Uniqueness and Complexity of Inverse MDP Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

What action sequence $aa'a''$ was likely responsible for reaching state s''' (from state s) in 3 steps? Addressing such questions is important in causal reasoning and in reinforcement learning. Inverse "MDP" models $p(aa'a''|ss''')$ can be used to answer them. In the traditional "forward" view, transition "matrix" $p(s'|sa)$ and policy $\pi(a|s)$ uniquely determine "everything": the whole dynamics $p(as'a's''a''...|s)$, and with it, the action-conditional state process $p(s's''...|saa'a'')$, the multi-step inverse models $p(aa'a''...|ss^i)$, etc. If the latter is our primary concern, a natural question, analogous to the forward case is to which extent 1-step inverse model $p(a|ss')$ plus policy $\pi(a|s)$ determine the multi-step inverse models or even the whole dynamics. In other words, can forward models be inferred from inverse models or even be side-stepped. This work addresses this question and variations thereof, and also whether there are efficient decision/inference algorithms for this.

Keywords

inverse models; reinforcement learning; causality; theory; multi-step models; reasoning.

1 Introduction

Consider an MDP with actions $a\in\{0,..,k-1\}$ and states $s\in\{1,...,d\}$. Rewards play no role in our analysis, so *controlled Markov process* [DY79] or *conditional Markov chain* may be a more apt naming. Transition "matrix" $p(s'|sa)$ (*"Forward model"*) and policy $\pi(a|s)$ uniquely determine the whole *dynamics*

$$p(as'a's''a''...|s) \;=\; \pi(a|s)\cdot p(s'|sa)\cdot\pi(a'|s')\cdot p(s''|s'a')\cdot... \tag{1}$$

and also determines the action-conditional state process ("Multi-Step Forward Model"):

$$p(s's''...|saa'a'') \;=\; p(as'a's''a''...|s)/\sum_{s's''...} p(as'a's''a''...|s) \tag{2}$$

Here we consider *Inverse Model* $p(a|ss')$ and *Multi-Step Inverse Models* $p(aa'a''...|ss's''s'''...)$ and $p(a|ss^i)$ and variations thereof. (Inverse MDP models should not be confused with inverse reinforcement learning [AD21], which infers rewards, which play no role here.)

Motivation. One motivation to consider inverse models is causal inference: An inverse model captures the likelihood that an action a was the cause of the transition from state s to state s'. A multi-step inverse model captures the likelihood that a first action a or action sequence $aa'...a^{i-1}$ was the cause of the state sequence $ss'...s^i$ or the cause of the transition from state s to state s^i. The latter is the primary goal in (automatic/stochastic) planning [HSHB99]: to find an action sequence that leads to a desired goal state $s^i=s_{\text{goal}}$. The shortest path, i.e. smallest i, that reaches s_{goal} (with high probability in the stochastic case) can easily be found via a trivial search over $i=1,2,3,...$ if the fixed-i planning problem can be solved efficiently.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Another machine-learning motivation is that inverse models may be substantially smaller than forward models. For instance, an action-independent Markov process $p(s'|sa) = p(s'|s)$ may be very complex for large d, but for a state-independent (known) policy $\pi(a|s) = \pi(a)$, the inverse model $p(aa'...|s..s''..) = \pi(a)\pi(a')...$ is trivial (and known). Of course this extreme case is uninteresting, but a partial similar simplification happens if state s decomposes into $s = (\dot{s},\ddot{s})$ [EMK+22]. In this case, if the forward model $p(s'|sa)$ factors into a (simple) controlled $p(\dot{s}'|\dot{s}a)$ and (complex) uncontrolled $p(\ddot{s}'|\ddot{s})$, and the policy $\pi(a|s) = \pi(a|\dot{s})$ only depends on (small) $\dot{s}$, then $p(aa'...|s..s''..) = p(aa'...|\dot{s}..\dot{s}''..)$ is independent of (large) $\ddot{s}$. Note that this simplification happens "automatically". We do not need to know the factorization structure, say $(\dot{s},\ddot{s}) = f(s)$ for some unknown f. Appendix I contains a bit of practical context/motivation/application.

Main questions.

The main question we consider here is:
to which extent do inverse model $p(a|ss')$ plus policy $\pi(a|s)$
determine the multi-step inverse model or even the whole dynamics.

For instance, do $p(a|ss')$ plus $\pi(a|s)$ determine

(i) the full dynamics (1),
(ii) the full dynamics, if also $p(aa'|ss'')$ is provided,
(iii) the multi-step inverse model $p(aa'...|ss^i)$ (or $p(aa'...|ss's''...)$),
(iv) the multi-step inverse model $p(aa'...|ss^i)$ (or $p(aa'...|ss's'')$), if also $p(a|ss'')$ is provided,
(v) just the initial action $p(a|ss'')$ from just final state s'',
(vi) $p(a|ss^i)$ if also $p(a|ss'')$ is provided,

and variations thereof? Also, is there an efficient algorithm that can decide whether the solution is unique and/or computes any or all of them?

Unlike in the "forward" case (1), the answer to all these questions is 'complicated' and 'sometimes'. For instance, (i) is true iff $k \geq d$ and $p(s'|sa)$ has full rank. (ii) seems true for "most" transition matrices. (iii-vi) can fail, but (iv) and (vi) seem to hold for interesting cases. In some situations there are efficient algorithms which sometimes work.

Related work. There is of course abundant literature on causal reasoning in general [PGJ16], and in the modern context of Deep Learning in particular [OKD+21], but to the best of our knowledge, the setup and questions we are asking are novel, at least in this generality and rigor.

A special case of our setup is considered in [EMK+22]. The authors consider Exogenous Block MDPs (EX-BMDPs) which correspond to the motivating decomposition example above, and formalized in Section 3 as tensor-product MDPs. Additionally they assume episodic MDPs with near-deterministic dynamics. Their PPE algorithm finds action sequences of high inverse probability $p(aa'...a^{i-1}|ss^i)$ in polynomial time in $\dot{s}$ rather than s, while our aim is to infer higher- from lower-step inverse models for general MDPs.

In the context of Deep Learning, there is ample empirical work that would benefit from a positive answer to our main question: Variational Intrinsic Control [GRW17] and Diversity is All You Need [EGIL18] are representative of a broad class of methods that learn diverse options (policies / action sequences) that are inferrable from their effects on the environment. This relies on inverse modelling, as their mutual information objective is decomposed into maximizing skill/policy entropy and minimizing the entropy of an inverse model:

$$I(s^i;a...a^{i-1}|s) \equiv H(a...a^{i-1}|s) - H(aa'...a^{i-1}|ss^i)$$

This is akin to finding all action sequences of sufficiently high probability $p(aa'...a^{i-1}|ss^i)$, or all skills when the policy space is captured by an auxiliary variable $p(z|ss^i)$. The EDDICT algorithm by [HDB+21] also maximizes this objective, and parameterizes the requisite inverse models such that they yield forward predictions, but as detailed in Section 4 its unlikely that such models would compose in general. Dynamics-Aware Unsupervised Discovery of Skills [SGL+19] decomposes the mutual information in the opposite direction, so as to avoid learning an inverse model and instead relies on a conventional forward model. Uniting all of the above mentioned methods is that the action sequence/skill horizon i must be fixed a priori. Inferring long horizon inverse models from shorter ones (the topic of the present work) would allow all of these methods to circumvent this constraint.

A second stream of empirical work uses single-step inverse models for representation learning [BEP+18]. Agent57 is arguably the most prominent of these methods [BPK+20], and therein the authors note that this choice of representation limits the generality of their approach, as multi-step effects can be aliased over. Despite this being a known limitation, multi-step inverse models are not used as they are too cumbersome to effectively learn online. A positive result to our questions (iii) or (iv) would allow such methods to leverage multi-step inverse predictions despite only learning a single-step model.

Contents. In Section 2 we will formalize questions (i)-(vi) in matrix/tensor notation. Section 3 gives a first probe into these questions by considering various degenerate cases. In Section 4 we study the solvability and uniqueness questions (i),(iii),(v), when only B^a is given, i.e. the case $i=1$, in preparation for and showing the necessity of considering $i>1$. In Section 5 we provide a polynomial-time algorithm via linear relaxation that works under certain conditions. Section 6 provides some validation experiments on toy domains. Section 7 concludes. Appendix A contains a list of notation. The other appendices contain further considerations: an alternative derivation of characterizing when (i) holds (B), an explicit SQE representation (C) with empirical rank analysis (D), counter-examples to (ii),(iv),(vi), and indeed for all $i>1$, which is surprising given the severe over-determined nature of the problem (E,F,G), a brief discussion of some applications (I), deterministic cases (H), systems of quadratic matrix equations (J), a simple and self-contained instantiation (L) of the main open problem (M), the possible NP-hardness of this problem, even setting uniqueness aside (K), and further experiments (N).

2 Problem Formalization and Preliminaries

We now formalize our questions (i)-(vi) from the introduction, and for this purpose introduce some useful matrix notation. We are not aware of prior work addressing these questions, so quite some ground-work to suitably formalize the various question is needed, and many little results are derived or mentioned in passing to give better insight into the structure of the problem. To avoid clutter, we will not constantly point out edge cases or domain constraints. For instance quantities that represent probabilities are obviously non-negative and sum to one. The reader worried about divisions by 0 here and there should best assume that all probabilities are strictly positive, but most considerations/results naturally generalize with some care, e.g. adding "almost surely" w.r.t. to the joint distribution (1).

Notation. Capital letters $A,B,M,W,I,D,V,...$ are used for $d\times d$ matrices over $[0;1]\subset\mathbb{R}$ and tensors by adding further upper indices, e.g. $M^{\cdot}_{\cdot\cdot}$ is an order-3 tensor, and $M^a_{\cdot\cdot}$ a matrix for each $a\in\{0:k-1\}:=\{0,...,k-1\}$. We define Id to be the identity (eye) matrix $\mathrm{Id}_{ss'}:=\delta_{ss'}:=[\![s=s']\!]\ \forall s,s'\in\{1:d\}$, and I to be the all-one matrix $I_{ss'}=1\ \forall ss'$. We drop all-quantifiers $\forall s,s',...$ if clear from context. Let $\odot$ denote element-wise (Hadamard) multiplication ($[A\odot B]_{ss'}=A_{ss'}B_{ss'}$), and similarly $\oslash$, while (no) $\cdot$ represents (conventional) matrix multiplication and has operator preference over $\odot$ and $\oslash$. Matrices form a ring under conventional $(+,\cdot)$ and a commutative ring under $(+,\odot)$, but $(A\cdot B)\odot C\neq A\cdot(B\odot C)$. A diagonal matrix D has the property $D=D\odot\mathrm{Id}$, i.e. $D_{ss'}=D_{ss}[\![s=s']\!]$. $V:=I\cdot D$ is a matrix with D_{ss} in the whole of column s ($V_{ss'}=V_{*s'}=D_{s's'}$). Note that $A\cdot D=A\odot V$ ($[A\cdot D]_{ss''}=\sum_{s'}A_{ss'}D_{s's''}=A_{ss''}D_{s''s''}=A_{ss''}V_{*s''}=[A\odot V]_{ss''}$). Similar left-right reversed identities hold. $\perp$ denotes 'undefined'. See Appendix A for a full List of Notation.

Matrix/tensor formalization. We define

$$M^a_{ss'}\ :=\ p(as'|s)\ =\ \pi(a|s)p(s'|sa)$$

Marginalizing out the action, gives

$$p(s'|s)\ =\ \sum_a p(as'|s)\ =\ \sum_a M^a_{ss'}\ =:\ M^+_{ss'}$$

Marginalizing out the next-state, gives back

$$\pi(a|s)\ =\ \sum_{s'} p(as'|s)\ =\ \sum_{s'} M^a_{ss'}\ =:\ M^a_{s+}$$

For instance, the multi-step dynamics can be written as

$$p(as'a's''...|s)\ =\ p(as'|s)\cdot p(a's''|a')\cdot...\ =\ M^a_{ss'}M^{a'}_{s's''}...$$

Marginalizing out the intermediate states gives

$$p(aa'...a^{i-1}s^i|s) = [M^a \cdot M^{a'} ... \cdot M^{a^{i-1}}]_{ss^i}$$

The inverse MDP model can then be expressed as

$$B^a_{ss'} := p(a|ss') = p(as'|s)/p(s'|s) = M^a_{ss'}/M^+_{ss'} = [M^a \oslash M^+]_{ss'}$$

The multi-step inverse model given the whole state sequence becomes

$$p(aa'...|ss's''...) = \frac{p(as'|s)p(a's''|s')...}{p(s'|s)p(s''|s')...} = \frac{M^a_{ss'}M^{a'}_{s's''}...}{M^+_{ss'}M^+_{s's''}...} = p(a|ss')p(a'|s's'')... \tag{3}$$

and can easily be computed from the 1-step inverse models. To answer the primary question: which action sequence can lead to (desired) state s^i from state s, we need to marginalize out $s'...s^{i-1}$. For instance, the two-step inverse model from s to s'' with s' marginalized out becomes

$$B^{aa'}_{ss''} := p(aa'|ss'') = \frac{\sum_{s'} M^a_{ss'}M^{a'}_{s's''}}{\sum_{s'} M^+_{ss'}M^+_{s's''}} = [M^a \cdot M^{a'} \oslash (M^+)^2]_{ss''} \tag{4}$$

Note that unlike the forward case, $B^{aa'} \neq B^a \cdot B^{a'}$, which is responsible for all the problems we will face. Also $B^{a+} \neq B^a$ but $B^+ = 1 = B^{++}$. We always use brackets to denote and disambiguate (matrix) powers $()^2$ from upper indices M^a. The initial-action 2-step (and similarly i-step) inverse models follow from further marginalizing $a'a''...$:

$$\begin{aligned} B^{a+}_{ss''} &= p(a|ss'') = [M^a M^+ \oslash (M^+)^2]_{ss''}, \\ B^{a+^{i-1}}_{ss^i} &= p(a|ss^i) = [M^a (M^+)^{i-1} \oslash (M^+)^i]_{ss^i} \end{aligned} \tag{5}$$

With this notation, questions (i-vi) in the introduction can formally be written as

(i) Can M be inferred from $B^a := M^a \oslash M^+$?
(ii) Can M be inferred from B^a and $B^{aa'} := M^a M^{a'} \oslash (M^+)^2$?
(iii) Can $B^{aa'...a^i} := M^a M^{a'} ... M^{a^i} \oslash (M^+)^i$ be inferred from B^a?
(iv) Can $B^{aa'...a^i}$ be inferred from B^a and $B^{aa'}$?
(v) Can $B^{a+} := M^a M^+ \oslash (M^+)^2$ be inferred from B^a?
(vi) Can $B^{a++} := M^a (M^+)^2 \oslash (M^+)^3$ be inferred from B^a and B^{a+}?

Each question comes in two versions, given also π, or not knowing π. We mainly consider the former version, i.e. knowing M^a_{s+}:

$$\text{Constraint on } M \text{ for known } \pi\text{:} \quad M^a_{s+} = \pi(a|s) \quad \text{and in particular} \quad M^+_{s+} = 1 \tag{6}$$

Questions (i)-(vi) also have multiple variations:

(I) Assume some arbitrary B^a (and $B^{aa'}$) is given, but not defined via M. Is there no, exactly one, or multiple M consistent with these B?
(II) Is there an efficient algorithm that can decide the previous question?
(III) Is there an efficient algorithm that can compute any/all solutions if one/many exist, and halts/loops if not (4 non-trivial combinations of $/$).

Formulation of the uniqueness questions. Abstractly, these questions ask whether M (in case of (i-ii)) or $g(M)$ for some function g (in case of (iii-vi)) can be inferred from some other function $f(M)$. Let us define another MDP $q(s'|sa)$ with same policy $\pi(s|a)$ and shorthand

$$W^a_{ss'} := \pi(a|s)q(s'|sa)$$

(In applications, B^a would be learned from data, and W or $B^{aa'\cdots}$ inferred from B^a in the hope that $W \approx M$.) One way to rephrase the questions is whether $f(M) = f(W)$ implies $M = W$ or $g(M) = g(W)$ for all (or most or some) M and W. The condition that π is the same for p and q, translates to

$$\text{Constraint on } M \text{ and } W\text{:} \quad M^a_{s+} = \pi(a|s) = W^a_{s+} \quad \text{and in particular} \quad M^+_{s+} = 1 = W^+_{s+} \tag{7}$$

We name the two most interesting equation versions as follows:

$$\text{EqIM}(ia)\text{:} \quad B^{aa'...a^i} := M^a M^{a'} ... M^{a^i} \oslash (M^+)^i \stackrel{?}{=} W^a W^{a'} ... W^{a^i} \oslash (W^+)^i \tag{8}$$

$$\text{EqIM}(i+)\text{:} \quad B^{a+\cdots+} := M^a (M^+)^{i-1} \oslash (M^+)^i \stackrel{?}{=} W^a (W^+)^{i-1} \oslash (W^+)^i \tag{9}$$

We allow $M^+_{ss'} = 0$ and keep probabilistic convention that $p(a|ss') = \pi(a|s)p(s'|sa)/p(s'|s)$ is undefined iff $p(s'|s) = 0$ (see end of Appendix F for some more discussion). Formally, $B^a_{ss'} = \perp = 0/0$ iff $M^+_{ss'} = 0$, also $W^+_{ss'} = 0$ iff $M^+_{ss'} = 0$, and similarly for larger i.

3 Degenerative Cases

To get some feeling about why these questions are so more intricate than analogous ones in forward models, we consider some simple examples and special cases first. Some further special cases (deterministic planning, deterministic reachability, and deterministic inverse models) are considered in Appendix H.

Example violating (i,iii,v). A specific example for M and W which satisfy EqIM(1) but violate EqIM$(2+)$ and hence EqIM$(2a)$ is as follows:

$$M^0 = \tfrac{1}{4}\begin{pmatrix}0&2\\1&1\end{pmatrix},\quad M^1 = \tfrac{1}{4}\begin{pmatrix}2&0\\1&1\end{pmatrix},\quad W^0 = \tfrac{1}{2}\begin{pmatrix}0&1\\1&0\end{pmatrix},\quad W^1 = \tfrac{1}{2}\begin{pmatrix}1&0\\1&0\end{pmatrix}$$

which satisfies (7) ($M^a_{s+} = \frac{1}{2} = W^a_{s+}$). In this example, $M^+ = \frac{1}{2}\begin{pmatrix}1&1\\1&1\end{pmatrix}$ and $W^+ = \frac{1}{2}\begin{pmatrix}1&1\\2&0\end{pmatrix}$, which shows $M^a \oslash M^+ = W^a \oslash W^+$, except that $W^+_{22} = 0 \neq 1 = M^+_{22}$, hence there is one "dubious" $1 \stackrel{?}{=} 0/0$ case. A simple calculation shows that EqIM(2+) is violated (w/o any division by 0). The division by 0 can easily avoided by mixing $U^a_{ss'} \equiv \frac{1}{4}$ into M and W, e.g. $M \rightsquigarrow \frac{1}{2}(M+U)$ and $W \rightsquigarrow \frac{1}{2}(W+U)$. This means that the 1-step inverse model B^a does not always uniquely determine the 2-step inverse model $B^{aa'}$, i.e. (i,iii,v) can fail.

$M = W$. This trivially implies $g(M) = g(W)$. This means if (i) is true, then trivially also (iii) and (v), and if (ii) is true, then trivially also (iv) and (vi).

M and W are independent a. Note that $M^a_{ss'} \equiv p(s'|sa)$ independent a implies M^a_{s+} independent a, hence $\pi(a|s) = M^a_{s+} = 1/k$ independent a as well, hence $M^a = \frac{1}{k}M^+$. The latter implies $M^a M^{a'} ... M^{a^i} \oslash (M^+)^i = k^{-i}$ is independent M hence is the same as for W. Since we can choose $M \neq W$, this shows that (i) and (ii) and higher order analogues fail for these degenerate M and W.

M and W are nearly independent a. The above degeneracy generalizes to $M^a_{ss'} = M_{ss'}\pi_a$ and $W^a_{ss'} = W_{ss'}\pi_a$, i.e. action-independent dynamics, and state-independent actions, which in turn is a special case of the tensor product below (with $s = \ddot{s}$ and $\dot{s} \equiv 0$).

M and W are independent s'. In this case, $M^a_{ss'} = \frac{1}{d}M^a_{s+} = \frac{1}{d}\pi(a|s) = W^a_{ss'}$, hence is a special case of case $M = W$ above.

M and W are independent s. In this case, $[M^a M^{a'}]_{ss''} = \sum_{s'} M^a_{*s'} M^{a'}_{*s''} = \pi(a|*) M^{a'}_{*s''}$. Also the policy $\pi(a|s) = M^a_{s+}$ is independent s. If we assume EqIM(1), this implies

$$[M^a M^{a'} \oslash (M^+)^2]_{ss''} \;=\; \frac{\pi(a|*)M^{a'}_{*s''}}{\pi(+|*)M^+_{*s''}} \;=\; \frac{\pi(a|*)W^{a'}_{*s''}}{\pi(+|*)W^+_{*s''}} \;=\; [W^a W^{a'} \oslash (W^+)^2]_{ss''}$$

hence EqIM(2) holds and similarly EqIM$(i)\forall i$. As an example, consider

$$M^0 := \tfrac{1}{2}\begin{pmatrix}0&1\\0&1\end{pmatrix},\quad M^1 := \tfrac{1}{2}\begin{pmatrix}1&0\\1&0\end{pmatrix},\quad W^0 := \tfrac{1}{3}\begin{pmatrix}0&1\\0&1\end{pmatrix},\quad W^1 := \tfrac{2}{3}\begin{pmatrix}1&0\\1&0\end{pmatrix}$$

These $M \neq W$ satisfy EqIM(1) ($M^a \oslash M^+ = 2M^a = W^a \oslash W^+$), hence constitute another failure case of (i) and (ii).

Block-diagonal M and W. For $M = \begin{pmatrix}\dot{M}&0\\0&\ddot{M}\end{pmatrix}$ and $W = \begin{pmatrix}\dot{W}&0\\0&\ddot{W}\end{pmatrix}$, all operations $(+-\times/\odot\oslash)$ preserve the block structure, so the above degenerative cases can be combined, one for the upper-left block and another for the lower-right block.

Tensor-product M and W. Let $[\dot{M}\otimes\ddot{M}]_{ss'} := \dot{M}_{\dot{s}\dot{s}'}\ddot{M}_{\ddot{s}\ddot{s}'}$ with $s := (\dot{s},\ddot{s})$ and $s' := (\dot{s}',\ddot{s}')$ be the tensor product of $\dot{M}$ and $\ddot{M}$ (not to be confused with the element-wise product $\odot$). Assume $M^a = \dot{M}^a \otimes \ddot{M}$, where the second factor is action-independent. In this case, $M^a M^{a'} ... = (\dot{M}^a \dot{M}^{a'} ...) \otimes (\ddot{M}\ddot{M} ...)$, and similarly if $a,a',...$ is replaced by $+$, hence $M^a M^{a'} ... M^{a^i} \oslash (M^+)^i = \dot{M}^a \dot{M}^{a'} ... \dot{M}^{a^i} \oslash (\dot{M}^+)^i$ is independent of $\ddot{M}$, and similarly for $W^a = \dot{W}^a \otimes \ddot{W}$. That means, EqIM$(i)$ hold if $\dot{M}^a = \dot{W}^a$, whatever $\ddot{M}$ and $\ddot{W}$ are. This formalizes our motivating example that if some part of the state ($\ddot{s}$) is not controlled (by a) and the dynamics factorizes ($p(s'|sa) = p(\dot{s}'|\dot{s}a)p(\ddot{s}'|\ddot{s})$) and the policy is independent $\ddot{s}$ ($\pi(a|s) = \pi(a|\dot{s})$), then the multi-step inverse models (3-5) become much simpler than the forward model (2), namely independent $\ddot{s}$. This case has been studied in [EMK+22] for episodic near-deterministic M.

4 (Non)Uniqueness of Inverse MDP Models

We will now consider EqIM(1) and EqIM(2). We first provide a dimensional analysis which provides some insight and tentative answers about the solution space for W (given B or M): No, one, finitely many, or a manifold (of some dimension) of solutions. We then consider EqIM(1) only and characterize M and W for which it holds. This will be used to provide an algorithm that can determine a (and in some sense all) solution for W and hence $B^{aa'\cdots}$, given only B^a. EqIM(1) was quite simple, since it is effectively linear, but EqIM(2) is quadratic in W, which is where the difficulties start.

Dimensional analysis / counting solutions. Assume $k \leq d$ and $B^{\cdot}$ or $M^{\cdot}$ are given. The kd^2 equations EqIM(1) in W constitute $(k-1)d^2$ (linear) constraints on (the kd^2 real entries in) W. It's only $(k-1)d^2$, since summing over a gives d^2 vacuous equations $B^+ = 1 = W^+ \oslash W^+$. There are kd further (linear) constraints $W^a_{s+} = \pi(a|s)$. Assuming no further (missed/accidental) redundancies, this leads to a $kd^2 - (k-1)d^2 - kd = d(d-k)$ dimensional (linear) solution space for W. This is consistent with the algorithm below inferring $B^{aa'}$ from B^a if all B^a have full rank. Hence the set of solutions for $B^{aa'}$ forms a non-linear manifold of dimension at least $d(d-k)$.

If also B^{a+} is given, EqIM(2+) provides $(k-1)d^2$ further (quadratic) constraints (EqIM(ia) even provides $(k^i-1)d^2$ constraints). Since $d(d-k) < (k-1)d^2$, this now gives an over-determined system which generally has no solution. But by assumption, M is a solution, which gives hope that there may be only one or a finite number of solutions.

We can use the $kd + (k-1)d^2$ linear equations to eliminate this number of variables in W, which leaves $(k-1)d^2$ quadratic equations, now in only $d(d-k)$ variables, and no further equality constraints. By Bézout's bound [FW89], such a System of Quadratic Equations (SQE), either has a continuum number of solutions (as in the counter-example of Appendix G) or at most $2^{d(d-k)}$ solutions (as possibly in the counter-example in Appendix F). Multiple discrete solutions are often caused by symmetries, so for random B^a and B^{a+} consistent with M, the solution may indeed be unique.

Inferring *some* $B^{aa'}$ from B^a. Even if B^a does not uniquely determine $B^{aa'}$, we can ask for an *algorithm* inferring *some* consistent $B^{aa'}$ from B^a. Indeed this was our primary goal before realizing that the answer is not always unique. We know that $B^a = W^a \oslash W^+$ for *some* W. This implies $W^a = B^a \odot W^+$. So $W^a = B^a \odot J$ for some J. We need to ensure proper normalization $W^a_{s+} = \pi(a|s)$, i.e. $[B^a \odot J]_{s+} = \pi(a|s)$. This leads to the following algorithm to produce some (and indeed all) $B^{aa'}$:

- Given inverse 1-step model $B^a_{ss'} := p(a|ss')$ and policy $\pi(a|s)$
- For each s, choose *some* d-vector $J_{s\cdot}$
  satisfying the k linear equations $\sum_{s'} B^a_{ss'} J_{ss'} = \pi(a|s)$
- Compute forward model $W^a := B^a \odot J$
- Compute 2-step inverse model $B^{aa'} := W^a W^{a'} \oslash (W^+)^2$
- Then $p(aa'|ss'') \equiv B^{aa'}_{ss''}$ is *some* solution.

If for every s, matrix $B^{\cdot}_{s\cdot}$ has rank d, then $B^{aa'}$ is unique. The equations have no solution *iff* B is invalid in the sense that no underlying MDP M could have produced such B. This can only happen for $k > d$, i.e. B based on M have some intrinsic constraints beyond $B^+ = 1$ for $k > d$. For instance $B^0 = \frac{1}{2}\left(\begin{smallmatrix}1 & 0\\ 1 & 0\end{smallmatrix}\right)$, $B^1 = \frac{1}{2}\left(\begin{smallmatrix}0 & 1\\ 0 & 1\end{smallmatrix}\right)$, $B^2 = \frac{1}{2}\left(\begin{smallmatrix}1 & 1\\ 1 & 1\end{smallmatrix}\right)$ is inconsistent with $\pi(a|s) = \frac{1}{3}$. For unknown π, any J with $J_{s+} = 1$ will do. In general, the valid J span a linear subspace, but the set of all consistent $B^{aa'}$ is "parabolic". Noting that the ranks of $M^{\cdot}_{s\cdot}$ and $W^{\cdot}_{s\cdot}$ are the same, this gives the precise conditions under which (i) is true:

Proposition 1 (Conditions under which (i) is true)

$$M^a \oslash M^+ = W^a \oslash W^+ \textit{ implies } M = W \quad \textit{iff} \quad M^{\cdot}_{s\cdot} \textit{ has rank} \geq d \textit{ for every } s.$$

For this to be possible at all, we need $k \geq d$, i.e. more actions than states. This is typically not the most interesting regime. See Appendix B for an alternative derivation of this result without an intermediary algorithm.

We will next show that EqIM(2) removes this limitation, but we do not know of any exact *algorithm* for inferring (some) $B^{aa'a''}$ from B^a and $B^{aa'}$. We cannot even rule out that finding approximate solutions is NP-complete.

(Non)Uniqueness of Inverse MDP Models for $i \geq 2$. Above we have established that B^a does not uniquely determine $B^{aa'}$ for the interesting regime of $k < d$. From the dimensional analysis, providing 2-step inverse model $B^{aa'}$ in addition, has the potential of uniquely determining forward model W and/or multi-step inverse models $B^{aa'a''\cdots}$. We have numerically verified that this is indeed the case for B^a and $B^{aa'}$ based on random M^a. A more detailed analysis of the linear/quadratic structure of the problem is provided in Appendix C and a rank analysis in Appendix D. Unfortunately, even providing B^a *and* $B^{aa'}$ does not always uniquely determine M^a, nor higher B, (ii) (iv), (vi) fail for some M^a. Furthermore this remains true for higher i-versions, i.e. even EqIM(1)...EqIM(i) do not always uniquely determine EqIM($i+1$). We provide (potential) counter-examples in Appendices E and F, but the first is only conjectured and the second involves "bad" 0/0. We discuss what this means at the end of Appendix F. We provide a fully satisfactory counter-example in Appendix G.

5 Linear Relaxation

In Section 4 we provided an algorithm if only B^a is given. Here we consider the $i > 1$ case, and derive an algorithm for $k^i \geq d$, provided the solution is unique and further conditions on B are met. That is, we require $i \geq \log_k(d)$, which is greater than the minimum necessary in theory $i = 2$ from the dimensional analysis. E.g. for $i = 1$ we recover $k \geq d$, and $i = 2$ improves this to $k \geq \sqrt{d}$, and $i = \lceil \log_2(d) \rceil$ works for all k.

Recursive formulation. From EqIM(1) we know that $W^a = B^a \odot W^+$. Plugging this into EqIM(ia) and abbreviating $a^{:i} := aa'...a^i$ and $a^{<i} := aa'...a^{i-1}$ and $j := i+1$, this gives

$$B^{a^{:i}} \odot (W^+)^i \;=\; (B^a \odot W^+) \cdot ... \cdot (B^{a^i} \odot W^+) \tag{10}$$

If we plug EqIM($(i-1)a$) into EqIM(ia) and abbreviate $V := (W^+)^{i-1}$ this simplifies to

$$B^{a^{:i}} \odot (V \cdot W^+) \;=\; (B^{a^{<i}} \odot V) \cdot (B^{a^i} \odot W^+)$$

which written out becomes

$$\sum_{s^i} B^{a^{:i}}_{ss^j} V_{ss^i} W^+_{s^i s^j} \;=\; \sum_{s^i} B^{a^{<i}}_{ss^i} V_{ss^i} B^{a^i}_{s^i s^j} W^+_{s^i s^j} \tag{11}$$

Linear relaxation. We can consider a linear relaxation of this System of Polynomial Equations (SPE) by introducing new variables $U_{ss^i s^j}$ (aiming at $U_{ss^i s^j} = V_{ss^i} W^+_{s^i s^j}$):

$$\sum_{s^i} A^{a^{:i}}_{ss^i s^j} U_{ss^i s^j} \;=\; 0 \quad \text{with} \quad A^{a^{:i}}_{ss^i s^j} \;:=\; B^{a^{:i}}_{ss^j} - B^{a^{<i}}_{ss^i} B^{a^i}_{s^i s^j} \tag{12}$$

These are $k^i d^2$ potentially independent linear equations in d^3 unknowns U. The solution can only be unique if $k^i \geq d$. For random B, for each fixed (s,s^j), the $k^i \times d$ matrix $A^{\cdots}_{s\cdot s^j}$ has indeed full rank $\min\{k^i, d\} \geq d$, hence $U_{ss^i s^j} \equiv 0$ is the only solution. This is inconsistent with the constraints (7), and hence shows that (unrestricted random) B do not come from some M. This makes the validity of the B's sometimes semi-decidable in time $O(d^4(d+k^i))$ or typically/randomized time $O(d^5)$. For the B's originating from some M, $\hat{U}_{ss^i s^j} = (M^+)^{i-1}_{ss^i} M^+_{s^i s^j}$ solves (12). Since for different ss^j the equations in (12) are independent, $U_{ss^i s^j} := \hat{U}_{ss^i s^j} K_{ss^j}$ also solves (12) for any K. In other words, the rank of $A^{\cdots}_{s\cdot s^j}$ is bounded by $\min\{k^i, d-1\}$, and achieved e.g. for random matrices B *consistent* with M. Since the solution is not unique, for many solutions U there will be no W^+ satisfying $U_{ss^i s^j} = (W^+)^{i-1}_{ss^i} W^+_{s^i s^j}$, not to speak of M^+, even if the original problem (10)+(7) has a unique solution.

Unique solution by lifted constraints. So we must (and at least for random M can) make the solution unique by taking into account the linear constraints (7). Applying them to $s \rightsquigarrow s^i, s' \rightsquigarrow s^j, a \rightsquigarrow a^i$ and multiplying from the left with V_{ss^i} and using $V_{ss^i} = U_{ss^i +}$ we lift them to

$$\sum_{s^j} B^{a^i}_{s^i s^j} U_{ss^i s^j} \;=\; U_{ss^i +} \pi(a^i | s^i) \quad \text{and} \quad U_{s++} \;=\; 1 \tag{13}$$

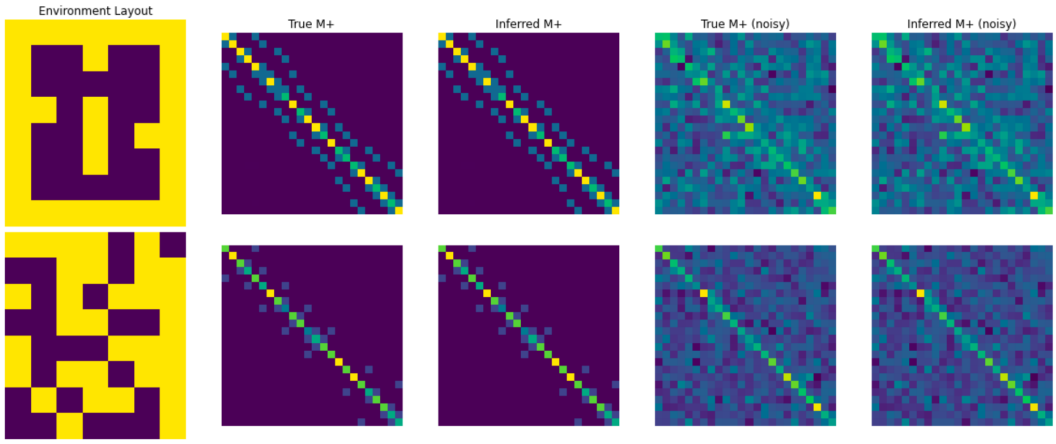

Figure 1: Environments, their transition matrices (i.e. M^+) and the matrices inferred by the algorithm (i.e. W^+). Results shown on the most and least noisiest variants of each environment. *Top* 'four-rooms' grid-world. *Bottom* One of the randomly generated grid-worlds.

These kd^2+d further linear constraints have the potential to make the solution of (12) unique, i.e. resolve the d^2 degeneracy K_{ss^i}. If so, we can recover $M^+_{s^is^j}=W^+_{s^is^j}=U_{ss^is^j}/V_{ss^i}$ (and finally $M^a=W^a=B^a\odot W^+$) in polynomial time. It actually suffices to solve (12) and (13) for one fixed s, e.g. $s=1$, which with some care can be done in time $O(d^4)$. In practice, for approximate B one would solve a least-squares problem using all equations or a random projection for speed.

Algorithm. Putting pieces together, we have the following algorithm for computing W^a and hence $B^{a^{:j}}$ for all j via EqIM(ja) from B^a and $B^{a^{<i}}$ and $B^{a^{:i}}$

- Given: Policy $\pi(a|s)$ and for $j-1:=i\geq 2$, inverse $1,i-1,i$-step models $B^a_{ss'}=p(a|ss')$ and $B^{a^{<i}}_{ss^j}=p(a^{<i}|ss^i)$ and $B^{a^{:i}}_{ss^j}=p(a^{:i}|ss^j)$
- Do the following calculations for one s (e.g. $s=1$), or a few or all s or some random linear combinations of s:
- For each s^j, let $\hat U_{ss^is^j}$ be a solution of (12) with $\hat U_{s+s^j}=1$
- If a non-zero solution does not exist, set $\hat U_{ss^is^j}=0\;\forall s^i$.
- Optional: If multiple solutions exist, return "W *may* not be unique"
- If $\hat U_{s++}=0$, return "B is not consistent with any M"
- Solve $\sum_{s^j}C^{a^i}_{ss^is^j}K_{ss^j}=0$ and $K_{s+}=1$ for K_{s*}, where $C^{a^i}_{ss^is^j}:=(B^{a^i}_{s^is^j}-\pi(a^i|s^i))\hat U_{ss^is^j}$
- If no solution, return "B is not consistent with any M"
- Optional: If multiple solutions exist, return "W *may* not be unique"
- $\tilde U_{ss^is^j}:=\hat U_{ss^is^j}K_{ss^j}$, $\quad U_{ss^is^j}:=\tilde U_{ss^is^j}/\tilde U_{s++}$, $\quad V_{ss^i}:=U_{ss^i+}$, $\quad W^+_{s^is^j}:=U_{ss^is^j}/V_{ss^i}$
- Optional: If different s lead to different W^+ or $V\neq(W^+)^{i-1}$, return "W *may* not be unique"
- Return forward model $W^a:=B^a\odot W^+$ and other inverse $B^{\cdots}$ computed via (8)

Variations that don't work. For unknown π, we only have d lifted constraints $U_{s++}=1$, which are not sufficient to make the solution unique, also resulting in too many solutions for the relinearization trick [CKPS00] to work. The same is true if we had relaxed $U_{ss's^j}=W^+_{ss'}V_{s's^j}$. If we had applied linear relaxation directly to EqIM(ia), this would have led to order-$i+1$ tensors and require $k\geq d^{1-1/i}$, which is much worse than $k\geq d^{1/i}$ for $i>2$. Including $B^{a^{:j}}$ and EqIM(ja) for some or all $j<i-1$ is not only unhelpful but even counter-productive.

6 Experiments

The algorithm described in Section 5 was motivated by properties of random matrices. Namely, that $A^{\cdots}_{s\cdot s^j}$ is likely to be "full" rank, and thus yield a unique solution. In order to explore the plausibility of this assumption in practice, we have evaluated the algorithm on a set of toy (but structured)

environments. This includes the canonical 'four-rooms' grid-world and samples from the distribution over all grid-worlds of that size. All environments have $k=5$ (local movement on the grid) and $d=24$, thus satisfying the $k \geq d^2$ constraint which permits solving EqIM(2).

As detailed in Appendix N, for all environments tested the algorithm yielded a unique solution (recovering M^a) up to a reasonable level of numerical precision. This remained true even after injecting noise (across several orders of magnitude) into the environmental transition dynamics. This is in contrast to related methods which rely on near-deterministic environments [EMK+22].

This result is non-trivial, as the statistics of these environments differ significantly from those produced by random matrices. For example, grid-world dynamics are both local and sparse, unlike random matrix dynamics which almost always have non-zero probability for all transitions. It remains to be seen whether or not larger-scale environments yield similar results, but it is at least non-obvious what additional environmental properties would break the constraints of the algorithm.

Further experiments and discussion can be found in Appendix N.

7 Conclusion

Summary. We have shown that the 1-step inverse model $p(a|ss')$ does not uniquely determine the 2-step probabilities $p(a|ss'')$ if there are less actions than states ($k < d$). Even for $k \geq d$, the implication can fail, e.g. if the extra actions are ineffective, but if $p(s'|sa) = M^a_{ss'}$ considered as matrices in a and s' for each s have full rank, the implication holds. Even providing $p(aa'...a^j|ss^j)$ for all $j < i$ not necessarily determines $p(a|ss^i)$. Since the involved SPE is (heavily) over-determined, we expect the failure cases to be sparse/rare in some sense. For (B based on) random M, we provided evidence that $a=2$ suffices to determine M and hence $p(aa'...|ss's''...)$ from $p(a|ss')$ and $p(a|ss'')$. For low-rank M the implication may fail.

Open Problems. Maybe characterizing all M for which EqIM(1) and EqIM(2) uniquely determine W is hopeless, not to speak of finding some or all W in case not. More formally, we can ask the question of whether there exists an efficient algorithm that can decide whether EqIM(i) has a unique solution.

Conjecture 2 (NP-hardness) *Deciding (ii), (iv), (vi) is NP-hard. Deciding whether B^a and $B^{aa'}$ are consistent with some M is also NP-hard. Computing some solution is FNP-hard.*

In Appendix K we provide some weak preliminary evidence, why this problem may be NP-hard. Appendix M contains fully self-contained a few versions of this open problem in their simplest instantiation and most elegant form.

Discussion. Given our analysis, we would expect that in practice, B^a and $B^{aa'}$ determines $B^{aa'a''...}$ and W sufficiently well. Sufficiently well in case of W means all and only those aspects of the forward model relevant for the inverse model. Then of course the question remains how to compute the/an answer. While the linear relaxation developed in Section 5 fails for $k < d^{1/i}$ as an exact method, it might still lead to useful approximate solutions [Stu02] without formal guarantees. Indeed, EqIM(ia) is heavily over-determined for $i \geq 2$, and heuristic solvers often work well in this regime.

Handling unlikely but potential non-uniqueness: In practice, the state space is very often infinite, and no finite amount of data will determine even B^a uniquely without further structural assumptions. Neural networks intrinsically restrict the solution space, but this may not suffice for modern over-parametrized deep networks. Aiming for the maximum-entropy distribution consistent with the (constraints from) data is popular, and could make the solution unique, as well as any other optimization constraint.

References

[AD21] Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500, August 2021.

[BEP+18] Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*, 2018.

[BPK+20] Adrià Puigdomènech Badia, Bilal Piot, Steven Kapturowski, Pablo Sprechmann, Alex Vitvitskyi, Zhaohan Daniel Guo, and Charles Blundell. Agent57: Outperforming the atari human benchmark. In *International Conference on Machine Learning*, pages 507–517. PMLR, 2020.

[CKPS00] Nicolas Courtois, Alexander Klimov, Jacques Patarin, and Adi Shamir. Efficient Algorithms for Solving Overdefined Systems of Multivariate Polynomial Equations. In Gerhard Goos, Juris Hartmanis, Jan van Leeuwen, and Bart Preneel, editors, *Advances in Cryptology — EUROCRYPT 2000*, volume 1807, pages 392–407. Springer Berlin Heidelberg, Berlin, Heidelberg, 2000.

[DY79] E. B. Dynkin and A. A. Yushkevich. *Controlled Markov processes*. Number 235 in Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin ; New York, 1979.

[EGIL18] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

[EMK+22] Yonathan Efroni, Dipendra Misra, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Provable RL with Exogenous Distractors via Multistep Inverse Dynamics. *arXiv:2110.08847 [cs]*, March 2022.

[FW89] William Fulton and Richard Weiss. *Algebraic Curves: An Introduction to Algebraic Geometry*. Addison-Wesley, 1989.

[GRW17] Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational Intrinsic Control. In *Workshop*, February 2017.

[HDB+21] Steven Stenberg Hansen, Guillaume Desjardins, Kate Baumli, David Warde-Farley, Nicolas Heess, Simon Osindero, and Volodymyr Mnih. Entropic Desired Dynamics for Intrinsic Control. In *Advances in Neural Information Processing Systems*, May 2021.

[HL13] Christopher J. Hillar and Lek-Heng Lim. Most Tensor Problems Are NP-Hard. *Journal of the ACM*, 60(6):1–39, November 2013.

[HSHB99] Jesse Hoey, Robert St-Aubin, Alan Hu, and Craig Boutilier. SPUDD: Stochastic planning using decision diagrams. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 279–288, 1999.

[KF09] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2009.

[OKD+21] Pedro A. Ortega, Markus Kunesch, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Joel Veness, Jonas Buchli, Jonas Degrave, Bilal Piot, Julien Perolat, Tom Everitt, Corentin Tallec, Emilio Parisotto, Tom Erez, Yutian Chen, Scott Reed, Marcus Hutter, Nando de Freitas, and Shane Legg. Shaking the foundations: Delusions in sequence models for interaction and control. *arXiv:2110.10819 [cs]*, October 2021.

[PGJ16] Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. *Causal Inference in Statistics: A Primer*. Wiley, Chichester, West Sussex, 2016.

[Pre00] Doina Precup. Temporal Abstraction in Reinforcement Learning, 2000.

[SGL+19] Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-Aware Unsupervised Discovery of Skills. In *International Conference on Learning Representations*, September 2019.

[SP02] Martin Stolle and Doina Precup. Learning Options in Reinforcement Learning. In Sven Koenig and Robert C. Holte, editors, *Abstraction, Reformulation, and Approximation*, Lecture Notes in Computer Science, pages 212–223, Berlin, Heidelberg, 2002. Springer.

[Stu02] Bernd Sturmfels. *Solving Systems of Polynomial Equations*. Number 97 in Regional Conference Series in Mathematics. American Mathematical Society, Providence, RI, 2002.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]