Extracting World Models From Q-values

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

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

Introduction

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 is a safety threat, preventing us from repurposing them for new goals and safety constraints without retraining from scratch, or querying their beliefs about the consequences of actions independently from the goal. We prove the conditions under which it is possible to reverse-engineer P from Q-values associated to one or many known reward functions, introduce practical methods to do so, and empirically demonstrate this in (1) stochastic gridworld environments, enabling agents to plan for novel safety constraints without retraining, and (2) continuous control tasks, where P can be recovered with high fidelity even when Q-values are inaccurate.

Full paper

Theoretical Results

Our main results are Theorems 1, 2, 3 below. Further results are summarised in Table 1 (for finite MDPs) and 2 (for continuous MDPs), where P ⇔ Q means that P can be extracted from Q.

Theorem 1 (Stochastic Finite MDPs)
The world model P can almost surely be extracted from exact Q-values if |G| >= |S|. With ε-approximate Q-values, recovery error is O(ε).

Theorem 2 (Deterministic Finite MDPs)
The world model can almost surely be extracted using a **single** generic goal |G| = 1. Dense rewards implicitly force Q-learners to encode world models, while sparse rewards may not.

Theorem 3 (Deterministic Continuous MDPs)
For $S \subseteq \mathbb{R}^d$, the world model can almost surely be extracted using a **finite** number of Gaussian goals |G| = 2d+1, despite the infinite state space.

Table 1: Number of goals for which $P \Leftrightarrow Q$ in finite MDPs (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

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

	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	$\lvert\mathcal{G}\rvert \geq 2d+1$ (Theorem 4)	$\mathcal{G} \supseteq \mathcal{S}$	$\mathcal{G} = \mathcal{S}$

Method

We rewrite the goal-conditioned Bellman equation as

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

to obtain, in tabular settings with invertible m, the **inverse Bellman equation**

$$P(\cdot \mid s,a) = (M^\pi)^{-1} Q_\cdot^\pi(s,a)$$

In practice (for large/continuous spaces), we parameterise a world model as a neural network $P_\varphi : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ and minimise the loss

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

In many environments of interest – particularly robotics and continuous control – we can take advantage of dynamics being deterministic. In this case, P is a successor function f : S x A → S and the loss reduces to

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

Experiment 1 (Finite MDP)

We train a model-free (PQN) agent on a gridworld of four rooms with 68 states, and extract the world model from their Q-values based on 30, 35 and 40 goals. The recovered world model matches the ground truth perfectly for 40 goals, 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 fewer goals, and using it for planning results in dangerous (for 30 goals) or suboptimal behaviour (for 35 goals).

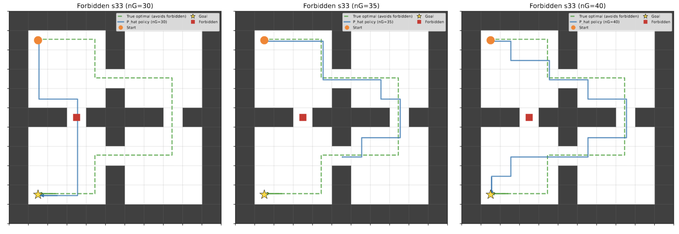
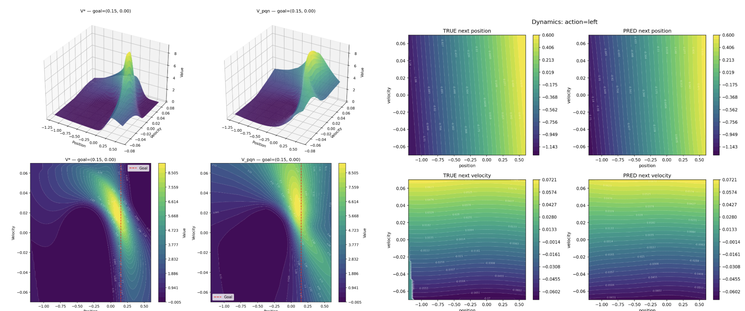


Figure 2: Gridworld – trained agent planning for novel safety constraints using the world model extracted from its Q-values based on $L = 30$ (left), $L = 35$ (middle) and $L = 40$ (right). The agent plans optimally using the extracted world model for $L = 40$, reaching the goal in the optimal number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

Experiment 2 (continuous MDP)

We train a model-free (PQN) agent the continuous control environment MountainCar and extract the world model their Q-values based on 2d+1=5 goals, in accordance with our theoretical results. 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 environment boundaries $p = -1.2$ and $p = 0.6$). The shape is approximately correct but the mean squared error is relatively large, MSE $= 1.79$ all 5 goals. (b) Recovered world model (right) vs true dynamics (left) for one of the three actions. The recovered dynamics match ground truth with high fidelity, MSE $= 0.00013$ across all actions.

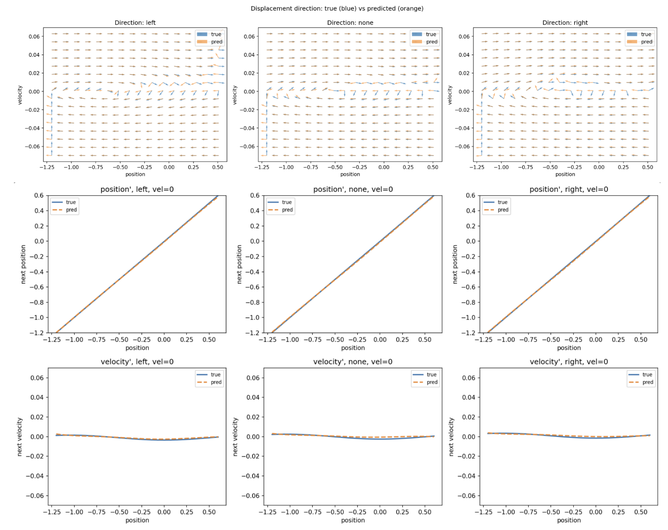


Figure 4: MountainCar results at the end of training (10M steps). (Top) Quiver plot showing the predicted vs true *direction* of the next state (ignoring magnitude) given current state (position on x-axis, velocity on y-axis). (Middle + Bottom) Comparison of predicted vs true position (middle) / velocity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