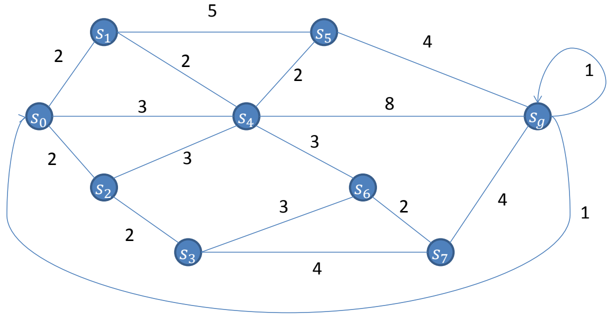

Figure 4. A traffic network. A car starts in s_0 and seeks to reach s_g by the quickest path. The time to traverse each edge under ideal conditions is shown. Exogenous traffic can add delays to these times.

A. Supplementary Materials

We experimented with several additional test problems that did not fit into the main body of the paper.

A.1. Problem 1: Route Planning with Traffic

Consider training a self-driving car to minimize the time it takes to get to the office every morning. It is natural to reward the car for minimizing the travel time, but the primary factor determining the reward is the exogenous traffic of all the other drivers on the road. This is similar in many ways to the cellular network management problem. The purpose of this first problem is to see if our methods work on a simple version of this problem. Figure 4 shows a road network MDP. The endogenous states are the nodes of the network. The exogenous state X_t is the level of traffic in the network. It evolves according to the linear system $X_{t+1} = 0.9X_t + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$. The reward function is:

$$r_t = \frac{1}{cost(s_t \to s_{t+1})} + X_t.$$

The actions at each node consist of choosing one of the outbound edges to traverse. To make the task easier, we restrict the set of actions to move only rightward (i.e., toward states with higher subscripts) except that s_g can return to s_0. The cost of traversing an edge is shown by the weights in Figure 4. For example, if the agent moves from s_0 to s_4, the $cost(s_0 \to s_4) = 3$.

The Q function is represented as a neural network with a 1-hot encoding of the 9 states plus a tenth input unit for X_t and an eleventh input unit for A_t. The PCC threshold $\epsilon = 0.05$.

Figure 5 confirms that both Endo-Q learners are able to learn much faster than the Q-learner that is given the full MDP reward and that they are able to match the performance of the oracle.

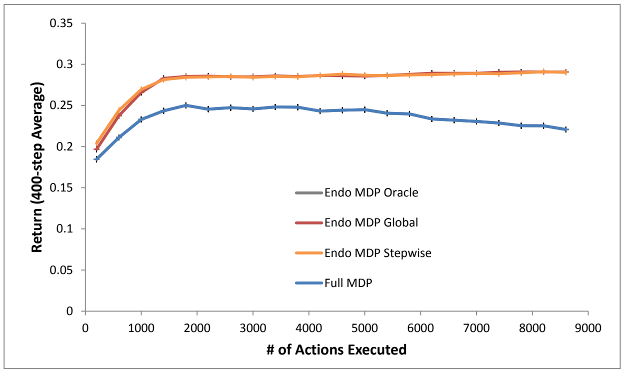

Figure 5. Comparison of Q Learning applied to the Full MDP and to the Endogenous MDP for the traffic network problem ($N = 200, T = 400$).

A.2. Problem 2: Linear system with 2-d exogenous state

Let $X_{1,t}$ and $X_{2,t}$ be the exogenous state variables and E_t be the endogenous state. The state transition function is defined as

$$\begin{aligned} X_{1,t+1} &= 0.9X_{1,t} + \epsilon_1, \\ X_{2,t+1} &= 0.7X_{2,t} + \epsilon_2, \\ E_{t+1} &= 0.4E_t + A_t + 0.1X_{1,t} + 0.1X_{2,t} + \epsilon_3, \end{aligned} \tag{14}$$

where $\epsilon_1 \sim \mathcal{N}(0, 0.16)$ and $\epsilon_2 \sim \mathcal{N}(0, 0.04)$ and $\epsilon_3 \sim \mathcal{N}(0, 0.04)$.

The observed state vector S_t is a linear mixture of the hidden exogenous and endogenous states $S_t = M[X_{1,t}, X_{2,t}, E_t]^\top$,where

$$M = \begin{bmatrix} 0.3, & 0.6, & 0.7 \\ 0.3, & -0.7, & 0.2 \\ 0.6, & 0.3, & 0.2 \end{bmatrix}$$

The reward at time t is defined as $R_t = R_{x,t} + R_{e,t}$, where $R_{x,t}$ is the exogenous reward $R_{x,t} = -X_{1,t} - X_{2,t}$ and $R_{e,t}$ is the endogenous reward $R_{e,t} == \exp[|E_t - 3|/4]$. Figure 7 shows that with the exception of a few extreme states, the learned W_x successfully reconstructs the values of X_1 and X_2.

A.3. Problem 3: 5-d linear system with 3-d exogenous state

Let $X_{1,t}$, $X_{2,t}$, $X_{3,t}$ be the exogenous state variables and $E_{1,t}$, $E_{2,t}$ be the endogenous state variables. The state

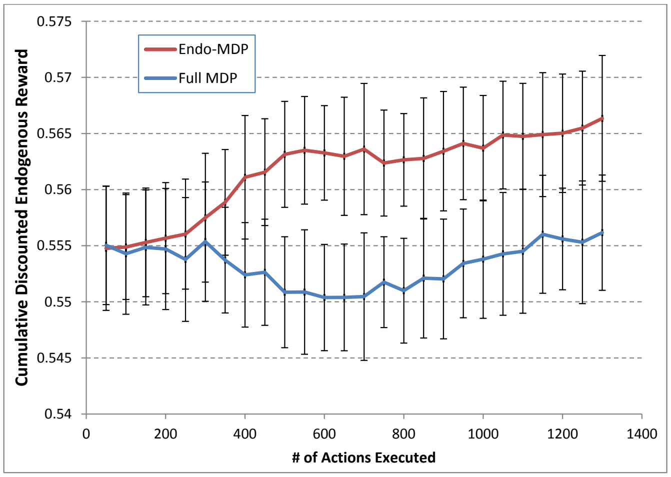

Figure 6. Comparison of Q Learning applied to the Full MDP and to the Endogenous MDP for a 3-d linear system with two coupled exogenous dimensions

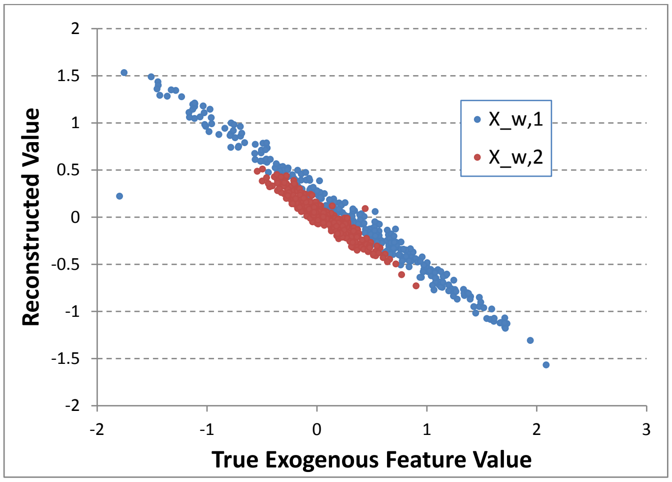

Figure 7. Comparison of Q Learning applied to the Full MDP and to the Endogenous MDP for a 3-d linear system with two coupled exogenous dimensions

transition function is defined as:

$$X_{1,t+1} = \frac{3}{5}X_{1,t} + \frac{9}{50}X_{2,t} + \frac{3}{10}X_{3,t} + \epsilon_1,$$

$$X_{2,t+1} = 7/15X_{2,t} + \frac{7}{50}X_{3,t} + \frac{7}{30}X_{1,t} + \epsilon_2,$$

$$X_{3,t+1} = \frac{8}{15}X_{3,t} + \frac{8}{50}X_{1,t} + \frac{7}{30}X_{2,t} + \epsilon_3,$$

$$E_{1,t+1} = \frac{13}{20}E_{1,t} + \frac{13}{40}E_{2,t} + A_t + 0.1X_{1,t} + 0.1X_{2,t} + \epsilon_4,$$

$$E_{2,t+1} = \frac{13}{20}E_{2,t} + \frac{13}{40}E_{1,t} + A_t + 0.1X_{2,t} + 0.1X_{3,t} + \epsilon_5,$$

where $\epsilon_1 \sim \mathcal{N}(0, 0.16)$, $\epsilon_2 \sim \mathcal{N}(0, 0.04)$, $\epsilon_3 \sim \mathcal{N}(0, 0.09)$, $\epsilon_4 \sim \mathcal{N}(0, 0.04)$, and $\epsilon_5 \sim \mathcal{N}(0, 0.04)$.

The observed state vector S_t is a linear mixture of the hidden exogenous and endogenous states:

$$S_t = \begin{bmatrix} 0.3 & 0.3 & 0.6 & 0.2 & -0.4 \\ 0.6 & -0.7 & 0.3 & 0.5 & -0.3 \\ 0.7 & 0.2 & 0.2 & -0.8 & 0.6 \\ 0.4 & -0.2 & -0.1 & -0.2 & 0.9 \\ 0.9 & 0.3 & -0.2 & 0.7 & -0.2 \end{bmatrix} \cdot \begin{bmatrix} X_{3,t} \\ X_{2,t} \\ X_{1,t} \\ E_{2,t} \\ E_{1,t} \end{bmatrix}.$$

The reward at time t is defined as $R_t = R_{x,t} + R_{e,t}$, where $R_{x,t}$ is the exogenous reward $R_{x,t} = -1.4X_{1,t} - 1.7X_{2,t} - 1.8X_{3,t}$ and $R_{e,t}$ is the endogenous reward $R_{e,t} = \exp[-\frac{|E_{1,t}+1.5E_{2.t}-1|}{5}]$. The action A_t can take the discrete values $\{-1.0, -0.9, \ldots, 0.9, 1.0\}$.

The PCC threshold was set to 0.1 for this problem.

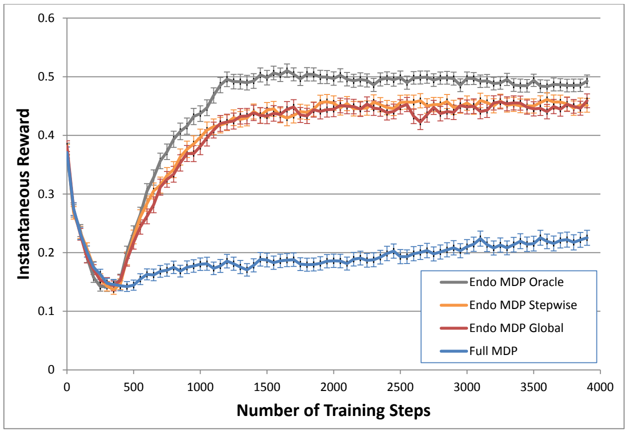

Figure 8. Comparison of Q Learning applied to the Full MDP and to the Endogenous MDP for a 5-d linear system with three coupled exogenous dimensions ($T = 50$, $N = 1000$).

Figure 8 shows the performance of four Q Learning algorithms: "endo oracle" is trained on the true endogenous reward, "endo stepwise" and "endo global" are trained on the estimated endogenous reward after applying either the stepwise or the global optimization methods to estimate W_x, and "full MDP" is trained on the original MDP. Q learning on "full MDP" is very slow, whereas both "endo stepwise" and "endo global" are able to learn nearly as quickly as "endo oracle". There is no apparent difference between the two optimization methods.