f(s_i, a_j)$ for the true successor. Equation (5) gives $Q_{ij} = M^\pi_{\cdot,k^*}$, so it suffices to show that $\hat{Q}_{ij}$ is closer to the true column k^* than to any other column k. The approximate matrix $\hat{M}$ satisfies $\hat{M}_{lk} - M_{lk} = \gamma\sum_{j'} \pi_{lkj'}(\hat{Q}_{lkj'} - Q_{lkj'})$, so for each column k:

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

By the triangle inequality, the distance from $\hat{Q}_{ij}$ to the true column satisfies

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

For any $k \neq k^*$, we use both standard and reverse triangle inequalities as follows:

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

where $\|M_{\cdot,k^*} - M_{\cdot,k}\|_1 \geq \Delta_1$ by definition of column separation. Recovery succeeds when the true column is strictly closer than any wrong column, ie $\epsilon(1 + \gamma m) < \Delta_1 - \epsilon(1 + \gamma m)$, 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 the improved bound $\epsilon < \Delta_1/2(1 + \gamma)$. □

A.4 Proof of Corollary 1

Corollary 1. *In a deterministic MDP with $|\mathcal{G}| = 1$, the following reward families guarantee unique recovery of the world dynamics via column-matching.*

(a) ***Dense / noisy rewards.*** *For any reward drawn from a density (including a standard state-goal $r_g(s) = \delta_{sg} + \epsilon_s$ initialised with any noise ϵ_s), $P \Leftrightarrow Q$ almost surely.*

(b) ***Gaussian rewards.*** *For any embedding $\varphi : \mathcal{S} \hookrightarrow \mathbb{R}^d$ (eg a gridworld), the reward $r_g(s) = \exp(-\|\varphi(s) - g\|^2/2\sigma^2)$ gives $P \Leftrightarrow Q$ for almost every $(g, \sigma) \in \mathbb{R}^d \times (0, \infty)$.*

Proof in Appendix A.4.

Proof. Since $|\mathcal{G}| = 1$, the policy is unconditional, and Equation (9) reduces to

$$M^\pi = r^\top G\,, \quad \text{where} \quad G \coloneqq (I - \gamma \bar{P}^\pi)^{-1}\,.$$

Column-injectivity requires $r^\top G_{\cdot,s} \neq r^\top G_{\cdot,s'}$ for all $s \neq s'$. Since $\bar{P}^\pi$ is column-stochastic with spectral radius 1, the matrix $I - \gamma \bar{P}^\pi$ is invertible for all $\gamma \in [0, 1)$, so G is invertible. In particular, the columns of G are pairwise distinct: if $G_{\cdot,s} = G_{\cdot,s'}$ for some $s \neq s'$, then $G(e_s - e_{s'}) = 0$ with $e_s - e_{s'} \neq 0$, contradicting invertibility. Defining $v_{ss'} \coloneqq G_{\cdot,s} - G_{\cdot,s'} \neq 0$, column-matching fails if and only if r lies in the union

$$\bigcup_{s \neq s'} H_{s,s'}\,, \quad \text{where} \quad H_{s,s'} \coloneqq \{r \in \mathbb{R}^n : r^\top v_{ss'} = 0\}\,.$$

Part (a). Each $H_{s,s'}$ is a hyperplane of codimension 1 (since $v_{ss'} \neq 0$), which has Lebesgue measure zero. A finite union of measure-zero sets has measure zero, so

$$\lambda\Big(\bigcup_{s \neq s'} H_{s,s'}\Big) \leq \sum_{s \neq s'} \lambda(H_{s,s'}) = 0\,.$$

Since $r \sim p$ is absolutely continuous with respect to λ and $\lambda\big(\bigcup_{s \neq s'} H_{s,s'}\big) = 0$,

$$\mathbb{P}_{r \sim p}\Big(P \not\Leftrightarrow Q\Big) = \mathbb{P}_{r \sim p}\left(r \in \bigcup_{s \neq s'} H_{s,s'}\right) = 0\,.$$

Part (b). For each pair $s \neq s'$, define $h_{ss'} : \mathbb{R}^d \times (0, \infty) \to \mathbb{R}$ by

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

Column-matching fails if and only if $h_{ss'}(g, \sigma) = 0$ for some $s \neq s'$. Since $\|\varphi(s_k) - g\|^2$ is a polynomial in g and $1/\sigma^2$ is real-analytic on $(0, \infty)$, each summand is a composition of real-analytic functions, so $h_{ss'}$ is real-analytic on $\mathbb{R}^d \times (0, \infty)$. To prove $h_{ss'} \not\equiv 0$, note that $v_{ss'} \neq 0$ implies there exists $k_0 \in [n]$ with $G_{k_0 s} \neq G_{k_0 s'}$. Set $g = \varphi(s_{k_0})$. By injectivity of φ, $\|\varphi(s_k) - \varphi(s_{k_0})\| > 0$ for all $k \neq k_0$, so $e^{-\|\varphi(s_k) - \varphi(s_{k_0})\|^2/2\sigma^2} \to 0$ as $\sigma \to 0^+$ (the exponent diverges to $-\infty$), while the k_0-th term has exponent 0. Therefore

$$h_{ss'}(\varphi(s_{k_0}), \sigma) \to G_{k_0 s} - G_{k_0 s'} \neq 0 \quad \text{as } \sigma \to 0^+\,,$$

so $h_{ss'} \not\equiv 0$. The zero set of a non-zero real-analytic function has Lebesgue measure zero, so $\lambda(h_{ss'}^{-1}(0)) = 0$ for each pair $s \neq s'$. Column-matching fails only if (g, σ) lies in $\bigcup_{s \neq s'} h_{ss'}^{-1}(0)$, a finite union of measure-zero sets, hence

$$\lambda\Big(\{(g, \sigma) \in \mathbb{R}^d \times (0, \infty) \mid P \not\Leftrightarrow Q\}\Big) \leq \sum_{s \neq s'} \lambda\big(h_{ss'}^{-1}(0)\big) = 0\,. \qquad \square$$

A.5 Proof of Theorem 3

Theorem 3 (Sparse). *Consider the class of N-sparse finite MDPs, and assume $|\mathcal{G}| \geq N-1$ or $|\mathcal{G}| \geq \min(2N-1, n-1)$.*

(a) Known support: For any $|\mathcal{G}| \geq N-1$, the set of rewards $R \in \mathbb{R}^{|\mathcal{G}|\times n}$ for which $P \Leftrightarrow Q$ has full Lebesgue measure, for any policy and discount factor.

(b) Unknown support: For any $|\mathcal{G}| \geq \min(2N-1, n-1)$, the set of rewards $R \in \mathbb{R}^{|\mathcal{G}|\times n}$ for which $P \Leftrightarrow Q$ has full Lebesgue measure, for any policy and discount factor.

Proof in Appendix A.5.

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

Part (a). When $\mathrm{supp}(i,j)$ is known, the Bellman equation reduces to a small linear system per state-action pair. Writing

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

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

$$\mathbf{M}^\pi_{ij} P_{ij} = \mathbf{Q}_{ij} \tag{13}$$

for all i,j, where $P_{ij} \coloneqq P_{i,j,\mathrm{supp}(i,j)} \in \Delta^N$. In particular, P is uniquely determined if $\mathrm{rank}(\mathbf{M}^\pi_{ij}) \geq N$. The proof now follows similarly to that of Theorem 1**(b)**. First recall Equation (9):

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

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 is uniquely determined by Equation (13) if $\mathrm{rank}(\mathbf{M}^\pi_{ij}) \geq N$ for all i,j. Taking the converse statement,

$$\begin{aligned}
\{R \in \mathbb{R}^{L\times n} \mid P \not\Leftrightarrow Q\} &\subseteq \{R \in \mathbb{R}^{L\times n} \mid \exists i,j \text{ s.t. } \mathrm{rank}(\mathbf{M}^\pi_{ij}) < N\} \\
&= F^{-1}\Big(\{M \in \mathbb{R}^{L\times n} \mid \exists i,j \text{ s.t. } \mathrm{rank}(\mathbf{M}_{ij}) < N\}\Big) \\
&= F^{-1}\Big(\bigcup_{i,j} \{M \in \mathbb{R}^{L\times n} \mid \mathrm{rank}(\mathbf{M}_{ij}) < N\}\Big).
\end{aligned}$$

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 $\mathrm{rank}(\mathbf{M}_{ij}) = \min(L+1, N) = N$ almost surely, for all $L \geq N-1$. In particular,

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

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

$$\lambda\left(\{R \in \mathbb{R}^{L\times n} \mid P \not\Leftrightarrow Q\}\right) = 0.$$

Part (b). P 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 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 $\mathrm{rank}(\mathbf{M}_{:,K}) \geq 2N$. 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 is uniquely determined if $\operatorname{rank}(\mathbf{M}^{\pi}_{:,K}) \geq 2N$ for all $K \subset [n]$ with $|K| = 2N$. Taking the converse statement,

$$\begin{aligned}\{R \in \mathbb{R}^{L\times n} \mid P \not\Leftrightarrow Q\} &\subseteq \{R \in \mathbb{R}^{L\times n} \mid \exists |K| = 2N \text{ s.t. } \operatorname{rank}(\mathbf{M}^{\pi}_{:,K}) < 2N\} \\ &= F^{-1}\Big(\{M \in \mathbb{R}^{L\times n} \mid \exists |K| = 2N \text{ s.t. } \operatorname{rank}(\mathbf{M}_{:,K}) < 2N\}\Big) \\ &= F^{-1}\left(\bigcup_{|K|=2N} \{M \in \mathbb{R}^{L\times n} \mid \operatorname{rank}(\mathbf{M}_{:,K}) < 2N\}\right).\end{aligned}$$

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},

$$\lambda\left(\{R \in \mathbb{R}^{L\times n} \mid P \not\Leftrightarrow Q\}\right) = 0\,. \qquad \square$$

A.6 Proof of Theorem 4

Theorem 4. *Assume the MDP is deterministic and consider the family of Gaussian rewards $r_g(s') = \exp(-\|s'-g\|^2/(2\sigma^2))$ with fixed $\sigma > 0$. For any $\mathcal{S} \subseteq \mathbb{R}^d$, $\gamma \in [0,1)$, unconditional π and any **finite** number of goals $|\mathcal{G}| \geq 2d+1$, the set of tuples $(g_1, \ldots, g_{|\mathcal{G}|}) \in \mathcal{S}^{|\mathcal{G}|}$ for which the Bellman family $\{m^{\pi}_{g_i}\}$ is measure determining on $\mathcal{S}$ has full Lebesgue measure in $\mathcal{S}^{|\mathcal{G}|}$. In particular, the transition function f is almost surely identified from (π, Q^{π}).* [Add GC conjecture.] *Proof in Appendix A.6.*

For deterministic environments, measure determination reduces to point separation.

Definition 2 (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 $f(x) \neq f(y)$.

If the MDP is deterministic and the family $\{m^{\pi}_g : g \in \mathcal{G}\}$ is point separating on $\mathcal{S}$, then the transition function f is uniquely identified by (π, Q^{π}): any f, f' satisfying the Bellman equation

$$m^{\pi}_g\big(f(s,a)\big) = Q^{\pi}(s,a,g) = m^{\pi}_g\big(f'(s,a)\big) \qquad \forall g, s, a$$

must have $f(s,a) = f'(s,a)$ for every (s,a), so f is unique. We now prove Theorem 4.

Proof. **(i) Integral representation.** Let $P^{\pi}(A \mid s) \coloneqq \int \mathbf{1}[f(s,a) \in A]\pi(\mathrm{d}a \mid s)$ for measurable $A \subseteq \mathbb{R}^d$ denote the transition kernel with respect to π. Since π is unconditional, P^{π} is independent of g. For each $x \in \mathcal{S}$, we define the discounted occupancy measure

$$\mu_x(\cdot) \coloneqq \sum_{t=0}^{\infty} \gamma^t (P^{\pi})^t(x, \cdot)\,,$$

a finite non-negative Borel measure with total mass $1/(1-\gamma)$. Since $0 \leq r_g \leq 1$, we apply Tonelli's theorem to express $m^{\pi}_g(x)$ as:

$$m^{\pi}_g(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_g(s_t) \,\middle|\, s_0 = x\right] = \sum_{t=0}^{\infty} \gamma^t \int r_g(s)\,(P^{\pi})^t(x, \mathrm{d}s) = \int r_g(s)\mu_x(\mathrm{d}s)\,. \tag{14}$$

(ii) Real-analyticity in g. For each $s' \in \mathbb{R}^d$, the map $g \mapsto r_g(s') = \exp(-\|s'-g\|^2/(2\sigma^2))$ extends to an entire function of $g \in \mathbb{C}^d$. For any $z \in \mathbb{C}^d$ with $\|\operatorname{Im}(z)\|_{\infty} \leq \sigma$, we have $\operatorname{Re}(\|s'-z\|^2) = \|s' - \operatorname{Re}(z)\|^2 - \|\operatorname{Im}(z)\|^2 \geq -d\sigma^2$, so $|r_z(s')| \leq e^{d/2}$ uniformly over all $s' \in \mathbb{R}^d$. Applying Cauchy's estimate on a polydisc of radius σ around any $g \in \mathbb{R}^d$,

$$\sup_{s' \in \mathbb{R}^d} \left|\partial^{\alpha}_g r_g(s')\right| \leq \frac{\alpha!}{\sigma^{|\alpha|}} e^{d/2}$$

for all multi-indices $\alpha \in \mathbb{N}^d$. Since this bound is uniform in s', dominated convergence justifies differentiating Equation (14) under the integral, and the triangle inequality gives $|\partial_g^\alpha m_g^\pi(x)| \leq A_K B_K^{|\alpha|} \alpha!/(1-\gamma)$. Since a smooth function whose derivatives satisfy $|\partial^\alpha f| \leq AB^{|\alpha|}\alpha!$ on every compact set has a convergent Taylor series, $g \mapsto m_g^\pi(x)$ is real-analytic on $\mathbb{R}^d$ for each $x \in \mathcal{S}$.

(iii) Point separation (fixed pairs). For each $(x, y) \in \mathcal{S} \times \mathcal{S}$ with $x \neq y$, define $h_{x,y}(g) \coloneqq m_g^\pi(x) - m_g^\pi(y)$, which is real-analytic in g by **(ii)**. By Equation (14), $h_{x,y}(g) = \int r_g(s)\,(\mu_x - \mu_y)(\mathrm{d}s)$. Now assume for contradiction that $h_{x,y} \equiv 0$. Then $\nu \coloneqq \mu_x - \mu_y$ is a finite signed Borel measure, and write $\rho(u) \coloneqq \exp(-\|u\|^2/(2\sigma^2))$, so that $r_g(s) = \rho(g - s)$ and thus $(\nu * \rho)(g) = \int \rho(g - s)\nu(\mathrm{d}s) = 0$ for all $g \in \mathbb{R}^d$. Taking Fourier transforms gives

$$\widehat{\nu}(\omega)\,\widehat{\rho}(\omega) = 0\,, \qquad \widehat{\rho}(\omega) = (2\pi\sigma^2)^{d/2} \exp(-\sigma^2\|\omega\|^2/2) > 0\,,$$

so $\widehat{\nu} \equiv 0$, and by injectivity of the Fourier transform on finite Borel measures, $\nu \equiv 0$, i.e. $\mu_x = \mu_y$. But the Neumann series $\mu_x = \sum_{t\geq 0} \gamma^t (P^\pi)^t(x, \cdot)$ satisfies $\mu_x - \gamma\mu_x P^\pi = (P^\pi)^0(x, \cdot) = \delta_x$ by telescoping (and similarly $\mu_y - \gamma\mu_y P^\pi = \delta_y$), so $\mu_x = \mu_y$ gives $\delta_x = \delta_y$, hence $x = y$, a contradiction. Therefore $h_{x,y} \not\equiv 0$, and the zero set

$$Z_{x,y} \coloneqq \{g \in \mathbb{R}^d : h_{x,y}(g) = 0\}$$

is a proper real-analytic subset of $\mathbb{R}^d$ with Hausdorff dimension at most $d - 1$.

(iv) Point separation (all pairs). The Bellman family $\{m_{g_i}^\pi\}_{i=1}^L$ fails to be point separating iff there exist $x \neq y$ in $\mathcal{S}$ with $h_{x,y}(g_i) = 0$ for all i, or equivalently, $g \in Z_{x,y}^L$. In particular, it is sufficient to show that the bad set

$$\mathcal{B} \coloneqq \{g \in (\mathbb{R}^d)^L \mid \exists\, x \neq y \in \mathcal{S} \text{ with } g \in Z_{x,y}^L\}$$

has Lebesgue measure zero for $L \geq 2d+1$. To handle the uncountable union over pairs (x, y), define

$$M_n \coloneqq \big\{(x, y) \in \mathcal{S} \times \mathcal{S} : \|x - y\| \geq 1/n\big\}\,,$$

so that $\mathcal{B} = \bigcup_{n\geq 1} \mathcal{B}_n$ with $\mathcal{B}_n \coloneqq \{g \in (\mathbb{R}^d)^L \mid \exists\,(x, y) \in M_n \text{ with } g \in Z_{x,y}^L\}$. Now define the incidence set

$$E_n \coloneqq \big\{((x, y), g) \in M_n \times (\mathbb{R}^d)^L : g \in Z_{x,y}^L\big\}\,,$$

whose fiber over (x, y) is $Z_{x,y}^L$ and whose projection onto $(\mathbb{R}^d)^L$ is $\mathcal{B}_n$. Since $\dim_{\mathrm{H}}(Z_{x,y}) \leq d - 1$ by **(iii)** and $\dim_{\mathrm{H}}(A \times B) \leq \dim_{\mathrm{H}}(A) + \dim_{\mathrm{H}}(B)$, we have $\dim_{\mathrm{H}}(Z_{x,y}^L) \leq L(d - 1)$. By the Hausdorff dimension fiber bound [ref],

$$\dim_{\mathrm{H}}(E_n) \leq \dim_{\mathrm{H}}(M_n) + \sup_{(x,y)\in M_n} \dim_{\mathrm{H}}(Z_{x,y}^L) \leq 2d + L(d - 1)\,.$$

Since projection is 1-Lipschitz and Hausdorff dimension does not increase under Lipschitz maps,

$$\dim_{\mathrm{H}}(\mathcal{B}_n) \leq \dim_{\mathrm{H}}(E_n) \leq 2d + L(d - 1) < Ld = \dim((\mathbb{R}^d)^L)$$

for all $L \geq 2d + 1$, so $\mathcal{B}_n$ has Lebesgue measure zero. Since $\mathcal{B} = \bigcup_{n\geq 1} \mathcal{B}_n$ is a countable union of measure-zero sets, $\lambda(\mathcal{B}) = 0$. For any $g \in \mathcal{S}^L \setminus \mathcal{B}$, the family $\{m_{g_i}^\pi\}_{i=1}^L$ is point separating, so f is uniquely identified, and we conclude

$$\lambda\big(\{g \in \mathcal{S}^L \mid \{m_{g_i}^\pi\} \text{ is not point separating}\}\big) \leq \lambda(\mathcal{B}) = 0\,. \qquad \square$$