Importance Sampling Policy Evaluation with an Estimated Behavior Policy

Josiah Hanna, Scott Niekum, and Peter Stone
Department of Computer Science
The University of Texas at Austin

LARG Learning Agents Research Group The University of Texas at Austin

PeARL Personal Autonomous Robotics Lab

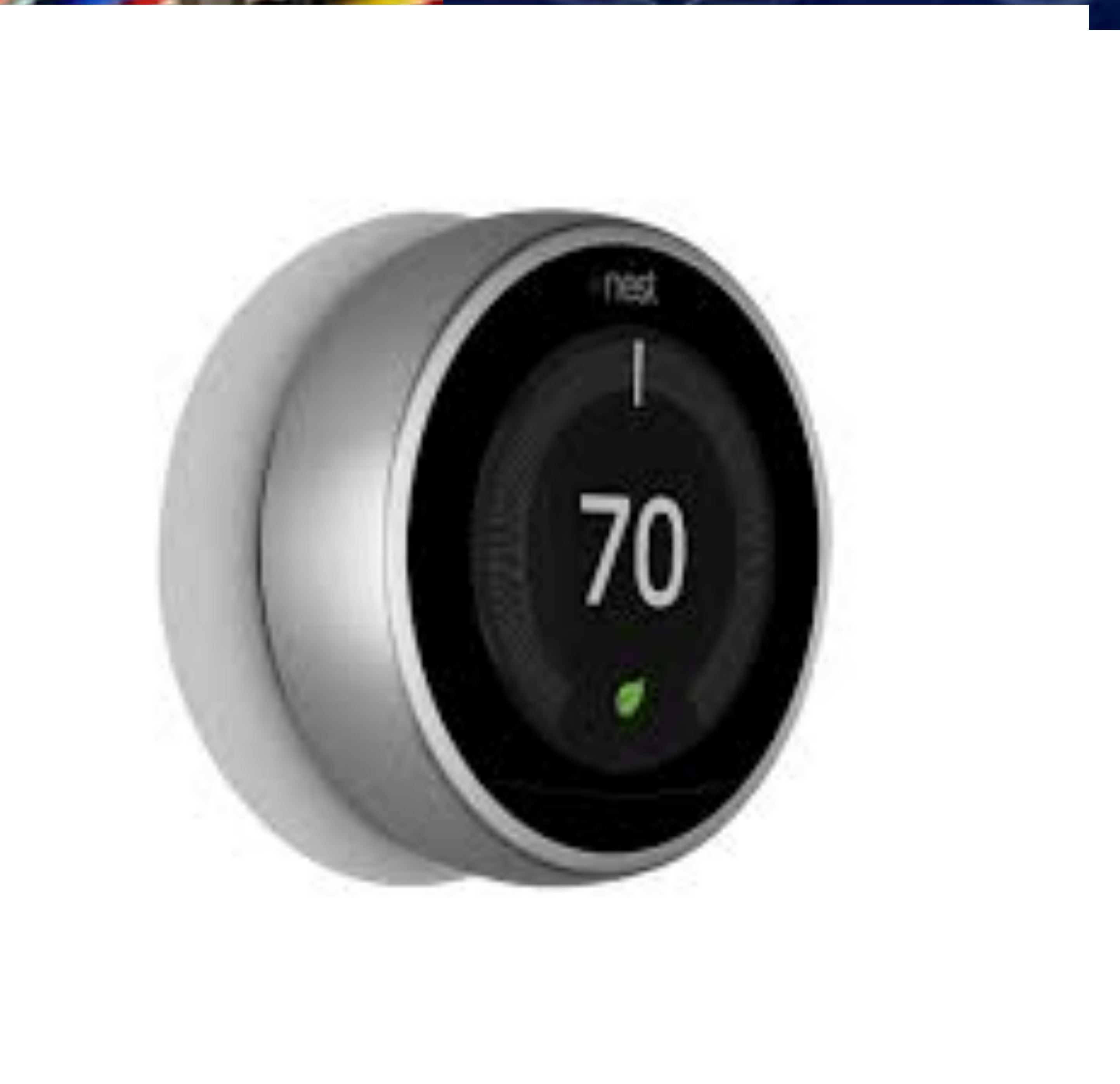

900

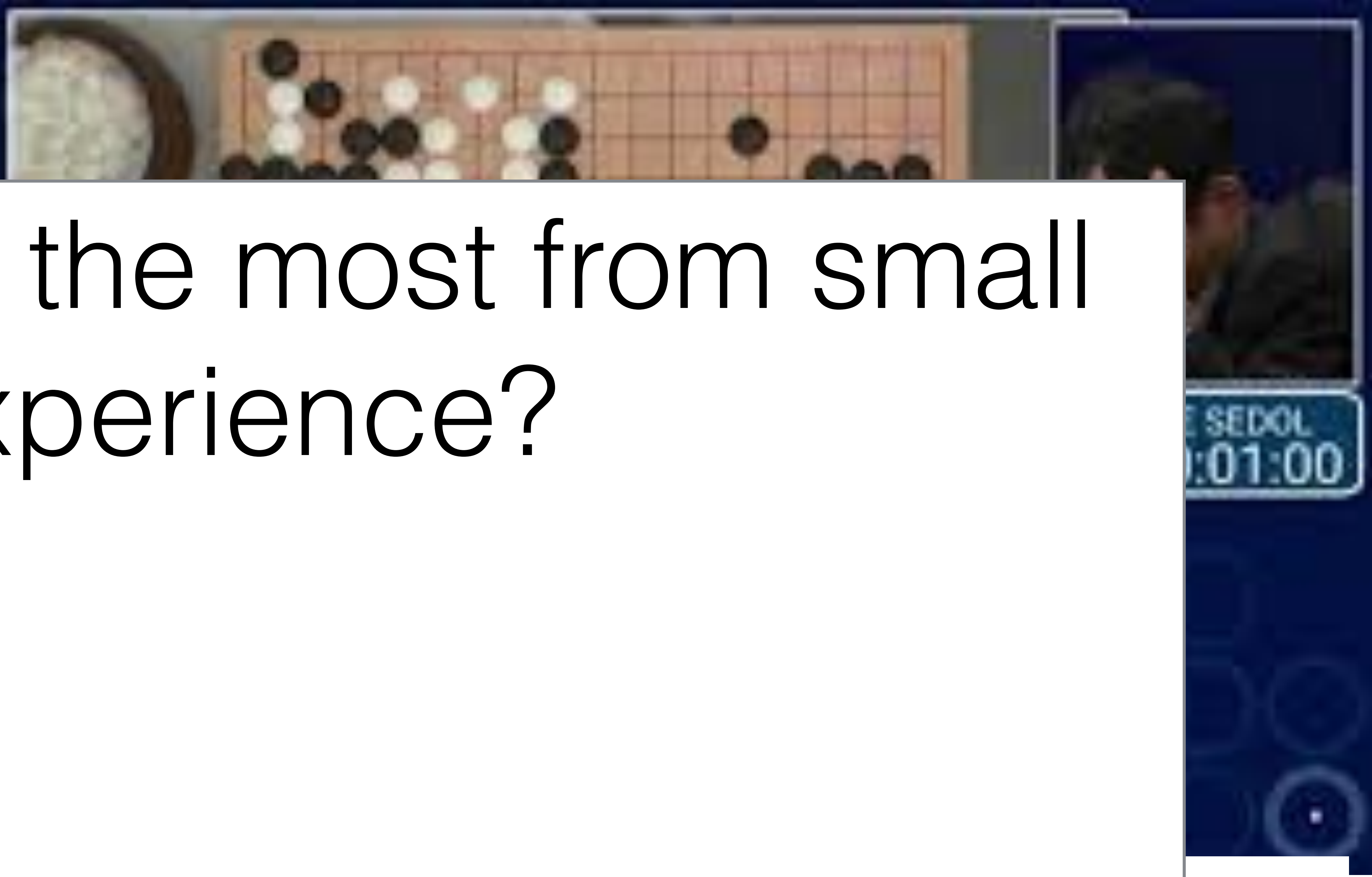

ALPHAGO
00:10:29
phaGo
Google DeepMind
LEE SEDOL
00:01:00


900

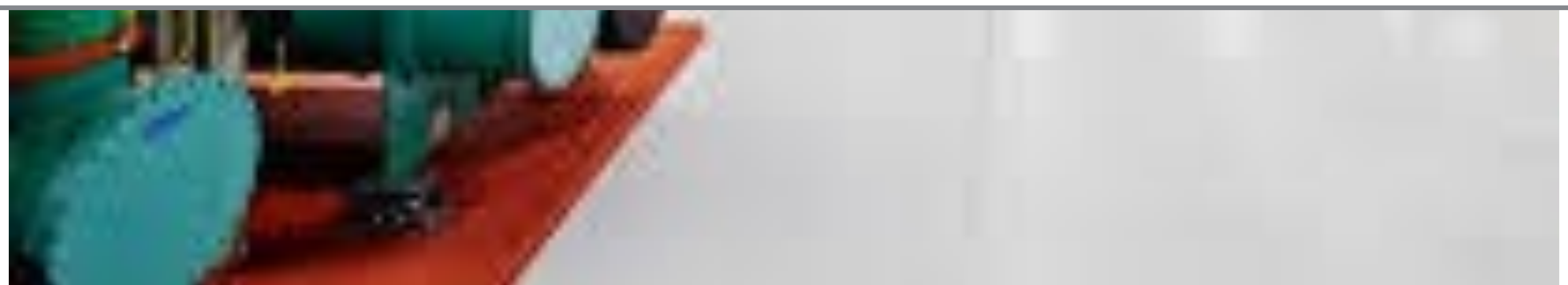
ALPHAGO
00:10:29
LEE SEDOL
00:01:00
phaGo

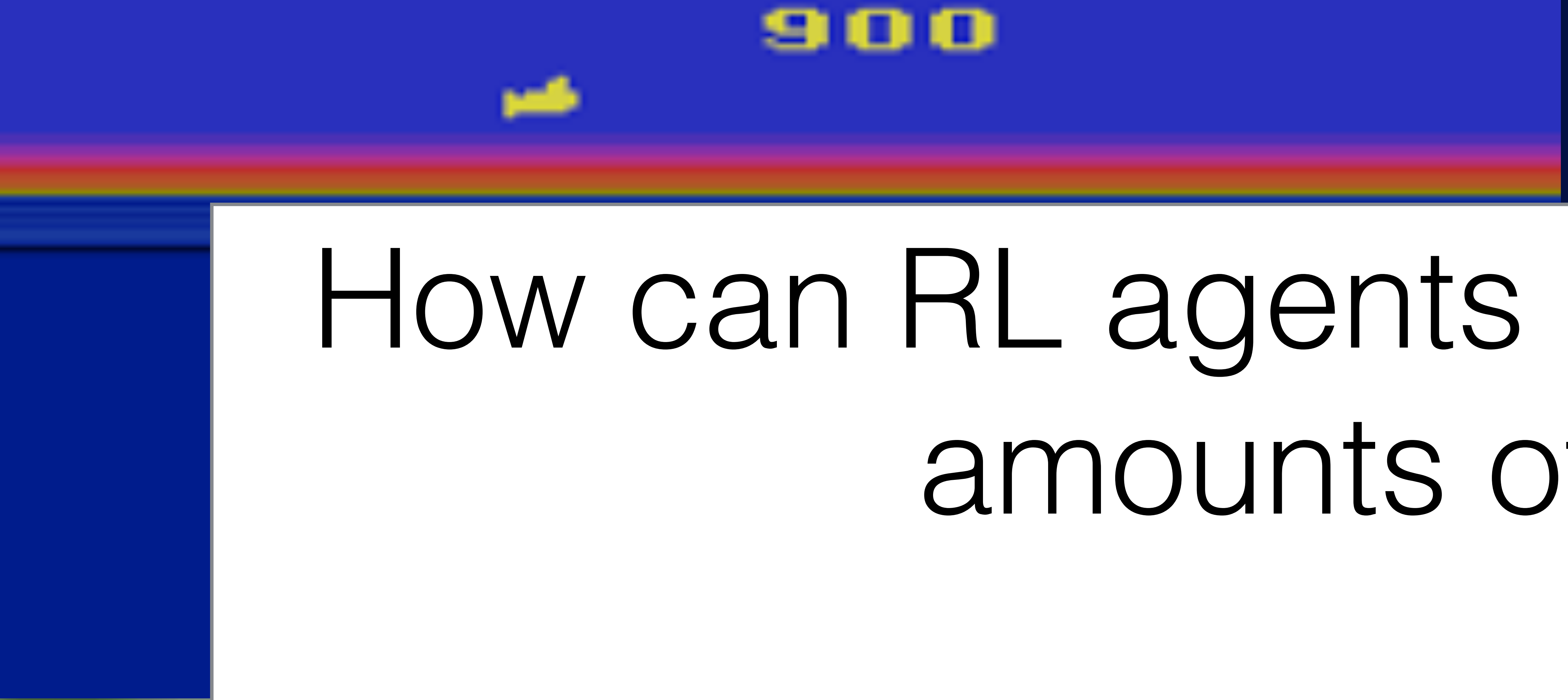

INTERNET
ADVERTISING
ENGINE
WEB
MARKETING
MARKETERS
COMPANIES
ONLINE
PRODUCT
BUSINESS
PRODUCTS
SERVICES
AFFILIATE
CONSUMERS
WEBSITE
INFORMATION
PER
PAY
CONSUMER
SPECIFIC
SALES
SEARCH

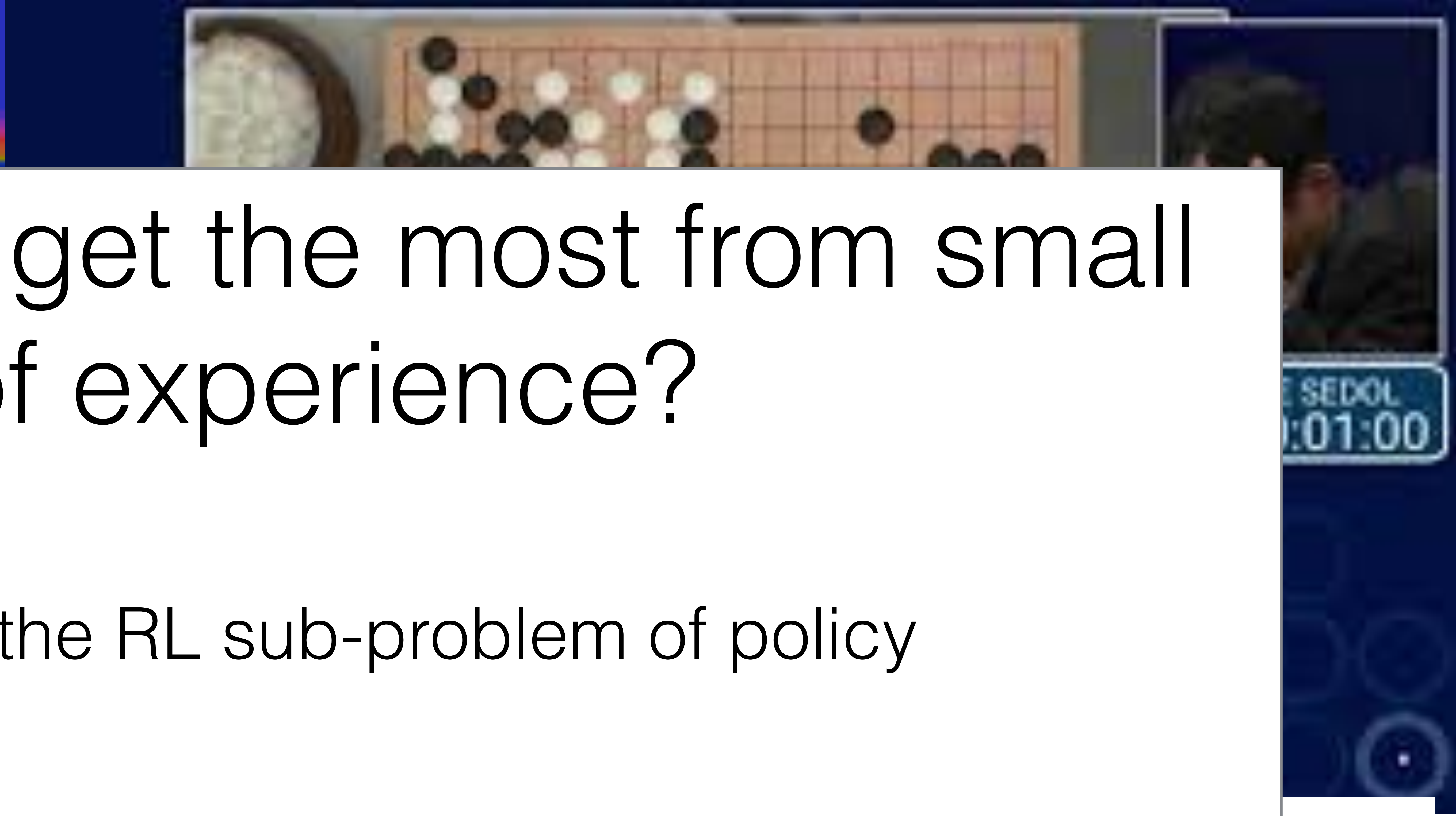

nest
70




How can RL agents get the most from small amounts of experience?

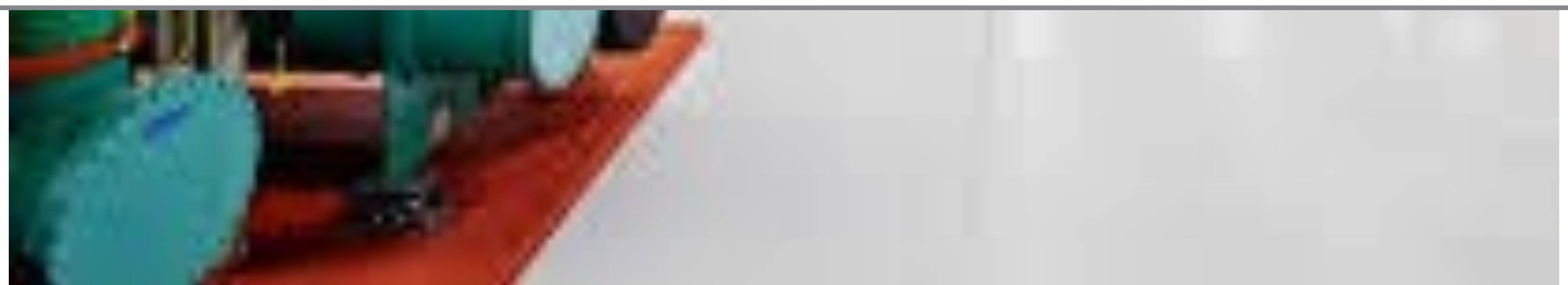


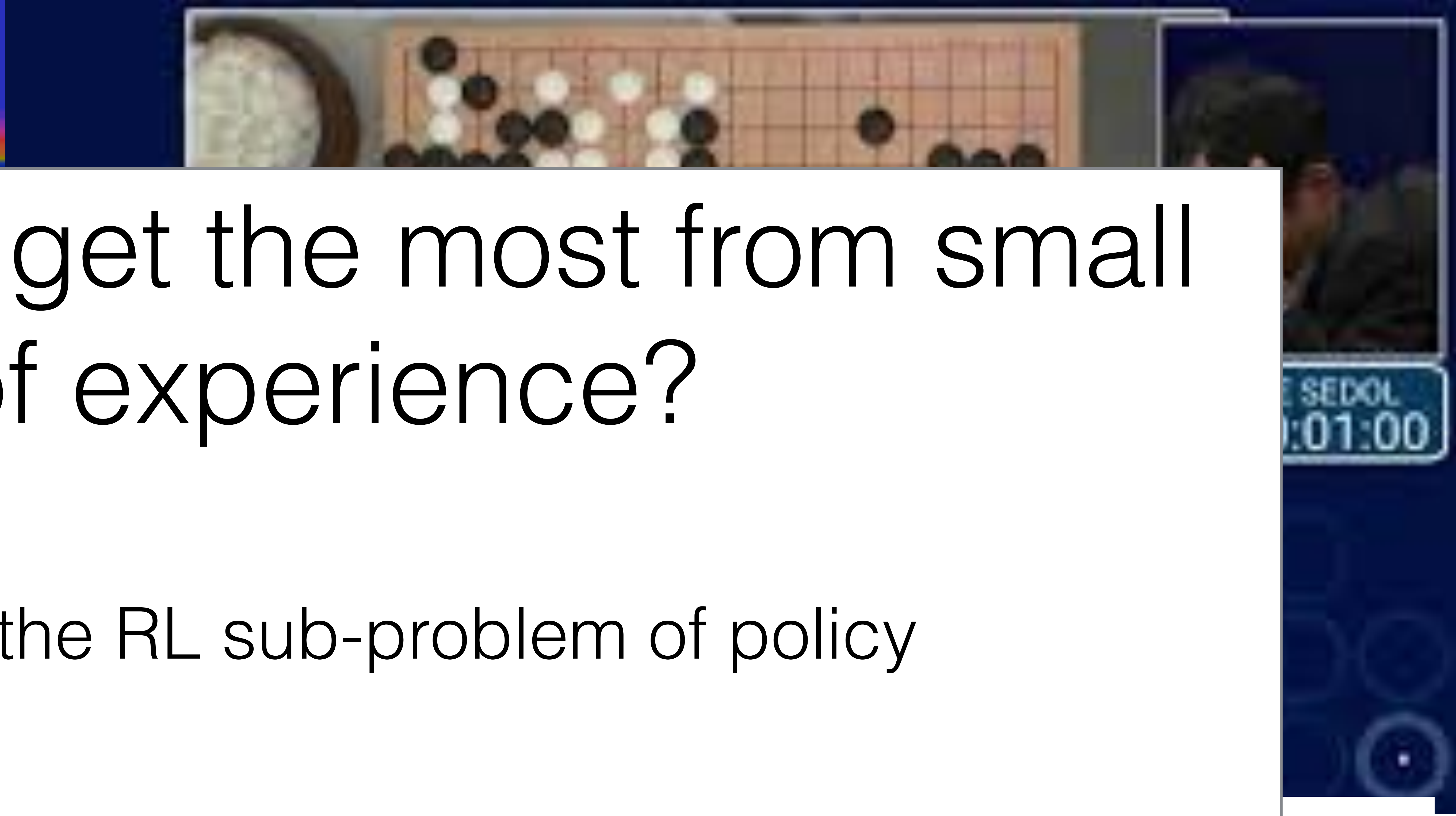




How can RL agents get the most from small amounts of experience?

Study importance sampling for the RL sub-problem of policy evaluation.

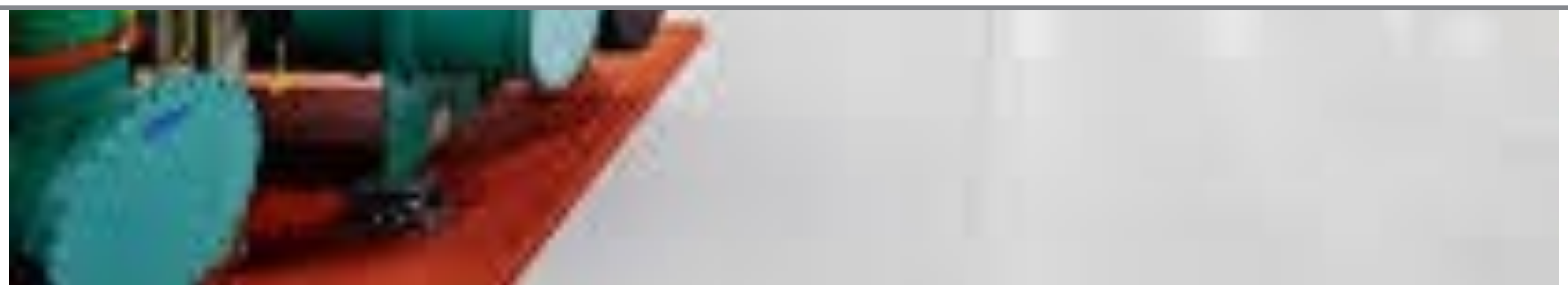


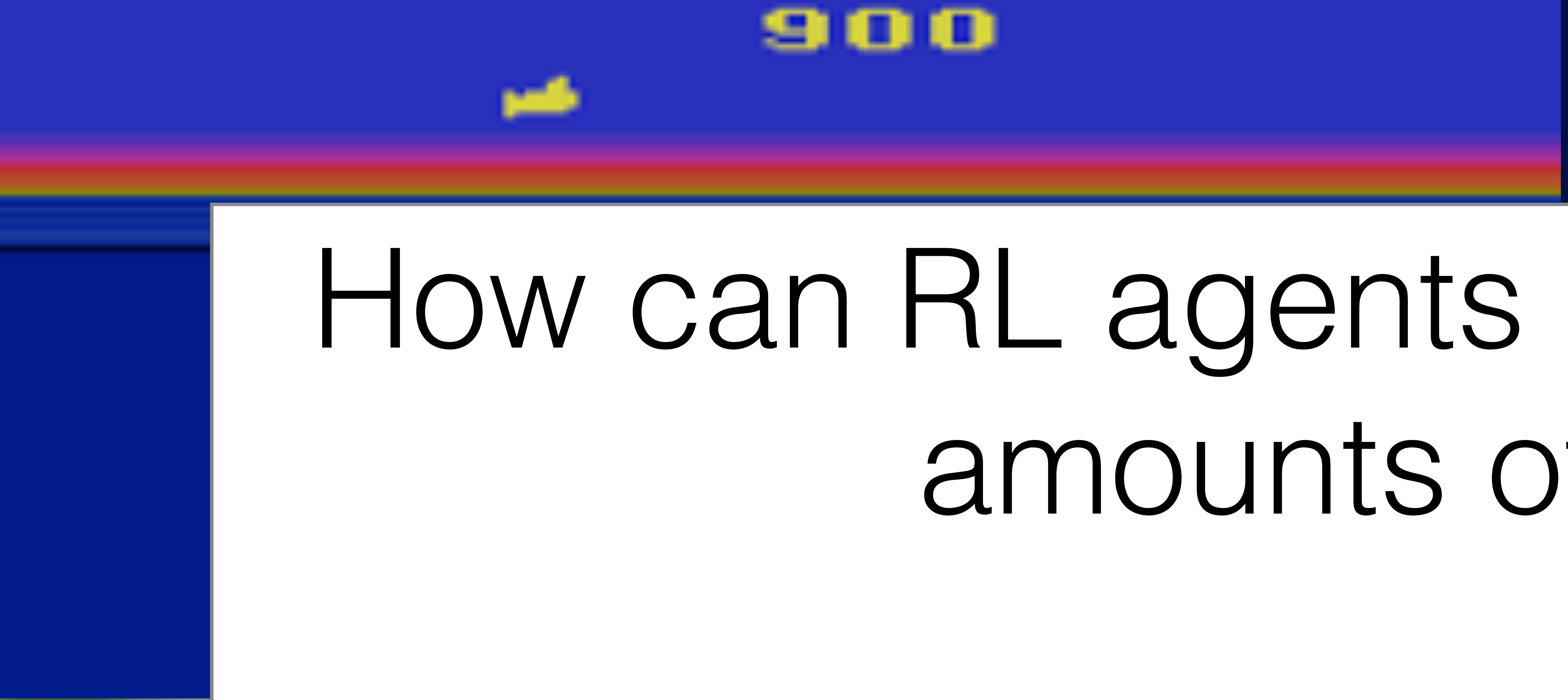


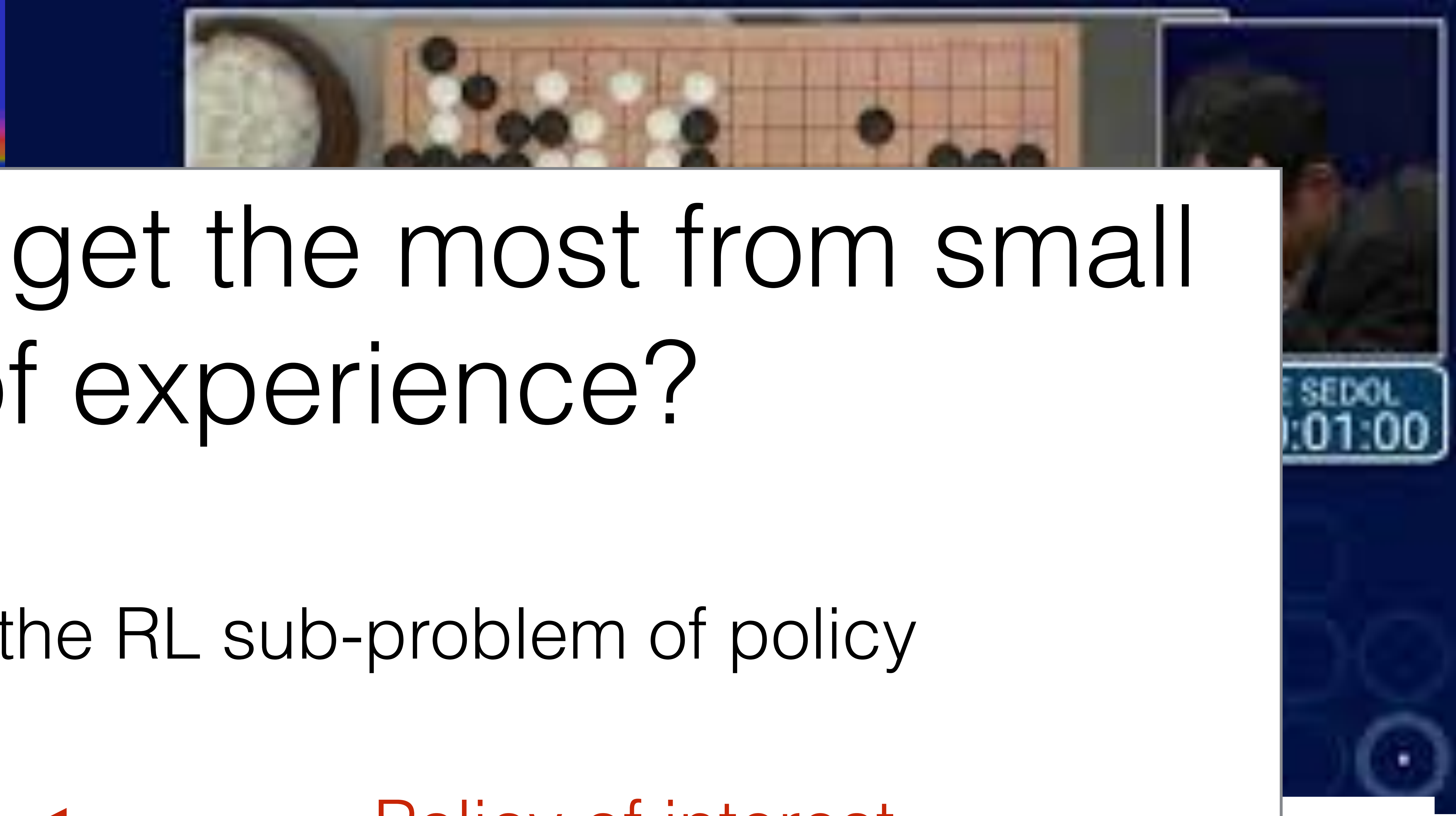
How can RL agents get the most from small amounts of experience?

Study importance sampling for the RL sub-problem of policy evaluation.

$$\frac{\pi(a|s)}{\pi_b(a|s)}$$



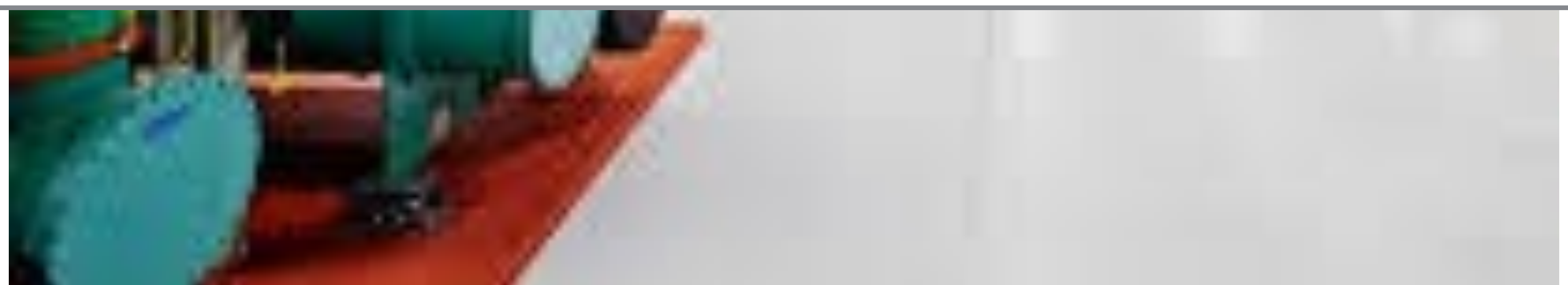


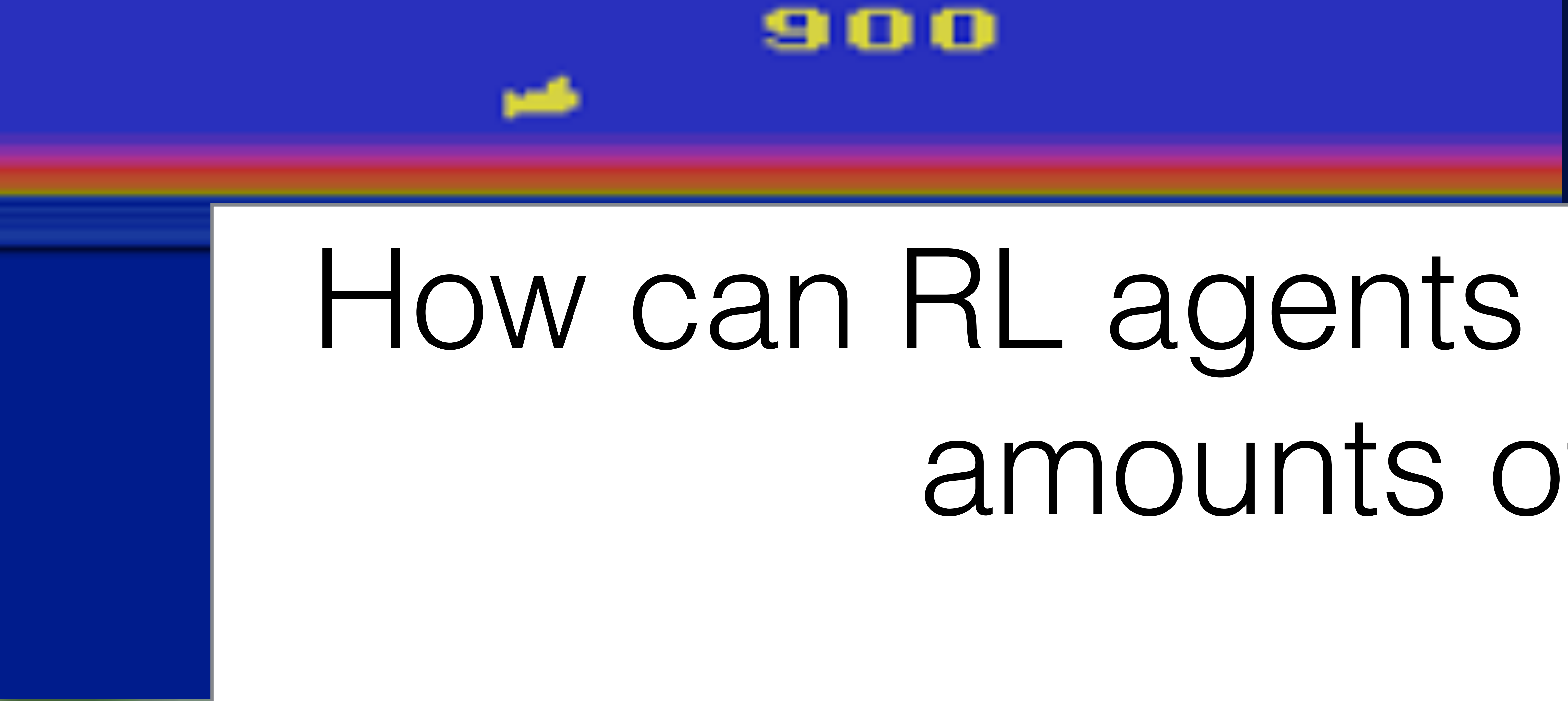
How can RL agents get the most from small amounts of experience?

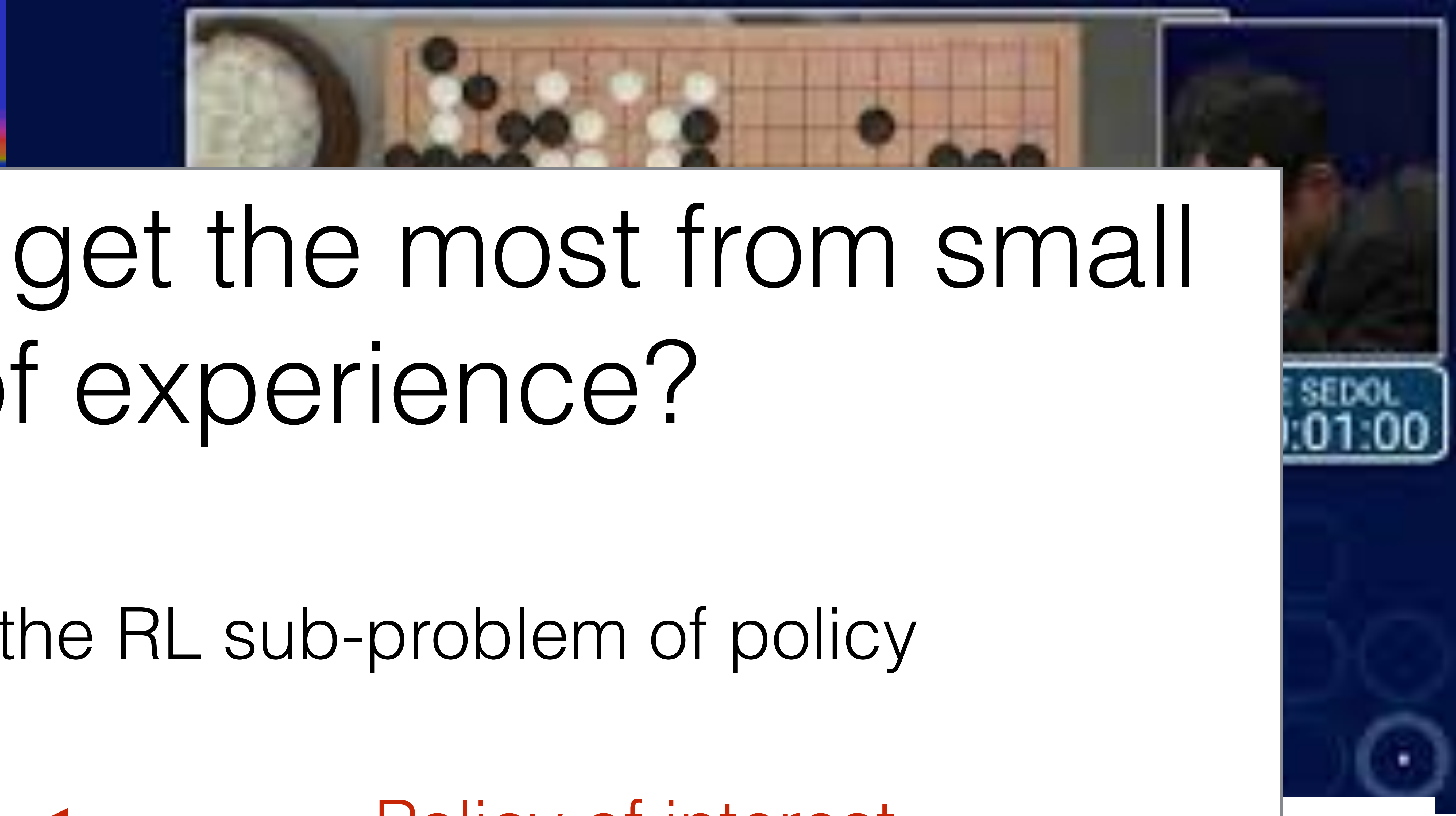
Study importance sampling for the RL sub-problem of policy evaluation.

$$\frac{\pi(a|s)}{\pi_b(a|s)}$$

Policy of interest







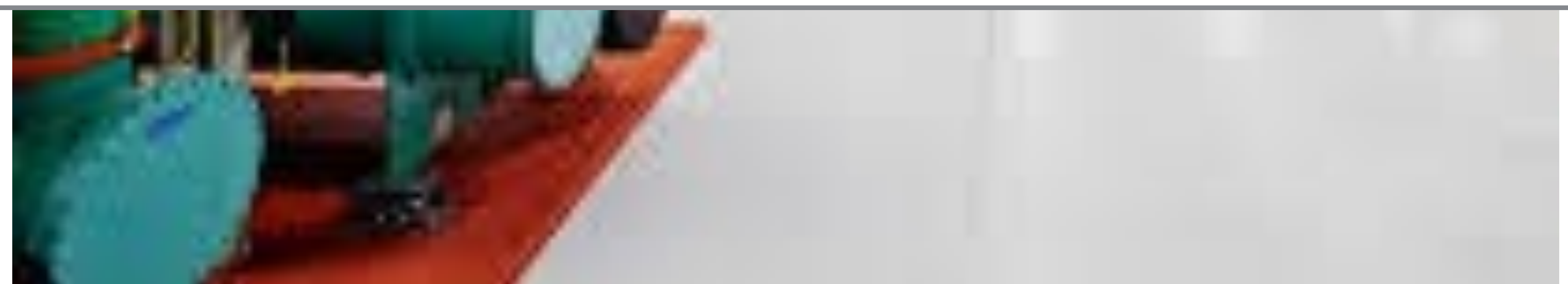
How can RL agents get the most from small amounts of experience?

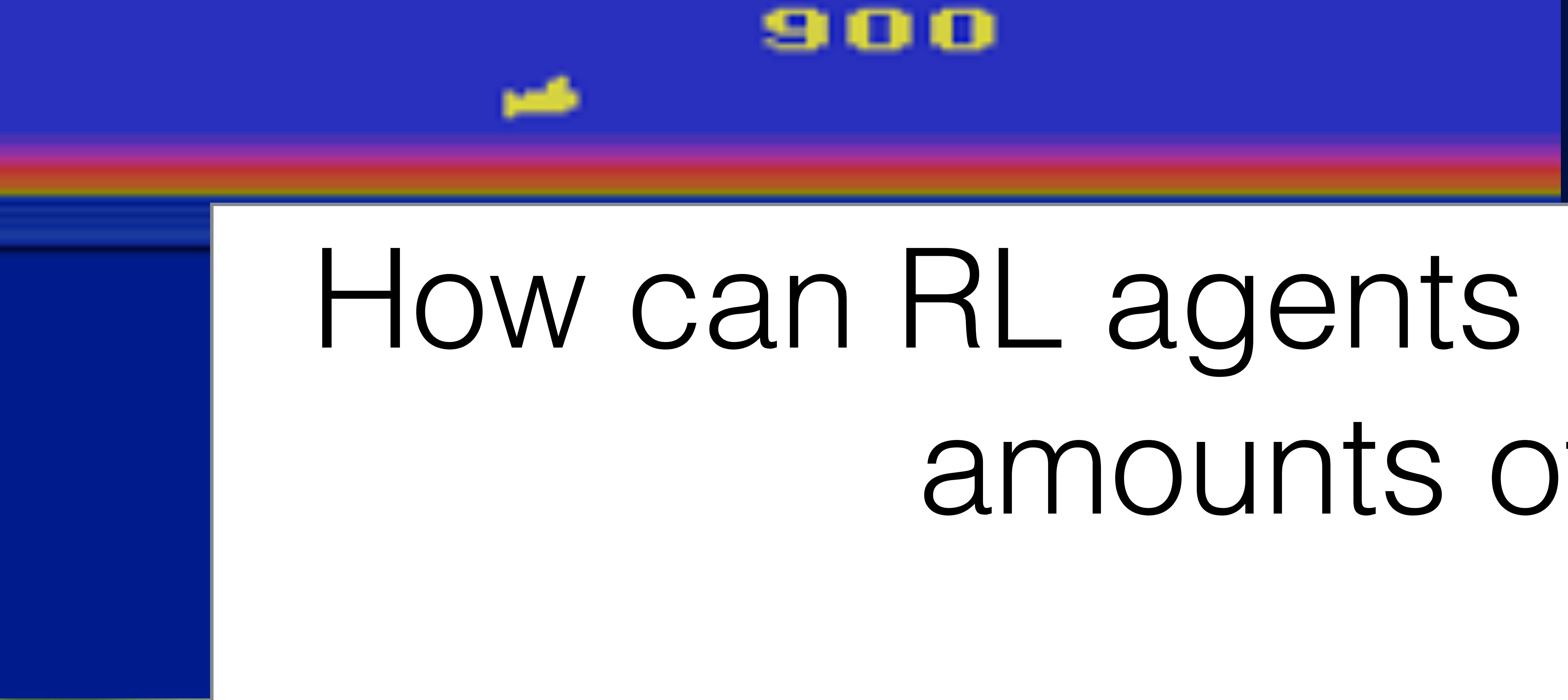
Study importance sampling for the RL sub-problem of policy evaluation.

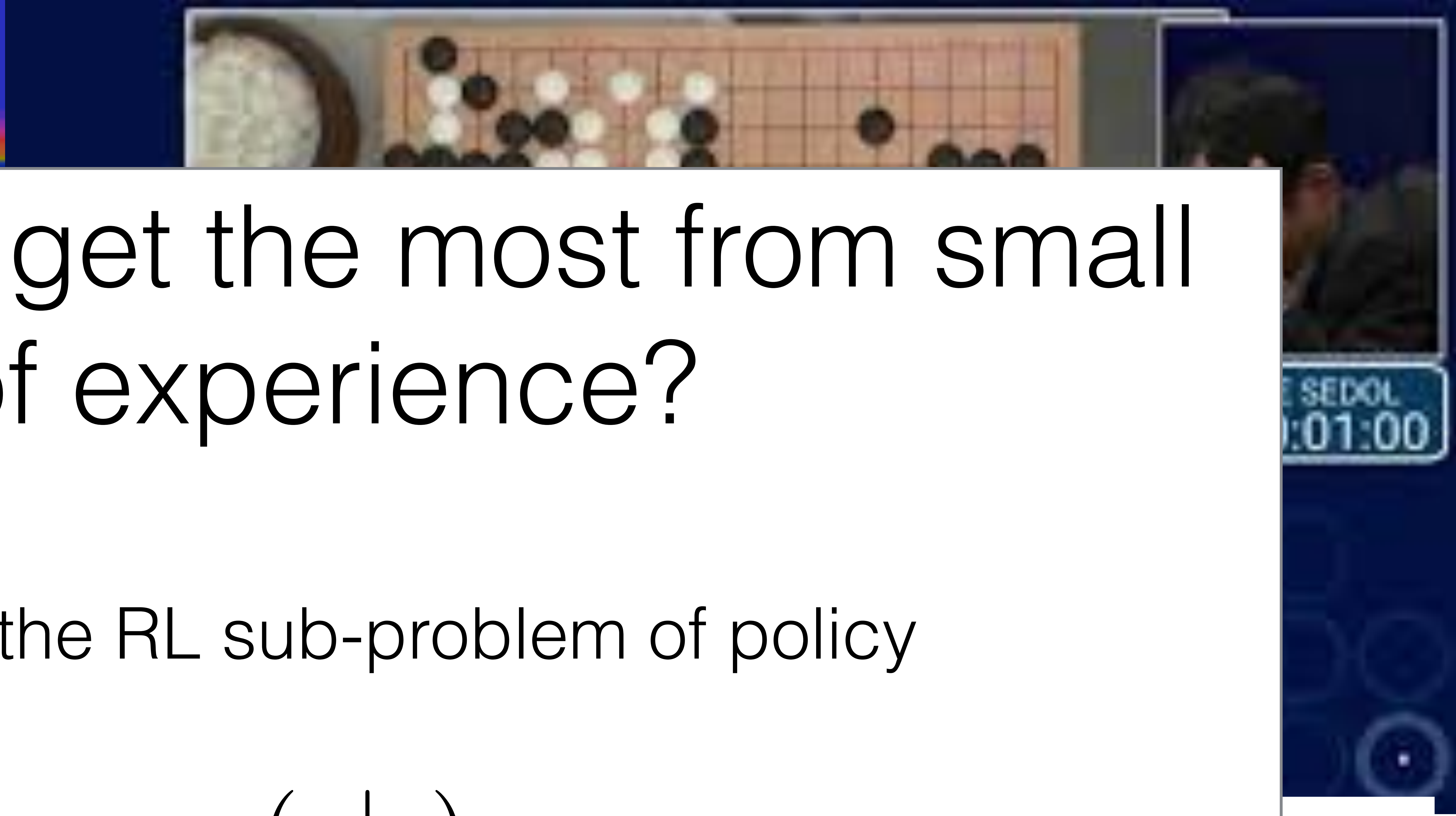
$$\frac{\pi(a|s)}{\pi_b(a|s)}$$

Policy of interest

Data collection policy (behavior policy)



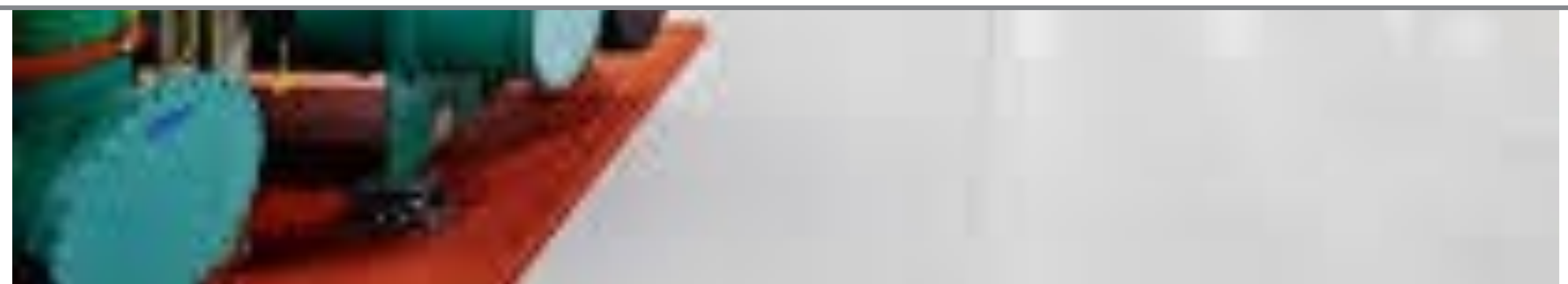


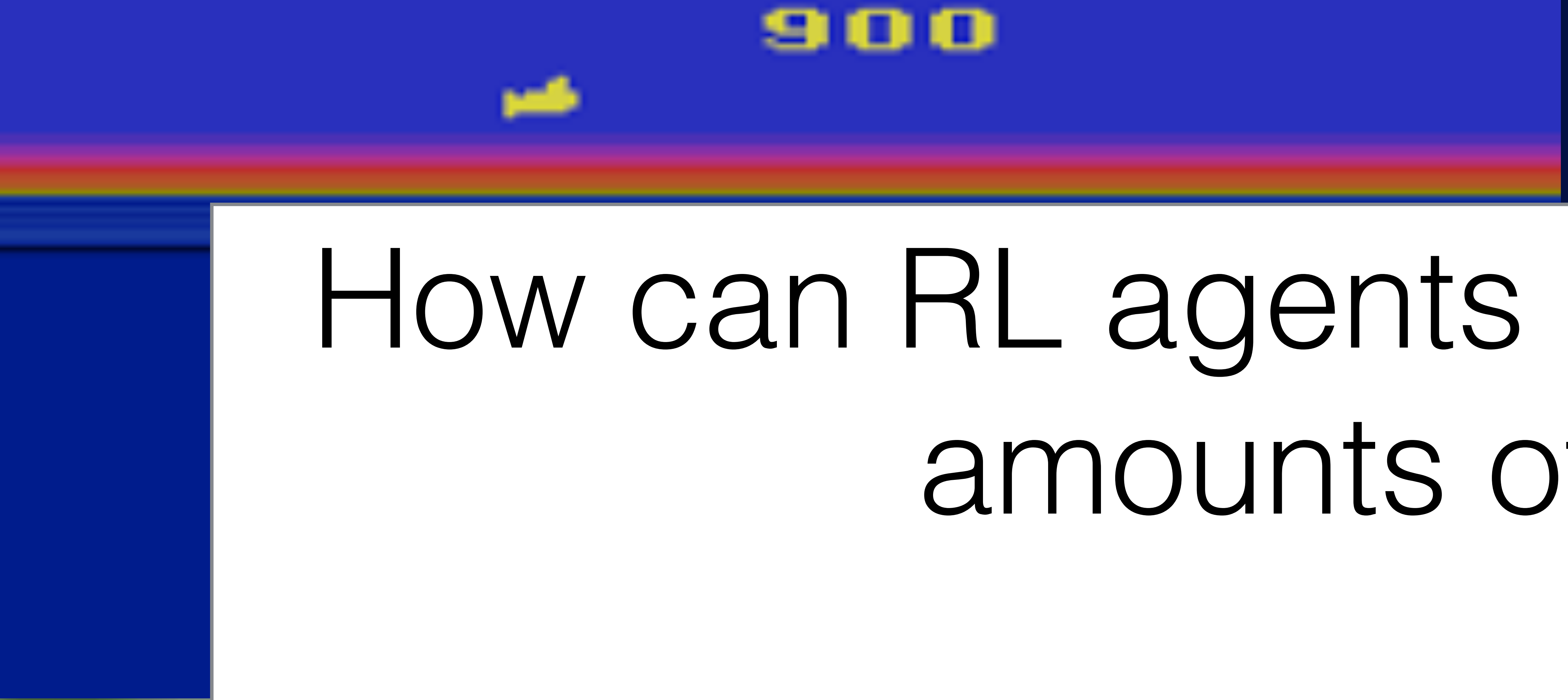


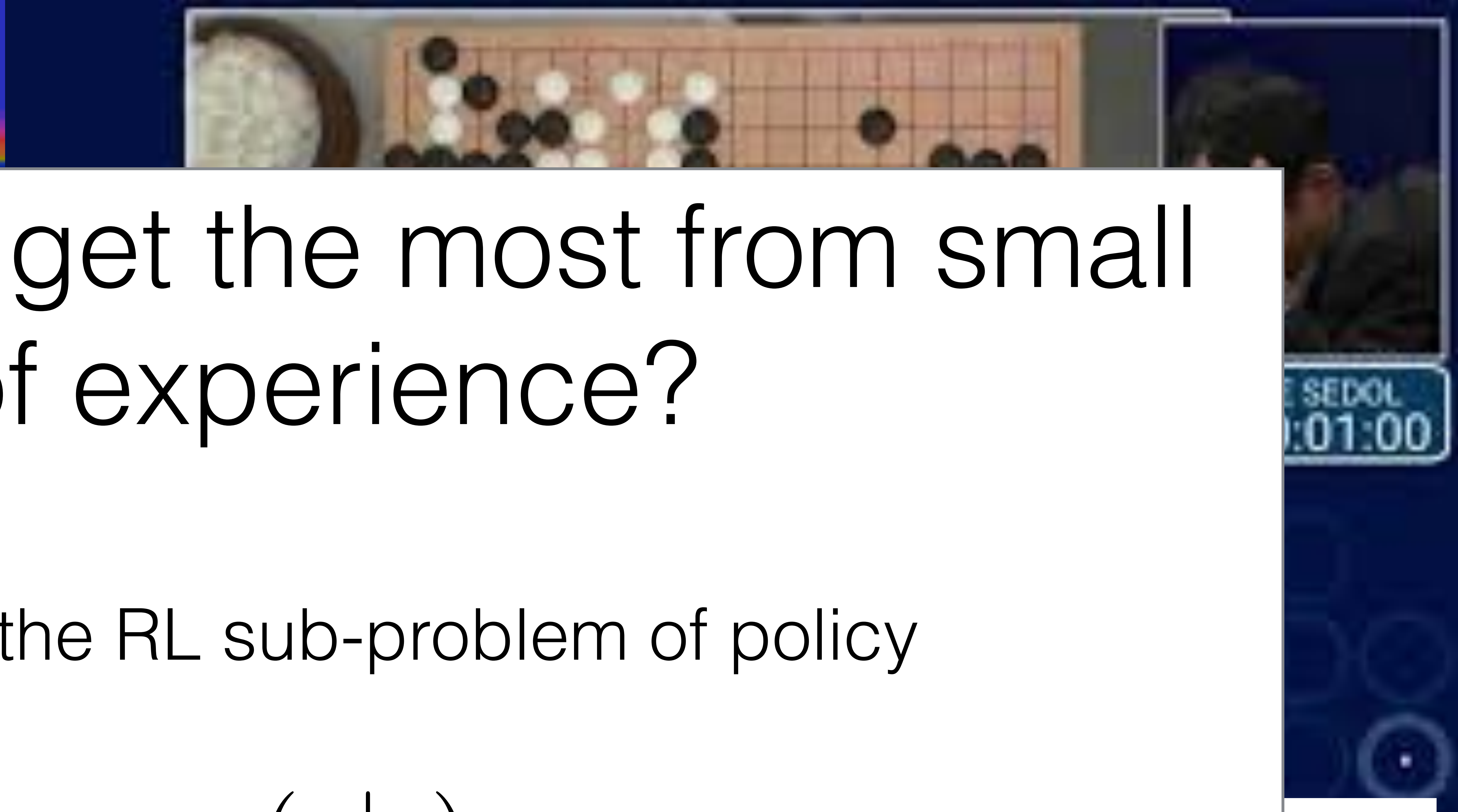

How can RL agents get the most from small amounts of experience?

Study importance sampling for the RL sub-problem of policy evaluation.

$$\frac{\pi(a|s)}{\pi_b(a|s)} \rightarrow \frac{\pi(a|s)}{\hat{\pi}_b(a|s)}$$





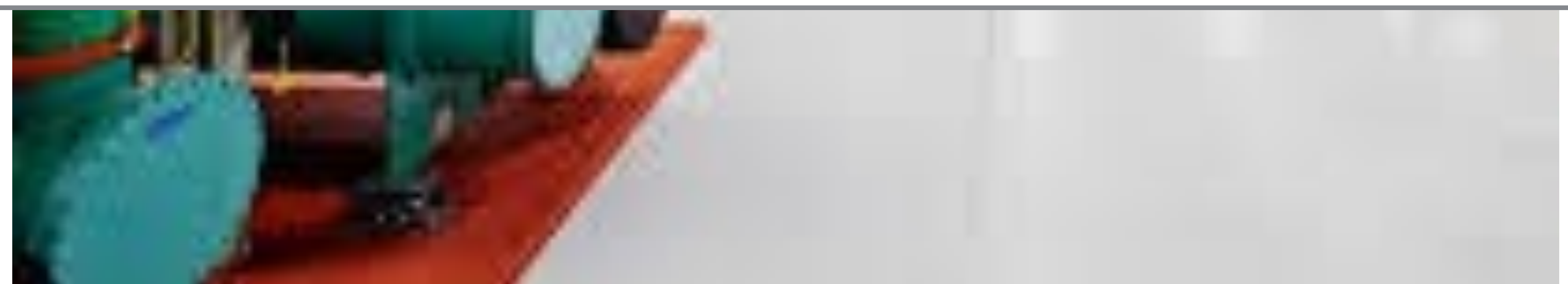


How can RL agents get the most from small amounts of experience?

Study importance sampling for the RL sub-problem of policy evaluation.

$$\frac{\pi(a|s)}{\pi_b(a|s)} \rightarrow \frac{\pi(a|s)}{\hat{\pi}_b(a|s)}$$

Provide empirical and theoretical support that estimating the behavior policy improves importance sampling for policy evaluation.



Batch Policy Evaluation

Batch Policy Evaluation

Given batch of trajectory data:

$$\{(S_0^i, A_0^i, R_0^i, ..., S_L^i, A_L^i, R_L^i)\}_{i=1}^m$$

Batch Policy Evaluation

Given batch of trajectory data:

$$\{(S_0^i, A_0^i, R_0^i, ..., S_L^i, A_L^i, R_L^i)\}_{i=1}^m$$

Given a target policy:

$$\pi : \mathcal{S} \rightarrow \mathbb{P}(\mathcal{A})$$

Batch Policy Evaluation

Given batch of trajectory data:

$$\{(S_0^i, A_0^i, R_0^i, ..., S_L^i, A_L^i, R_L^i)\}_{i=1}^m$$

Given a target policy:

$$\pi : \mathcal{S} \to \mathbb{P}(\mathcal{A})$$

Estimate:

$$v(\pi) := \mathbf{E}\left[\sum_{t=0}^{L} \gamma^t R_t\right]$$

Ordinary Importance Sampling in RL

Ordinary Importance Sampling in RL

$$\texttt{OIS}(\pi, \mathcal{D}) = \frac{1}{m} \sum_{i=1}^{m} \prod_{t=0}^{L} \frac{\pi(a_t|s_t)}{\pi_b(a_t|s_t)} \sum_{t=0}^{L} \gamma^t R_t$$

Ordinary Importance Sampling in RL

$$\texttt{OIS}(\pi, \mathcal{D}) = \frac{1}{m} \sum_{i=1}^{m} \prod_{t=0}^{L} \frac{\pi(a_t|s_t)}{\pi_b(a_t|s_t)} \boxed{\sum_{t=0}^{L} \gamma^t R_t}$$

Discounted sum of rewards

Ordinary Importance Sampling in RL

$$\texttt{OIS}(\pi, \mathcal{D}) = \frac{1}{m}\sum_{i=1}^{m} \boxed{\prod_{t=0}^{L} \frac{\pi(a_t|s_t)}{\pi_b(a_t|s_t)}} \boxed{\sum_{t=0}^{L} \gamma^t R_t}$$

Correction from behavior policy to target policy

Discounted sum of rewards

Regression Importance Sampling

$$\texttt{RIS}(n)(\pi, \mathcal{D}) = \frac{1}{m}\sum_{i=1}^{m}\prod_{t=0}^{L}\frac{\pi(a_t|s_t)}{\pi_{\mathcal{D}}(a_t|s_{t-n}, a_{t-n}, ..., s_t)}\sum_{t=0}^{L}\gamma^t R_t$$

Regression Importance Sampling

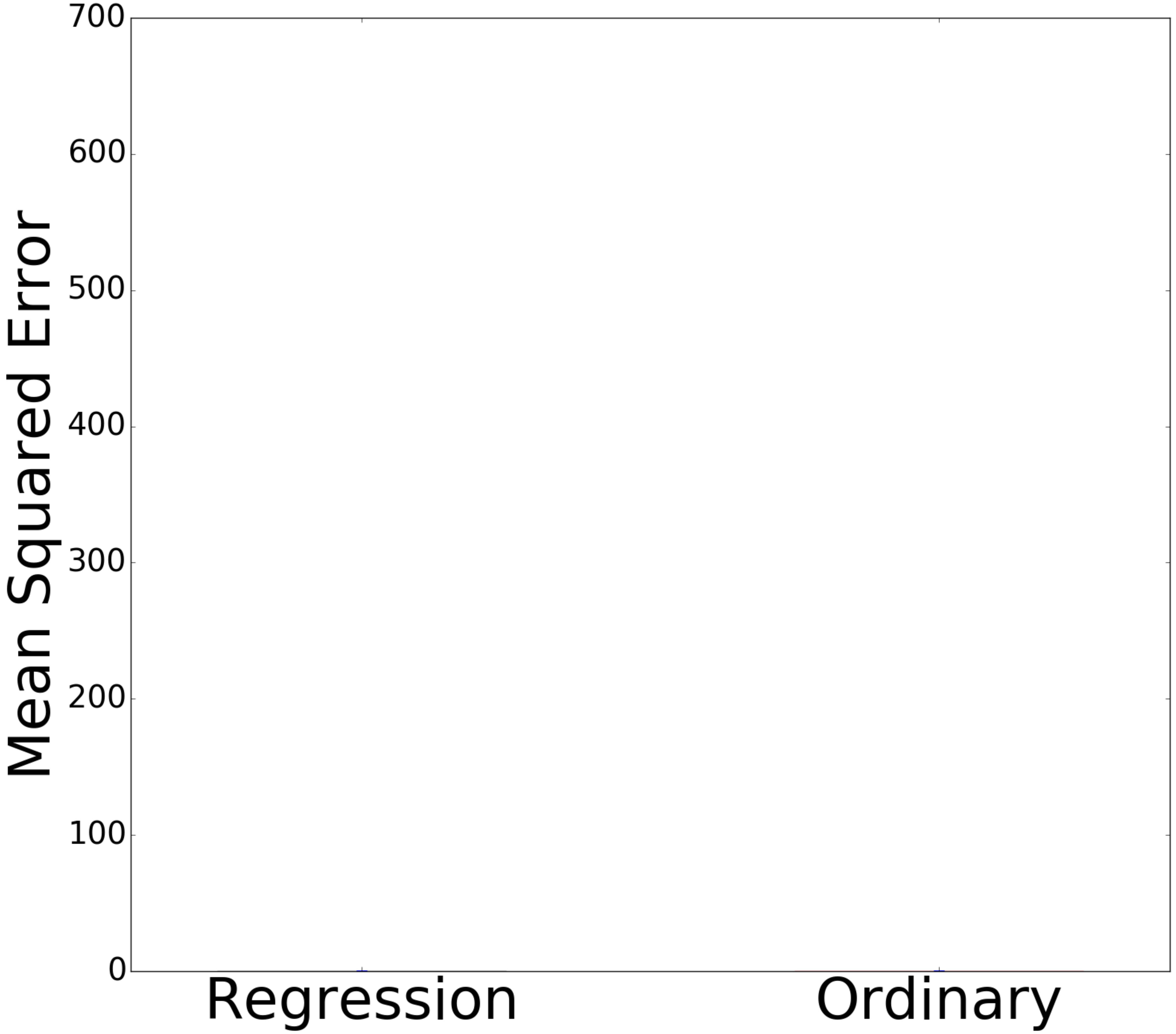
$$\texttt{RIS}(n)(\pi, \mathcal{D}) = \frac{1}{m} \sum_{i=1}^{m} \prod_{t=0}^{L} \boxed{\frac{\pi(a_t|s_t)}{\pi_{\mathcal{D}}(a_t|s_{t-n}, a_{t-n}, ..., s_t)}} \sum_{t=0}^{L} \gamma^t R_t$$

Maximum likelihood
behavior policy estimate.

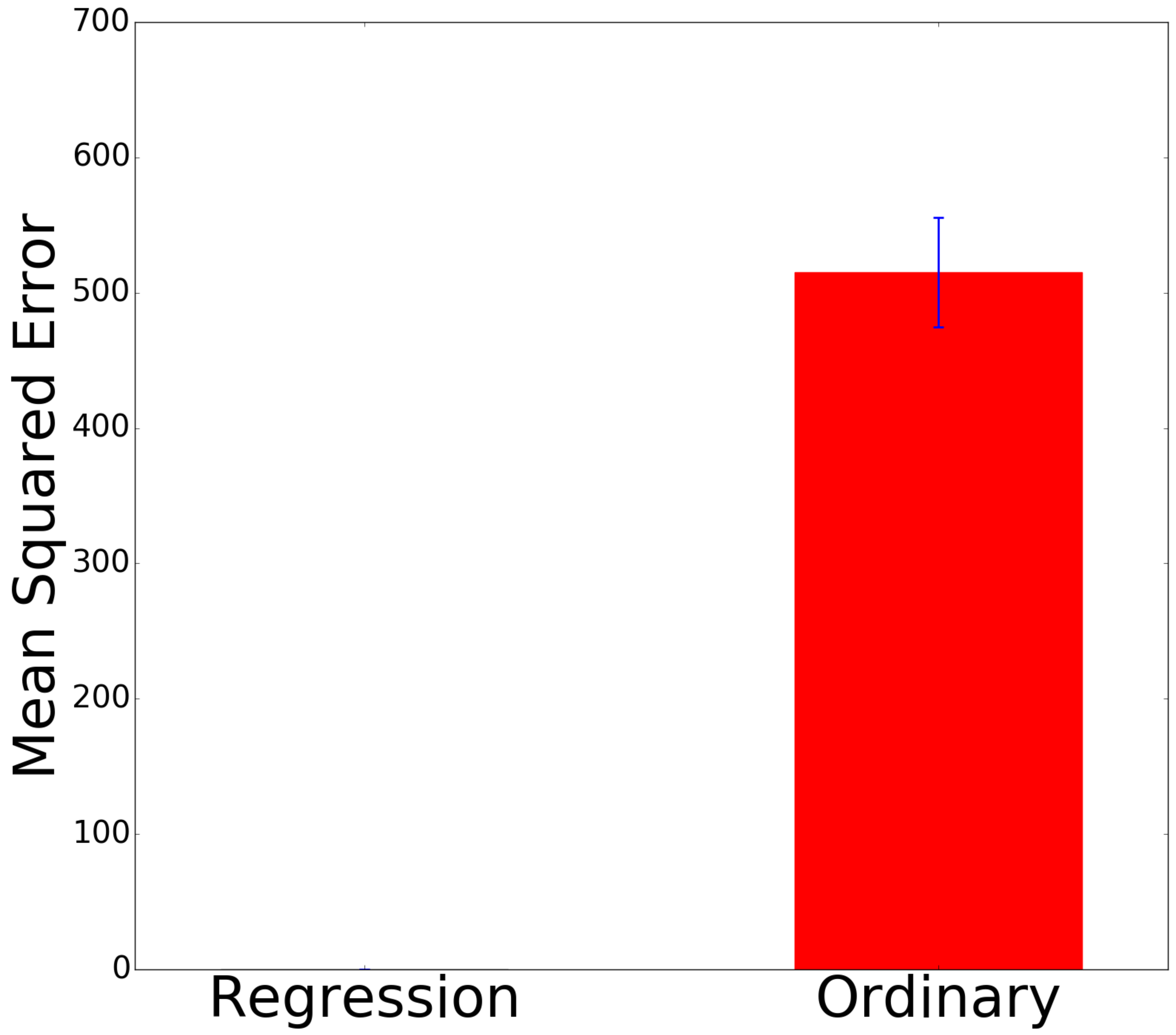
Regression Importance Sampling

$$\texttt{RIS}(n)(\pi, \mathcal{D}) = \frac{1}{m} \sum_{i=1}^{m} \boxed{\prod_{t=0}^{L} \frac{\pi(a_t|s_t)}{\pi_{\mathcal{D}}(a_t|s_{t-n}, a_{t-n}, \ldots, s_t)}} \sum_{t=0}^{L} \gamma^t R_t$$

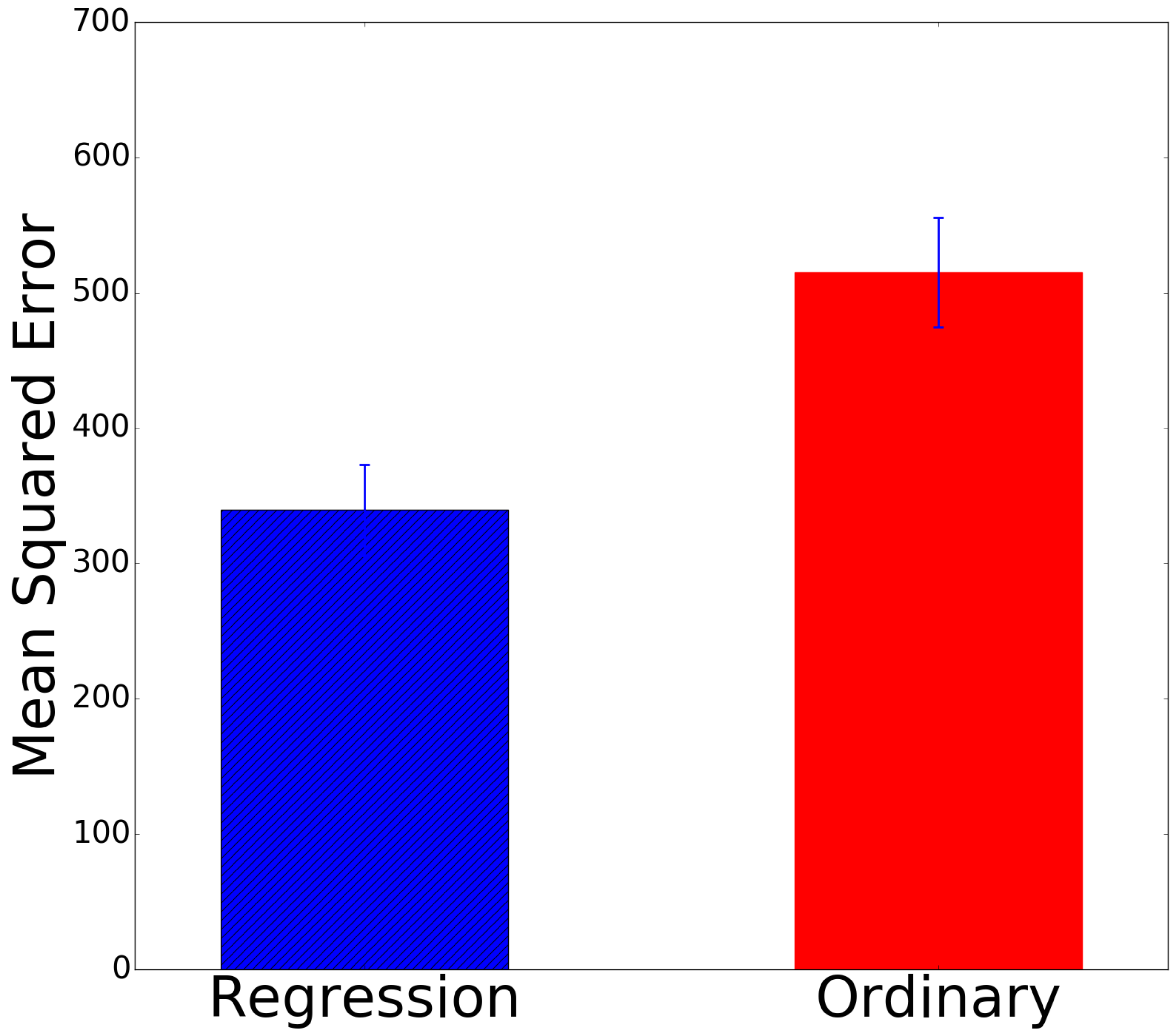
Correction from empirical distribution to target policy.



OpenAI's RoboschoolHopper-v1

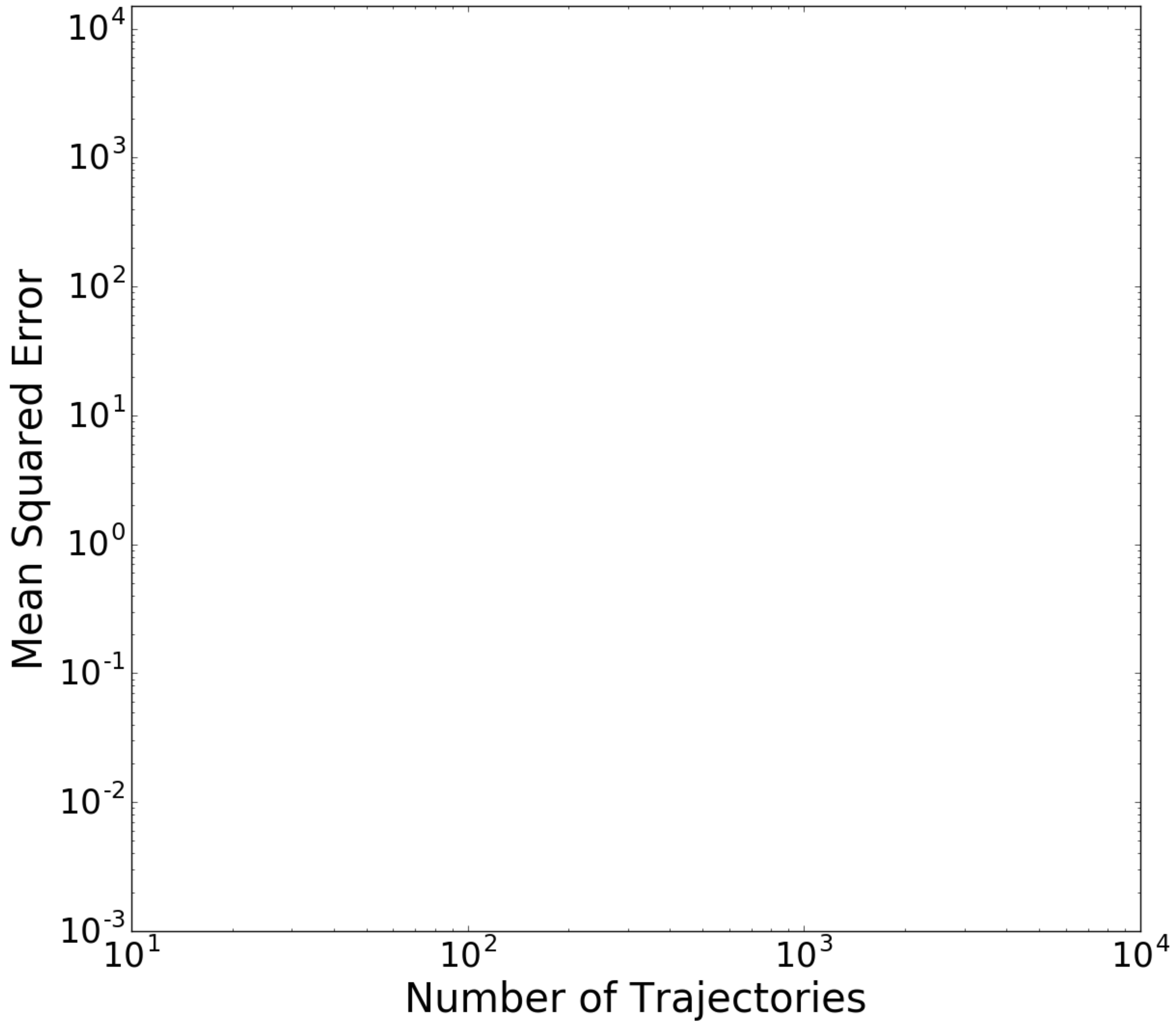


OpenAI's RoboschoolHopper-v1



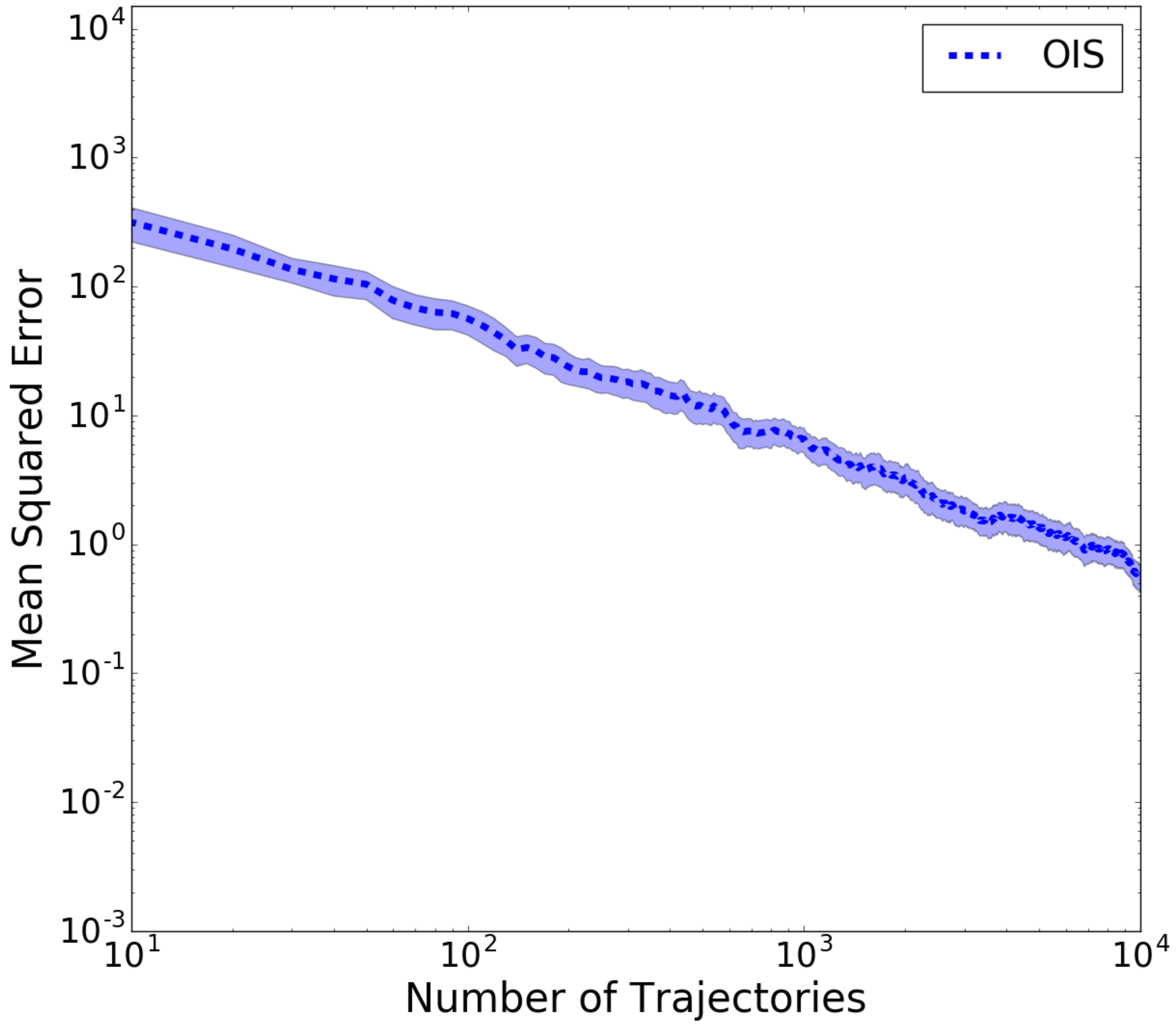
OpenAI's RoboschoolHopper-v1

Empirical Results



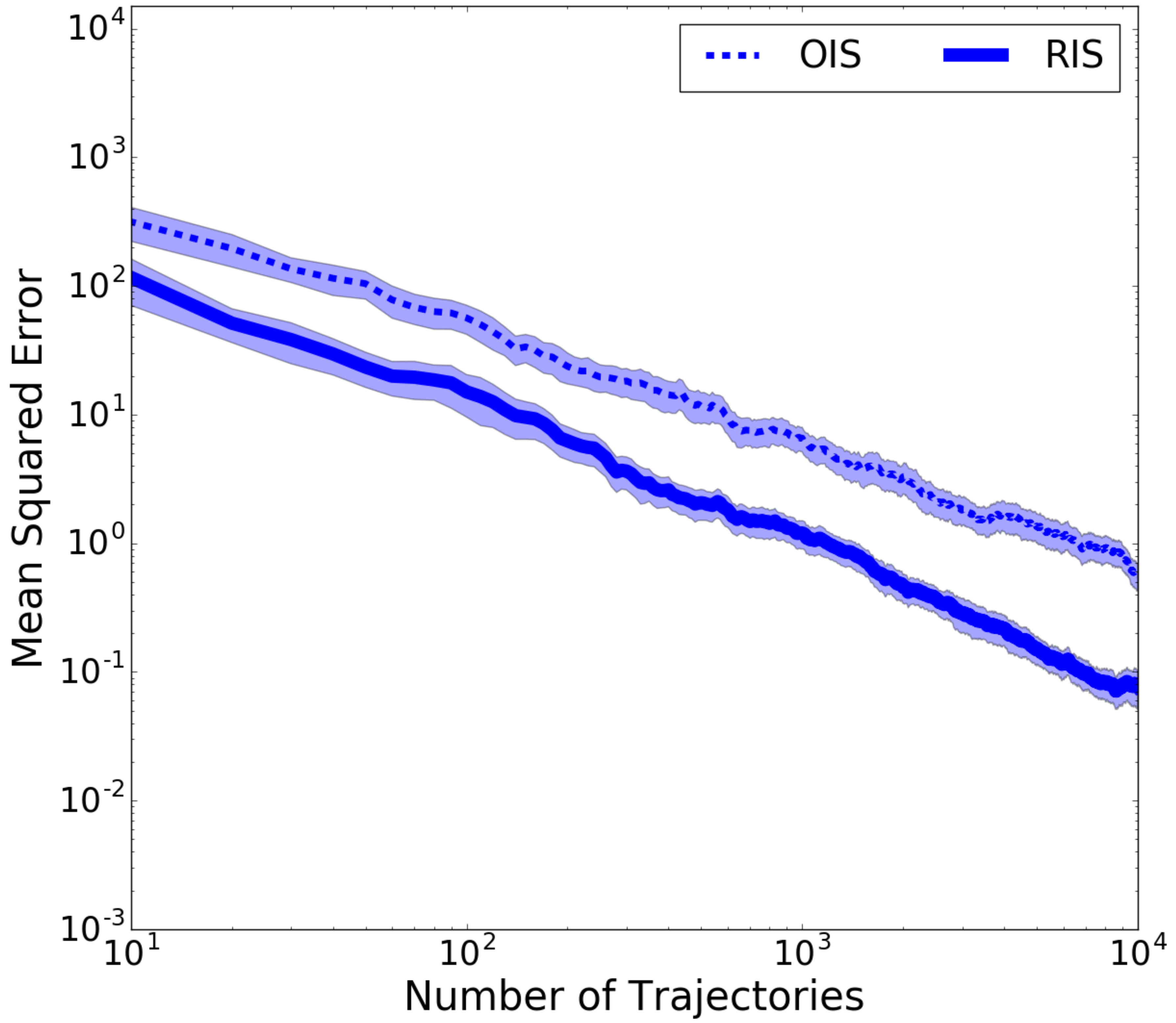
Gridworld

Empirical Results



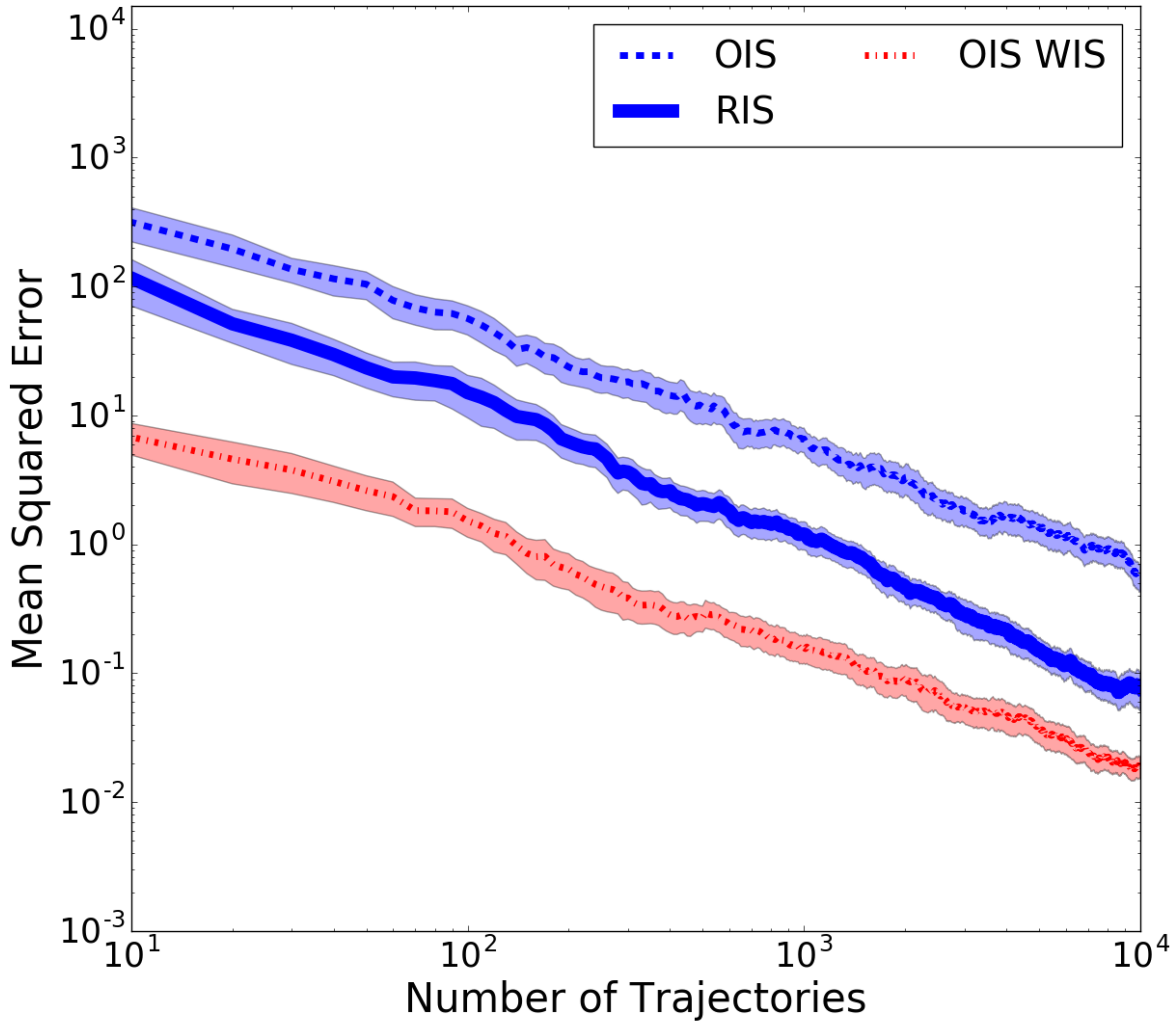
Gridworld

Empirical Results



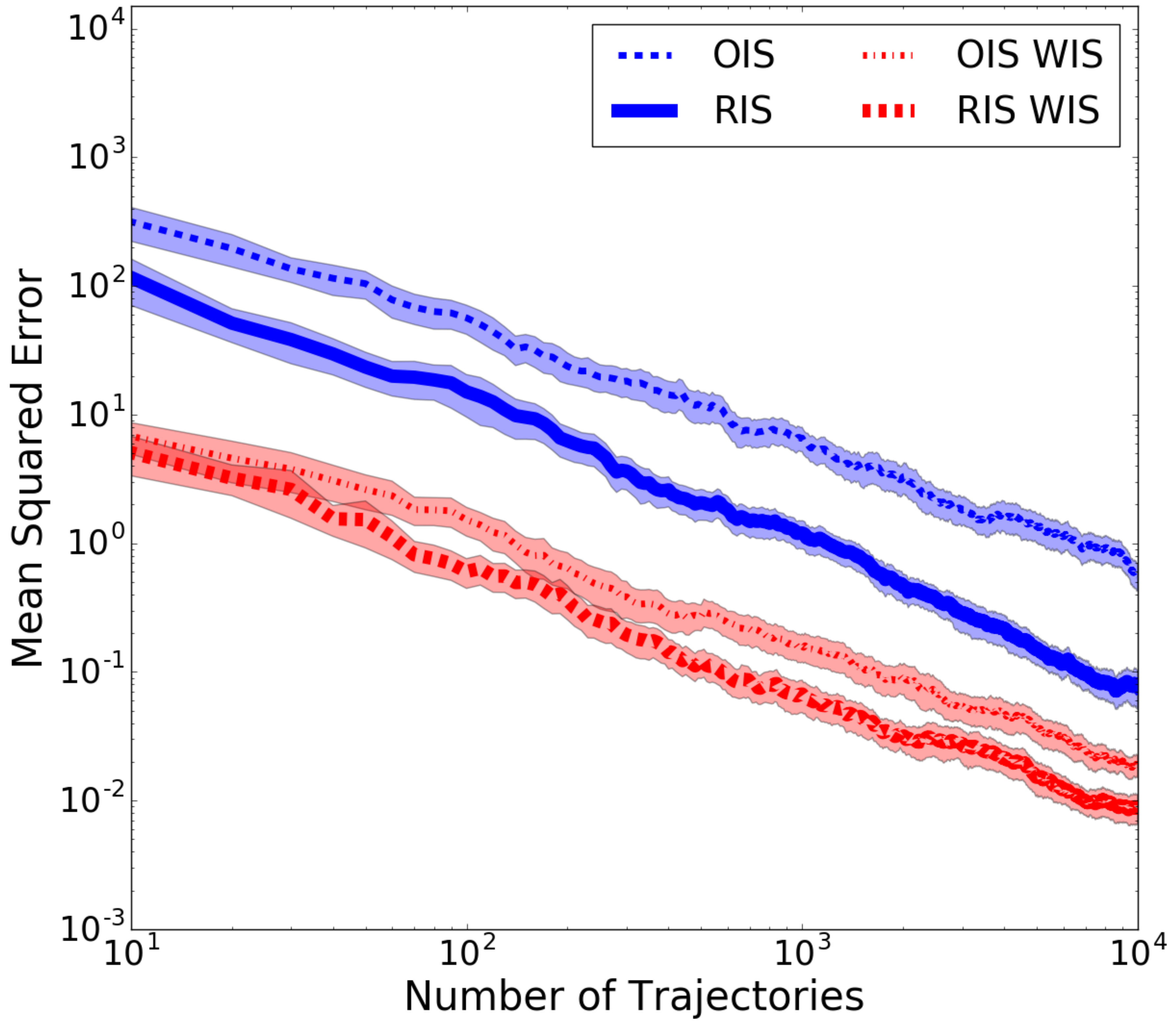
Gridworld

Empirical Results



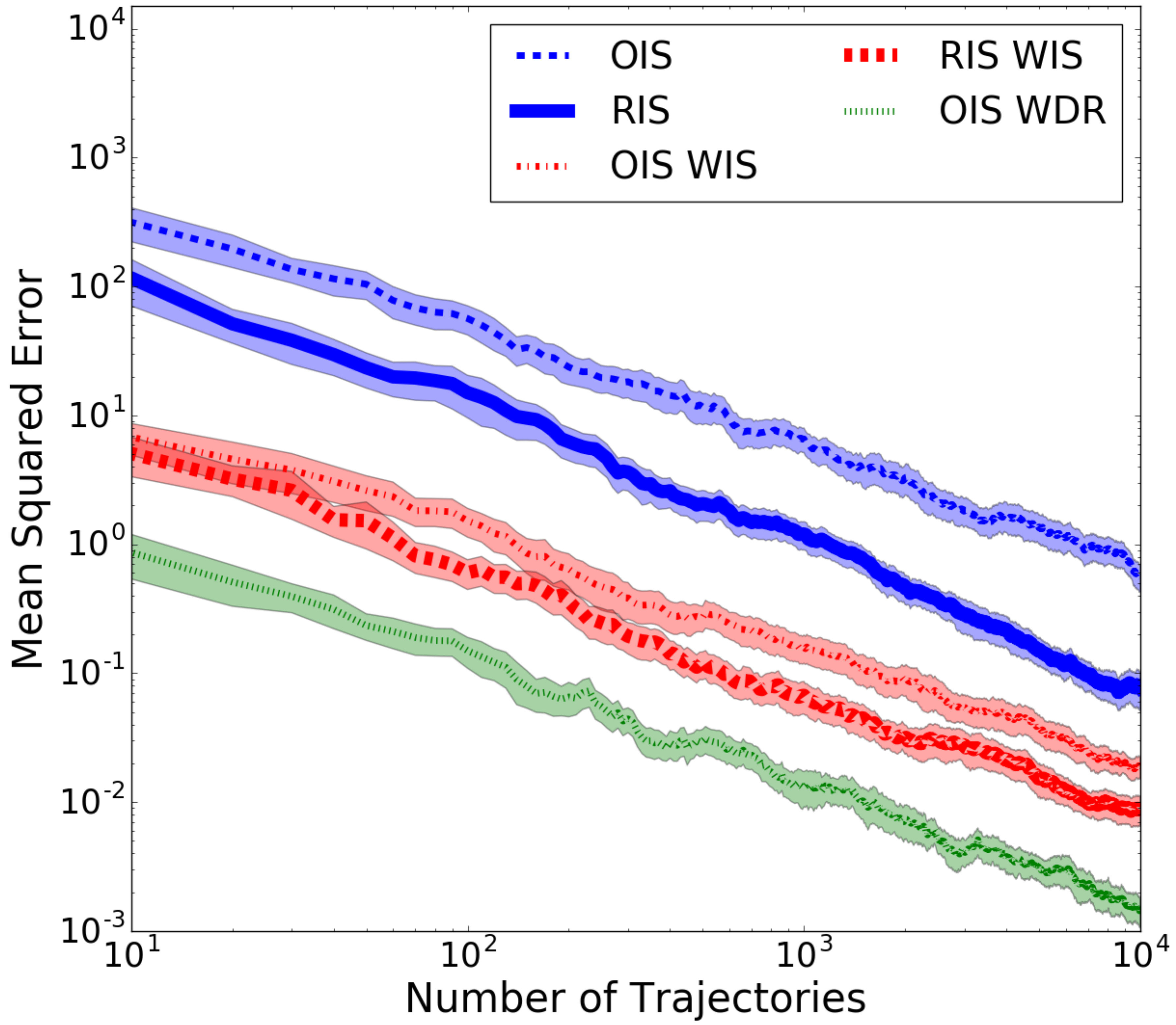
Gridworld

Empirical Results



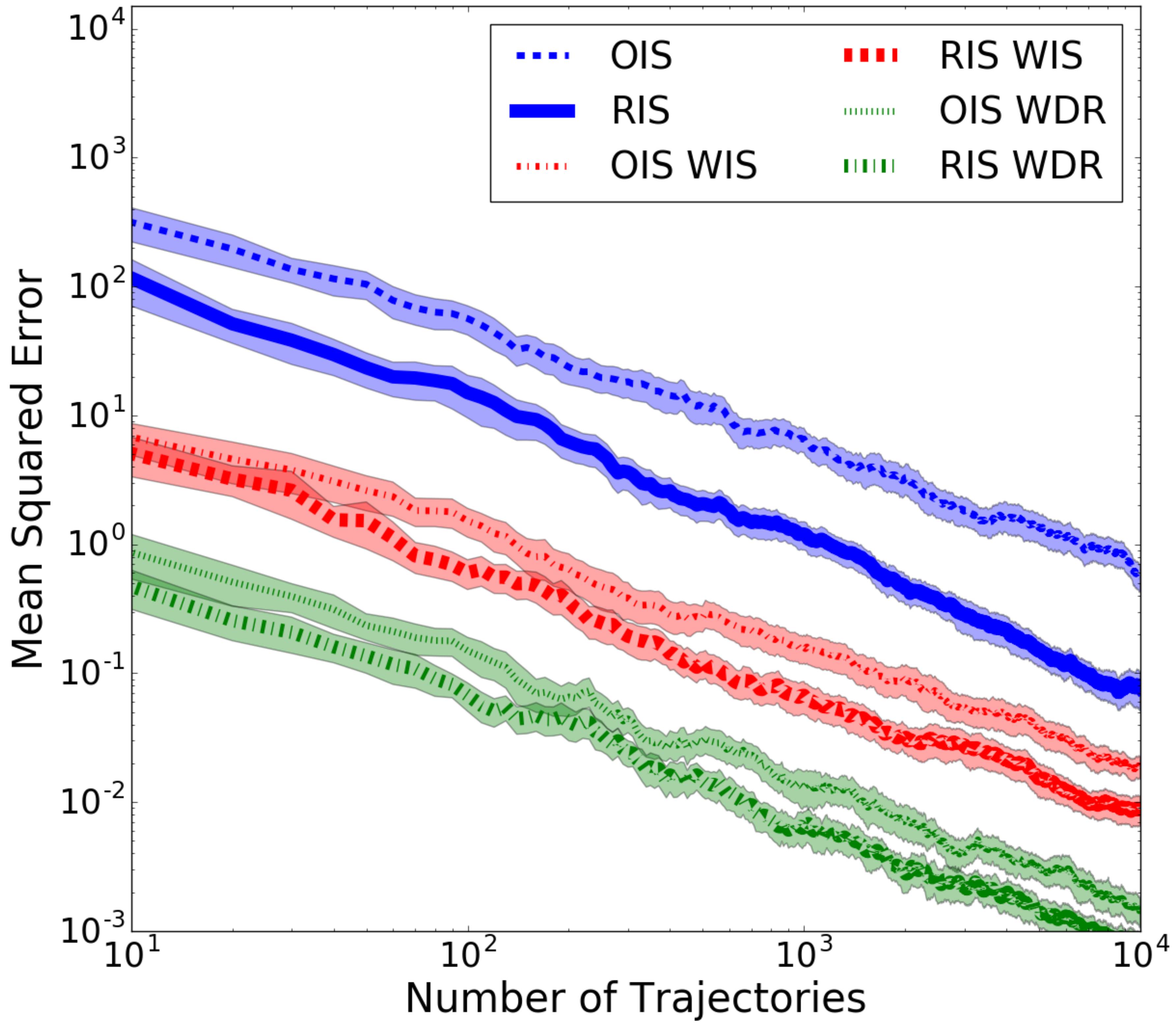
Gridworld

Empirical Results



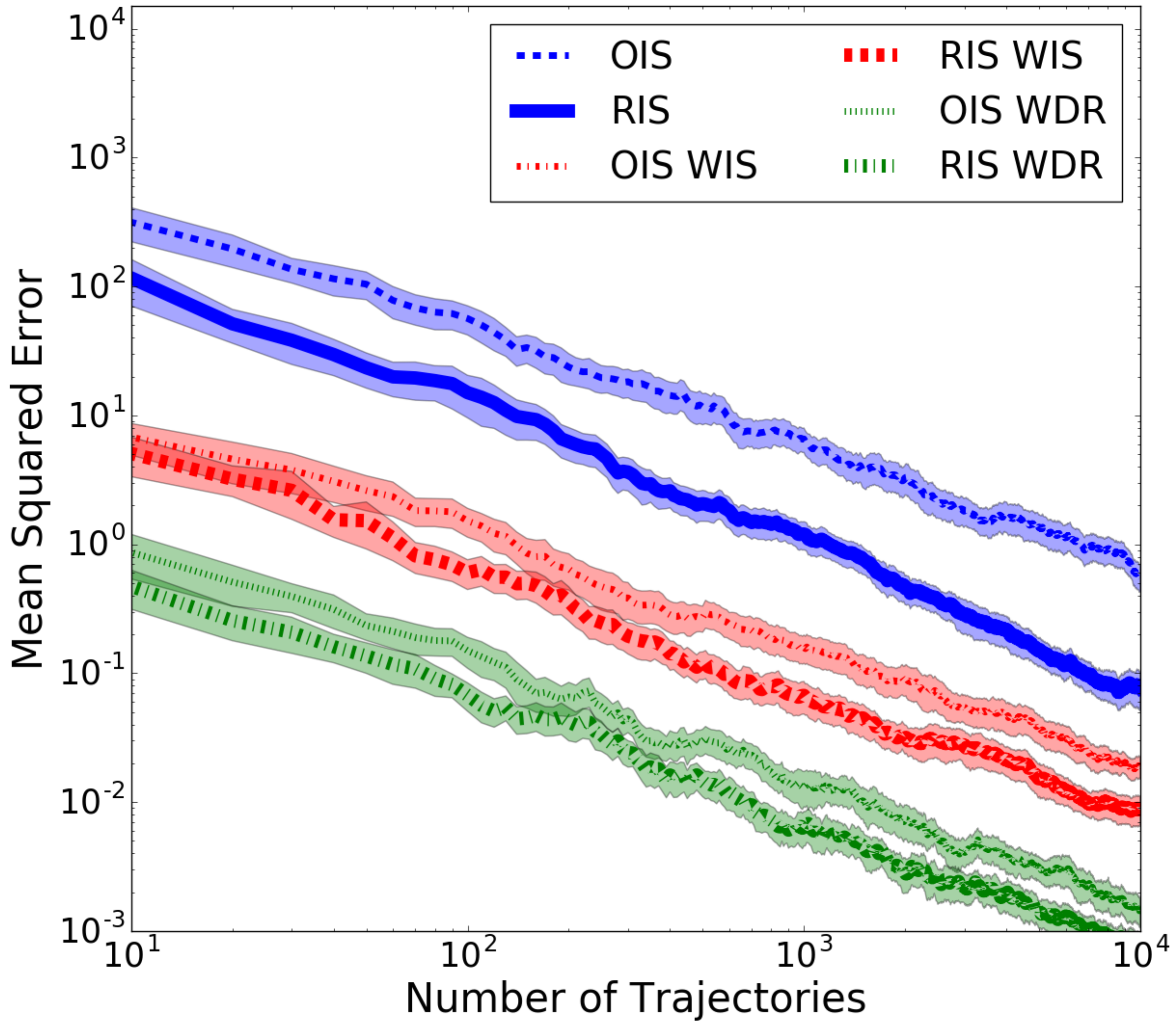
Gridworld

Empirical Results

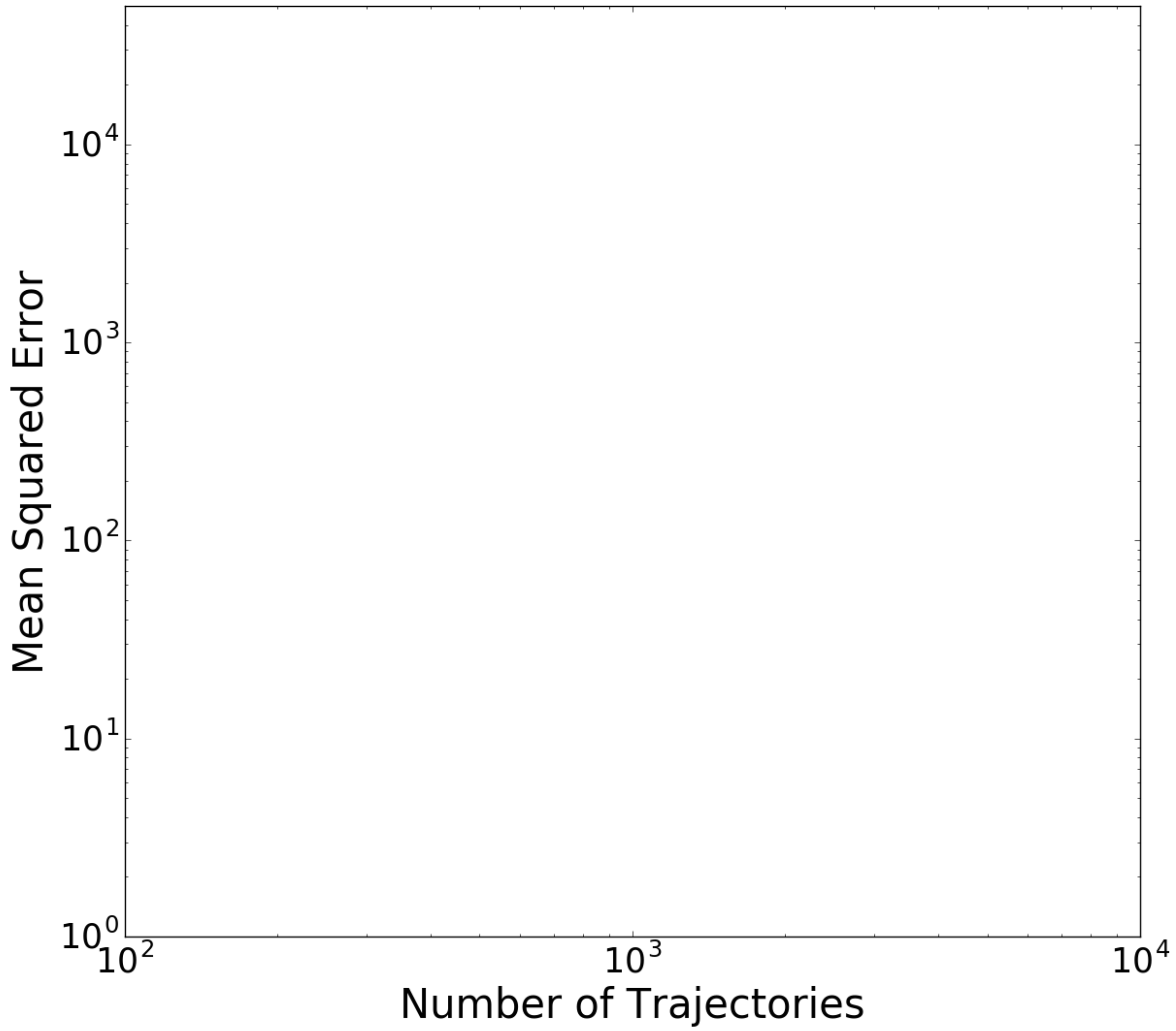


Gridworld

Empirical Results

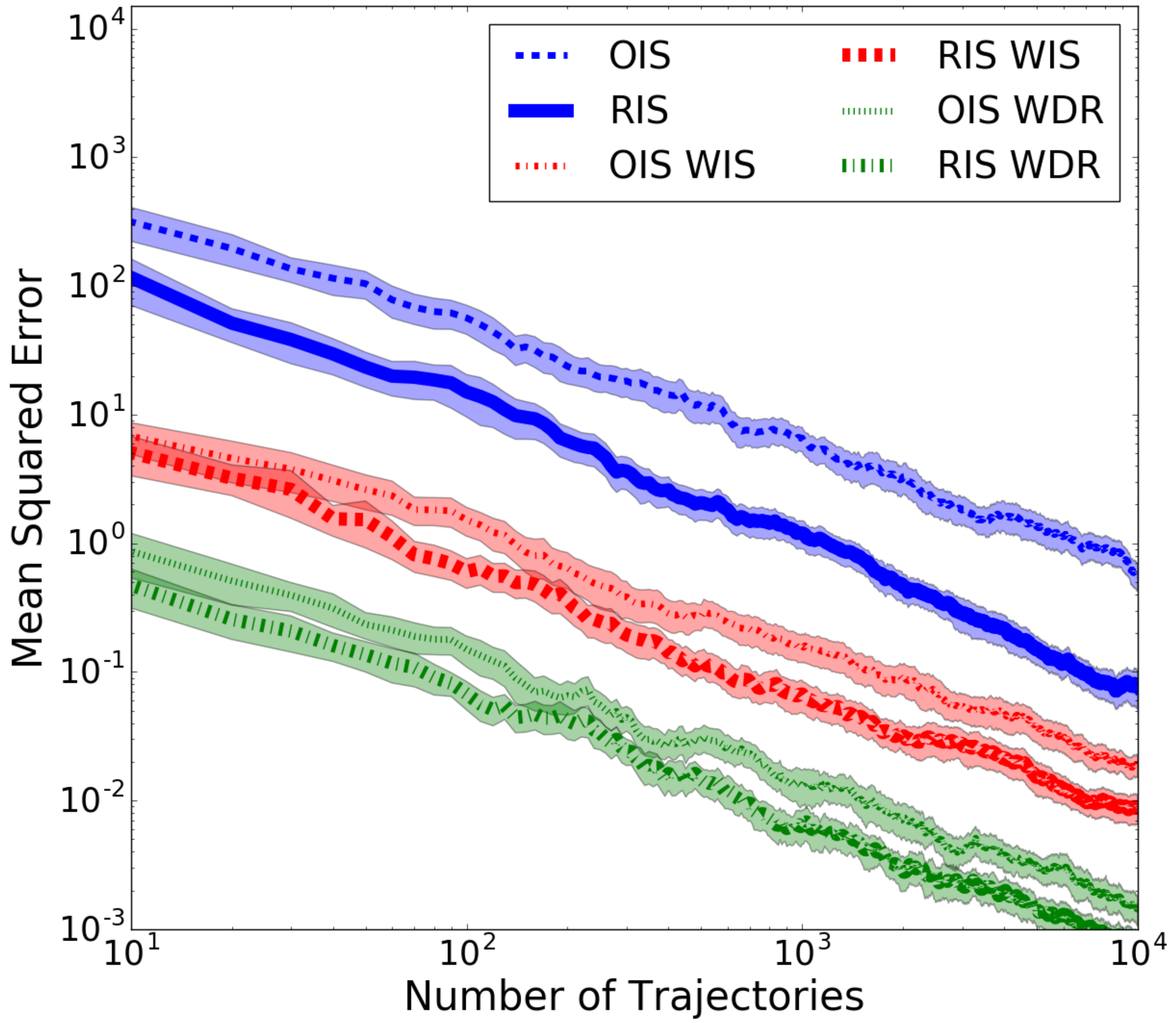


Gridworld

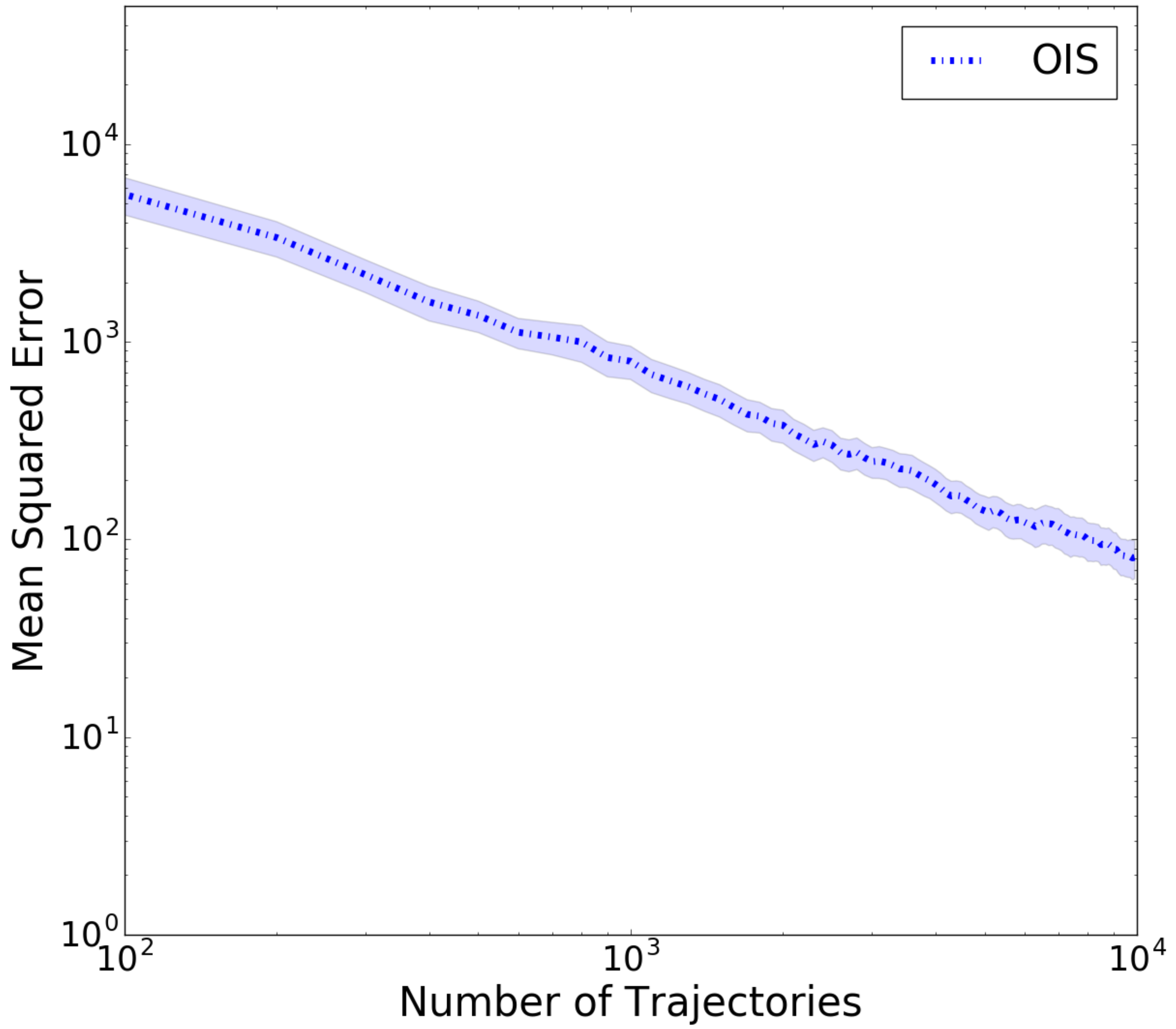


Linear Dynamical System

Empirical Results

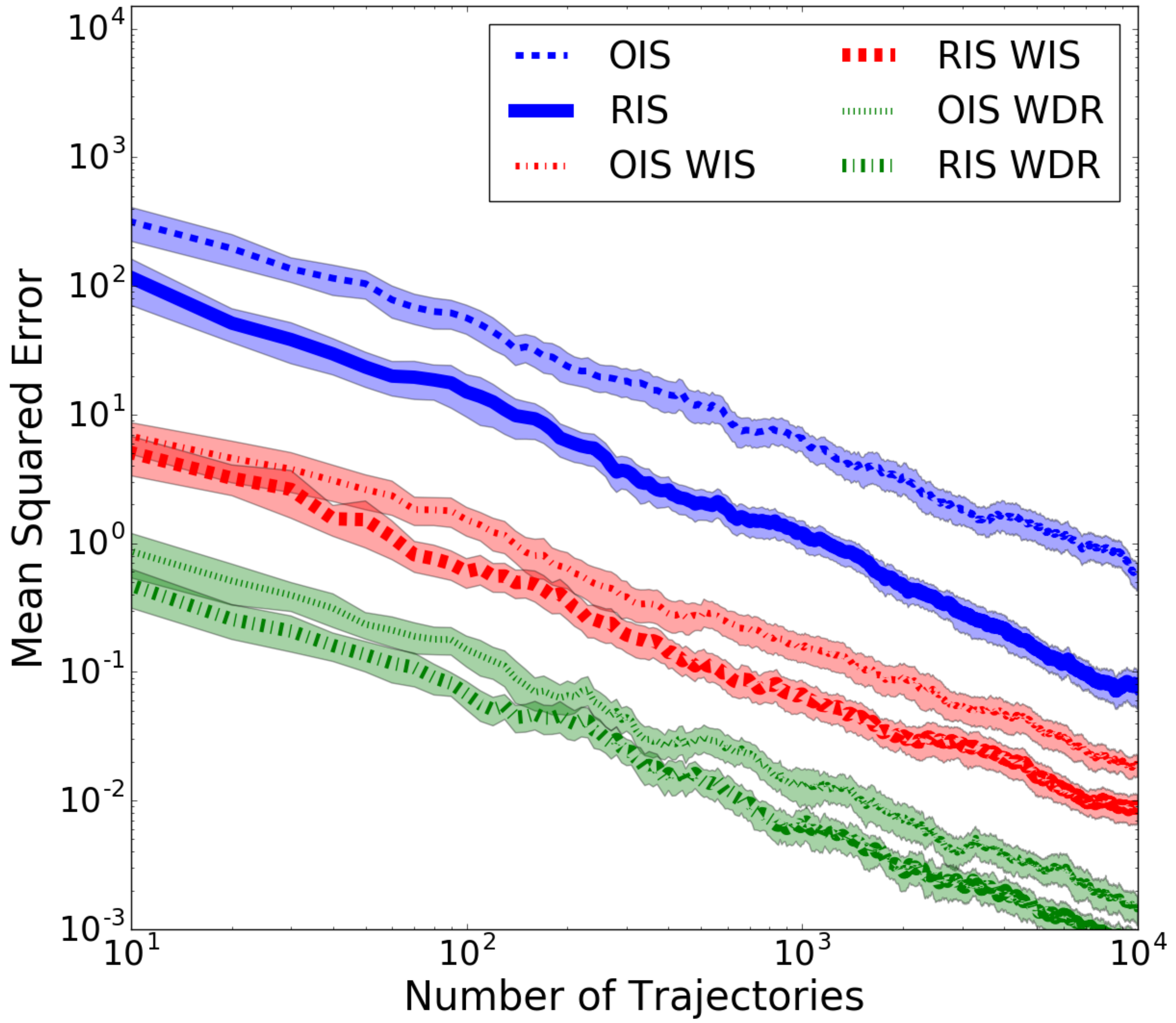


Gridworld

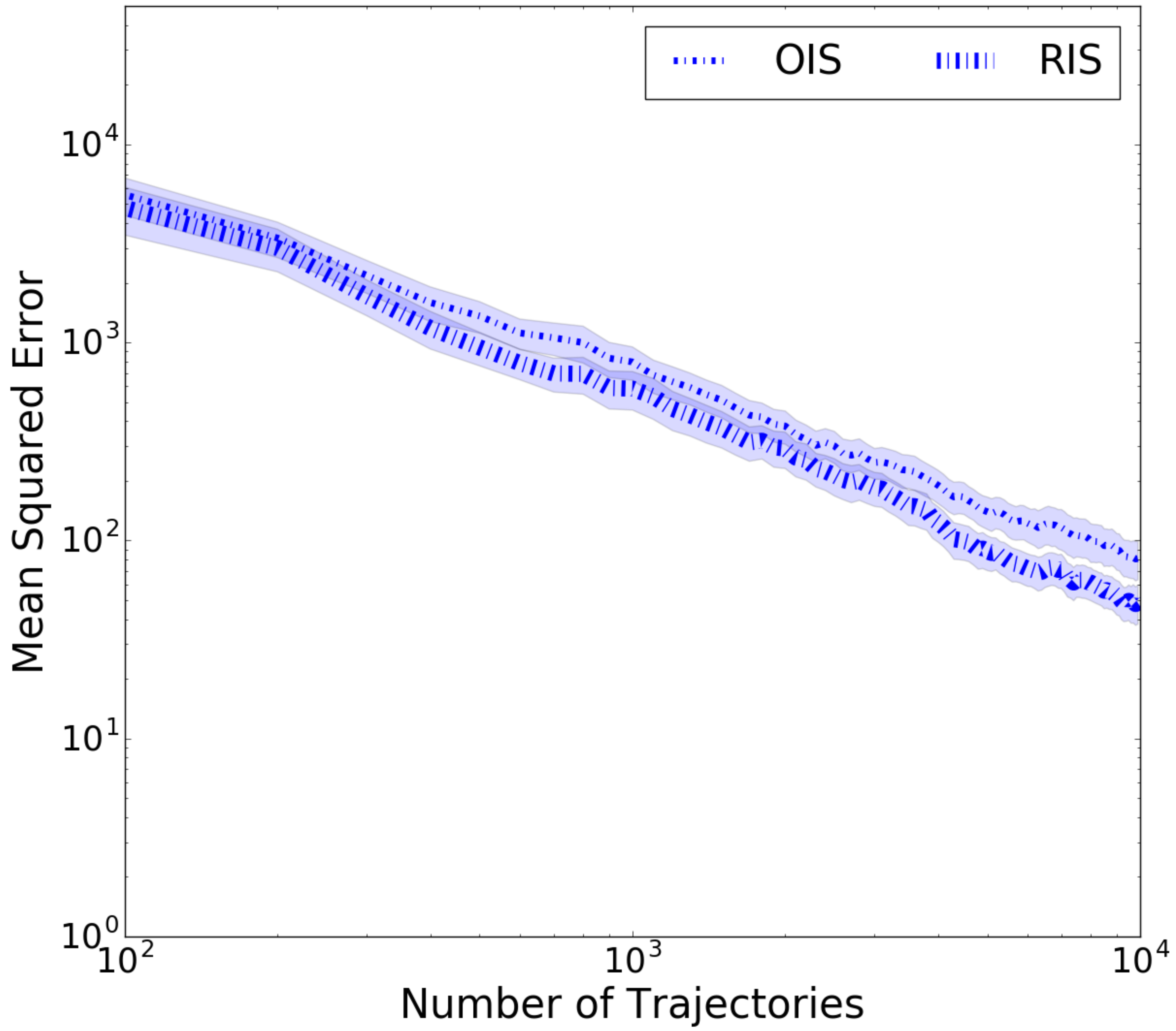


Linear Dynamical System

Empirical Results

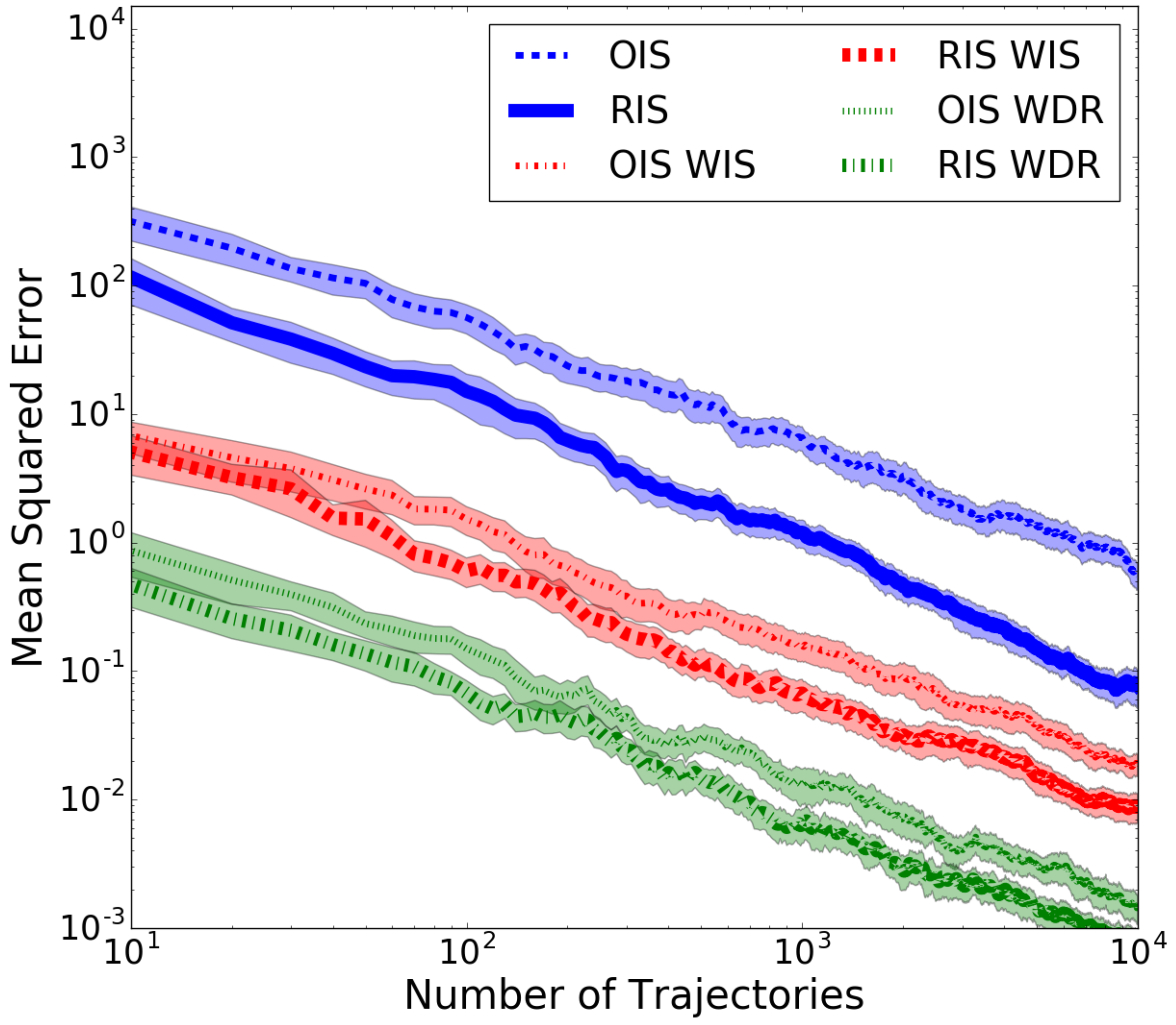


Gridworld

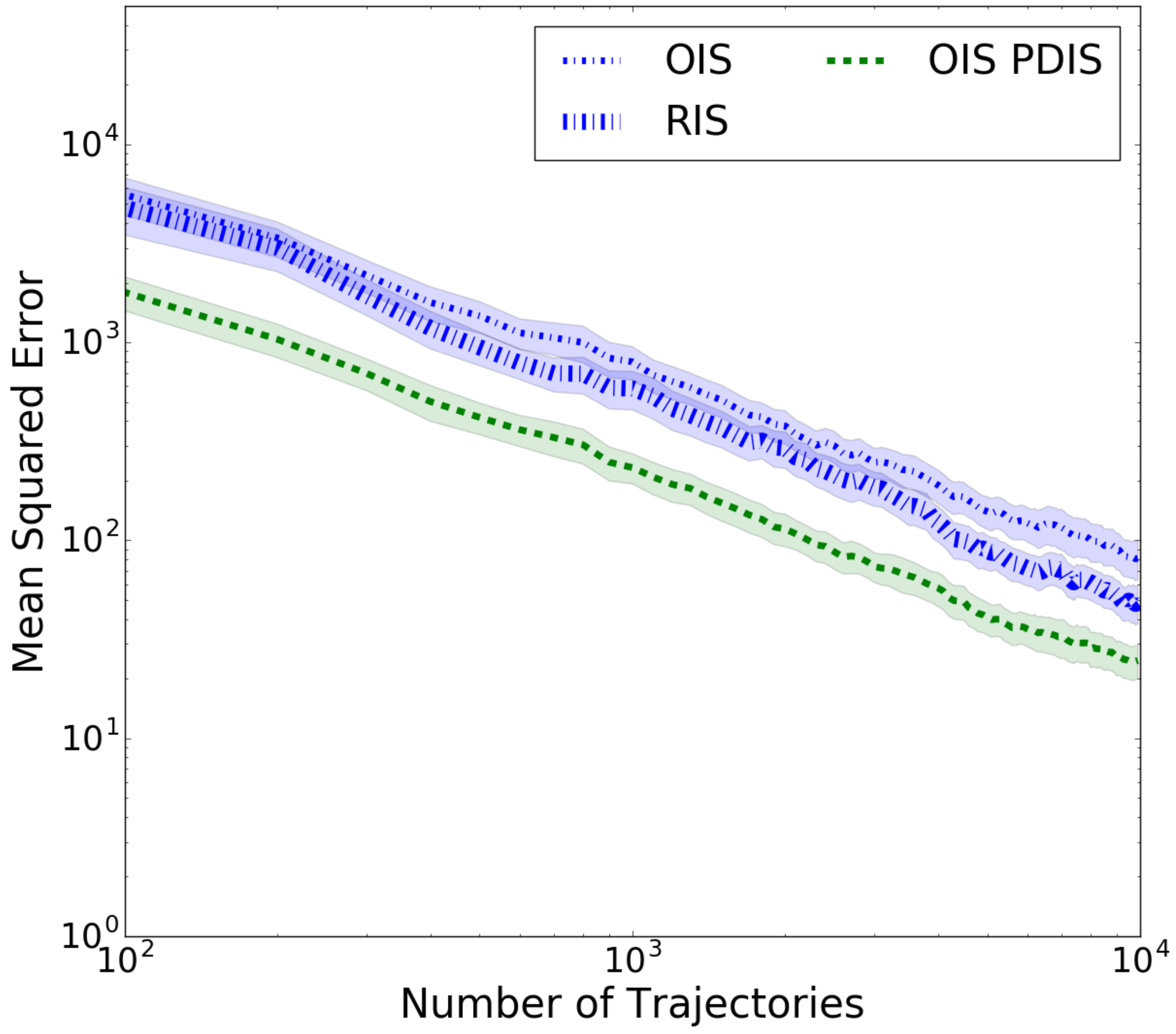


Linear Dynamical System

Empirical Results

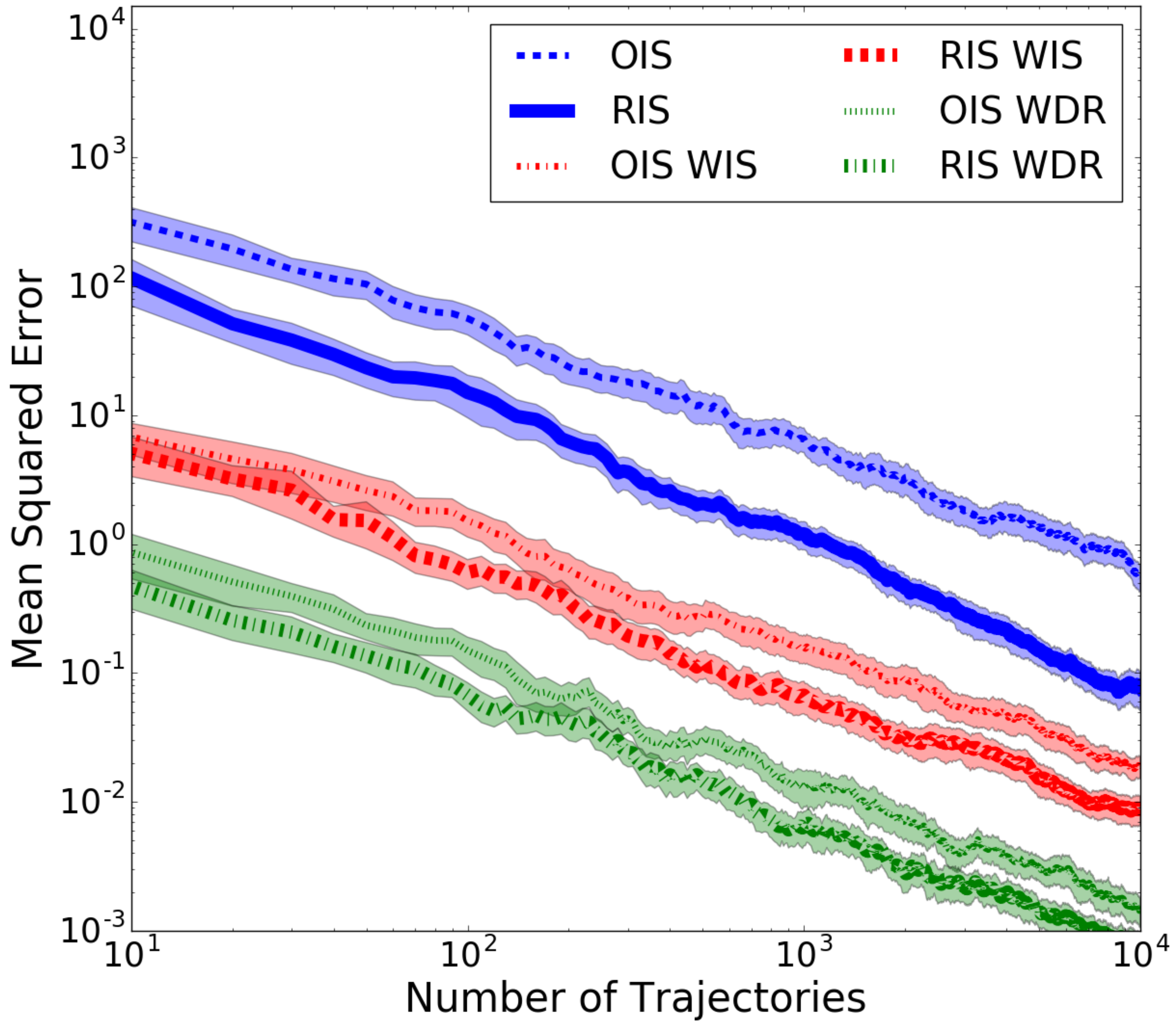


Gridworld

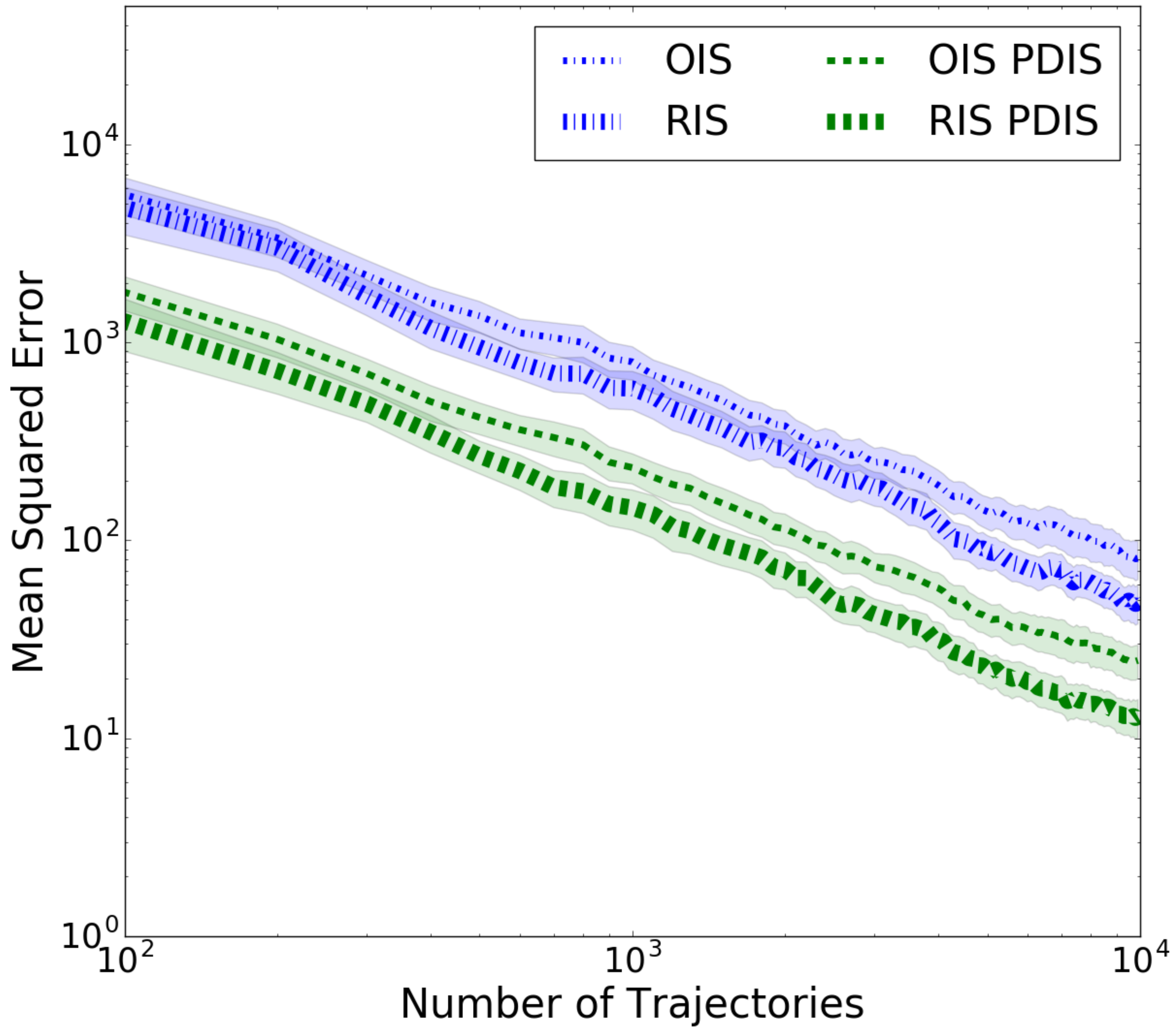


Linear Dynamical System

Empirical Results

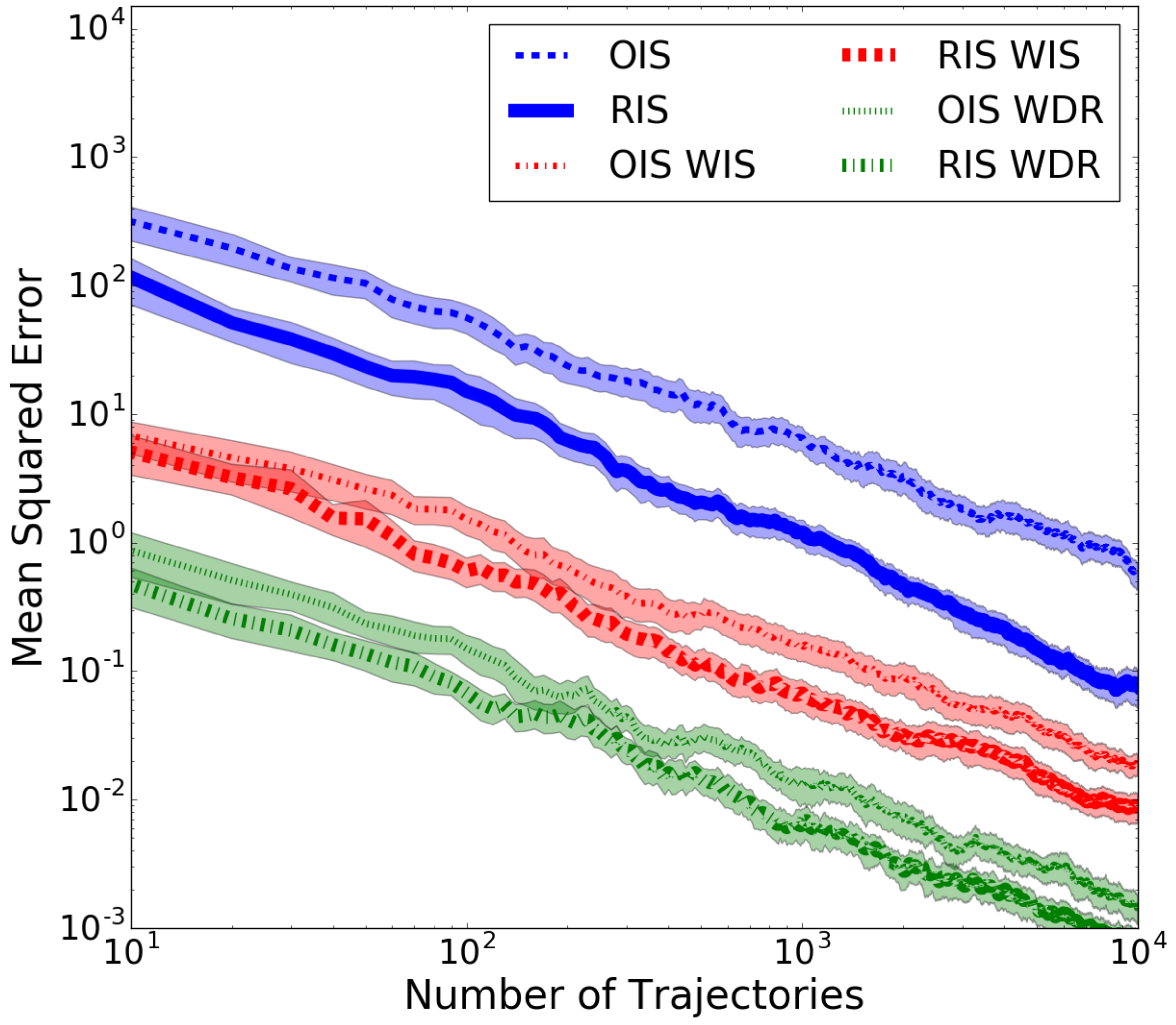


Gridworld

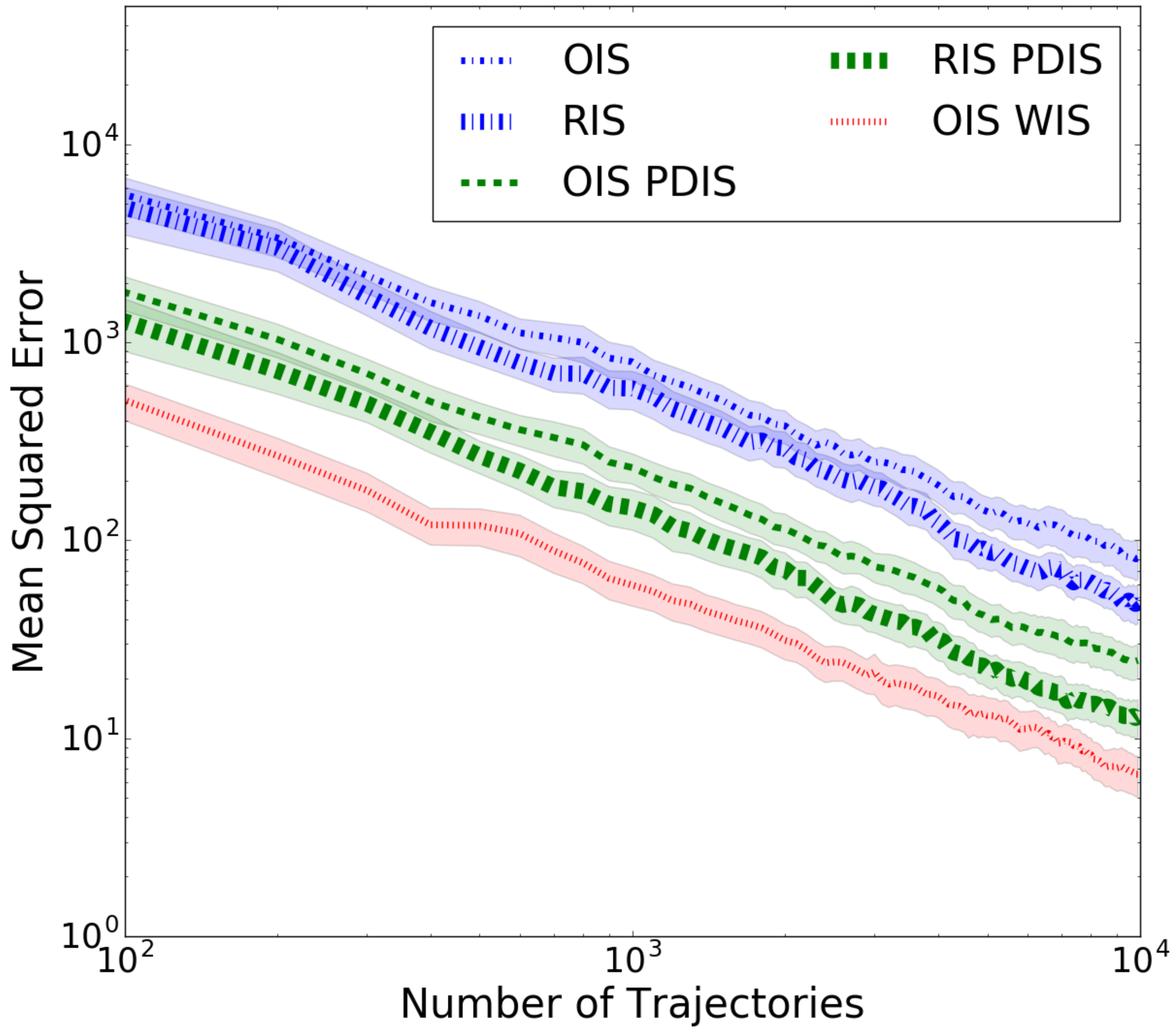


Linear Dynamical System

Empirical Results

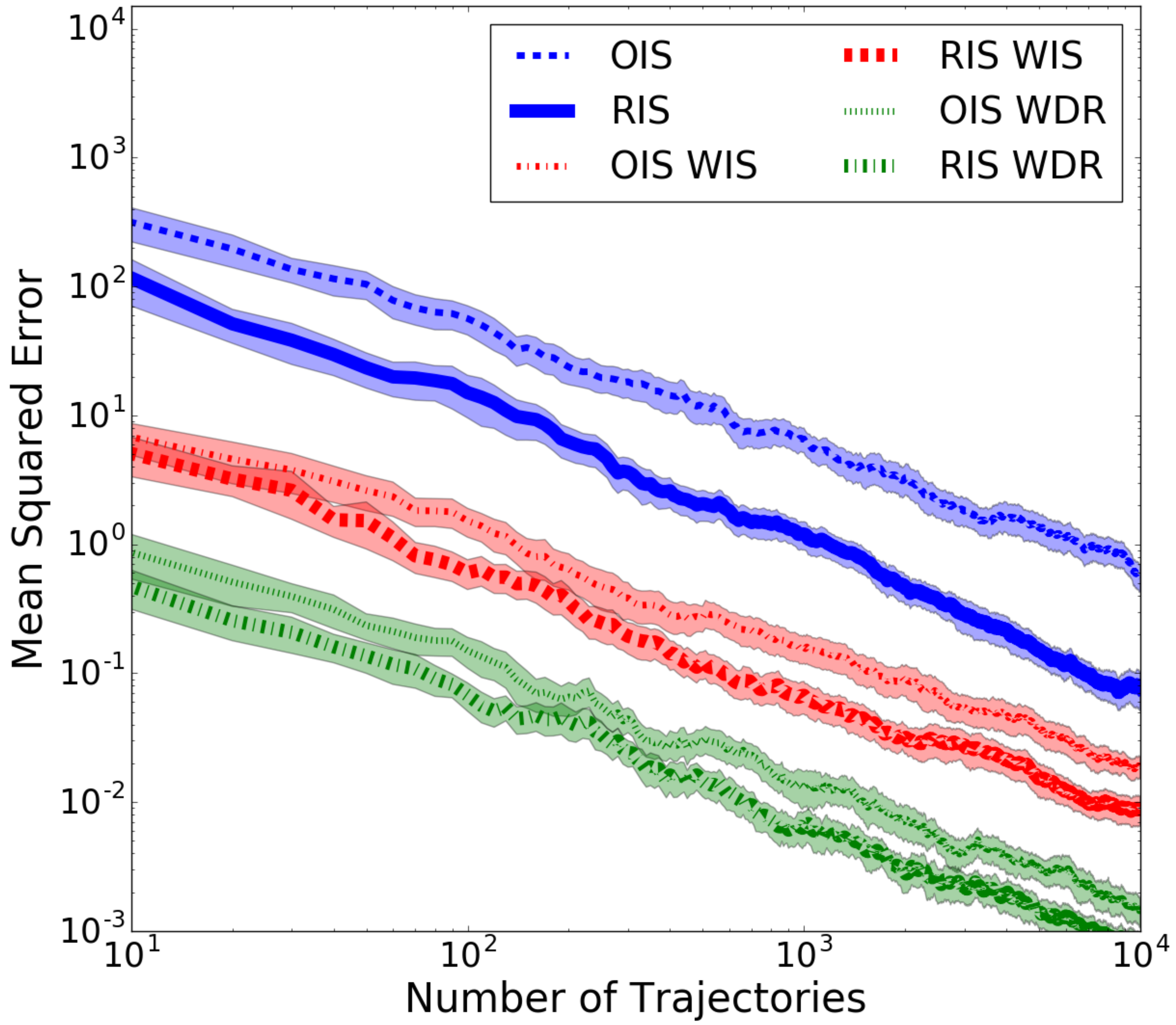


Gridworld

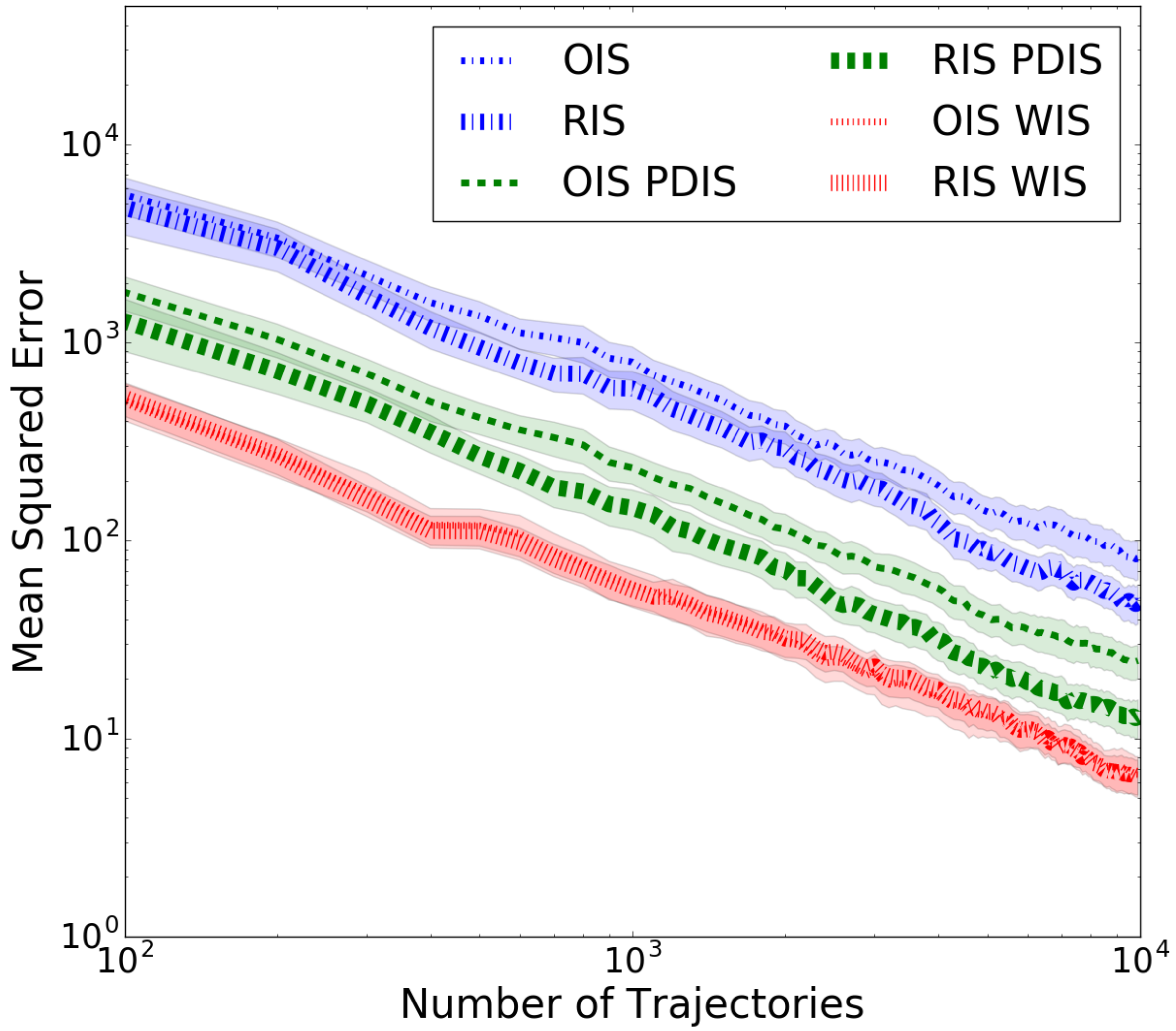


Linear Dynamical System

Empirical Results



Gridworld



Linear Dynamical System

Related Work

Related Work

1. Estimated Propensity Scores (Hirano et al. 2003, Li et al. 2015).

Related Work

1. Estimated Propensity Scores (Hirano et al. 2003, Li et al. 2015).
2. Learning in bandits + MDPs (Xie et al. 2019, Hanna and Stone 2019, Narita et al. 2019)

Related Work

1. Estimated Propensity Scores (Hirano et al. 2003, Li et al. 2015).
2. Learning in bandits + MDPs (Xie et al. 2019, Hanna and Stone 2019, Narita et al. 2019)

We are the first to show using an estimated behavior policy improves importance sampling in multi-step environments.

Tuesday 6:30-9, Pacific Ballroom #109

Tuesday 6:30-9, Pacific Ballroom #109

1. Off-policy importance sampling methods typically use the known behavior policy action probabilities.

Tuesday 6:30-9, Pacific Ballroom #109

1. Off-policy importance sampling methods typically use the known behavior policy action probabilities.

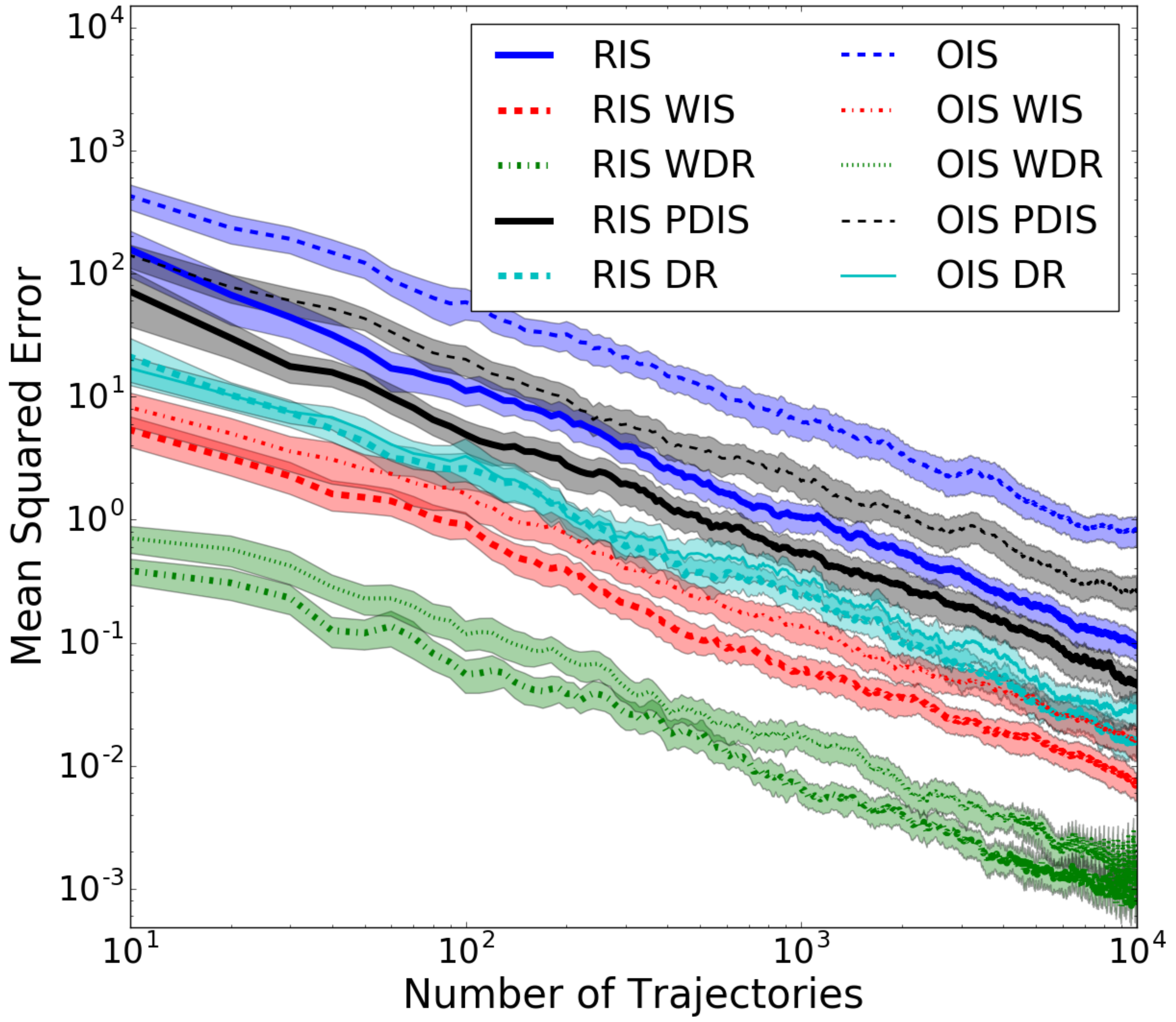
2. Replacing the true behavior policy action probabilities with their empirical estimate increases the effectiveness of importance sampling.

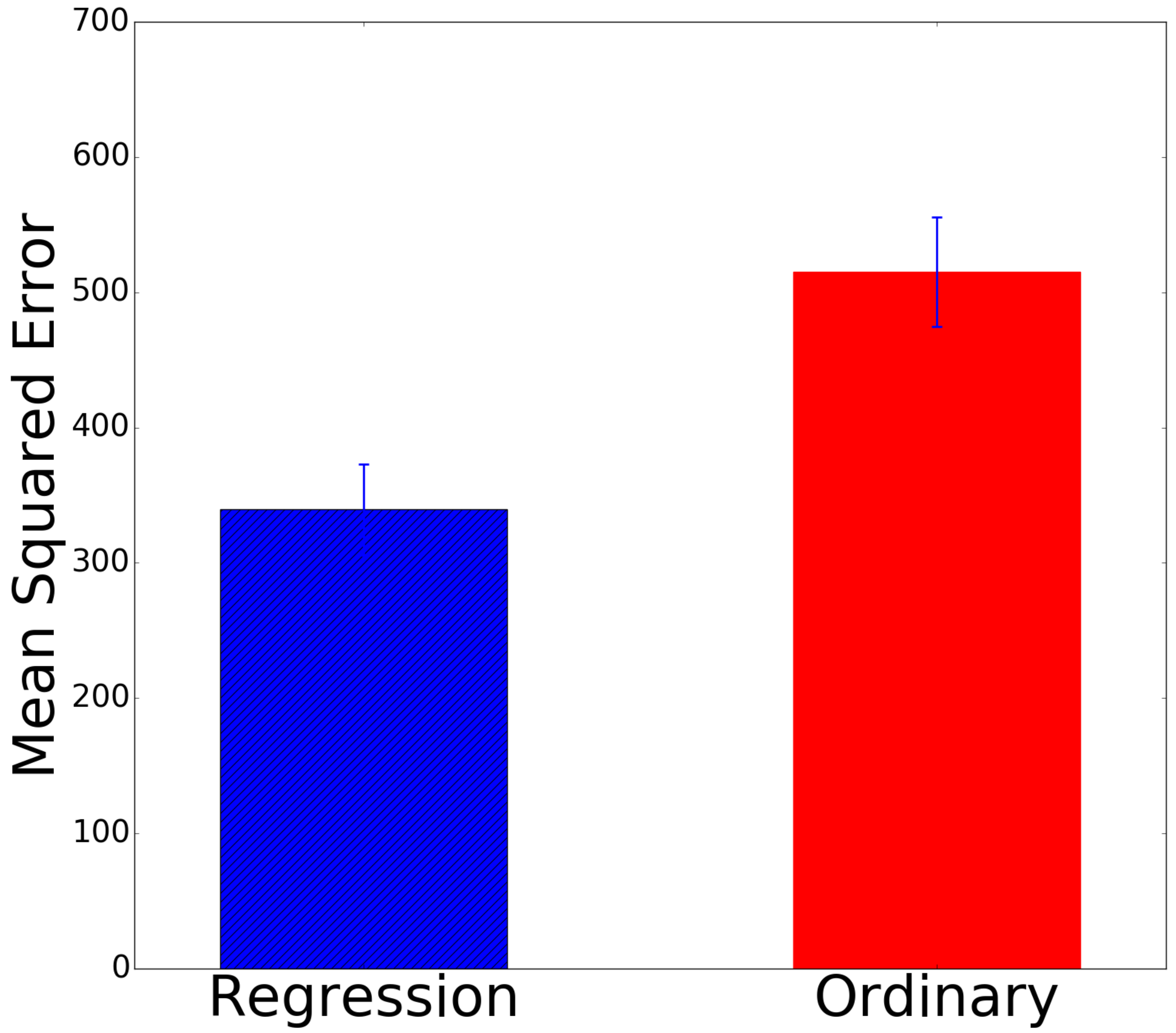
Tuesday 6:30-9, Pacific Ballroom #109

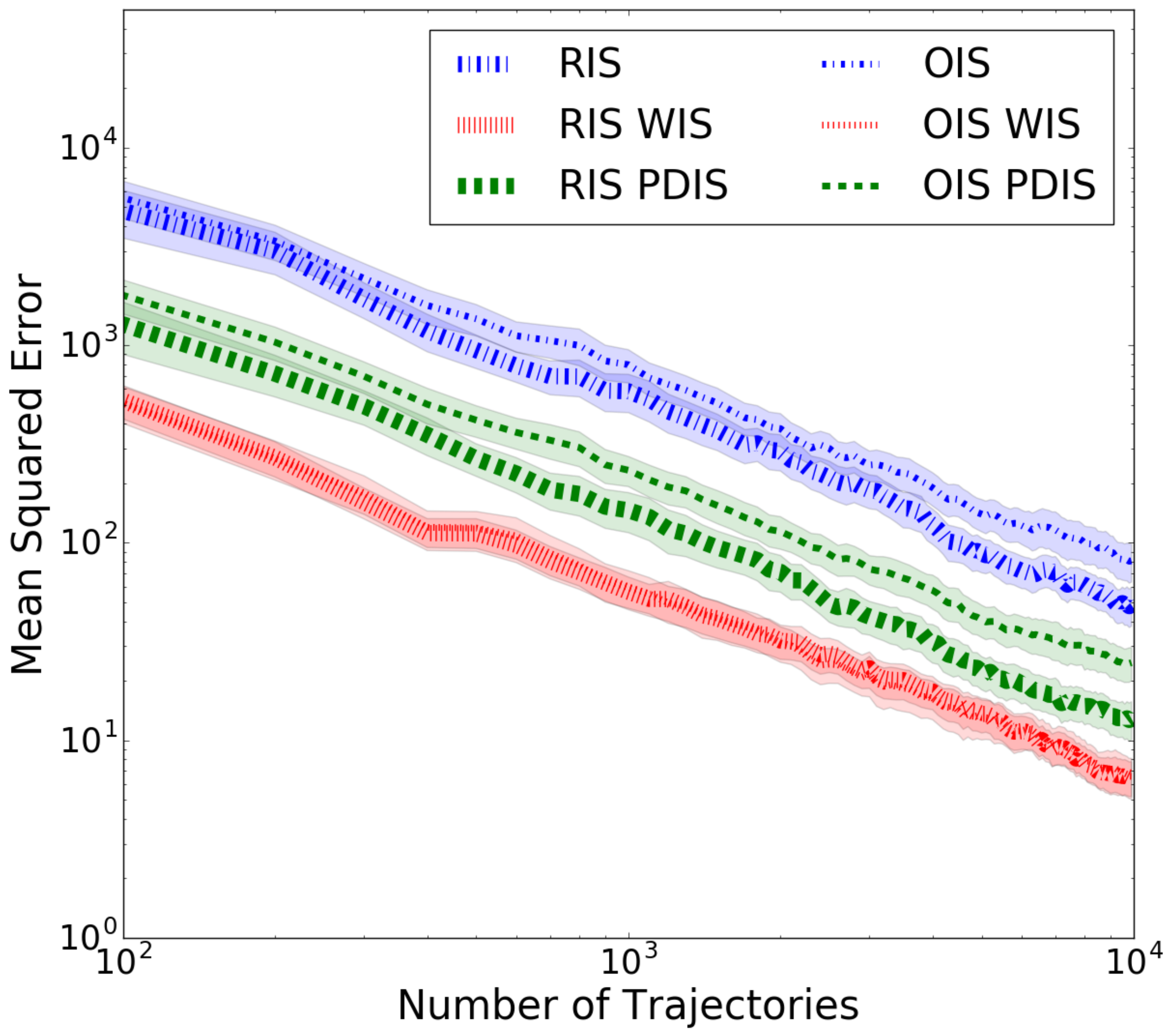
1. Off-policy importance sampling methods typically use the known behavior policy action probabilities.

2. Replacing the true behavior policy action probabilities with their empirical estimate increases the effectiveness of importance sampling.

3. We introduced the regression importance sampling and show it improves batch policy evaluation in a wide range of RL tasks.







Tuesday 6:30-9, Pacific Ballroom #109

jphanna@cs.utexas.edu