

Learning Compact Models for Planning with Exogenous Processes

Rohan Chitnis, Tomás Lozano-Pérez

MIT Computer Science and Artificial Intelligence Laboratory
{ronuchit, tlp}@mit.edu

Abstract: We address the problem of approximate model minimization for MDPs in which the state is partitioned into *endogenous* and (much larger) *exogenous* components. An exogenous state variable is one whose dynamics are independent of the agent’s actions. We formalize the *mask-learning problem*, in which the agent must choose a subset of exogenous state variables to reason about when planning; doing planning in such a reduced state space can often be significantly more efficient than planning in the full model. We then explore the various value functions at play within this setting, and describe conditions under which a policy for a reduced model will be optimal for the full MDP. The analysis leads us to a tractable approximate algorithm that draws upon the notion of mutual information among exogenous state variables. We validate our approach in simulated robotic manipulation domains where a robot is placed in a busy environment, in which there are many other agents also interacting with the objects. Visit <http://tinyurl.com/chitnis-exogenous> for a supplementary video.

Keywords: learning to plan efficiently, exogenous events, model minimization

1 Introduction

Most aspects of the world are *exogenous* to us: they are not affected by our actions. However, though they are exogenous, these processes (e.g., weather and traffic) will often play a major role in the way we should choose to perform a task at hand. Despite being faced with such a daunting space of processes that are out of our control, humans are extremely adept at quickly reasoning about *which* aspects of this space they need to concern themselves with, for the particular task at hand.

Consider the setting of a household robot tasked with doing laundry inside a home. It should not get bogged down by reasoning about the current weather or traffic situation, because these factors are irrelevant to its task. If, instead, it were tasked with mowing the lawn, then good decision-making would require it to reason about the time of day and weather (so it can finish the task by sunset, say).

In this work, we address the problem of approximately solving Markov decision processes (MDPs), without too much loss in solution quality, by leveraging the structure of their exogenous processes. We begin by formalizing the *mask-learning problem*. An autonomous agent is given a generative model of an MDP that is partitioned into an endogenous state (i.e., that which can be affected by the agent’s actions) and a much higher-dimensional exogenous state. The agent must choose a *mask*, a subset of the exogenous state variables, that yields a policy *not too much worse* than a policy that would be obtained by reasoning about the entire exogenous state.

After formalizing the mask-learning problem, we discuss how we can leverage exogeneity to quickly learn transition models for only the relevant variables from data. Then, we explore the various value functions of interest within the problem, and discuss the conditions under which a policy for a particular mask will be optimal for the full MDP. Our analysis lends theoretical credence to the idea that a good mask should contain not only the exogenous state variables that directly influence the agent’s reward function, but also ones whose dynamics are correlated with theirs. This idea leads to a tractable approximate algorithm for the mask-learning problem that leverages the structure of the MDP, drawing upon the notion of mutual information among exogenous state variables.

We experiment in simulated robotic manipulation domains where a robot is put in a busy environment, among many other agents that also interact with the objects. We show that 1) in small domains where we can plan directly in the full MDP, the masks learned by our approximate algorithm yield competitive returns; 2) our approach outperforms strategies that do not leverage the structure of the MDP; and 3) our algorithm can scale up to planning problems with large exogenous state spaces.

2 Related Work

The notion of an *exogenous event*, one that is outside the agent’s control, was first introduced in the context of planning problems by Boutilier et al. [1]. The work most similar to ours is that by Dietterich et al. [2], who also consider the problem of model minimization in MDPs by removing exogenous state variables. Their formulation of an MDP with exogenous state variables is similar to ours, but the central assumption they make is that the reward decomposes additively into an endogenous state component and an exogenous state component. Under this assumption, the value function of the full MDP decomposes into two parts, and any policy that is optimal for the endogenous MDP is also optimal for the full MDP. On the other hand, we do not make this reward decomposition assumption, and so our value function does not decompose; instead, our work focuses on a different set of questions: 1) what are the conditions under which an optimal policy in a given reduced model is optimal in the full MDP? 2) can we build an algorithm that leverages exogeneity to efficiently (approximately) discover such a reduced model?

Model minimization of factored Markov decision processes is often defined using the notion of *stochastic bisimulation* [3, 4], which describes an equivalence relation among states based upon their transition dynamics. Other prior work in state abstraction tries to remove irrelevant state variables in order to form reduced models [5, 6]. Our approach differs from these techniques in two major ways: 1) we consider only reducing the exogenous portion of the state, allowing us to develop algorithms which leverage the computational benefits enjoyed by the exogeneity assumption; 2) rather than trying to build a reduced model that is faithful to the full MDP, we explicitly optimize a different objective (Equation 1), which tries to find a reduced model yielding high rewards in the full MDP. Recent work in model-free reinforcement learning has considered how to exploit exogenous events for better sample complexity [7, 8], whereas we tackle the problem from a model-based perspective.

If we view the state as a vector of features, then another perspective on our approach is that it is a technique for *feature selection* [9] applied to MDPs with exogenous state variables. This is closely related to the typical subset selection problem in supervised learning [10, 11, 12, 13], in which a learner must determine a subset of features that suffices for making predictions with little loss.

3 Problem Formulation

In this section, we introduce the notion of exogeneity in the context of a Markov decision process, and use this idea to formalize the mask-learning problem.

3.1 Markov Decision Processes with Exogenous Variables

We will formalize our setting as a discrete Markov decision process (MDP) with special structure. An MDP is a tuple $(\mathcal{S}, \mathcal{A}, T, R, \gamma)$ where: \mathcal{S} is the state space; \mathcal{A} is the action space; $T(s_t, a_t, s_{t+1}) = P(s_{t+1} | s_t, a_t)$ is the state transition distribution with $s_t, s_{t+1} \in \mathcal{S}$ and $a_t \in \mathcal{A}$; $R(s_t, a_t)$ is the reward function; and γ is the discount factor in $(0, 1]$. On each timestep, the agent chooses an action, transitions to a new state sampled from T , and receives a reward as specified by R . The solution to an MDP is a *policy* π , a mapping from states in \mathcal{S} to actions in \mathcal{A} , such that the expected discounted sum of rewards over trajectories resulting from following π , which is $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(s_t))]$, is maximized. Here, the expectation is with respect to the stochasticity in the initial state and state transitions. The *value* of a state $s \in \mathcal{S}$ under policy π is defined as the expected discounted sum of rewards from following π , starting at state s : $V_{\pi}(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(s_t)) | s_0 = s]$. The *value function* for π is the mapping from \mathcal{S} to \mathbb{R} defined by $V_{\pi}(s), \forall s \in \mathcal{S}$.

In this work, we assume that the agent is given a *generative* model of an MDP, in the sense of Kearns et al. [14]. Concretely, the agent is given the following:

- Knowledge of \mathcal{S} , \mathcal{A} , and γ .

- A black-box sampler of the transition model, which takes as input a state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, and returns a next state $s' \sim T(s, a, s')$.
- A black-box reward function, which takes as input a state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, and returns the reward $R(s, a)$ for that state-action pair.
- A black-box sampler of an initial state s_0 from some distribution $P(s_0)$.

We note that this assumption of having a generative model of an MDP lies somewhere in between that of only receiving execution traces (as in the typical reinforcement learning setting, assuming no ability to reset the environment) and that of having knowledge of the full analytical model. One can also view this assumption as saying that a simulator is available. Generative models are a natural way to specify a large MDP, as it is typically easier to produce samples from a transition model than to write out the complete next-state distributions.

We assume that the state of our MDP decomposes into an *endogenous* component and a (much larger) *exogenous* component, and that the agent knows this partition. Thus, we write $s_t = [n_t \ x_t]$, where n_t is the endogenous component of the state, whose transitions the agent can affect through its actions, and x_t is the exogenous component of the state. This assumption means that $P(x_{t+1} \mid x_t, a_t) = P(x_{t+1} \mid x_t)$, and therefore $P(s_{t+1} \mid s_t, a_t) = P(n_{t+1} \mid n_t, a_t, x_t) \cdot P(x_{t+1} \mid x_t)$.

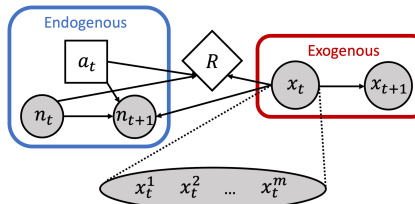


Figure 1: A Markov decision process with exogenous state variables $x_t^1, x_t^2, \dots, x_t^m$.

Though unaffected by the agent’s actions, the exogenous state variables influence the agent through the rewards and endogenous state transitions. Therefore, to plan, the agent will need to reason about future values of the x_t .

We will focus on the setting where the exogenous state x_t is made up of m (not necessarily independent) state variables $x_t^1, x_t^2, \dots, x_t^m$, with m large. The fact that m is large precludes reasoning about the entire exogenous state. We will define the mask-learning problem (Section 3.2) an optimization problem for deciding which variables x^1, x^2, \dots, x^m the agent should reason about.

To be able to unlink the effects of each exogenous state variable on the agent’s reward, we will need to make an assumption on the form of the reward function $R(s_t, a_t) = R(n_t, x_t, a_t)$. This assumption is necessary for the agent to be able to reason about the effect of dropping a particular exogenous state variable from consideration. Specifically, we assume that $R(n_t, x_t, a_t) = \sum_{i=1}^m R^i(n_t, x_t^i, a_t)$. In words, this says that the reward function decomposes into a sum over the individual effects of each exogenous state variable x^i . Although this means that the computation of the agent’s reward cannot include logic based on any combinations of x^i , this assumption is not too restrictive: one could always construct “super-variables” encompassing state variables coupled in the reward. Note that despite this assumption, the value function might depend non-additively on the exogenous variables.

3.2 The Mask-Learning Problem

Our central problem of focus is that of finding a *mask*, a subset of the m exogenous state variables, that is “good enough” for planning, in the sense that we do not lose too much by ignoring the others. Before being able to formalize the problem, we must define precisely what it means to plan with only a subset of the m exogenous variables. Suppose we have an MDP M as described in Section 3.1. Let $x = [x^1 \ x^2 \ \dots \ x^m]$, and $\tilde{x} \subseteq x$ be an arbitrary subset (a *mask*).

We define the *reduced model* \tilde{M} corresponding to mask \tilde{x} and MDP M as another MDP:

- \mathcal{A} and γ are the same as in M .
- \mathcal{S} is reduced by removing the dimensions corresponding to any $x^i \notin \tilde{x}$.
- $P(\tilde{s}_{t+1} \mid \tilde{s}_t, a_t) = P(n_{t+1} \mid n_t, a_t, \tilde{x}_t) \cdot P(\tilde{x}_{t+1} \mid \tilde{x}_t)$. Here, $\tilde{s}_t = [n_t \ \tilde{x}_t]$.¹
- $R(\tilde{s}_t, a_t) = R(n_t, \tilde{x}_t, a_t) = \sum_{x^i \in \tilde{x}} R^i(n_t, x_t^i, a_t)$. Here, we are leveraging the assumption that the reward function decomposes as discussed at the end of Section 3.1.

¹The \tilde{x} might, in actuality, not be Markov, since we are ignoring the variables $x \setminus \tilde{x}$. Nevertheless, this expression is well-formed, and estimating it from data marginalizes out these ignored variables. If they were not exogenous, such estimates would depend on the data-gathering policy (and thus could be very error-prone).

Since the agent only has access to a generative model of the MDP M , planning in \tilde{M} will require estimating the reduced dynamics and reward models, $\hat{T}(\tilde{s}_t, a_t, \tilde{s}_{t+1})$ and $\hat{R}(\tilde{s}_t, a_t)$.

Formally, the *mask-learning problem* is to determine:

$$\tilde{x}^* = \operatorname{argmax}_{\tilde{x} \subseteq x} J(\tilde{x}) = \operatorname{argmax}_{\tilde{x} \subseteq x} \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t R(n_t, x_t, \tilde{\pi}(n_t, \tilde{x}_t)) \right] - \lambda \cdot \operatorname{Cost}(\tilde{x}), \quad (1)$$

where $\tilde{\pi}$ is the policy (mapping reduced states to actions) that is obtained by planning in \tilde{M} . In words, we seek the mask \tilde{x}^* that yields a policy maximizing the expected total reward *accrued in the actual environment* (the complete MDP M), minus a cost on \tilde{x} . Note that if $\lambda = 0$, then the choice $\tilde{x} = x$ is always optimal, and so the λ serves as a regularization hyperparameter that balances the expected reward with the complexity of the considered mask. Reasonable choices of $\operatorname{Cost}(\tilde{x})$ include $|\tilde{x}|$ or the amount of time needed to produce the policy $\tilde{\pi}$ corresponding to \tilde{x} .

4 Leveraging Exogeneity

The agent has only a generative model of the MDP M (Section 3.1). In order to build a reduced MDP \tilde{M} , it must estimate $\hat{T}(\tilde{s}_t, a_t, \tilde{s}_{t+1}) = \hat{P}(n_{t+1} | n_t, a_t, \tilde{x}_t) \cdot \hat{P}(\tilde{x}_{t+1} | \tilde{x}_t)$ and $\hat{R}(\tilde{s}_t, a_t)$.

Recall that we are considering the setting where the space of exogenous variables is much larger than the space of endogenous variables. At first glance, then, it seems challenging to estimate $\hat{P}(\tilde{x}_{t+1} | \tilde{x}_t)$ using only the generative model for $P(x_{t+1} | x_t)$. However, this estimation problem is in fact greatly simplified due to the exogeneity of the x^i . To see why, consider the typical strategy for estimating a transition model from data: generate trajectories starting from some initial state x_0 , then fit a one-step prediction model to this data. Now, if the x^i were endogenous, then we would need to commit to some policy π_{rollout} in order to generate these trajectories, and the reduced transition model we learn would depend heavily on this π_{rollout} . However, since the x^i are exogenous, we can roll out simulations conditioned *only on the initial state* x_0 , *not on a policy*, and use this data to efficiently estimate the transition model $P(\tilde{x}_{t+1} | \tilde{x}_t)$ of the reduced MDP \tilde{M} .²

Because we can efficiently estimate the reduced model dynamics of exogenous state variables, we can not only tractably construct the reduced MDP, but also allow ourselves to explore algorithms that depend heavily on estimating these variables' dynamics, as we will do in Section 5.3.

5 Algorithms for Mask-Learning

In this section, we explore some value functions induced by the mask-learning problem, and use this analysis to develop a tractable but effective approximate algorithm for finding good masks.

5.1 Preliminaries

Observe that the expectation in the objective $J(\tilde{x})$ (Equation 1) is the value $V_{\tilde{\pi}}(s_0)$ of an initial state under the policy $\tilde{\pi}$, corresponding to mask \tilde{x} . Because the full MDP M is very large, computing this value (the expected discounted sum of rewards) exactly will not be possible. Instead, we can use rollouts of $\tilde{\pi}$ to produce an estimate $\hat{V}_{\tilde{\pi}}$, which in turns yields an estimate of the objective, $\hat{J}(\tilde{x})$:

Procedure ESTIMATE-OBJECTIVE($M, \tilde{x}, n_{\text{rollouts}}$)

- 1 | Construct estimated reduced MDP \tilde{M} defined by mask \tilde{x} and full MDP M .
- 2 | Solve \tilde{M} to get policy $\tilde{\pi}$.
- 3 | **for** $i = 1, 2, \dots, n_{\text{rollouts}}$ **do**
- 4 | | Execute $\tilde{\pi}$ in the full MDP M , obtain total discounted returns r_i .
- 5 | $\hat{J}(\tilde{x}) = \frac{1}{n_{\text{rollouts}}} \sum_i r_i - \lambda \cdot \operatorname{Cost}(\tilde{x})$

As discussed in Section 4, Line 1 is tractable due to the exogeneity of the x variables, which make up most of the dimensionality of the state. With ESTIMATE-OBJECTIVE in hand, we can write down some very simple strategies for solving the mask-learning problem:

²In high dimensions, we may still need a lot of data, unless the state factors nicely.

- **MASK-LEARNING-BRUTE-FORCE:** Evaluate $\hat{J}(\tilde{x})$ for every possible mask $\tilde{x} \subseteq x$. Return the highest-scoring mask.
- **MASK-LEARNING-GREEDY:** Start with an empty mask \tilde{x} . While $\hat{J}(\tilde{x})$ increases, pick a variable x^i at random, add it into \tilde{x} , and re-evaluate $\hat{J}(\tilde{x})$.

While optimal, **MASK-LEARNING-BRUTE-FORCE** is of course intractable for even medium-sized values of m , as it will not be feasible to evaluate all 2^m possible subsets of x . Unfortunately, even **MASK-LEARNING-GREEDY** will likely be ineffective for medium and large MDPs, as it does not leverage the structure of the MDP whatsoever. To develop a better algorithm, we will start by exploring the connection between the value functions of the reduced MDP \tilde{M} and the full MDP M .

5.2 Analyzing the Value Functions of Interest

It is illuminating to outline the various value functions at play within our problem. We have:

- $V^*(s)$: the value function under an optimal policy; unknown and difficult to compute exactly.
- $V_{\tilde{\pi}}(s)$: given a mask \tilde{x} , the value function under the policy $\tilde{\pi}$; unknown and difficult to compute exactly. Note that $V^*(s) \geq V_{\tilde{\pi}}(s) \forall s \in \mathcal{S}$, by the definition of an optimal value function.
- $\hat{V}_{\tilde{\pi}}(s)$: given a mask \tilde{x} , the empirical estimate of $V_{\tilde{\pi}}(s)$, which the agent can obtain by rolling out $\tilde{\pi}$ in the environment many times, as was done in the **ESTIMATE-OBJECTIVE** procedure.
- $\tilde{V}_{\tilde{\pi}}(\tilde{s})$: given a mask \tilde{x} , the value function of policy $\tilde{\pi}$ *within the reduced* MDP \tilde{M} . Here, \tilde{s} is the reduced form of state s (i.e., the endogenous state n concatenated with \tilde{x}).

Intuitively, $\tilde{V}_{\tilde{\pi}}(\tilde{s})$ corresponds to the expected reward that the agent *believes* it will receive by following $\tilde{\pi}$, which typically will not match $V_{\tilde{\pi}}(s)$, the *actual* expected reward. In general, we cannot say anything about the ordering between $\tilde{V}_{\tilde{\pi}}(\tilde{s})$ and $V_{\tilde{\pi}}(s)$. For instance, if the mask \tilde{x} ignores some negative effect in the environment, then the agent will expect to receive *higher* reward than it actually receives during its rollouts. On the other hand, if the mask \tilde{x} ignores some positive effect in the world, then the agent will expect to receive *lower* reward than it actually receives.

It is now natural to ask: under what conditions would an optimal policy $\tilde{\pi}$ for the reduced MDP \tilde{M} also be optimal for the full MDP M ? The following theorem describes sufficient conditions:

Theorem 1. *Consider an MDP M as defined in Section 3.1, with exogenous state variables $x = [x^1 \ x^2 \ \dots \ x^m]$, and a mask $\tilde{x} \subseteq x$. Let $\bar{x} = x \setminus \tilde{x}$ be the variables not included in the mask. If the following conditions hold: (1) $R^i(n_t, x_t^i, a_t) = 0 \forall x_t^i \in \bar{x}$, (2) $P(n_{t+1} | n_t, a_t, x_t) = P(n_{t+1} | n_t, a_t, \tilde{x}_t)$, (3) $P(\tilde{x}_{t+1}, \bar{x}_{t+1} | \tilde{x}_t, \bar{x}_t) = P(\tilde{x}_{t+1} | \tilde{x}_t) \cdot P(\bar{x}_{t+1} | \bar{x}_t)$; then $\tilde{V}_{\tilde{\pi}}(\tilde{s}) = V_{\tilde{\pi}}(s) \forall s \in \mathcal{S}$. If $\tilde{\pi}$ is optimal for the reduced MDP \tilde{M} , then it must also be true that $\tilde{V}_{\tilde{\pi}}(\tilde{s}) = V^*(s) \forall s \in \mathcal{S}$.*

Proof: See Appendix A. □

The conditions in Theorem 1 are very intuitive: (1) all exogenous variables *not* in the mask do not influence the agent’s reward, (2) the endogenous state transitions do not depend on variables not in the mask, and (3) the variables in the mask transition independently of the ones not in the mask. If these conditions hold, and we use an optimal planner for the reduced model, then we will obtain a policy that is optimal *for the full* MDP, *not just the reduced one*. Based on these conditions, it is clear that our mask-learning algorithm should be informed by two things: 1) the agent’s reward function and 2) the degree of correlation among the dynamics of the various state variables.

Hoeffding’s inequality [15] allows us to bound the difference between $V_{\tilde{\pi}}(s)$ and the empirical estimate $\hat{V}_{\tilde{\pi}}(s)$. For rewards in the range $[0, R_{\max}]$, discount factor γ , number of rollouts n , and policy $\tilde{\pi}$, we have that for any state s , $|V_{\tilde{\pi}}(s) - \hat{V}_{\tilde{\pi}}(s)| \leq \lambda$ with probability at least $1 - 2e^{-2n\lambda^2(1-\gamma)^2/R_{\max}^2}$. This justifies the use of $\hat{V}_{\tilde{\pi}}(s)$ in place of $V_{\tilde{\pi}}(s)$, as was done in the **ESTIMATE-OBJECTIVE** procedure. Next, we describe a tractable algorithm for mask-learning that draws on Theorem 1.

5.3 A Correlational Algorithm for Mask-Learning

Of course, it is very challenging to directly search for a low-cost mask \tilde{x} that meets all the conditions of Theorem 1, if one even exists. However, we can use that intuition to develop an approximate

algorithm based upon greedy forward selection techniques for feature selection [9], which at each iteration add a single variable that most improves some performance metric.

Algorithm MASK-LEARNING-CORRELATIONAL($M, \tau_{\text{correl}}, \tau_{\text{variance}}, n_1, n_2$)

```

1   $\tilde{x} \leftarrow \text{ESTIMATE-REWARD-VARIABLES}(M, \tau_{\text{variance}}, n_1, n_2)$  // Initial mask for MDP  $M$ .
2  while  $\hat{J}(\tilde{x})$  increases do // Iteratively add to initial mask.
3       $x^i = \operatorname{argmax}_{x^j \notin \tilde{x}} D_{\text{KL}}(\hat{T}_{\tilde{s}, x^j} \parallel \hat{T}_{\tilde{s}} \otimes \hat{T}_{x^j})$  // Measure mutual informations.
4      if  $D_{\text{KL}}(\hat{T}_{\tilde{s}, x^i} \parallel \hat{T}_{\tilde{s}} \otimes \hat{T}_{x^i}) < \tau_{\text{correl}}$  then
5          break // All mutual informations below threshold, terminate.
6           $\tilde{x} = \tilde{x} \cup x^i$ 
7           $\hat{J}(\tilde{x}) \leftarrow \text{ESTIMATE-OBJECTIVE}(M, \tilde{x})$ 
8  return  $\tilde{x}$ 

```

Subroutine ESTIMATE-REWARD-VARIABLES($M, \tau_{\text{variance}}, n_1, n_2$)

```

8  for each variable  $x^i \in x$  do
9      for  $n_1$  random samples of endog. state, exog. variables  $x \setminus x^i$ , action  $a$  do
10         Randomly sample  $n_2$  settings of variable  $x^i$ , construct full states  $s_1, s_2, \dots, s_{n_2}$ .
11         Compute  $\sigma_j^2 \leftarrow \text{Var}(\{R(s_1, a), \dots, R(s_{n_2}, a)\})$ .
12         if  $\frac{1}{n_1} \sum_{j=1}^{n_1} \sigma_j^2 > \tau_{\text{variance}}$  then // Mean variance above threshold, accept.
13             emit  $x^i$ 

```

Algorithm description. In line with Condition (1) of Theorem 1, we begin by estimating the set of exogenous state variables relevant to the agent’s reward function, which we can do by leveraging the generative model of the MDP. As shown by the ESTIMATE-REWARD-VARIABLES subroutine, the idea is to compute, for each variable x^i , the average amount of variance in the reward across different values of x^i when the remainder of the state is held fixed, and threshold this at τ_{variance} . We define our initial mask to be this set. The efficiency-accuracy tradeoff of this subroutine can be controlled by tuning the number of samples, n_1 and n_2 . In practice, we can make a further improvement by heuristically biasing the sampling to favor states that are more likely to be encountered by the agent (such heuristics can be computed, for instance, by planning in a relaxed version of the problem).

Then, we employ an iterative procedure to approximately build toward Conditions (2) and (3): that the variables not in the mask transition independently of those in the mask and the endogenous state. To do so, we greedily add variables into the mask based upon the mutual information between the empirical transition model of the reduced state and that of each remaining variable. Intuitively, this quantitatively measures: “How much better would we be able to predict the dynamics of \tilde{s} if \tilde{s} included variable x^i , versus if it didn’t?” This mutual information will be 0 if x^i transitions independently of all variables in \tilde{s} . To calculate the mutual information between \tilde{s} and variable x^i , we must first learn the empirical transition models $\hat{T}_{\tilde{s}, x^i}$, $\hat{T}_{\tilde{s}}$, and \hat{T}_{x^i} from data. The exogeneity of most of the state is critical here: not only does it make learning these models much more efficient, but also, without exogeneity, we could not be sure whether two variables actually transition independently of each other or we just happened to follow a data-gathering policy that led us to believe so.

6 Experiments

Our experiments are designed to answer the following questions: (1) In small domains, are the masks learned by our algorithm competitive with the optimal masks? (2) Quantitatively, how well do the learned masks perform in large, complicated domains? (3) Qualitatively, do the learned masks properly reflect different goals given to the robot? (4) What are the limitations of our approach?

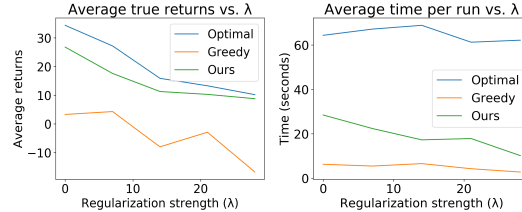
We experiment in simulated robotic manipulation domains in which a robot is placed in a busy environment with objects on tables, among many other agents that are also interacting with objects. The robot is rewarded for navigating to a given goal object (which changes on each episode) and penalized for crashing into other agents. The exogenous variables are the states of the other agents³,

³It is true that typically, the other agents will react to the robot’s actions, but it is well worth making the exogeneity assumption in these complicated domains we are considering. This assumption is akin to how we treat traffic patterns as exogenous to ourselves, even though technically we can slightly affect them.

a binary-valued occupancy grid discretizing the environment, the object placements on tables, and information specifying the goal. We plan using value iteration with a timeout of 60 seconds. Empirical value estimates are computed as averages across 500 independent rollouts. Each result reports an average across 50 independent trials. We regularize masks by setting $\text{Cost}(\tilde{x}) = |\tilde{x}|$; our initial experimentation suggested that other mask choices, such as planning time, perform similarly.

6.1 In small domains, are our learned masks competitive with the optimal masks?

Finding the optimal mask \tilde{x}^* is only possible in small models. Therefore, to answer this question we constructed a small gridworld representation of our experimental domain, with only ~ 600 states and 5 exogenous variables, in which we can plan exactly. We find \tilde{x}^* using the MASK-LEARNING-BRUTE-FORCE strategy given in Section 5.1. We compare the optimal mask with both 1) *Ours*, the mask returned by our algorithm in Section 5.3; and 2) *Greedy*, the mode mask chosen across 10 independent trials of the MASK-LEARNING-GREEDY strategy given in Section 5.1.

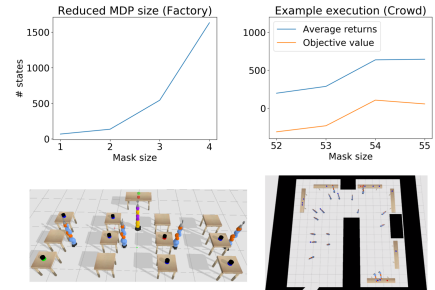


Discussion. For higher values of λ especially, our algorithm performs on par with the optimal brute-force strategy, which is only viable in small domains. The gap widens as λ decreases because as the optimal mask gets larger, it becomes harder to find using a forward selection strategy such as ours. In practical settings, one should typically set λ quite high so that smaller masks are preferred, as these will yield the most compact reduced models. Meanwhile, the greedy strategy does not perform well because it disregards the structure of the MDP and the conditions of Theorem 1. We also observe that our algorithm takes significantly less time than the optimal brute-force strategy; we should expect this gap to widen further in larger domains. Our algorithm’s stopping condition is that the score function begins to decrease, and so it tends to terminate more quickly for higher values of λ , as smaller masks become increasingly preferred.

6.2 Quantitatively, how well do the learned masks perform in large, complicated domains?

To answer this question, we consider two large domains: *Factory*, in which a robot must fulfill manipulation tasks being issued in a stream; and *Crowd*, in which a robot must navigate to target objects among a crowd of agents executing their own fixed stochastic policies. Even after discretization, these domains have 10^{11} and 10^{80} states respectively; *Factory* has 22 exogenous variables and *Crowd* has 124. In either domain, both planning exactly in the full MDP and searching for \tilde{x}^* by brute force are prohibitively expensive. All results use $n_{\text{rollouts}} = 500$, $\tau_{\text{correl}} = 10^{-5}$, $\tau_{\text{variance}} = 0$, $n_1 = 250$, and $n_2 = 5$. We compare our algorithm with the *Greedy* one described earlier, which disregards the structure of the MDP. We also compare to only running our algorithm’s first phase, which chooses the mask to be the estimated set of variables directly influencing the reward.

Algorithm	Domain	Average True Returns	Time / Run (sec)
Greedy	Factory	0	13.4
Ours (first phase)	Factory	186	—
Ours (full)	Factory	226	53.8
Greedy + heuristics	Crowd	188	43.7
Ours (first phase)	Crowd	260	—
Ours (full)	Crowd	545	123.5



Screenshots of *Factory* domain (L) and *Crowd* domain (R).

Discussion. In the *Factory* domain, the baseline greedy algorithm did not succeed even once at the task. The reason is that in this domain, several exogenous variables directly influence the reward function, but the greedy algorithm starts with an empty mask and only adds one variable at a time, and so cannot detect the improvement arising from adding in several variables at once. To give the

baseline a fair chance, we initialized it by hand to a better mask for the *Crowd* domain, which is why it sees success there (*Greedy + heuristics* in the table). The results suggest that our algorithm, explicitly framed around *all* conditions of Theorem 1, performs well. The graph of the example execution in the *Crowd* domain shows that adding a 55th variable to the mask yields a decrease in the estimated objective even though the average returns slightly increase, due to the regularizer $|\tilde{x}|$.

6.3 Qualitatively, do the learned masks properly reflect different goals given to the robot?

Please see the [supplementary video](#) for more qualitative results, examples, and explanations.

An important characteristic of our algorithm is its ability to learn different masks based on what the goal is; we illustrate this concept with an example. Let us explore the masks resulting from two different goals in the *Crowd* domain: Goal (1) is for the robot to navigate to an object that cannot be moved by any of the other agents, and Goal (2) is for the robot to navigate to an object that is manipulable by the other agents. In either case, the variables that directly affect the reward function are the object placements on the tables (which tell the robot where it must navigate to) and the occupancy grid (which helps the robot avoid crashing). However, for Goal (2), there is another variable that is important to consider: the states of any other agents that can manipulate the objects. This desideratum gets captured by the second phase of our algorithm: reasoning about the states of the other agents will allow the robot to better predict the dynamics of the object placements, enabling it to succeed at its task more efficiently and earn higher rewards. See Figure 2 for a visualization of this concept in our experimental domain, simulated using *pybullet* [16].



Figure 2: *Left*: Example of the full world. We control the singular gold-and-purple robot in the environment; the others follow fixed policies and are exogenous. *Medium, Right*: Examples of reduced models learned by the robot for Goals (1) and (2). The red squares on the ground describe which locations, within a discretization of the environment, the robot has learned to consider within its masked occupancy grid, while green circles denote currently occupied locations. Because all goals require manipulating objects on the tables, the robot recognizes that it does not need to consider occupancies in the lower-left quarter of the environment. For Goal (1), in which the other agents cannot manipulate the objects, the robot recognizes that it does not need to consider the states of any other agents. For Goal (2), the robot considers one of the other agents (here, holding the green object) within its reduced model, since this helps it better predict the dynamics of the objects.

Discussion. In a real-world setting, all of the exogenous variables could potentially be relevant to solving *some* problem, but typically only a small subset will be relevant to a *particular* problem. Under this lens, our method gives a way of deriving option policies in lower-dimensional subspaces.

6.4 What are the limitations of our approach?

Our experimentation revealed some limitations of our approach that are valuable to discuss. If the domains of the exogenous variables are large (or continuous), then it is expensive to compute the necessary mutual information quantities. To remedy this issue, one could turn to techniques for estimating mutual information, such as MINE [17]. Another limitation is that the algorithm as presented greedily adds one variable to the mask at a time, after the initial mask is built. In some settings, it can be useful to instead search over groups of variables to add in all at once, since these may contain information for better predicting dynamics that is not present in any single variable.

7 Conclusion and Future Work

We have formalized and given a tractable algorithm for the *mask-learning problem*, in which an agent must choose a subset of the exogenous state variables of an MDP to reason about when planning. An important avenue for future work is to remove the assumption that the agent knows the partition of endogenous versus exogenous aspects of the state. An interesting fact to ponder is that the agent can actually *control* this partition by choosing its actions appropriately. Thus, the agent can commit to a particular choice of exogenous variables in the world, and plan under the constraint of never influencing these variables. Another avenue for future work is to develop an incremental, real-time version of the algorithm, necessary in settings where the agent’s task constantly changes.

Acknowledgments

We gratefully acknowledge support from NSF grants 1523767 and 1723381; from AFOSR grant FA9550-17-1-0165; from ONR grant N00014-18-1-2847; from Honda Research; and from the MIT-Sensetime Alliance on AI. Rohan is supported by an NSF Graduate Research Fellowship. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of our sponsors.

References

- [1] C. Boutilier, T. Dean, and S. Hanks. Decision-theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research*, 11:1–94, 1999.
- [2] T. Dietterich, G. Trimonias, and Z. Chen. Discovering and removing exogenous state variables and rewards for reinforcement learning. In *International Conference on Machine Learning*, pages 1261–1269, 2018.
- [3] T. Dean and R. Givan. Model minimization in Markov decision processes. In *AAAI/IAAI*, pages 106–111, 1997.
- [4] R. Givan, T. Dean, and M. Greig. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence*, 147(1-2):163–223, 2003.
- [5] N. K. Jong and P. Stone. State abstraction discovery from irrelevant state variables. In *IJCAI*, volume 8, pages 752–757. Citeseer, 2005.
- [6] N. Mehta, S. Ray, P. Tadepalli, and T. Dietterich. Automatic discovery and transfer of MAXQ hierarchies. In *Proceedings of the 25th international conference on Machine learning*, pages 648–655. ACM, 2008.
- [7] H. Mao, S. B. Venkatakrisnan, M. Schwarzkopf, and M. Alizadeh. Variance reduction for reinforcement learning in input-driven environments. *arXiv preprint arXiv:1807.02264*, 2018.
- [8] J. Choi, Y. Guo, M. L. Moczulski, J. Oh, N. Wu, M. Norouzi, and H. Lee. Contingency-aware exploration in reinforcement learning. 2019.
- [9] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.
- [10] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2):273–324, 1997.
- [11] A. Miller. *Subset selection in regression*. Chapman and Hall/CRC, 2002.
- [12] G. H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In *Machine Learning Proceedings 1994*, pages 121–129. Elsevier, 1994.
- [13] J. Yang and V. Honavar. Feature subset selection using a genetic algorithm. In *Feature extraction, construction and selection*, pages 117–136. Springer, 1998.
- [14] M. Kearns, Y. Mansour, and A. Y. Ng. A sparse sampling algorithm for near-optimal planning in large Markov decision processes. *Machine learning*, 49(2-3):193–208, 2002.
- [15] W. Hoeffding. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pages 409–426. Springer, 1994.
- [16] E. Coumans, Y. Bai, and J. Hsu. Pybullet physics engine. 2018. URL <http://pybullet.org/>.
- [17] M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm. MINE: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.
- [18] M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Appendix A: Proof of Theorem 1

Theorem 1. Consider an MDP M as defined in Section 3.1, with exogenous state variables $x = [x^1 \ x^2 \ \dots \ x^m]$, and a mask $\tilde{x} \subseteq x$. Let $\bar{x} = x \setminus \tilde{x}$ be the variables not included in the mask. If the following conditions hold: (1) $R^i(n_t, x_t^i, a_t) = 0 \ \forall x^i \in \bar{x}$, (2) $P(n_{t+1} \mid n_t, a_t, x_t) = P(n_{t+1} \mid n_t, a_t, \tilde{x}_t)$, (3) $P(\tilde{x}_{t+1}, \bar{x}_{t+1} \mid \tilde{x}_t, \bar{x}_t) = P(\tilde{x}_{t+1} \mid \tilde{x}_t) \cdot P(\bar{x}_{t+1} \mid \bar{x}_t)$; then $\tilde{V}_{\tilde{\pi}}(\tilde{s}) = V_{\tilde{\pi}}(s) \ \forall s \in \mathcal{S}$. If $\tilde{\pi}$ is optimal for the reduced MDP \tilde{M} , then it must also be true that $\tilde{V}_{\tilde{\pi}}(\tilde{s}) = V^*(s) \ \forall s \in \mathcal{S}$.

Proof: Consider an arbitrary state $s \in \mathcal{S}$, with corresponding reduced state \tilde{s} . We begin by showing that under the stated conditions, $\tilde{V}_{\tilde{\pi}}(\tilde{s}) = V_{\tilde{\pi}}(s)$. The recursive form of these value functions is:

$$\begin{aligned} V_{\tilde{\pi}}(s) &= R(s, \tilde{\pi}(\tilde{s})) + \gamma \sum_{s'} P(s' \mid s, \tilde{\pi}(\tilde{s})) \cdot V_{\tilde{\pi}}(s'), \\ \tilde{V}_{\tilde{\pi}}(\tilde{s}) &= R(\tilde{s}, \tilde{\pi}(\tilde{s})) + \gamma \sum_{\tilde{s}'} P(\tilde{s}' \mid \tilde{s}, \tilde{\pi}(\tilde{s})) \cdot \tilde{V}_{\tilde{\pi}}(\tilde{s}'). \end{aligned}$$

Now, consider an iterative procedure for obtaining these value functions, which repeatedly applies the above equations starting from $V_{\tilde{\pi}}^0(s) = \tilde{V}_{\tilde{\pi}}^0(\tilde{s}) = 0 \ \forall s \in \mathcal{S}$. Let the value functions at iteration k be denoted as $V_{\tilde{\pi}}^k(s)$ and $\tilde{V}_{\tilde{\pi}}^k(\tilde{s})$. We will show by induction on k that $V_{\tilde{\pi}}^k(s) = \tilde{V}_{\tilde{\pi}}^k(\tilde{s}) \ \forall k$.

The base case is immediate. Suppose $V_{\tilde{\pi}}^k(s) = \tilde{V}_{\tilde{\pi}}^k(\tilde{s}) \ \forall s \in \mathcal{S}$, for some value of k . We compute:

$$\begin{aligned} V_{\tilde{\pi}}^{k+1}(s) &= R(s, \tilde{\pi}(\tilde{s})) + \gamma \sum_{s'} P(s' \mid s, \tilde{\pi}(\tilde{s})) \cdot V_{\tilde{\pi}}^k(s') \\ &= \sum_{i=1}^m R^i(n, x^i, \tilde{\pi}(\tilde{s})) + \gamma \sum_{s'} P(n' \mid n, \tilde{\pi}(\tilde{s}), x) \cdot P(x' \mid x) \cdot V_{\tilde{\pi}}^k(s') && \text{Model assumptions.} \\ &= \sum_{x^i \in \tilde{x}} R^i(n, x^i, \tilde{\pi}(\tilde{s})) + \gamma \sum_{s'} P(n' \mid n, \tilde{\pi}(\tilde{s}), x) \cdot P(x' \mid x) \cdot V_{\tilde{\pi}}^k(s') && \text{Condition (1).} \\ &= R(\tilde{s}, \tilde{\pi}(\tilde{s})) + \gamma \sum_{s'} P(n' \mid n, \tilde{\pi}(\tilde{s}), x) \cdot P(x' \mid x) \cdot V_{\tilde{\pi}}^k(s') && \text{Defn. of reduced } R. \\ &= R(\tilde{s}, \tilde{\pi}(\tilde{s})) + \gamma \sum_{s'} P(n' \mid n, \tilde{\pi}(\tilde{s}), \tilde{x}) \cdot P(x' \mid x) \cdot V_{\tilde{\pi}}^k(s') && \text{Condition (2).} \\ &= R(\tilde{s}, \tilde{\pi}(\tilde{s})) + \gamma \sum_{n', \tilde{x}'} \sum_{\bar{x}'} P(n' \mid n, \tilde{\pi}(\tilde{s}), \tilde{x}) \cdot P(\tilde{x}', \bar{x}' \mid \tilde{x}, \bar{x}) \cdot V_{\tilde{\pi}}^k(s') && \text{Split up sum.} \\ &= R(\tilde{s}, \tilde{\pi}(\tilde{s})) + \gamma \sum_{n', \tilde{x}'} \sum_{\bar{x}'} P(n' \mid n, \tilde{\pi}(\tilde{s}), \tilde{x}) \cdot P(\tilde{x}' \mid \tilde{x}) \cdot P(\bar{x}' \mid \bar{x}) \cdot V_{\tilde{\pi}}^k(s') && \text{Condition (3).} \\ &= R(\tilde{s}, \tilde{\pi}(\tilde{s})) + \gamma \sum_{n', \tilde{x}'} \sum_{\bar{x}'} P(n' \mid n, \tilde{\pi}(\tilde{s}), \tilde{x}) \cdot P(\tilde{x}' \mid \tilde{x}) \cdot P(\bar{x}' \mid \bar{x}) \cdot \tilde{V}_{\tilde{\pi}}^k(n', \tilde{x}') && \text{Inductive assumption.} \\ &= R(\tilde{s}, \tilde{\pi}(\tilde{s})) + \gamma \sum_{n', \tilde{x}'} P(n' \mid n, \tilde{\pi}(\tilde{s}), \tilde{x}) \cdot P(\tilde{x}' \mid \tilde{x}) \cdot \tilde{V}_{\tilde{\pi}}^k(n', \tilde{x}') \cdot \cancel{\sum_{\bar{x}'} P(\bar{x}' \mid \bar{x})} && \text{Rearrange.} \\ &= R(\tilde{s}, \tilde{\pi}(\tilde{s})) + \gamma \sum_{\tilde{s}'} P(\tilde{s}' \mid \tilde{s}, \tilde{\pi}(\tilde{s})) \cdot \tilde{V}_{\tilde{\pi}}^k(\tilde{s}') && \text{Defn. of reduced } T. \\ &= \tilde{V}_{\tilde{\pi}}^{k+1}(\tilde{s}). \end{aligned}$$

We have shown that $V_{\tilde{\pi}}^k(s) = \tilde{V}_{\tilde{\pi}}^k(\tilde{s}) \ \forall k$. By standard arguments (e.g., in [18]), this iterative procedure converges to the true $V_{\tilde{\pi}}(s)$ and $\tilde{V}_{\tilde{\pi}}(\tilde{s})$ respectively. Therefore, we have that $\tilde{V}_{\tilde{\pi}}(\tilde{s}) = V_{\tilde{\pi}}(s)$.

Now, if $\tilde{\pi}$ is optimal for \tilde{M} , then it is optimal for the full MDP M as well. This is because Condition (1) assures us that the variables not considered in the mask do not affect the reward, implying that $\tilde{\pi}$ optimizes the expected reward in not just \tilde{M} , but also M . Therefore, under this assumption, we have that $\tilde{V}_{\tilde{\pi}}(\tilde{s}) = V_{\tilde{\pi}}(s) = V^*(s) \ \forall s \in \mathcal{S}$. \square