DISCOVERING AND REMOVING EXOGENOUS STATE VARIABLES AND REWARDS FOR REINFORCEMENT LEARNING

Tom Dietterich (Oregon State)
George Trimponias (Huawei Noah's Ark Lab)
Zhitang Chen (Huawei Noah's Ark Lab)

Motivation

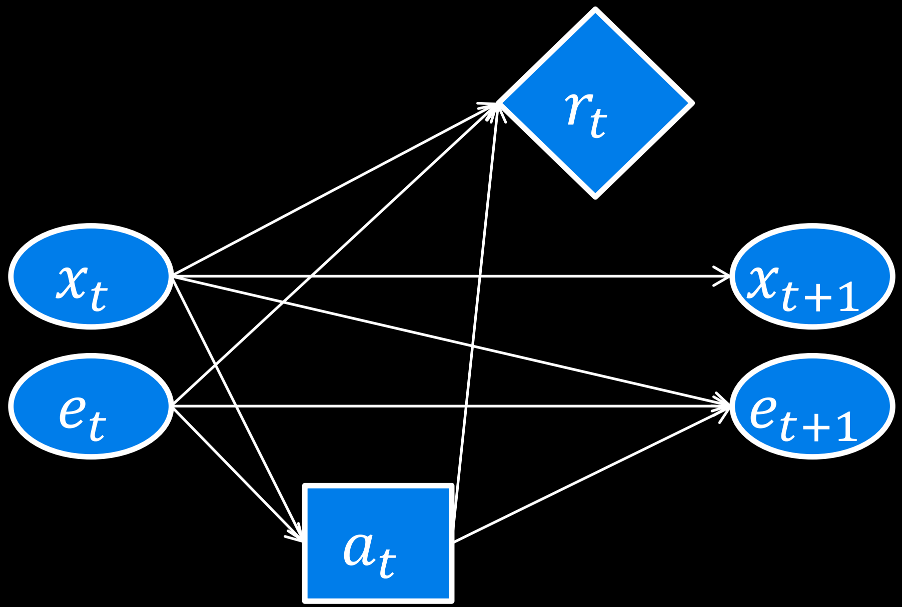
- Consider training your car to drive you to work every day
- MDP
  - states: car location + traffic
  - actions: turns to make
  - cost: total time to reach the office
- Problem:
  - Your actions only control part of the cost. Most of the cost is determined by what other drivers are doing

Consequences

- The cost of any policy π will have high variance
- This will require smaller learning rates and larger training samples
- Policy gradient will require tiny step sizes
  - Needs to average over many trajectories to estimate $\nabla_\theta V^\pi(s_0; \theta)$
- Q learning will require tiny learning rate
  - Needs to average over many transitions to estimate $Q(s, a)$

Exogenous State MDP

- MDP state can be partitioned into $s = (x, e)$, where x is exogenous and e is endogeneous



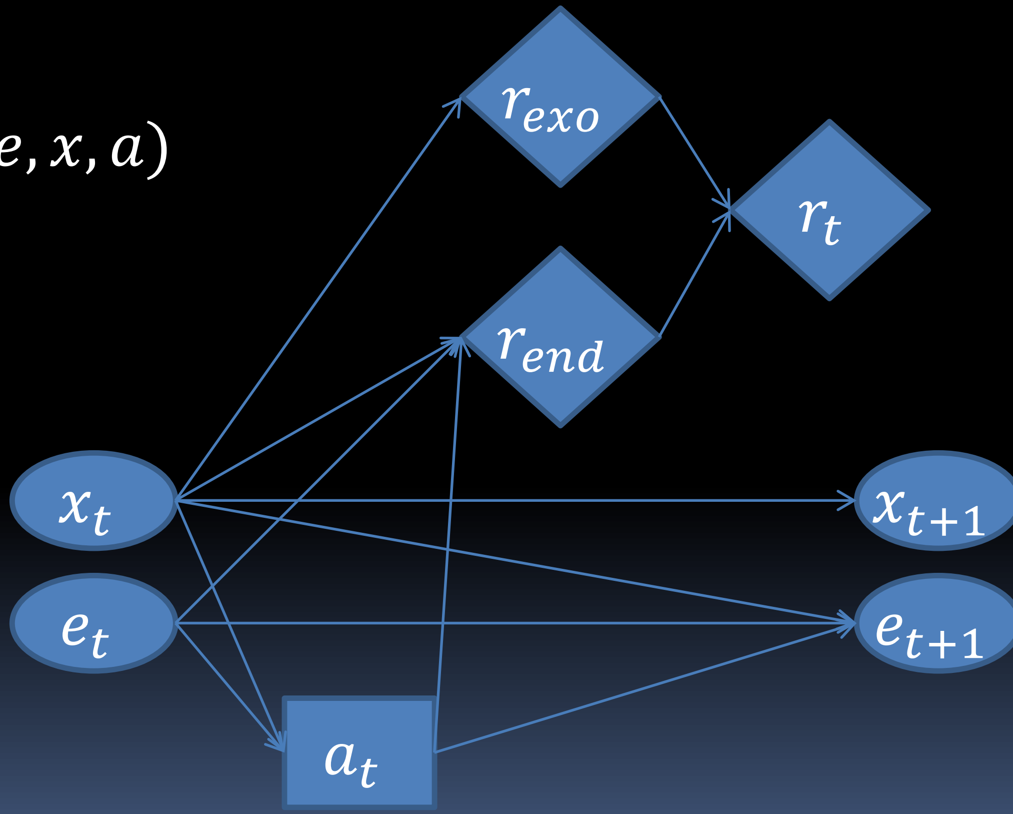
- Transitions:
  - $P(x_{t+1}, e_{t+1} | x_t, e_t, a_t) = P(e_{t+1} | x_t, e_t, a_t) P(x_{t+1} | x_t)$

Actions only affect e_{t+1}
x evolves independently but is still Markov

Analysis

- Assumption: Reward Decomposes Additively

$$r(e, x, a) = r_{exo}(x) + r_{end}(e, x, a)$$



Exo-Endo Decomposition

- Theorem 1: Any exogenous MDP can be decomposed into an exogenous Markov Reward Process and an endogenous MDP

$$V^*(e,x) = V^*_{exo}(x) + V^*_{end}(e,x)$$

$$V^*_{exo}(x) = r_{exo}(x) + \gamma \mathbb{E}_{x' \sim P(x'|x)}[V^*_{exo}(x')]$$

$$V^*_{end}(e,x) = \max_a r_{end}(e,x,a) + \gamma \mathbb{E}_{x' \sim P(x'|x)} \mathbb{E}_{e' \sim P(e'|e,x,a)}[V^*_{end}(e',x')]$$

Corollary

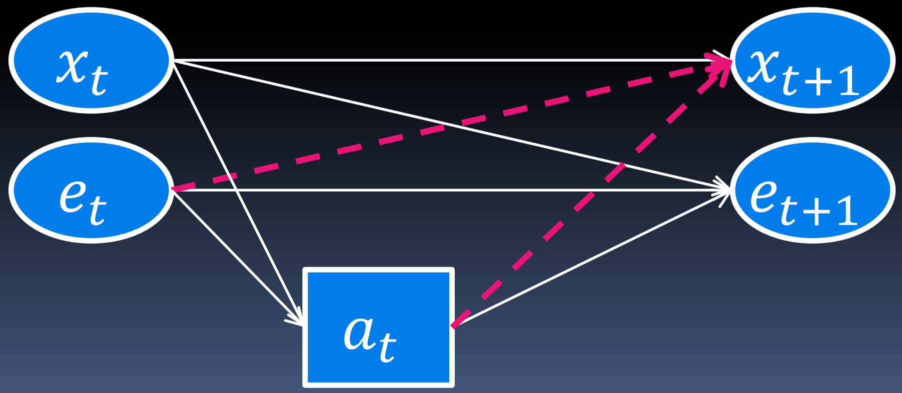
- Corollary: Any optimal policy for the endogenous MDP is an optimal policy for the original MDP

When is it easier to solve the Endogenous MDP?

- Answer: When the variance of the return of the Endogenous MDP is less than the variance of the return of the original MDP
- Covariance Condition:
  - Let $B(\tau)$ denote the cumulative discounted reward along trajectory τ
  - $\mathbb{V}ar[B_{exo}(\tau)] > -2\mathbb{C}ov[B_{end}(\tau), B_{exo}(\tau)]$
- Paper derives Bellman updates for variance and covariance of the return

Estimating the Endo-Exo Decomposition

- Suppose we have a database of transitions $\{(s_i, a_i, r_i, s_i')\}_{i=1}^{n}$ gathered by executing one or more exploration policies on the MDP
- Linear case ⇒ additive decomposition:
$$x = W^\top s;\ e = s - WW^\top s$$
- Find W to satisfy $I(x_{t+1}; (e_t, a_t) | x_t) = 0$



Two Algorithms

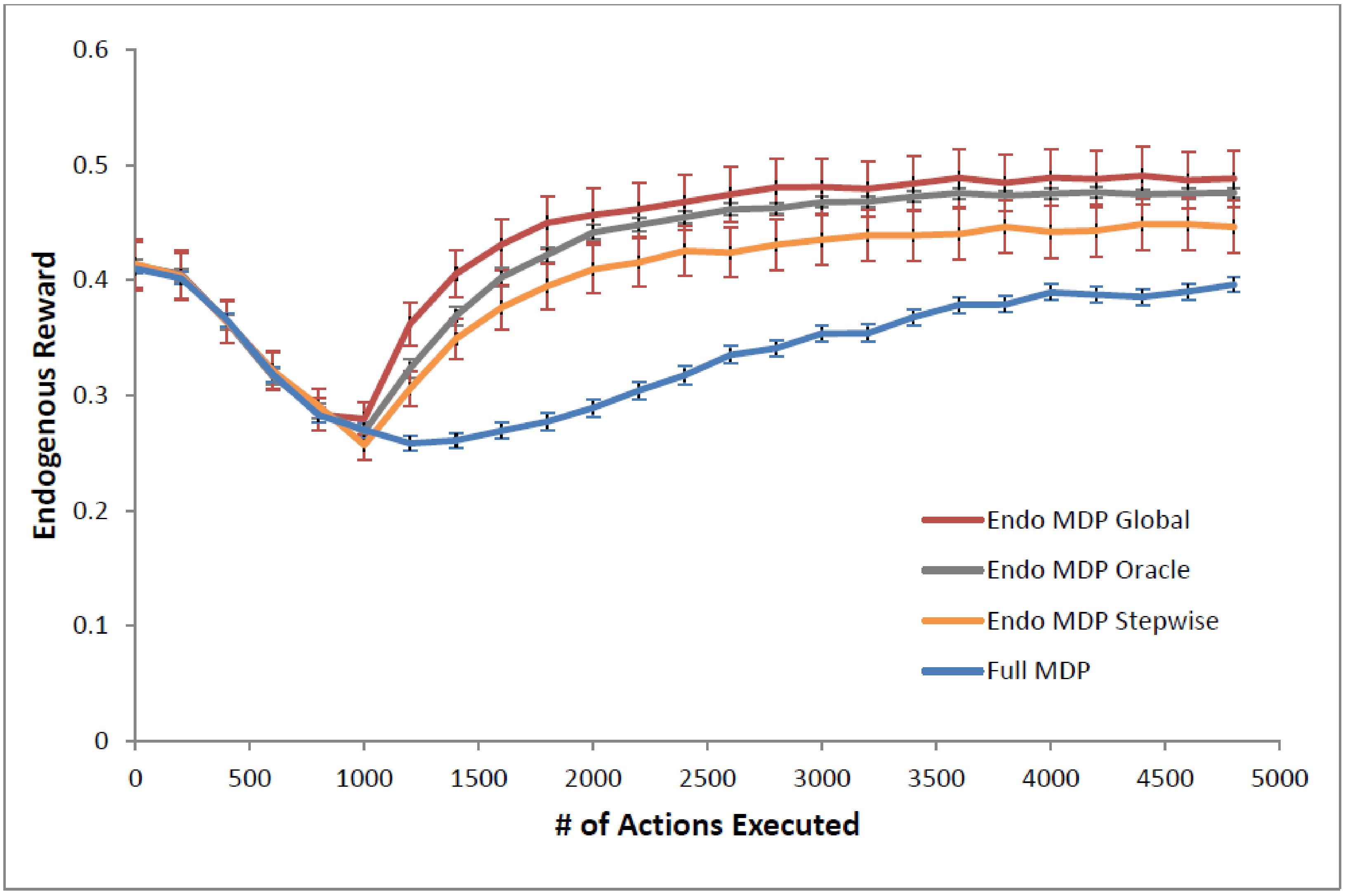
- Approximate $I(x_{t+1}; (e_t, a_t) | x_t)$ by the Partial Correlation Coefficient
- Global Algorithm
  - For each $1 \leq d_x \leq d$, compute a d-dimensional W
  - Solves d Steiffel Manifold optimizations of increasing size
- Stepwise Algorithm
  - Similar to PCA
  - Compute one column of W in each iteration
  - Solves d 1-dimensional Steiffel Manifold optimizations
- Matlab Manopt

Toy Problem 1: 30 Dimensions

- 15 dimensions are exogenous
- 15 dimensions are endogenous
- $X_{t+1} = M_x X_t + \mathcal{E}_x$
- $E_{t+1} = M_e \begin{bmatrix} E_t \\ X_t \\ A_t \end{bmatrix} + \mathcal{E}_e$
- $\mathcal{E}_x \sim \mathcal{N}(0,0.09)$; $\mathcal{E}_e \sim \mathcal{N}(0,0.04)$
- $S_t = M \begin{bmatrix} E_t \\ X_t \end{bmatrix}$
- $R_x = -3\, avg(X)$; $R_e = \exp[-|avg(E_t) - 1|]$
- M, M_x, M_e are random matrices with elements $\sim\mathcal{N}(0,1)$. Rows normalized to sum to 0.99.

- $\beta = 1$, learning rate = 0.05. 2 hidden layers w/ 40 tanh units

Results

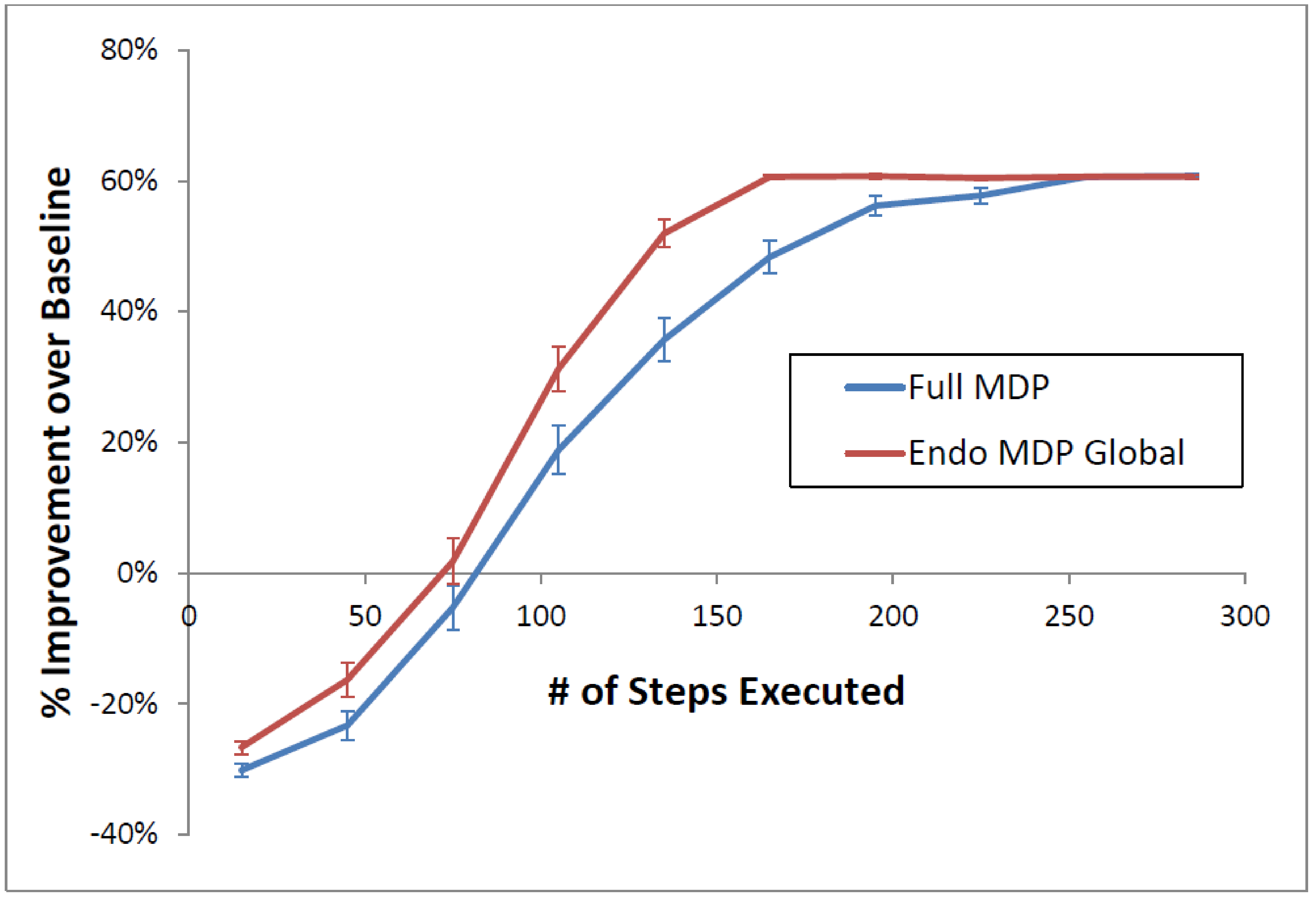

0.6
0.5
0.4
0.3
0.2
0.1
0
Endogenous Reward
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
# of Actions Executed
Endo MDP Global
Endo MDP Oracle
Endo MDP Stepwise
Full MDP

Cell Network Optimization

- Adjust cell tower parameters to minimize # of users experiencing poor throughput
- Action: increase/reduce threshold on signal power for when to switch channel for a user
- Time step: 1 hour
- Data: 5 days, hourly, 105 cells, Huawei Customer
- Simulator: MFMC (Fonteneau et al 2012)
- discount factor 0.95
- features: # active users, avg # of users, channel quality index, small packets/total packets; small packet bytes / total packet bytes

- Reward function: $R_t = -P_t =$ fraction of customers with low bandwidth during period $(t, t + \Delta t)$
- Separate fixed horizon evaluation trials

Results



80%
60%
40%
20%
0%
-20%
-40%
% Improvement over Baseline
0
50
100
150
200
250
300
# of Steps Executed
Full MDP
Endo MDP Global

Summary

- Exogenous state can lead to high-variance rewards, which make RL slow
- An MDP with exogenous state can be decomposed into an exogenous MRP and an endogenous MDP
- Solving the endogenous MDP gives an optimal policy for the original MDP

Acknowledgments

- Dietterich's time was supported by a gift to OSU from Huawei, Inc.

Questions?