

---

# Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning

---

Cameron Voloshin  
Caltech

Hoang M. Le  
Caltech

Nan Jiang  
UIUC

Yisong Yue  
Caltech

## Abstract

Off-policy policy evaluation (OPE) is the problem of estimating the online performance of a policy using only pre-collected historical data generated by another policy. Given the increasing interest in deploying learning-based methods for safety-critical applications, many recent OPE methods have recently been proposed. Due to disparate experimental conditions from recent literature, the relative performance of current OPE methods is not well understood. In this work, we present the first comprehensive empirical analysis of a broad suite of OPE methods. Based on thousands of experiments and detailed empirical analyses, we offer a summarized set of guidelines for effectively using OPE in practice, and suggest directions for future research.

## 1 Introduction

Off-policy policy evaluation (OPE) aims to estimate a policy’s value using only pre-collected data generated by some other (possibly unknown) behavior policy (Sutton & Barto, 2018; Dann et al., 2014). For real-world reinforcement learning (RL) applications, such as robotics, autonomous vehicles, trading, advertising, drug trials, and traffic control, deploying a new policy without first assessing its performance can be costly, and sometimes dangerous (Li et al., 2011; Wiering, 2000; Bottou et al., 2013; Bang & Robins, 2005). It is critically important to generate accurate off-line counterfactual predictions of how a new policy performs.

The earliest OPE methods rely on classical importance sampling to handle the distribution mismatch between the target and behavior policies (Precup et al., 2000).

More advanced methods have since been proposed for both the contextual bandit (Dudík et al., 2011b; Bottou et al., 2013; Swaminathan et al., 2017; Wang et al., 2017) and RL setting (Jiang & Li, 2016; Dudík et al., 2011a; Farajtabar et al., 2018; Liu et al., 2018). Recent interest in OPE reflects the recognition that OPE is central to many off-policy learning algorithms (Degris et al., 2012; Munos et al., 2016; Le et al., 2019; Liu et al., 2019; Nie et al., 2019), in addition to being an important and challenging problem in its own right.

Managing the bias-variance trade-off is a recurring theme in OPE research. While many recent methods are built on sound mathematical principles, a practitioner is often faced with the non-trivial task of selecting the most appropriate estimator for their application. A notable gap in the current literature is a comprehensive empirical understanding of contemporary methods, due in part to the disparate testing environments and varying experimental conditions among prior work. Consequently, there is little holistic insight into where different methods particularly shine, nor a systematic summary of the challenges one may encounter when in different scenarios.

In this work, we provide a thorough empirical study of a wide range of OPE methods. Our study encompasses a variety of conditions to explore the success and failure modes of different methods. We synthesize high-level insights to guide practitioners, and suggest directions for future research. Finally, we provide a software package that can interface with different experimental platforms to run OPE experiments at scale.

## 2 Preliminaries

We adopt standard RL notations, where the environment is represented by a Markov Decision Process  $\langle X, A, P, R, \gamma \rangle$ .  $X$  is the state space (or observation space in the POMDP case),  $A$  is the action space with finite cardinality,  $P : X \times A \times X \rightarrow [0, 1]$  is the transition function,  $R : X \times A \rightarrow \mathbb{R}$  is the reward function, and  $\gamma \in (0, 1]$  is the discount factor. A policy  $\pi$  maps states to a distribution over actions, and  $\pi(a|x)$  de-

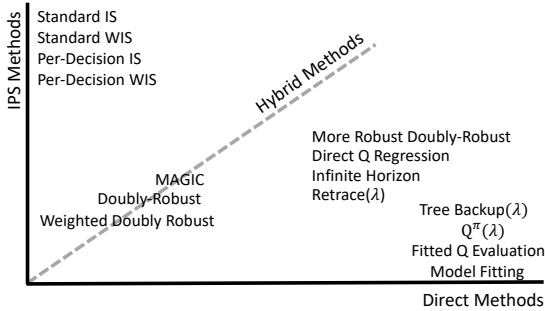


Figure 1: *Categorization of OPE methods. Some methods are direct but have IPS influence and thus fit slightly away from the direct methods axis.*

Table 1: IPS methods (Precup et al., 2000)

	STANDARD	PER-DECISION
IS	$\sum_{i=1}^N \frac{\rho_{0:T-1}^i}{N} \sum_{t=0}^{T-1} \gamma^t r_t$	$\sum_{i=1}^N \sum_{t=0}^{T-1} \gamma^t \frac{\rho_{0:t}^i}{N} r_t$
WIS	$\sum_{i=1}^N \frac{\rho_{0:T-1}^i}{w_{0:T-1}} \sum_{t=0}^{T-1} \gamma^t r_t$	$\sum_{i=1}^N \sum_{t=0}^{T-1} \gamma^t \frac{\rho_{0:t}^i}{w_{0:t}} r_t$

notes the probability of choosing  $a \in A$  in  $x \in X$ .

OPE is typically considered in the episodic RL setting. A behavior policy  $\pi_b$  generates a historical data set,  $D = \{\tau^i = (x_0^i, a_0^i, r_0^i, x_1^i, \dots)\}_{i=1}^N$ , of  $N$  trajectories (or episodes), where  $i$  indexes over trajectories, and  $\tau = (x_0, a_0, r_0, \dots, x_{T-1}, a_{T-1}, r_{T-1})$ . The episode length  $T$  is frequently assumed to be fixed, for notational and mathematical convenience. In practice, one can pad additional absorbing states to handle variable lengths. Given a desired evaluation policy  $\pi_e$ , the OPE problem is to estimate the value  $V(\pi_e)$ , defined as:

$$V(\pi_e) = \mathbb{E}_{x \sim d_0} \left[ \sum_{t=0}^{T-1} \gamma^t r_t | x_0 = x \right],$$

with  $a_t \sim \pi_e(\cdot | x_t)$ ,  $x_{t+1} \sim P(\cdot | x_t, a_t)$ ,  $r_t \sim R(x_t, a_t)$ , and  $d_0$  is the initial state distribution.

### 3 Overview of OPE Methods

OPE methods were historically categorized into importance sampling methods, direct methods, or doubly robust methods. This demarcation was first introduced for contextual bandits (Dudík et al., 2011a), and later extended to the RL setting (Jiang & Li, 2016). Some recent methods have blurred the boundary of these categories. Examples include Retrace( $\lambda$ ) (Munos et al., 2016) that uses a product of importance weights of multiple time steps for off-policy  $Q$  correction, and MAGIC (Thomas & Brunskill, 2016) that switches between importance weighting and direct methods.

In this paper, we propose to group OPE into three similar classes of methods, but with expanded definition for each category. Figure 1 provides an overview of OPE methods that we consider. The relative position-

ing of different methods reflects how close they are to being pure regression-based estimator vs. pure importance sampling-based estimator. Appendix B contains a full description of all methods under consideration.

#### 3.1 Inverse Propensity Scoring (IPS)

Inverse Propensity Scoring (IPS) has a rich history in statistics (Powell & Swann, 1966; Hammersley & Handscomb, 1964; Horvitz & Thompson, 1952), with successful crossover to RL (Precup et al., 2000). Let  $\rho_{j:j'}^i = \rho_{j:j'}(\tau^i, \pi_e, \pi_b) = \prod_{t=j}^{\min(j', T-1)} \frac{\pi_e(a_t^i | x_t^i)}{\pi_b(a_t^i | x_t^i)}$  be the cumulative importance weight between  $\pi_e$  and  $\pi_b$  (denote  $\rho_{t:t'}^i = 1$  for  $t' < t$ ). The key idea is to reweight the rewards in the historical data by the importance sampling ratio between  $\pi_e$  and  $\pi_b$ , i.e., how likely a reward is under  $\pi_e$  versus  $\pi_b$ . IPS methods yield consistent and (typically) unbiased estimates; however the product of importance weights can be unstable for long time horizons. Weighted IPS replaces the normalization factor  $N$  by  $w_{j:j'} = \frac{1}{N} \sum_{i=1}^N \rho_{j:j'}^i$ . The weighted versions are biased but strongly consistent.

Table 1 describes the four main IPS variants that we consider: Importance Sampling (IS), Per-Decision Importance Sampling (PDIS), Weighted Importance Sampling (WIS) and Per-Decision WIS (PDWIS). Note that other variants exist but are neither consistent nor unbiased (Thomas, 2015). IPS requires knowing  $\pi_b$ , which is often not possible – one approach is to estimate  $\pi_b$  from data (Hanna et al., 2019), resulting in a potentially biased estimator that can sometimes outperform traditional IPS methods.

#### 3.2 Direct Methods (DM)

While some direct methods make use of importance weight adjustments, a key distinction of direct methods is the focus on regression-based techniques to (more) directly estimate the value functions of the evaluation policy ( $Q^{\pi_e}$  or  $V^{\pi_e}$ ). We consider 8 different direct approaches. Similar to policy learning in RL, direct methods in OPE can also be viewed through the lens of model-based vs. model-free approaches.

*Model-based.* Perhaps the most commonly used DM is model-based (also called approximate model, denoted AM), where the transition dynamics, reward function and termination condition are directly estimated from historical data (Jiang & Li, 2016; Paduraru, 2013). The resulting learned MDP is then used to compute the value of  $\pi_e$ , e.g., by Monte-Carlo policy evaluation.

*Model-free.* Estimating the action-value function  $\hat{Q}(\cdot; \theta)$ , parametrized by some  $\theta$  is the focus of several model-free approaches. Estimating  $\hat{Q}$  yields  $\hat{V}(\pi_e) = \frac{1}{N} \sum_{i=1}^N \sum_{a \in A} \pi_e(a | s_0^i) \hat{Q}(x_0^i, a; \theta)$ . A simple example is Fitted Q Evaluation (FQE) (Le et al., 2019), which

Table 2: Environment parameters

Environment	Graph	Graph-MC	MC	Pix-MC	Enduro	Graph-POMDP	GW	Pix-GW
Is MDP?	yes	yes	yes	yes	yes	no	yes	yes
State desc.	position	position	[pos, vel]	pixels	pixels	position	position	pixels
$T$	4 or 16	250	250	250	1000	2 or 8	25	25
Stoch Env?	variable	no	no	no	no	no	no	variable
Stoch Rew?	variable	no	no	no	no	no	no	no
Sparse Rew?	variable	terminal	terminal	terminal	dense	terminal	dense	dense
$\hat{Q}$ Func. Class	tabular	tabular	linear/NN	NN	NN	tabular	tabular	NN
Initial state	0	0	variable	variable	gray img	0	variable	variable
Absorb. state	2T	22	[.5,0]	[.5,0]	zero img	2T	64	zero img
Frame height	1	1	2	2	4	1	1	1
Frame skip	1	1	5	5	1	1	1	1

is a model-free counterpart to AM, and is functionally a policy evaluation counterpart to batch Q learning.  $\hat{Q}(\cdot, \theta) = \lim_{k \rightarrow \infty} \hat{Q}_k$  where:

$$\hat{Q}_k = \min_{\theta} \frac{1}{N} \sum_{i=1}^N \sum_{t=0}^{T-1} (\hat{Q}_{k-1}(x_t^i, a_t^i; \theta) - y_t^i)^2,$$

$$y_t^i \equiv r_t^i + \gamma \mathbb{E}_{\pi_e} \hat{Q}_{k-1}(x_{t+1}^i, \cdot; \theta), \quad \hat{Q}_0 \equiv 0$$

Note:  $\mathbb{E}_{\pi_e} Q(x_{t+1}, \cdot) = \sum_{a \in A} \pi_e(a|x_{t+1}) Q(x_{t+1}, a)$ .

Several model-free methods originated from off-policy learning settings, but are also natural for OPE.  $Q^\pi(\lambda)$  (Harutyunyan et al., 2016) can be viewed as a generalization of FQE that looks to the horizon limit to incorporate the long-term value into the backup step. Retrace( $\lambda$ ) (Munos et al., 2016) and Tree-Backup( $\lambda$ ) (Precup et al., 2000) also use full trajectories, but additionally incorporate varying levels of clipped importance weights adjustment. The  $\lambda$ -dependent term mitigates instability in the backup step, and is selected based on experimental findings of Munos et al. (2016).

Direct Q Regression (Q-Reg) and More Robust Doubly-Robust (MRDR) (Farajtabar et al., 2018) are two recently proposed direct methods that make use of cumulative importance weights in deriving the regression estimate for  $Q^{\pi_e}$ , solved through a quadratic program. MRDR changes the objective of the regression to that of directly minimizing the variance of the Doubly-Robust estimator (see Section 3.3).

Liu et al. (2018) recently proposed a method for the infinite horizon setting (IH). While IH can be viewed as a Rao-Blackwellization of the IS estimator, we include it in the DM category because it essentially solves the Bellman equation for state distributions and requires function approximation, which are more characteristic of DM. IH shifts the focus from importance sampling over action sequences to estimating the importance ratio  $\omega$  between *state density distributions* induced by  $\pi_b$  and  $\pi_e$ . This ratio replaces all but the final importance weights  $\rho_{T-1}$  in the IH estimate, which resembles IS.

### 3.3 Hybrid Methods (HM)

Hybrid methods subsume doubly robust-like approaches, which combine aspects of both IPS and DM. Standard doubly robust OPE (denoted DR) (Jiang & Li, 2016) is an unbiased estimator that leverages DM to decrease the variance of the unbiased estimates produced by importance sampling techniques:

$$\sum_{i=1}^N \frac{\hat{V}(x_0^i)}{N} + \frac{1}{N} \sum_{i=1}^N \sum_{t=0}^{T-1} \gamma^t \rho_{0:t}^i [r_t^i - \hat{Q}(x_t^i, a_t^i) + \gamma \hat{V}(x_{t+1}^i)]$$

Other HM include Weighted Doubly-Robust (WDR) and MAGIC. WDR replaces the importance weights with self-normalized importance weights (similar to WIS). MAGIC introduces adaptive switching between DR and DM; in particular, one can imagine using DR to estimate the value for part of a trajectory and then using DM for the remainder. Using this idea, MAGIC (Thomas & Brunskill, 2016) finds an optimal linear combination among a set that varies the switch point between WDR and DM. Note that any DM that returns  $\hat{Q}^{\pi_e}(x, a; \theta)$  yields a set of corresponding DR, WDR, and MAGIC estimators. As a result, we consider 21 hybrid approaches in our experiments.

## 4 Experiments

**Protocol.** An experiment generally comprises a choice of environment, data-collecting policy ( $\pi_b$ ), evaluation policy ( $\pi_e$ ), and number of trajectories to collect ( $N$ ). For each experiment,  $\pi_b$  is rolled out  $N$  times to simulate the historical data  $D$ . The true on-policy value  $V(\pi_e)$  is the Monte-Carlo estimate via 10,000 rollouts of  $\pi_e$ . We repeat each experiment 10 times with different random seeds. We judge the quality of an OPE estimator by its relative mean squared error:  $\frac{1}{10} \sum_{i=1}^{10} \frac{(\hat{V}(\pi_e)_i - \frac{1}{10} \sum_{j=1}^{10} V(\pi_e)_j)^2}{(\frac{1}{10} \sum_{j=1}^{10} V(\pi_e)_j)^2}$ , which allows a fair comparison across different conditions.<sup>1</sup>

<sup>1</sup>The performance metric in prior OPE work is typically mean squared error  $\text{MSE} = \frac{1}{10} \sum_{i=1}^{10} (\hat{V}(\pi_e)_i - V(\pi_e)_i)^2$

Table 3: High-level Guidelines

Class	Recommended method	When to use	Prototypical example
Direct	FQE	Stochastic env, severe policy mismatch	Graph, MC, Pix-MC
	$Q(\lambda)$	Compute non-issue, moderate policy mismatch	GW/Pix-GW
	IH	Long horizon, mild policy mismatch, good kernel	Graph-MC
IPS	PDWIS	Short horizon, mild policy mismatch	Graph
Hybrid	MAGIC FQE	Severe model misspecification	Graph-POMDP, Enduro
	MAGIC $Q(\lambda)$	Compute non-issue, severe model misspecification	Graph-POMDP

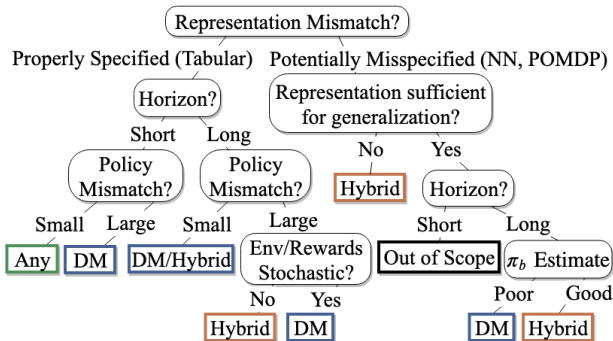


Figure 2: General Guideline Decision Tree

**Design.** We consider various domain characteristics (simple-complex, deterministic-stochastic, sparse-dense rewards, short-long horizon),  $\pi_b, \pi_e$  pairs (close-far), and data sizes  $N$  (small-large), to understand the effect of experimental conditions on OPE performance.

We use two standard RL benchmarks from OpenAI (Brockman et al., 2016): Mountain Car (MC) and Enduro Atari game. As many RL benchmarks are fixed and deterministic, we design 6 additional environments that allow control over various conditions: (i) Graph domain (tabular, varying stochasticity and horizon), (ii) Graph-POMDP (tabular, control for representation), (iii) Graph-MC (simplifying MC to tabular case), (iv) Pixel-MC (study MC in high-dimensional setting), (v) Gridworld (tabular, long horizon version) and (vi) Pixel-Gridworld (controlled Gridworld experiments with function approximation).

All together, our benchmark consists of 8 environments with characteristics summarized in table 2. Complete descriptions can be found in Appendix C. All environments have finite action spaces.

**Implementation.** With 33 different OPE methods considered, we run thousands of experiments across the above 8 domains. We create a software package to perform experiments at scale, accommodating both local and distributed computation. Our platform can be easily integrated with new domains for future research. Due to limited space, we show the results from selected environmental conditions in the next section. The full detailed results, with highlighted best method in each class, are available in the appendix.

## 5 Results and Discussion

### 5.1 High-Level Conclusions

The first important takeaway from our empirical results is that *there is no clear-cut winner*: no single method or method class is consistently the best performer. With that caveat in mind, we summarize the key general trends in Figure 2, where the recommendations are based on several key decision factors:

- *Horizon length*: Long horizons hurt all methods, but especially those dependent on importance weights (including IPS, HM and some DM).
- *Environment/Reward stochasticity*: Stochastic environments hurt the data efficiency of all methods, but favors DM over HM and IPS.
- *Unknown behavior policy*<sup>2</sup>:  $\pi_b$  estimation quality depends on the state and action dimensionality, and historical data size. Poor  $\pi_b$  estimates cause HM and IPS to underperform simple DM.
- *Policy mismatch*: Large divergence between  $\pi_b$  and  $\pi_e$  hurts all methods, but tends to favor DM in the small data regime relative to HM and IPS. HM will catch up with DM as data size increases.
- *Model misspecification*: Creates issues related to the representation power of function approximators, and partial observability. Model misspecification does not impact IPS. Severe misspecification favors HM and weakens DM.

These factors often affect performance in interdependent ways, and the impact varies for different method classes. Thus, it is important to be aware of the nuances when applying different methods. Figure 3 shows a typical comparison of the best performing method in each class, under a tabular setting (Graph domain) with both short and long horizons, and a large mismatch between  $\pi_b$  and  $\pi_e$ . We note that the best method in each class may change depending on the specific conditions. Within each class, a general guideline for method selection is summarized in Table 3. The appendix contains the full empirical results.

<sup>2</sup>Poor estimation of  $\pi_b$  can be viewed as an instance of model misspecification. We distinguish the representation issue of  $\pi_b$  from other representation issues related to DM

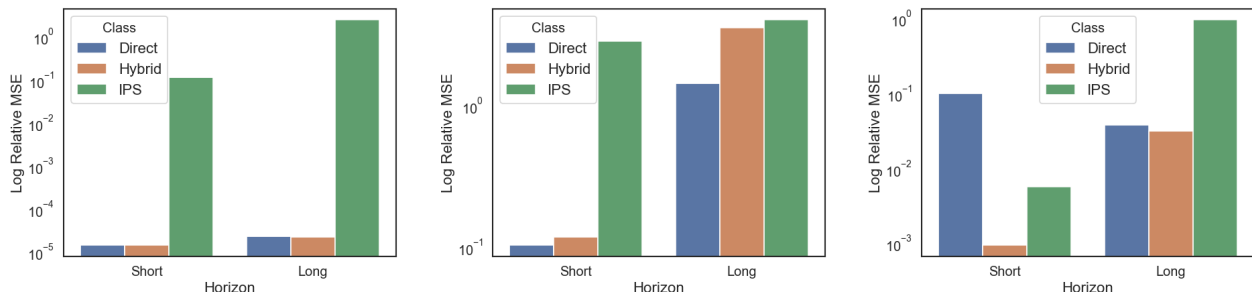


Figure 3: Comparing IPS vs. DM vs. HM under short and long horizon, large policy mismatch and large data. Left: (Graph domain) Deterministic environment. Center: (Graph domain) Stochastic environment and rewards. Right: (Graph-POMDP) Model misspecification (POMDP). Minimum error per class is shown.

## 5.2 Method Selection Guideline

**Inverse Propensity Scoring (IPS).** IPS methods generally only do well when the horizon is short, or when behavior and evaluation policies are close to one another. IPS has the advantage of being insensitive to state dimensionality, thus immune to function approximation errors when  $\pi_b$  is known. Other than the simple short horizon setting, traditional IPS methods almost always perform worse than other method classes.

Having unbiased estimates is a notable feature of basic importance sampling methods. In practice, weighted importance sampling, which is biased, tends to be more accurate and data-efficient. Among the four IPS-based methods, *PDWIS tends to perform best* (Figure 4 left). Non-weighted importance sampling should be preferred when the horizon is short and unbiasedness is required.

**Direct methods (DM).** Overall, DM are surprisingly competitive. In tabular MDPs, we often do not see significant benefits of hybrid methods (standard DR, WDR, MAGIC) over DM. When the data size is relatively small, in complex domains that require function approximation, have high stochasticity, or large policy mismatch, the best DM tends to outperform the all other methods.

Generally,  $Q^\pi(\lambda)$ , *FQE* and *IH* tend to perform the best among DM (Figure 5). *FQE* tends to be more data efficient and is the best method when data is limited.  $Q^\pi(\lambda)$ , which generalizes *FQE* to multi-step backup, works particularly well with large data set, but is computationally expensive in complex domains. *Retrace*( $\lambda$ ) and *Tree-Backup*( $\lambda$ ) are iterative backup methods (similar to  $Q^\pi(\lambda)$  and *FQE*), but with cumulative IS terms.  $Q^\pi(\lambda)$  can be unstable under severe policy mismatch (e.g., Graph-MC domain in Tables 367-370). *Retrace*( $\lambda$ ) uses clipped importance weight adjustment and is more stable, but generally does not perform better than  $Q^\pi(\lambda)$ . *Tree-Backup*( $\lambda$ ) is typically worse than other DM under small policy mis-

match (e.g., Tables 375-382). *IH*, on the other hand, is highly competitive in long horizon domains, with small policy mismatch. In pixel-based domains, however, choosing a good kernel function for *IH* is not straightforward. We provide a numerical comparison among direct methods for tabular (Figure 16) and complex settings (Figure 4 center).

While AM performs well in tabular setting in the large data case (Figure 16), it tends to perform poorly in high dimensional settings with function approximation (e.g., Figure 4 center). Fitting the transition model  $P(x'|x, a)$  is often more prone to small errors than directly approximating  $Q(x, a)$ . Model fitting errors also compound with long horizons.

Q-Reg and MRDR both require solving large linear systems even for modest horizons. Thus implementing Q-Reg and MRDR require extra care to avoid ill-conditioning, such as tuning with L1 and L2 regularization.<sup>3</sup> MRDR was designed to improve upon direct Q regression by minimizing the variance of a doubly-robust version of Q-Regression. In our experiments, the benefit of MRDR over the simpler direct Q regression method is not clear. It is arguable that both methods also suffer from the curse of horizon (like IPS). In fact, the direct versions of Q-Reg and MRDR underperform PDWIS in several tabular settings (e.g., Gridworld Tables 426-450).

**Hybrid methods (HM).** In a large data regime, HM typically outperform DM (Figure 30-36). Under severe model misspecification (such as for POMDPs), HM improve on the weakness of DM and also lowers variance of IPS. Knowing the behavior policy exactly guarantees unbiased estimates (for DR)(Thomas & Brunskill, 2016). Unfortunately, under high-dimensionality, long horizons, estimated behavior policies, and reward/environment stochasticity, HM can underperform simple DM, sometimes significantly (e.g., see Figure 17).

<sup>3</sup>From correspondence with the authors.

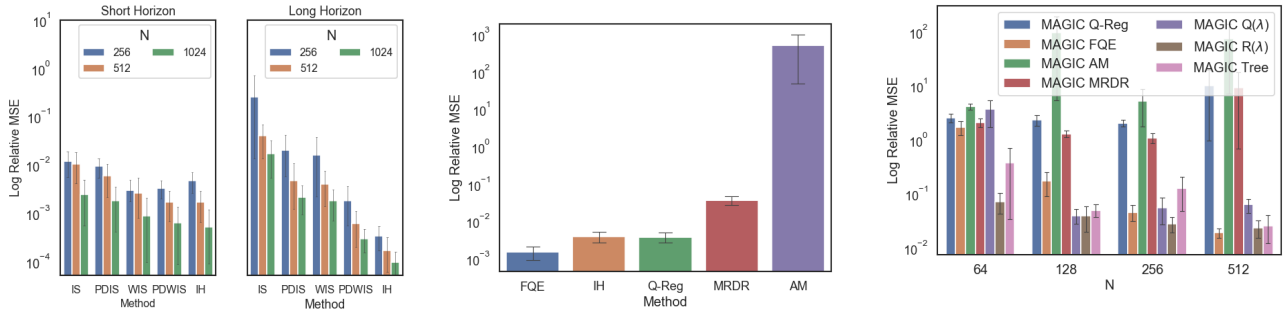


Figure 4: Left: (Graph domain) Comparing IPS (and IH) under short and long horizon. Mild policy mismatch setting. PDWIS is often best among IPS. But IH outperforms in long horizon. Center: (Pixel-MC) Comparing direct methods in high-dimensional, long horizon setting. Relatively large policy mismatch. FQE and IH tend to outperform. AM is significantly worse in complex domains. Retrace( $\lambda$ ),  $Q(\lambda)$  and Tree-Backup( $\lambda$ ) are very computationally expensive and thus excluded. Right: (Pixel Gridworld) Comparing MAGIC with different base DM and different data size. Large policy mismatch, deterministic environment, known  $\pi_b$ .

With the exception of IH, each DM corresponds to three hybrid methods: standard doubly robust (DR), weighted doubly robust (WDR), and MAGIC. For each DM, its WDR version often outperforms standard DR version. MAGIC can often do better than WDR and DR. However, MAGIC comes with additional hyperparameter tuning requirement, as one needs to specify the set of partial trajectory length to be considered. Unsurprisingly, their performance highly depends on the underlying DM. In our experiments, FQE and  $Q^\pi(\lambda)$  are typically the most reliable: *MAGIC with FQE or MAGIC with  $Q^\pi(\lambda)$  tend to be among the best hybrid methods* (see Figures 22 - 26).

In many situations with limited data, however, HM may perform worse than DM:

- Tabular domains with large policy mismatch, or under stochastic environments (Figure 17).
- Complex domains with long horizon and unknown behavior policy (Figure 27-29). Long horizon, high dimensional setting with good function approximation (see GW/Pix-GW tables)

When data is sufficient, or model misspecification is severe, HM do provide consistent improvement over DM. The Graph-POMDP environment illustrates the advantage of HM when model misspecification is certain to be an issue (see Figure 3 right).

### 5.3 Deeper Dive into Key Decision Factors

*Horizon length.* It is well-known that IPS-based methods are sensitive to trajectory length (Li et al., 2015). Long horizon leads to exponential blow-up of the importance sampling term, and is exacerbated by significant mismatch between  $\pi_b$  and  $\pi_e$ . This issue is inevitable for any unbiased estimator (Jiang & Li, 2016) (a.k.a., the curse of horizon (Liu et al., 2018)). Similar to IPS, DM also suffer from long horizon (Figure

16), though to a lesser degree. IH aims to bypass the effect of cumulative weighting in long horizons, and indeed performs substantially better than IPS methods in very long horizon domains (Figure 4 left).

A frequently ignored aspect in previous OPE work is a proper distinction between fixed, finite horizon tasks (IPS focus), infinite horizon tasks (IH focus), and indefinite horizon tasks, where the trajectory length is finite but varies depending on the policy. Many applications should properly belong to the indefinite horizon category<sup>4</sup>. Applying HM in this setting requires proper padding of the rewards (without altering the value function in the infinite horizon limit) as DR correction typically assumes fixed length trajectories.

*Environment stochasticity.* While stochasticity affects all methods by straining the data requirement, HM are more negatively impacted than DM (Figure 3 center, Figure 17). This can be justified by e.g., the variance analysis of DR, which shows that the variance of the value function with respect to stochastic transitions will be amplified by cumulative importance weights and then contribute to the overall variance of the estimator; see Jiang & Li (2016, Theorem 1) for further details. We empirically observe that DM frequently outperform their DR versions in the small data case (Figure 17). In a stochastic environment and tabular setting, HM do not provide significant edge over DM, even in short horizon case. The gap closes as the data size increases (Figure 3 center).

*Unknown behavior policy.* In many applications, the behavior policy may not be known exactly and requires estimation, which can introduce bias and cause HM to underperform simple DM, especially in low data

<sup>4</sup>Applying IH in the indefinite horizon case requires setting up a separate absorbing state that loops over itself with zero terminal reward.

regime (e.g., pixel gridworld appendix figure 27-29). Similar phenomenon was observed in the statistics literature (Kang et al., 2007). As the historical data size increases, HM regain the advantage as the quality of the  $\pi_b$  estimate improves.

*Policy mismatch.* Similar to IPS, the performance of DM is negatively correlated with the degree of policy divergence. Figure 5 shows the interplay of increasing policy mismatch and historical data size, on the top DM in the deterministic gridworld. We use  $(\sup_{a \in A, x \in X} \frac{\pi_e(a|x)}{\pi_b(a|x)})^T$  as an environment-independent metric of divergence between the two policies. The performance of the top DM (FQE,  $Q^\pi(\lambda)$ , IH) tend to hold up better than IPS methods (WIS) when the policy gap increases (Figure 18). FQE and IH are best in the small data regime, and  $Q^\pi(\lambda)$  performs better as data size increases (Figure 5). Increased policy mismatch weakens the DM that use importance weights (Q-Reg, MRDR, Retrace( $\lambda$ ) and Tree-Backup( $\lambda$ )).

*Model misspecification.* Model misspecification refers to the insufficient representation power of the function class used to approximate either the transition dynamics (AM), value function (other DM), or state distribution density ratio (in IH). We study the effect of misspecification via two controlled scenarios:

- Simple domains: Tabular representation for DM for partially observable environments causing poor generalization.
- Complex domains: Function approximation has good generalization capacity but (potentially) introduces inherent Bellman error<sup>5</sup> (Munos & Szepesvári, 2008; Le et al., 2019).

Tabular representation controls for one aspect of the misspecification by ensuring zero inherent Bellman error, for both MDPs and POMDPs. As tabular representation lacks a natural ability to generalize without sufficient historical data, the effect of misspecification is thus exposed in the partial observation case, unlike the fully observable case. HM substantially outperform DM in this setting (Figure 3 right vs. left).

In complex domains, function approximation with good generalization ability makes DM very competitive with HM, especially under limited data (pixel-Gridworld Figures 27-29, see also linear vs. neural networks comparison for Mountain Car in Figure 13). However, function approximation bias may cause serious problem for high dimensional and long horizon settings. In the extreme case of Enduro (very long horizon with sparse rewards), all direct methods fail to convincingly outperform a naïve average of behavior

<sup>5</sup>defined as  $\sup_{g \in \mathcal{F}} \inf_{f \in \mathcal{F}} \|f - \mathbb{T}^\pi g\|_{d_\pi}$ , where  $\mathcal{F}$  is function class chosen for approximation, and  $d_\pi$  is state distribution induced by evaluation policy  $\pi$

policy (Figure 12). Quantifying biasedness, such as inherent Bellman error, for different function classes is currently an open problem (Chen & Jiang, 2019).

#### 5.4 Other Considerations

*Sparsity (non-smoothness) of the rewards:* Methods that are dependent on cumulative importance weights are also sensitive to reward sparsity. If the rewards are sparse, then all IPS methods perform poorly. If the rewards are dense, then per-decision estimators can salvage some performance (Figure 19). An often over-looked aspect of using importance weighting is the need to normalize the rewards. As a rough guideline, zero-centering rewards often benefits the performance of IPS overall. This seemingly naïve practice can be actually viewed as a special case of DR using a constant DM component, and can yield improvements over vanilla IPS (Jiang & Li, 2016).

*Computational considerations.* DM are generally significantly more computationally demanding than IPS. In complex domains, model-free iterative methods can be expensive in training time. Iterative DM that incorporate rollouts until the end of trajectories during training (Retrace( $\lambda$ ),  $Q^\pi(\lambda)$ , Tree-Backup( $\lambda$ )) are the most computationally demanding.<sup>6</sup> They require order  $T$  times the number of  $\hat{Q}_{k-1}(x, a)$  lookups per gradient step compared to FQE. Model-based method (AM) are expensive at test time when coupled with HM, since rolling-out the learned model is required at every state along the trajectory.<sup>7</sup> HM versions of direct methods require  $T$  times more inference steps, which is often fast after training. In difficult tasks such as Atari games, running AM, Retrace( $\lambda$ ),  $Q^\pi(\lambda)$ , Tree-Backup( $\lambda$ ) can be prohibitively expensive. Q-Reg, MRDR are non-iterative methods and thus are the fastest to execute among DM. The run-time of IH is dependent on the batch size in building a kernel matrix to compute state similarity. The batch size for IH should be as large as possible, but could significantly slow the training.

*Hyperparameter tuning.* Different types of direct methods require different sets of hyperparameters. For example, the choice of specific function approximator varies for model-based (transition dynamics and rewards), IH (kernel function), and other model-free direct methods (Q parameterization). In episodic environments, iterative DM (FQE,  $Q^\pi(\lambda)$ , R( $\lambda$ ), Tree) need to be run at least  $T$  times for the entire backup to complete. Also, when using function approxima-

<sup>6</sup>Munos et al. (2016) limits the rolling-out horizon to 16 in Atari domains, but for policy learning scenario.

<sup>7</sup>Unlike iterative DM (e.g., FQE), model-based method AM does not benefit from stochastic gradient speedup. Parallelizing the rollouts of AM is highly recommended.

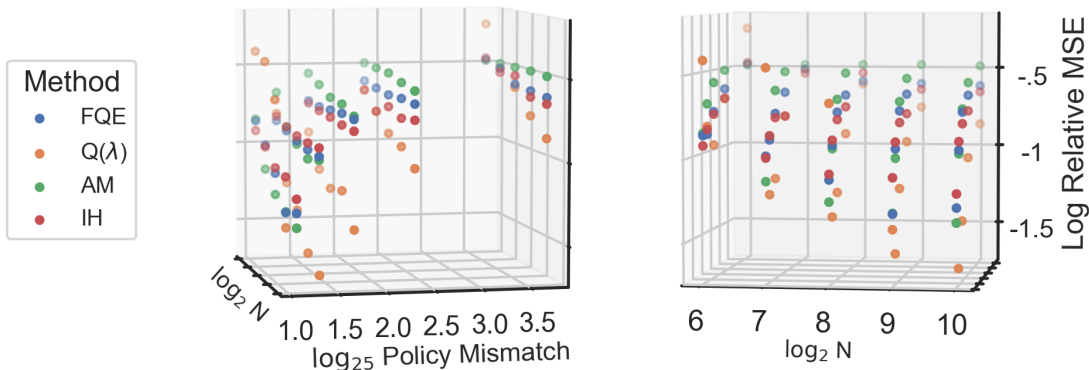


Figure 5: (Gridworld domain) Errors are directly correlated with policy mismatch but inversely correlated to the amount of data. We pick the best direct methods for illustration. The two plots represent the same figure from two different vantage points. See full figures in appendix.

tion, direct methods may not have satisfactory convergence, and require setting a reasonable termination threshold hyperparameter. In our experiments, we aim to maintain a consistent set of hyperparameters for each direct approach and each environment across experimental conditions (see Table 14 in appendix). Problem-dependent hyperparameter search, while technically possible, comes at the cost of an independent data requirement and extra computational cost. *In general, given the choice among different hybrid (or direct) methods, we suggest opting for simplicity as a guiding principle.*

## 6 Limitations and Future Directions

Atari games pose significant challenges for contemporary techniques due to very long horizon and high state space dimensionality. As the amount of data we collect in our Enduro experiments is much lower than the typical number of samples for policy learning setting, it is possible that substantially more historical data is required for current OPE methods to succeed. To overcome computational challenge in complex RL domains, it is important to identify principled ways to stabilize iterative methods such as FQE, Retrace( $\lambda$ ),  $Q(\lambda)$  when using function approximation, as convergence is typically not attainable. Similarly, the various choice of the kernel function for IH and the index set for hybrid method such as MAGIC have large impact on the performance. Future work should address the need for systematic hyperparameter tuning.

Validation of other complex RL tasks with short horizon is currently beyond the scope of our study, due to the lack of a proper benchmark. We refer to prior work on OPE for contextual bandits, which are RL problems with horizon 1 (Dudík et al., 2011b). For contextual bandits, it has been shown that while DR is highly competitive, it is sometimes substantially out-

performed by DM (Wang et al., 2017). New benchmark tasks should have longer horizon than contextual bandits, but shorter than Atari games. We also currently lack natural stochastic environments in high-dimensional RL benchmarks. A candidate for medium horizon, complex OPE domain is an NLP task such as dialogue, which has not been the focus of prior work.

Another drawback of recent literature on OPE is the exclusive focus on finite actions. OPE for continuous action domains will benefit continuous control applications. Currently, continuous action domains will not work with all IPS and HM (see IPS for continuous contextual bandits by (Kallus & Zhou, 2018)). Among DM, perhaps only FQE may reasonable work with continuous action tasks with some adaptation.

Finally, while we have identified a general guideline for selecting OPE method, often it is not easy to judge whether some decision criteria are satisfied (e.g., quantifying model misspecification, degree of stochasticity, or appropriate data size). As more OPE methods continue to be developed, an important missing piece is a systematic technique for model selection, given a high degree of variability among existing techniques.

## 7 Conclusion

We have presented a systematic study of contemporary methods for the problem of off-policy policy evaluation in reinforcement learning. For the first time, we gather comprehensive empirical evidence for the strengths and weaknesses of various techniques to guide researchers and practitioners. We design our empirical study to cover a wide range of experimental conditions that one may encounter in typical reinforcement learning tasks. Outside of the domains considered in this paper, our software package can integrate a new environment to allow further analysis at scale.



## References

- Bang, H. and Robins, J. M. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005. doi: 10.1111/j.1541-0420.2005.00377.x.
- Bottou, L., Peters, J., nonero Candela, J. Q., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research (JMLR)*, 14:3207–3260, 2013.
- Brockman, G., Cheung, V., Petteersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym, 2016.
- Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2019.
- Dann, C., Neumann, G., and Peters, J. Policy evaluation with temporal differences: A survey and comparison. *Journal of Machine Learning Research (JMLR)*, 15(1): 809–883, 2014.
- Degrís, T., White, M., and Sutton, R. S. Off-policy actor-critic. 2012.
- Dudík, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. In *International Conference on Machine Learning (ICML)*, 2011a.
- Dudík, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pp. 1097–1104. Omnipress, 2011b.
- Farajtabar, M., Chow, Y., and Ghavamzadeh, M. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning (ICML)*, 2018.
- Gauci, J., Conti, E., Liang, Y., Virochsiri, K., Chen, Z., He, Y., Kaden, Z., Narayanan, V., and Ye, X. Horizon: Facebook’s open source applied reinforcement learning platform. *arXiv preprint arXiv:1811.00260*, 2018.
- Hammersley, J. M. and Handscomb, D. C. Monte carlo methods. 1964.
- Hanna, J., Niekum, S., and Stone, P. Importance sampling policy evaluation with an estimated behavior policy. In *International Conference on Machine Learning (ICML)*, 2019.
- Harutyunyan, A., Bellemare, M. G., Stepleton, T., and Munos, R.  $Q(\lambda)$  with off-policy corrections. In *Conference on Algorithmic Learning Theory (ALT)*, 2016.
- Horvitz, D. G. and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260): 663–685, 1952.
- Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2016.
- Kallus, N. and Zhou, A. Policy evaluation and optimization with continuous treatments. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.
- Kang, J. D., Schafer, J. L., et al. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007.
- Le, H. M., Voloshin, C., and Yue, Y. Batch policy learning under constraints. In *International Conference on Machine Learning (ICML)*, 2019.
- Li, L., Chu, W., Langford, J., and Wang, X. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *ACM Conference on Web Search and Data Mining (WSDM)*, 2011.
- Li, L., Munos, R., and Szepesvari, C. Toward minimax off-policy value estimation. In *Artificial Intelligence and Statistics*, pp. 608–616, 2015.
- Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Neural Information Processing Systems (NeurIPS)*, 2018.
- Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. Off-policy policy gradient with state distribution correction. *arXiv preprint arXiv:1904.08473*, 2019.
- Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research (JMLR)*, 9(May):815–857, 2008.
- Munos, R., Stepleton, T., Harutyunyan, A., and Bellemare, M. Safe and efficient off-policy reinforcement learning. In *Neural Information Processing Systems (NeurIPS)*. 2016.
- Nie, X., Brunskill, E., and Wager, S. Learning when-to-treat policies. *arXiv preprint arXiv:1905.09751*, 2019.
- Paduraru, C. *Off-policy evaluation in Markov decision processes*. PhD thesis, McGill University Libraries, 2013.
- Powell, M. J. and Swann, J. Weighted uniform sampling monte carlo technique for reducing variance. *IMA Journal of Applied Mathematics*, 2(3):228–236, 1966.
- Precup, D., Sutton, R. S., and Singh, S. P. Eligibility traces for off-policy policy evaluation. In *International Conference on Machine Learning (ICML)*, 2000.
- Sherman, J. and Morrison, W. J. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Ann. Math. Statist.*, 21(1):124–127, 03 1950. doi: 10.1214/aoms/1177729893.
- Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.
- Swaminathan, A., Krishnamurthy, A., Agarwal, A., Dudík, M., Langford, J., Jose, D., and Zitouni, I. Off-policy evaluation for slate recommendation. In *Neural Information Processing Systems (NeurIPS)*. 2017.
- Thomas, P. Reinforcement learning: An introduction, 2015.
- Thomas, P. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2016.
- Wang, Y.-X., Agarwal, A., and Dudík, M. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning (ICML)*, 2017.
- Wiering, M. Multi-agent reinforcement learning for traffic light control. In *International Conference on Machine Learning (ICML)*, 2000.

<b>Contents</b>	C.1.8 Pixel-Gridworld (Pixel-GW) . . . 15
<b>1 Introduction</b>	<b>1</b>
<b>2 Preliminaries</b>	<b>1</b>
<b>3 Overview of OPE Methods</b>	<b>2</b>
3.1 Inverse Propensity Scoring (IPS) . . . . .	2
3.2 Direct Methods (DM) . . . . .	2
3.3 Hybrid Methods (HM) . . . . .	3
<b>4 Experiments</b>	<b>3</b>
<b>5 Results and Discussion</b>	<b>4</b>
5.1 High-Level Conclusions . . . . .	4
5.2 Method Selection Guideline . . . . .	5
5.3 Deeper Dive into Key Decision Factors . . . . .	6
5.4 Other Considerations . . . . .	7
<b>6 Limitations and Future Directions</b>	<b>8</b>
<b>7 Conclusion</b>	<b>8</b>
<b>A Glossary of Terms</b>	<b>11</b>
<b>B Methods</b>	<b>12</b>
B.1 Inverse Propensity Scoring (IPS) Methods	12
B.2 Hybrid Methods . . . . .	12
B.3 Direct Methods (DM) . . . . .	12
B.3.1 Model-Based . . . . .	12
B.3.2 Model-Free . . . . .	12
<b>C Environments</b>	<b>13</b>
C.1 Environment Descriptions . . . . .	13
C.1.1 Graph . . . . .	13
C.1.2 Graph-POMDP . . . . .	13
C.1.3 Graph Mountain Car (Graph-MC) . . . . .	13
C.1.4 Mountain Car (MC) . . . . .	14
C.1.5 Pixel-based Mountain Car (Pix-MC) . . . . .	14
C.1.6 Enduro . . . . .	14
C.1.7 Gridworld (GW) . . . . .	15
C.1.8 Pixel-Gridworld (Pixel-GW) . . . . .	15
<b>D Experimental Setup</b>	<b>15</b>
D.1 Description of the policies . . . . .	15
D.2 Enumeration of Experiments . . . . .	15
D.2.1 Graph . . . . .	15
D.2.2 Graph-POMDP . . . . .	15
D.2.3 Gridworld . . . . .	15
D.2.4 Pixel-Gridworld (Pix-GW) . . . . .	15
D.2.5 Graph-MC . . . . .	15
D.2.6 Mountain Car (MC) . . . . .	16
D.2.7 Pixel-Mountain Car (Pix-MC) . . . . .	16
D.2.8 Enduro . . . . .	16
D.3 Representation and Function Class . . . . .	16
D.4 Choice of hyperparameters . . . . .	17
<b>E Additional Supporting Figures</b>	<b>19</b>
<b>F Tables of Results, per Environment</b>	<b>25</b>
F.1 Detailed Results for Graph . . . . .	26
F.2 Detailed Results for Graph-POMDP . . . . .	90
F.3 Detailed Results for Graph Mountain Car (Graph-MC) . . . . .	114
F.4 Detailed Results for Mountain Car (MC) . . . . .	120
F.5 Detailed Results for Pixel-Based Mountain Car (Pix-MC) . . . . .	128
F.6 Detailed Results for Gridworld . . . . .	129
F.7 Detailed Results for Pixel Gridworld . . . . .	136
F.8 Detailed Results for Enduro . . . . .	149

## A Glossary of Terms

See Table 4 for a description of the terms used in this paper.

Table 4: Glossary of terms

Acronym	Term
OPE	Off Policy Evaluation
$X$	State Space
$A$	Action Space
$P$	Transition Function
$R$	Reward Function
$\gamma$	Discount Factor
$d_0$	Initial State Distribution
$D$	Dataset
$\tau$	Trajectory/Episode
$T$	Horizon/Episode Length
$N$	Number of episodes in $D$
$\pi_b$	Behavior Policy
$\pi_e$	Evaluation Policy
$V$	Value, ex: $V(\pi_e)$
$Q$	Action-Value, ex: $Q(\pi_e, a)$
$\rho_{j:j'}^i$	Cumulative Importance Weight, $\prod_{t=j}^{\min(j', T-1)} \frac{\pi_e(a_t^i x_t^i)}{\pi_b(a_t^i x_t^i)}$
IPS	Inverse Propensity Scoring
DM	Direct Method
HM	Hybrid Method
IS	Importance Sampling
PDIS	Per-Decision Importance Sampling
WIS	Weighted Importance Sampling
PDWIS	Per-Decision Weighted Importance Sampling
PDWIS	Per-Decision Weighted Importance Sampling
FQE	Fitted Q Evaluation (Le et al., 2019)
IH	Infinite Horizon (Liu et al., 2018)
Q-Reg	Q Regression (Farajtabar et al., 2018)
MRDR	More Robust Doubly Robst (Farajtabar et al., 2018)
AM	Approximate Model (Model Based)
$Q(\lambda)$	$Q^\pi(\lambda)$ (Harutyunyan et al., 2016)
$R(\lambda)$	Retrace( $\lambda$ ) (Munos et al., 2016)
Tree	Tree-Backup( $\lambda$ ) (Precup et al., 2000)
DR	Doubly-Robust (Jiang & Li, 2016; Dudík et al., 2011a)
WDR	Weighted Doubly-Robust (Dudík et al., 2011a)
MAGIC	Model And Guided Importance Sampling Combining (Estimator) (Thomas & Brunskill, 2016)
Graph	Graph Environment
Graph-MC	Graph Mountain Car Environment
MC	Mountain Car Environment
Pix-MC	Pixel-Based Mountain Car Environment
Enduro	Enduro Environment
Graph-POMDP	Graph-POMDP Environment
GW	Gridworld Environment
Pix-GW	Pixel-Based Gridworld Environment

## B Methods

Below we include a description of each of the methods we tested. Let  $\tilde{T} = T - 1$ .

### B.1 Inverse Propensity Scoring (IPS) Methods

Table 5: IPS methods. (Dudík et al., 2011a; Jiang & Li, 2016)

	STANDARD	PER-DECISION
IS	$\sum_{i=1}^N \frac{\rho_{0:\tilde{T}}^i}{N} \sum_{t=0}^{\tilde{T}} \gamma^t r_t$	$\sum_{i=1}^N \sum_{t=0}^{\tilde{T}} \gamma^t \frac{\rho_{0:t}^i}{N} r_t$
WIS	$\sum_{i=1}^N \frac{\rho_{0:\tilde{T}}^i}{w_{0:\tilde{T}}} \sum_{t=0}^{\tilde{T}} \gamma^t r_t$	$\sum_{i=1}^N \sum_{t=0}^{\tilde{T}} \gamma^t \frac{\rho_{0:t}^i}{w_{0:t}} r_t$

Table 5 shows the calculation for the four traditional IPS estimators:  $V_{IS}, V_{PDIS}, V_{WIS}, V_{PDWIS}$ . In addition, we include the following method as well since it is a Rao-Blackwellization (Liu et al., 2018) of the IPS estimators:

### B.2 Hybrid Methods

Hybrid rely on being supplied an action-value function  $\hat{Q}$ , an estimate of  $Q$ , from which one can also yield  $\hat{V}(x) = \sum_{a \in A} \pi(a|x) \hat{Q}(x, a)$ . Doubly-Robust (DR): (Thomas & Brunskill, 2016; Jiang & Li, 2016)

$$V_{DR} = \frac{1}{N} \sum_{i=1}^N \hat{V}(x_0^i) + \frac{1}{N} \sum_{i=1}^N \sum_{t=0}^{\infty} \gamma^t \rho_{0:t}^i [r_t^i - \hat{Q}(x_t^i, a_t^i) + \gamma \hat{V}(x_{t+1}^i)]$$

Weighted Doubly-Robust (WDR): (Thomas & Brunskill, 2016)

$$V_{WDR} = \frac{1}{N} \sum_{i=1}^N \hat{V}(x_0^i) + \sum_{i=1}^N \sum_{t=0}^{\infty} \gamma^t \frac{\rho_{0:t}^i}{w_{0:t}} [r_t^i - \hat{Q}(x_t^i, a_t^i) + \gamma \hat{V}(x_{t+1}^i)]$$

MAGIC: (Thomas & Brunskill, 2016) Given  $g_J = \{g^i | i \in J \subseteq \mathbb{N} \cup \{-1\}\}$  where

$$g^j(D) = \sum_{i=1}^N \sum_{t=0}^j \gamma^t \frac{\rho_{0:t}^i}{w_{0:t}} r_t^i + \sum_{i=1}^N \gamma^{j+1} \frac{\rho_{0:t}^i}{w_{0:t}} \hat{V}(x_{j+1}^i) - \sum_{i=1}^N \sum_{t=0}^j \gamma^t \left( \frac{\rho_{0:t}^i}{w_{0:t}} \hat{Q}(x_t^i, a_t^i) - \frac{\rho_{0:\tilde{T}}^i}{w_{0:\tilde{T}}} \hat{V}(x_t^i) \right),$$

then define  $dist(y, Z) = \min_{z \in Z} |y - z|$  and

$$\hat{b}_n(j) = dist(g_j^J(D), CI(g^\infty(D), 0.5))$$

$$\hat{\Omega}_n(i, j) = Cov(g_i^J(D), g_j^J(D))$$

then, for a  $|J|$ -simplex  $\Delta^{|J|}$  we can calculate

$$\hat{x}^* \in \arg \min_{x \in \Delta^{|J|}} x^T [\hat{\Omega}_n + \hat{b}\hat{b}^T] x$$

which, finally, yields

$$V_{MAGIC} = (\hat{x}^*)^T g_J.$$

MAGIC can be thought of as a weighted average of different blends of the DM and Hybrid. In particular, for some  $i \in J$ ,  $g^i$  represents estimating the first  $i$  steps of  $V(\pi_e)$  according to DR (or WDR) and then estimating the remaining steps via  $\hat{Q}$ . Hence,  $V_{MAGIC}$  finds the most appropriate set of weights which trades off between using a direct method and a Hybrid.

### B.3 Direct Methods (DM)

#### B.3.1 Model-Based

Approximate Model (AM): (Jiang & Li, 2016) An approach to model-based value estimation is to directly fit the transition dynamics  $P(x_{t+1}|x_t, a_t)$ , reward  $R(x_t, a_t)$ , and terminal condition  $P(x_{t+1} \in X_{terminal}|x_t, a_t)$  of the MDP using some form of maximum likelihood or function approximation. This yields a simulation environment from which one can extract the value of a policy using an average over rollouts. Thus,  $V(\pi) = \mathbb{E}[\sum_{t=1}^{\tilde{T}} \gamma^t r(x_t, a_t) | x_0 = x, a_0 = \pi(x_0)]$  where the expectation is over initial conditions  $x \sim d_0$  and the transition dynamics of the simulator.

#### B.3.2 Model-Free

Every estimator in this section will approximate  $Q$  with  $\hat{Q}(\cdot; \theta)$ , parametrized by some  $\theta$ . From  $\hat{Q}$  the OPE estimate we seek is

$$V = \frac{1}{N} \sum_{i=1}^N \sum_{a \in A} \pi_e(a|s) \hat{Q}(s_0^i, a; \theta)$$

Note that  $\mathbb{E}_{\pi_e} Q(x_{t+1}, \cdot) = \sum_{a \in A} \pi_e(a|x_{t+1}) Q(x_{t+1}, a)$ .

Direct Model Regression (Reg): (Farajtabar et al., 2018)

$$\hat{Q}(\cdot, \theta) = \min_{\theta} \frac{1}{N} \sum_{i=1}^N \sum_{t=0}^{\tilde{T}} \gamma^t \rho_{0:t}^i \left( R_{t:\tilde{T}}^i - \hat{Q}(x_t^i, a_t^i; \theta) \right)^2$$

$$R_{t:\tilde{T}}^i = \sum_{t'=t}^{\tilde{T}} \gamma^{t'-t} \rho_{(t+1):t'}^i r_{t'}^i$$

Fitted Q Evaluation (FQE): (Le et al., 2019)  $\hat{Q}(\cdot, \theta) = \lim_{k \rightarrow \infty} \hat{Q}_k$  where

$$\hat{Q}_k = \min_{\theta} \frac{1}{N} \sum_{i=1}^N \sum_{t=0}^{\tilde{T}} \left( \hat{Q}_{k-1}(x_t^i, a_t^i; \theta) - y_t^i \right)^2$$

$$y_t^i \equiv r_t^i + \gamma \mathbb{E}_{\pi_e} \hat{Q}_{k-1}(x_{t+1}^i, \cdot; \theta)$$

Retrace( $\lambda$ ) (R( $\lambda$ )), Tree-Backup (Tree),  $Q^\pi(\lambda)$ :  
 (Munos et al., 2016; Precup et al., 2000; Harutyunyan et al., 2016)  $\widehat{Q}(\cdot, \theta) = \lim_{k \rightarrow \infty} \widehat{Q}_k$  where

$$\widehat{Q}_k(x, a; \theta) = \widehat{Q}_{k-1}(x, a; \theta) + \mathbb{E}_{\pi_b} \left[ \sum_{t \geq 0} \gamma^t \prod_{s=1}^t c_s y_t | x_0 = x, a_0 = a \right]$$

and

$$y_t = r^t + \gamma \mathbb{E}_{\pi_e} \widehat{Q}_{k-1}(x_{t+1}, \cdot; \theta) - \widehat{Q}_{k-1}(x_t, a_t; \theta)$$

$$c_s = \begin{cases} \lambda \min(1, \frac{\pi_e(a_s | x_s)}{\pi_b(a_s | x_s)}) & R(\lambda) \\ \lambda \pi_e(a_s | x_s) & Tree \\ \lambda & Q^\pi(\lambda) \end{cases}$$

More Robust Doubly-Robust (MRDR): (Farajtabar et al., 2018) Given

$$\Omega_{\pi_b}(x) = \text{diag}[1/\pi_b(a|x)]_{a \in A} - ee^T$$

$$e = [1, \dots, 1]^T$$

$$R_{t:\bar{T}}^i = \sum_{j=t}^{\bar{T}} \gamma^{j-t} \rho_{(t+1):j}^i r(x_j^i, a_j^i)$$

and

$$q_\theta(x, a, r) = \text{diag}[\pi_e(a'|x)]_{a' \in A} [\widehat{Q}(x, a'; \theta)]_{a' \in A} - r[\mathbf{1}\{a' = a\}]_{a' \in A}$$

where  $\mathbf{1}$  is the indicator function, then

$$\widehat{Q}(\cdot, \theta) = \min_{\theta} \frac{1}{N} \sum_{i=1}^N \sum_{t=0}^{\bar{T}} \gamma^{2t} (\rho_{0:\bar{T}}^i)^2 \times \rho_{t:\bar{T}}^i q_\theta(x_t^i, a_t^i, R_{t:\bar{T}}^i)^T \Omega_{\pi_b}(x_t^i) q_\theta(x_t^i, a_t^i, R_{t:\bar{T}}^i)$$

State Density Ratio Estimation (IH): (Liu et al., 2018)

$$V_{IH} = \sum_{i=1}^N \sum_{t=0}^{\bar{T}} \frac{\gamma^t \omega(s_t^i) \rho_{t:t}^i}{\sum_{i'=0}^N \sum_{t'=1}^{\bar{T}} \gamma^{t'} \omega(s_{t'}^{i'}) \rho_{t':t'}^{i'}}$$

$$\omega(s_t^i) = \lim_{t \rightarrow \infty} \frac{\sum_{t=0}^T \gamma^t d_{\pi_e}(s_t^i)}{\sum_{t=0}^T \gamma^t d_{\pi_b}(s_t^i)}$$

where  $\pi_b$  is assumed to be a fixed data-generating policy, and  $d_\pi$  is the distribution of states when executing  $\pi$  from  $s_0 \sim d_0$ . The details for how to find  $\omega$  can be found in Algorithm 1 and 2 of (Liu et al., 2018).

## C Environments

For every environment, we initialize the environment with a fixed horizon length  $T$ . If the agent reaches a goal before  $T$  or if the episode is not over by step  $T$ , it will transition to an environment-dependent absorbing state where it will stay until time  $T$ . For a high level description of the environment features, see Table 2.

### C.1 Environment Descriptions

#### C.1.1 Graph

Figure 6 shows a visualization of the Toy-Graph environment. The graph is initialized with horizon  $T$  and with absorbing state  $x_{abs} = 2T$ . In each episode, the agent starts at a single starting state  $x_0 = 0$  and has two actions,  $a = 0$  and  $a = 1$ . At each time step  $t < T$ , the agent can enter state  $x_{t+1} = 2t + 1$  by taking action  $a = 0$ , or  $x_{t+1} = 2t + 2$  by taking action  $a = 1$ . If the environment is stochastic, we simulate noisy transitions by allowing the agent to slip into  $x_{t+1} = 2t + 2$  instead of  $x_{t+1} = 2t + 1$  and vice-versa with probability .25. At the final time  $t = T$ , the agent always enters the terminal state  $x_{abs}$ . The reward is +1 if the agent transitions to an odd state, otherwise is -1. If the environment provides sparse rewards, then  $r = +1$  if  $x_{T-1}$  is odd,  $r = -1$  if  $x_{T-1}$  is even, otherwise  $r = 0$ . Similarly to deterministic rewards, if the environment's rewards are stochastic, then the reward is  $r \sim N(1, 1)$  if the agent transitions to an odd state, otherwise  $r \sim N(-1, 1)$ . If the rewards are sparse and stochastic then  $r \sim N(1, 1)$  if  $x_{T-1}$  is odd, otherwise  $r \sim N(-1, 1)$  and  $r = 0$  otherwise.

#### C.1.2 Graph-POMDP

Figure 10 shows a visualization of the Graph-POMDP environment. The underlying state structure of Graph-POMDP is exactly the Graph environment. However, the states are grouped together based on a choice of Graph-POMDP horizon length,  $H$ . This parameter groups states into  $H$  observable states. The agent only is able to observe among these states, and not the underlying MDP structure. Model-Fail (Thomas & Brunskill, 2016) is a special case of this environment when  $H = T = 2$ .

#### C.1.3 Graph Mountain Car (Graph-MC)

Figure 7 shows a visualization of the Toy-MC environment. This environment is a 1-D graph-based simplification of Mountain Car. The agent starts at  $x_0 = 0$ , the center of the valley and can go left or right. There are 21 total states, 10 to the left of the starting position and 11 to the right of the starting position, and a terminal absorbing state  $x_{abs} = 22$ . The agent receives a reward of  $r = -1$  at every timestep. The reward becomes zero if the agent reaches the goal, which is state  $x = +11$ . If the agent reaches  $x = -10$  and continues left then the agent remains in  $x = -10$ . If the agent does not reach state  $x = +11$  by step  $T$  then the episode terminates and the agent transitions to the absorbing state.

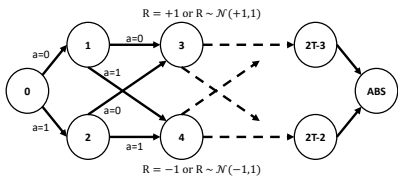


Figure 6: Graph Environment

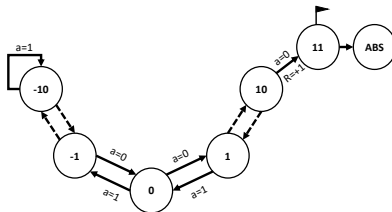


Figure 7: Graph-MC Environment

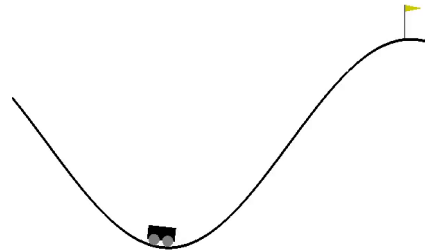


Figure 8: MC Environment, pixel-version. The non-pixel version involves representing the state of the car as the position and velocity.

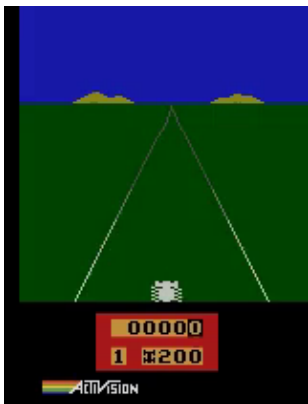


Figure 9: Enduro Environment

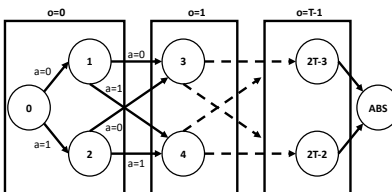


Figure 10: Graph-POMDP Environment. Model-Fail (Thomas & Brunskill, 2016) is a special case of this environment where  $T=2$ . We also extend the environment to arbitrary horizon which makes it a semi-mdp.

S	S	S	S	S	S	S	S
S		F		H			S
S				H			F
S	F				H		F
S				H			F
S	H	H		F		H	
S	H			H		H	
S				H		F	G

Figure 11: Gridworld environment. Blank spaces indicate areas of a small negative reward, S indicates the starting states, F indicates a field of slightly less negative reward, H indicates a hole of severe penalty, G indicates the goal of positive reward.

### C.1.4 Mountain Car (MC)

We use the OpenAI version of Mountain Car with a few simplifying modifications (Brockman et al., 2016; Sutton & Barto, 2018). The car starts in a valley and has to go back and forth to gain enough momentum to scale the mountain and reach the end goal. The state space is given by the position and velocity of the car. At each time step, the car has the following options: accelerate backwards, forwards or do nothing. The reward is  $r = -1$  for every time step until the car reaches the goal. While the original trajectory length is capped at 200, we decrease the effective length by applying every action  $a_t$  five times before observing  $x_{t+1}$ . Furthermore, we modify the random initial position from being uniformly between  $[-.6, -.4]$  to being one of  $\{-.6, -.5, -.4\}$ , with no velocity. The environment is initialized with a horizon  $T$  and absorbing state  $x_{abs} = [.5, 0]$ , position at  $.5$  and no velocity.

### C.1.5 Pixel-based Mountain Car (Pix-MC)

This environment is identical to Mountain Car except the state space has been modified from position and velocity to a pixel based representation of a ball, representing a car, rolling on a hill, see Figure 8. Each frame  $f_t$  is a  $80 \times 120$  image of the ball on the mountain. One cannot deduce velocity from a single frame, so we represent the state as  $x_t = \{f_{t-1}, f_t\}$  where  $f_{-1} = f_0$ , the initial state. Everything else is identical between the pixel-based version and the position-velocity version described earlier.

### C.1.6 Enduro

We use OpenAI’s implementation of Enduro-v0, an Atari 2600 racing game. We downsample the image to a grayscale of size  $(84,84)$ . We apply every action one time and we represent the state as  $x_t = \{f_{t-3}, f_{t-2}, f_{t-1}, f_t\}$  where  $f_i = f_0$ , the initial state, for  $i < 0$ . See Figure 9 for a visualization.

### C.1.7 Gridworld (GW)

Figure 11 shows a visualization of the Gridworld environment. The agent starts at a state in the first row or column (denoted S in the figure), and proceeds through the grid by taking actions, given by the four cardinal directions, for  $T = 25$  timesteps. An agent remains in the same state if it chooses an action which would take it out of the environment. If the agent reaches the goal state  $G$ , in the bottom right corner of the environment, it transitions to a terminal state  $x = 64$  for the remainder of the trajectory and receives a reward of +1. In the grid, there is a field (denoted F) which gives the agent a reward of  $-.005$  and holes (denoted H) which give  $-.5$ . The remaining states give a reward of  $-.01$ .

### C.1.8 Pixel-Gridworld (Pixel-GW)

This environment is identical to Gridworld except the state space has been modified from position to a pixel based representation of the position: 1 for the agent’s location, 0 otherwise. We use the same policies as in the Gridworld case.

## D Experimental Setup

### D.1 Description of the policies

Graph, Graph-POMDP and Graph-MC use static policies with some probability of going left and another probability of going right, ex:  $\pi(a = 0) = p, \pi(a = 1) = 1 - p$ , independent of state. We vary  $p$  in our experiments.

GW, Pix-GW, MC, Pixel-MC, and Enduro all use an  $\epsilon$ -Greedy policy. In other words, we train a policy  $Q^*$  (using value iteration or DDQN) and then vary the deviation away from the policy. Hence  $\epsilon$ -Greedy( $Q^*$ ) implies we follow a mixed policy  $\pi = \arg \max_a Q^*(x, a)$  with probability  $1 - \epsilon$  and uniform with probability  $\epsilon$ . We vary  $\epsilon$  in our experiments.

### D.2 Enumeration of Experiments

#### D.2.1 Graph

See Table 6 for a description of the parameters of the experiment we ran in the Graph Environment. The experiments are the Cartesian product of the table.

#### D.2.2 Graph-POMDP

See Table 7 for a description of the parameters of the experiment we ran in the Graph-POMDP Environment. The experiments are the Cartesian product of the table.

Table 6: Graph parameters

Parameters	
$\gamma$	.98
N	$2^{3:11}$
T	{4, 16}
$\pi_b(a = 0)$	{.2, .6}
$\pi_e(a = 0)$	.8
Stochastic Env	{True, False}
Stochastic Rew	{True, False}
Sparse Rew	{True, False}
Seed	{10 of random(0 : $2^{16}$ )}
ModelType	Tabular
Regress $\pi_b$	False

Table 7: Graph-POMDP parameters

Parameters	
$\gamma$	.98
N	$2^{8:11}$
(T,H)	{(2, 2), (16, 6)}
$\pi_b(a = 0)$	{.2, .6}
$\pi_e(a = 0)$	.8
Stochastic Env	{True, False}
Stochastic Rew	{True, False}
Sparse Rew	{True, False}
Seed	{10 of random(0 : $2^{16}$ )}
ModelType	Tabular
Regress $\pi_b$	False

#### D.2.3 Gridworld

See Table 8 for a description of the parameters of the experiment we ran in the Gridworld Environment. The experiments are the Cartesian product of the table.

Table 8: Gridworld parameters

Parameters	
$\gamma$	.98
N	$2^{6:11}$
T	25
$\epsilon$ -Greedy, $\pi_b$	{.2, .4, .6, .8, 1.}
$\epsilon$ -Greedy, $\pi_e$	.1
Stochastic Env	False
Stochastic Rew	False
Sparse Rew	False
Seed	{10 of random(0 : $2^{16}$ )}
ModelType	Tabular
Regress $\pi_b$	True

#### D.2.4 Pixel-Gridworld (Pix-GW)

See Table 9 for a description of the parameters of the experiment we ran in the Pix-GW Environment. The experiments are the Cartesian product of the table.

#### D.2.5 Graph-MC

See Table 10 for a description of the parameters of the experiment we ran in the TMC Environment. The

Table 9: Pix-GW parameters

Parameters	
$\gamma$	.96
N	$2^{6:9}$
T	25
$\epsilon - \text{Greedy}, \pi_b$	{.2, .4, .6, .8, 1.}
$\epsilon - \text{Greedy}, \pi_e$	.1
Stochastic Env	{True, False}
Stochastic Rew	False
Sparse Rew	False
Seed	{10 of random( $0 : 2^{16}$ )}
ModelType	NN
Regress $\pi_b$	{True, False}

experiments are the Cartesian product of the table.

Table 10: Graph-MC parameters

Parameters	
$\gamma$	.99
N	$2^{7:11}$
T	250
$(\pi_b(a=0), \pi_e(a=0))$	{(.45, .45), (.6, .6), (.45, .6), (.6, .45), (.8, .2), (.2, .8)}
Stochastic Env	False
Stochastic Rew	False
Sparse Rew	False
Seed	{10 of random( $0 : 2^{16}$ )}
ModelType	Tabular
Regress $\pi_b$	False

### D.2.6 Mountain Car (MC)

See Table 11 for a description of the parameters of the experiment we ran in the MC Environment. The experiments are the Cartesian product of the table.

Table 11: MC parameters

Parameters	
$\gamma$	.99
N	$2^{7:10}$
T	250
$\epsilon - \text{Greedy}, (\pi_b, \pi_e)$	{(.1, 0), (1, 0), (1, .1), (.1, 1)}
Stochastic Env	False
Stochastic Rew	False
Sparse Rew	False
Seed	{10 of random( $0 : 2^{16}$ )}
ModelType	{Tabular, NN}
Regress $\pi_b$	False

### D.2.7 Pixel-Mountain Car (Pix-MC)

See Table 12 for a description of the parameters of the experiment we ran in the Pix-MC Environment. The experiments are the Cartesian product of the table.

Table 12: Pix-MC parameters

Parameters	
$\gamma$	.97
N	512
T	500
$\epsilon - \text{Greedy}, (\pi_b, \pi_e)$	{(.25, 0), (.1, 0), (.25, .1)}
Stochastic Env	False
Stochastic Rew	False
Sparse Rew	False
Seed	{10 of random( $0 : 2^{16}$ )}
ModelType	{Tabular, NN}
Regress $\pi_b$	False

### D.2.8 Enduro

See Table 13 for a description of the parameters of the experiment we ran in the Enduro Environment. The experiments are the Cartesian product of the table.

Table 13: Enduro parameters

Parameters	
$\gamma$	.9999
N	512
T	500
$\epsilon - \text{Greedy}, (\pi_b, \pi_e)$	{(.25, 0), (.1, 0), (.25, .1)}
Stochastic Env	False
Stochastic Rew	False
Sparse Rew	False
Seed	{10 of random( $0 : 2^{16}$ )}
ModelType	{Tabular, NN}
Regress $\pi_b$	False

## D.3 Representation and Function Class

For the simpler environments, we use a tabular representation for all the methods. AM amounts to solving for the transition dynamics, rewards, terminal state, etc. through maximum likelihood. FQE, Retrace( $\lambda$ ),  $Q^\pi(\lambda)$ , and Tree-Backup are all implemented through dynamics programming with Q tables. MRDR and Q-Reg used the Sherman Morrison (Sherman & Morrison, 1950) method to solve the weighted-least square problem, using a basis which spans a table.

In the cases where we needed function approximation, we did not directly fit the dynamics for AM; instead, we fit on the difference in states  $T(x' - x|x, a)$ , which is common practice.

For the MC environment, we ran experiments with both a linear and NN function class. In both cases, the representation of the state was not changed and remained [position, velocity]. The NN architecture was dense with [16,8,4,2] as the layers. The layers had relu activations (except the last, with a linear activation) and were all initialized with truncated normal centered



at 0 with a standard deviation of 0.1.

For the pixel-based environments (MC, Enduro), we use a convolutional NN. The architecture is a layer of size 8 with filter (7,7) and stride 3, followed by max-pooling and a layer of size 16 with filter (3,3) and stride 1, followed by max pooling, flattening and a dense layer of size 256. The final layer is a dense layer with the size of the action space, with a linear activation. The layers had elu activations and were all initialized with truncated normal centered at 0 with a standard deviation of 0.1. The layers also have kernel L2 regularizers with weight 1e-6.

When using NNs for the IH method, we used the radial-basis function and a shallow dense network for the kernel and density estimate respectively.

#### D.4 Choice of hyperparameters

Many methods require selection of convergence criteria, regularization parameters, batch sizes, and a whole host of other hyperparameters. Often there is a trade-off between computational cost and the accuracy of the method. See Table 14 for a list of hyperparameters that were chosen for the experiments.

Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning

Table 14: Hyperparameters for each model by Environment

Method	Parameter	Graph	TMC	MC	Pix-MC	Enduro	Graph-POMDP	GW	Pix-GW
AM	Max Traj Len	T	T	50	50	-	T	T	T
	NN Fit Epochs	-	-	100	100	-	-	-	100
	NN Batchsize	-	-	32	32	-	-	-	25
	NN Train size	-	-	.8	.8	-	-	-	.8
	NN Val size	-	-	.2	.2	-	-	-	.2
NN Early Stop delta	-	-	1e-4	1e-4	-	-	-	1e-4	
Q-Reg	Omega regul.	1	1	-	-	-	1	1	-
	NN Fit Epochs	-	-	80	80	80	-	-	80
	NN Batchsize	-	-	32	32	32	-	-	32
	NN Train size	-	-	.8	.8	.8	-	-	.8
	NN Val size	-	-	.2	.2	.2	-	-	.2
NN Early Stop delta	-	-	1e-4	1e-4	1e-4	-	-	1e-4	
FQE	Convergence $\epsilon$	1e-5	1e-5	1e-4	1e-4	1e-4	1e-5	4e-4	1e-4
	Max Iter	-	-	160	160	600	-	50	80
	NN Batchsize	-	-	32	32	32	-	-	32
	Optimizer Clipnorm	-	-	1.	1.	1.	-	-	1.
IH	Quad. prog. regular.	1e-3	1e-3	-	-	-	1e-3	1e-3	-
	NN Fit Epochs	-	-	10001	10001	10001	-	-	1001
	NN Batchsize	-	-	1024	128	128	-	-	128
MRDR	Omega regul.	1	1	-	-	-	1	1	-
	NN Fit Epochs	-	-	80	80	80	-	-	80
	NN Batchsize	-	-	1024	1024	1024	-	-	32
	NN Train size	-	-	.8	.8	.8	-	-	.8
	NN Val size	-	-	.2	.2	.2	-	-	.2
NN Early Stop delta	-	-	1e-4	1e-4	1e-4	-	-	1e-4	
$R(\lambda)$	$\lambda$	.9	.9	.9	-	-	.9	.9	.9
	Convergence $\epsilon$	1e-3	2e-3	1e-3	-	-	1e-3	2e-3	1e-3
	Max Iter	500	500	-	-	-	500	50	-
	NN Fit Epochs	-	-	80	-	-	-	-	80
	NN Batchsize	-	-	4	-	-	-	-	4
	NN Train Size	-	-	.03	-	-	-	-	.03
NN ClipNorm	-	-	1.	-	-	-	-	1.	
$Q^\pi(\lambda)$	$\lambda$	.9	.9	.9	-	-	.9	.9	.9
	Convergence $\epsilon$	1e-3	2e-3	1e-3	-	-	1e-3	2e-3	1e-3
	Max Iter	500	500	-	-	-	500	50	-
	NN Fit Epochs	-	-	80	-	-	-	-	80
	NN Batchsize	-	-	4	-	-	-	-	4
	NN Train Size	-	-	.03	-	-	-	-	.03
NN ClipNorm	-	-	1.	-	-	-	-	1.	
Tree	$\lambda$	.9	.9	.9	-	-	.9	.9	.9
	Convergence $\epsilon$	1e-3	2e-3	1e-3	-	-	1e-3	2e-3	1e-3
	Max Iter	500	500	-	-	-	500	50	-
	NN Fit Epochs	-	-	80	-	-	-	-	80
	NN Batchsize	-	-	4	-	-	-	-	4
	NN Train Size	-	-	.03	-	-	-	-	.03
NN ClipNorm	-	-	1.	-	-	-	-	1.	

E Additional Supporting Figures

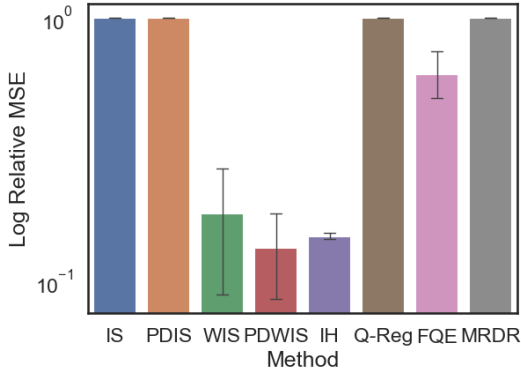


Figure 12: Enduro DM vs IPS.  $\pi_b$  is a policy that deviates uniformly from a trained policy 25% of the time,  $\pi_e$  is a policy trained with DDQN. *IH* has relatively low error mainly due to tracking the simple average, since the kernel function did not learn useful density ratio. The computational time required to calculate the multi-step rollouts of AM, *Retrace*( $\lambda$ ),  $Q^\pi(\lambda)$ , *Tree-Backup*( $\lambda$ ) exceeded our compute budget and were thus excluded.

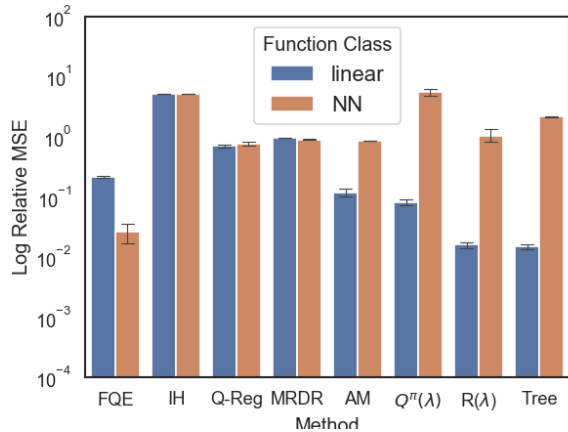


Figure 13: MC comparison.  $N = 256$ .  $\pi_b$  is a uniform random policy,  $\pi_e$  is a policy trained with DDQN

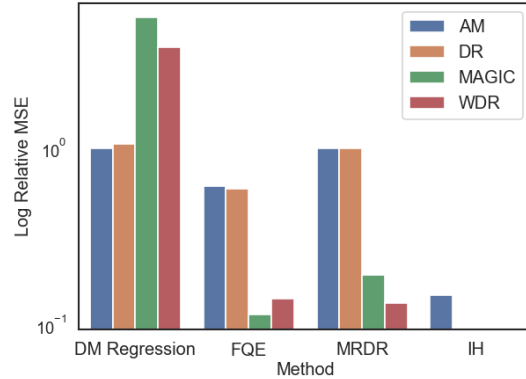


Figure 14: Enduro DM vs HM.  $\pi_b$  is a policy that deviates uniformly from a trained policy 25% of the time,  $\pi_e$  is a policy trained with DDQN.

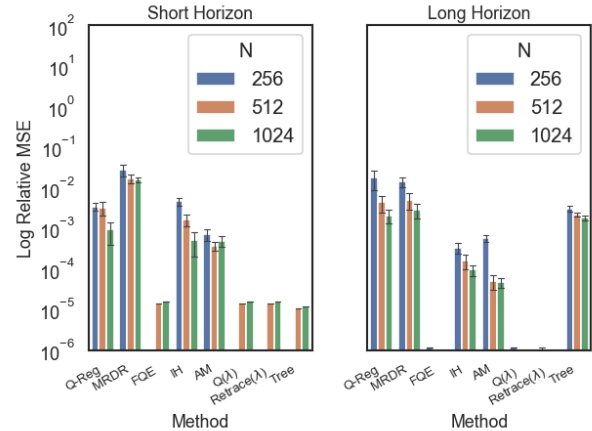


Figure 15: Comparison of Direct methods' performance across horizon and number of trajectories in the Toy-Graph environment. Small policy mismatch under a deterministic environment.

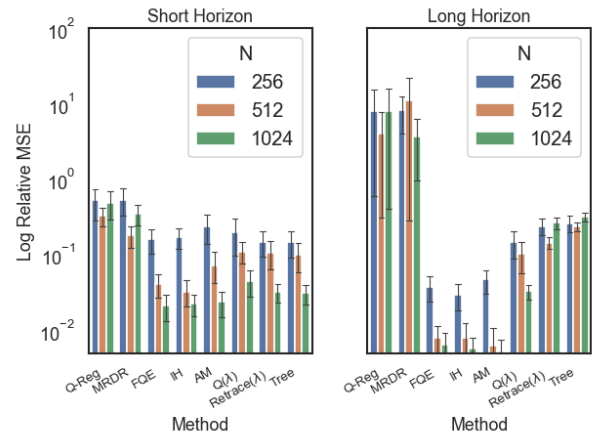


Figure 16: (Graph domain) Comparing DMs across horizon length and number of trajectories. Large policy mismatch and a stochastic environment setting.

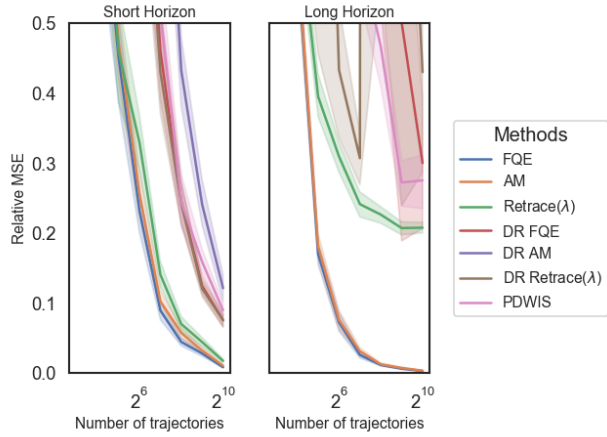


Figure 17: Comparing DM to DR in a stochastic environment with large policy mismatch. (Graph)

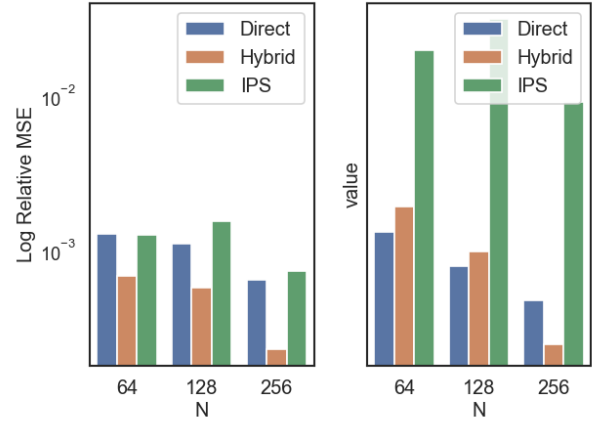


Figure 20: Exact vs Estimated  $\pi_b$ . Exact  $\pi_b = .2$ -Greedy(optimal),  $\pi_e = .1$ -Greedy(optimal). Min error per class. (Pixel Gridworld, deterministic)

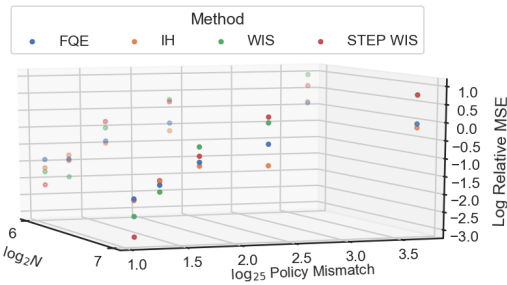


Figure 18: Comparison between FQE, IH and WIS in a low data regime. For low policy mismatch, IPS is competitive to DM in low data, but as the policy mismatch grows, the top DM outperform. Experiments ran in the Gridworld Environment.

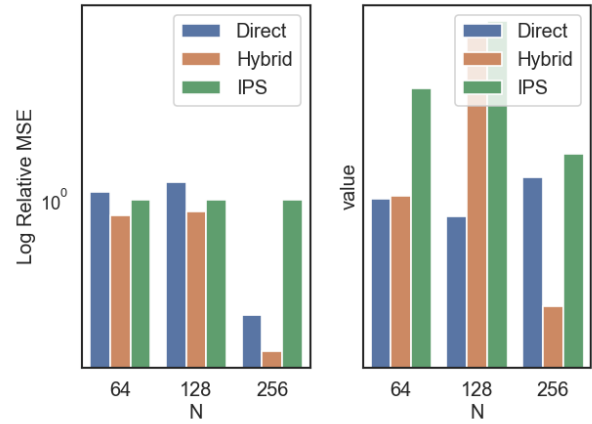


Figure 21: Exact vs Estimated  $\pi_b$ . Exact  $\pi_b = \text{uniform}$ ,  $\pi_e = .1$ -Greedy(optimal). Min error per class. (Pixel Gridworld, deterministic)

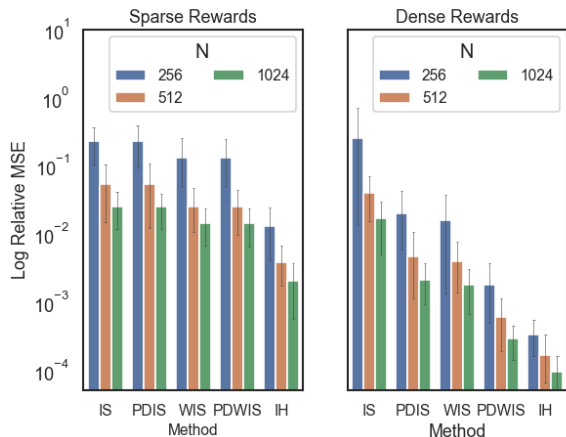


Figure 19: Comparison between IPS methods and IH with dense vs sparse rewards. Per-Decision IPS methods see substantial improvement when the rewards are dense. Experiments ran in the Toy-Graph environment with  $\pi(a = 0) = .6, \pi_e(a = 0) = .8$  See Tables 212, 213, 214, 116, 117, 118

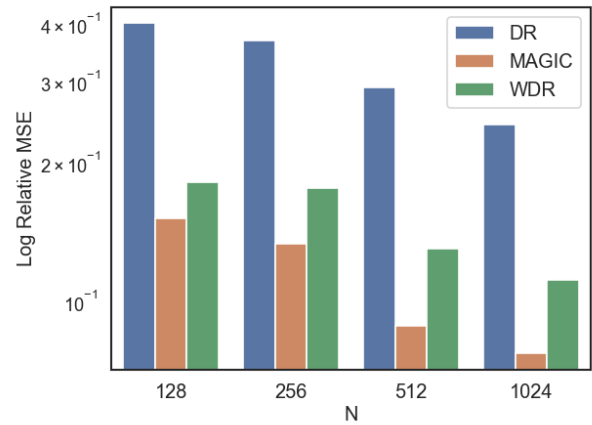


Figure 22: Hybrid Method comparison.  $\pi_b(a = 0) = .2, \pi_e(a = 0) = .8$ . Min error per class. (Graph-MC)

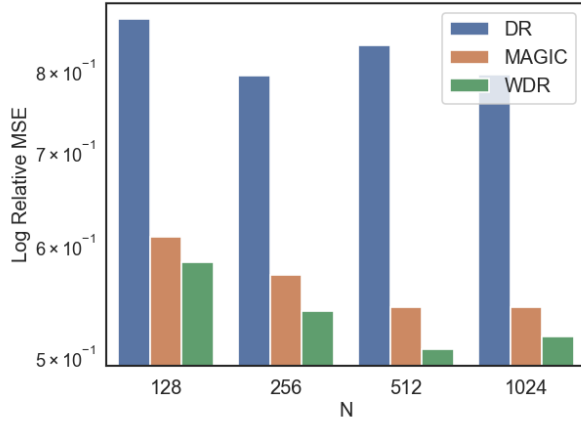


Figure 23: Hybrid Method comparison.  $\pi_b(a=0) = .8, \pi_e(a=0) = .2$ . Min error per class. (Graph-MC)

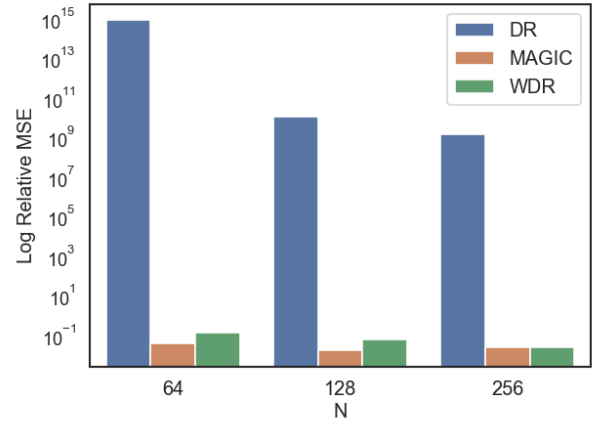


Figure 26: Hybrid Method comparison.  $\pi_b = .8$ -Greedy(optimal),  $\pi_e = .1$ -Greedy(optimal). Min error per class. (Pixel Gridworld)

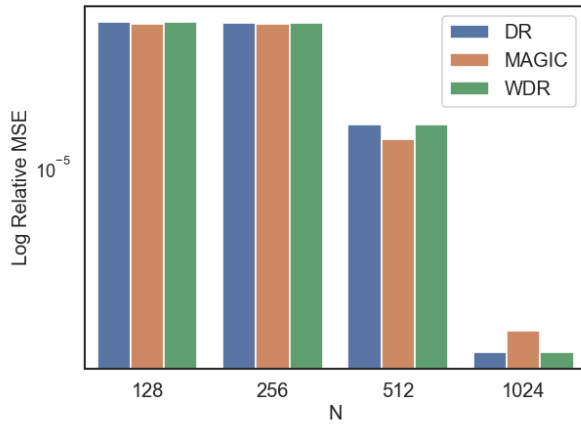


Figure 24: Hybrid Method comparison.  $\pi_b(a=0) = .6, \pi_e(a=0) = .6$ . Min error per class. (Graph-MC)

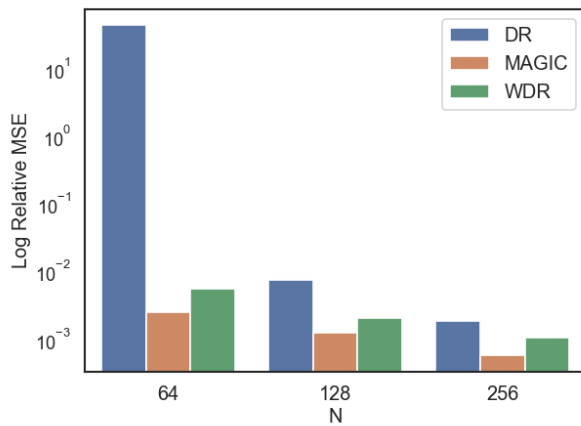


Figure 25: Hybrid Method comparison. Exact  $\pi_b = .2$ -Greedy(optimal),  $\pi_e = .1$ -Greedy(optimal). Min error per class. (Pixel Gridworld)

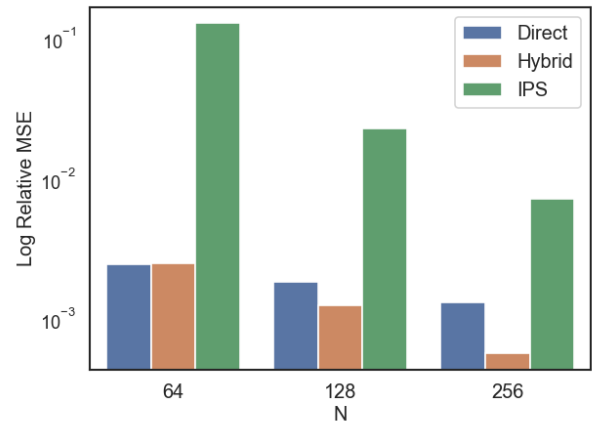


Figure 27: Class comparison with unknown  $\pi_b$ . At first, HM underperform DM because  $\pi_b$  is more difficult to calculate leading to imprecise importance sampling estimates. Exact  $\pi_b = .2$ -Greedy(optimal),  $\pi_e = .1$ -Greedy(optimal). Min error per class. (Pixel Gridworld, stochastic env with .2 slippage)

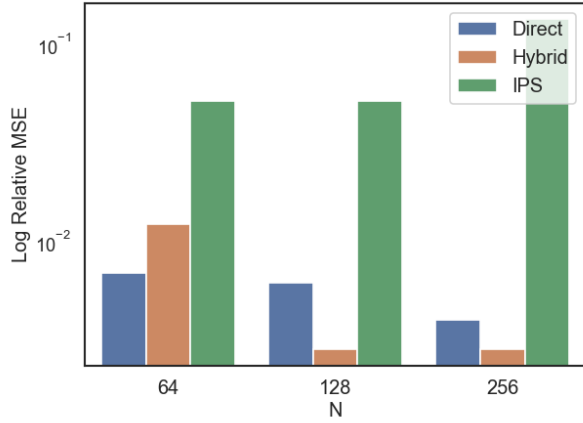


Figure 28: Class comparison with unknown  $\pi_b$ . At first, HM underperform DM because  $\pi_b$  is more difficult to calculate leading to imprecise importance sampling estimates. Exact  $\pi_b = .6$ -Greedy(optimal),  $\pi_e = .1$ -Greedy(optimal). Min error per class. (Pixel Gridworld, stochastic env with .2 slippage)

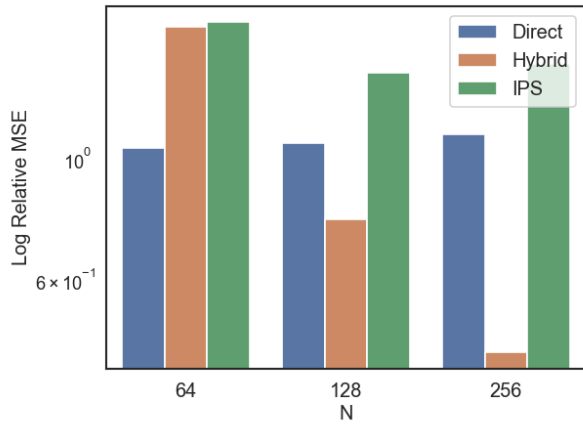


Figure 29: Class comparison with unknown  $\pi_b$ . At first, HM underperform DM because  $\pi_b$  is more difficult to calculate leading to imprecise importance sampling estimates. Exact  $\pi_b = \text{uniform}$ ,  $\pi_e = .1$ -Greedy(optimal). Min error per class. (Pixel Gridworld, stochastic env with .2 slippage)

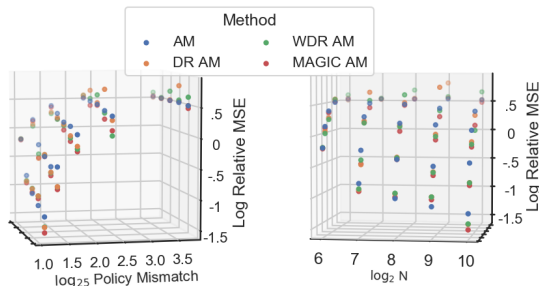


Figure 30: AM Direct vs Hybrid comparison for AM. (Gridworld)

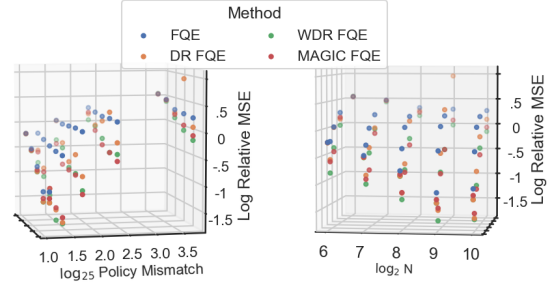


Figure 31: FQE Direct vs Hybrid comparison. (Gridworld)

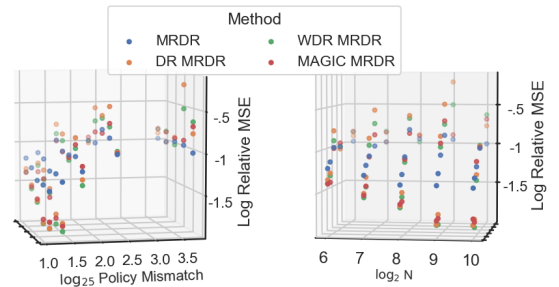


Figure 32: MRDR Direct vs Hybrid comparison. (Gridworld)

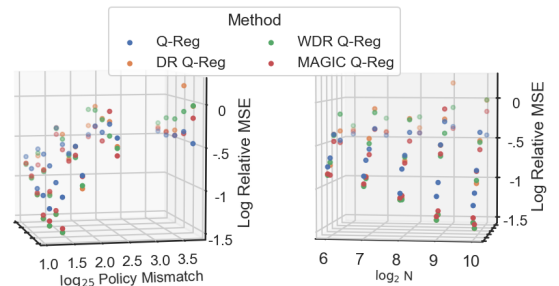


Figure 33: Q-Reg Direct vs Hybrid comparison. (Gridworld)

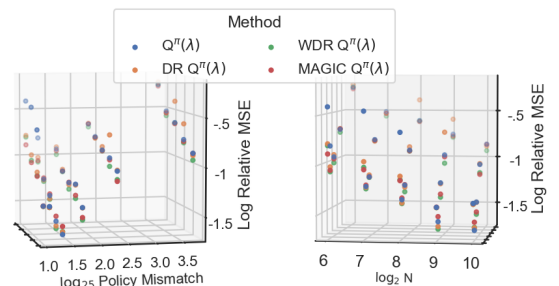


Figure 34:  $Q^\pi(\lambda)$  Direct vs Hybrid comparison. (Gridworld)

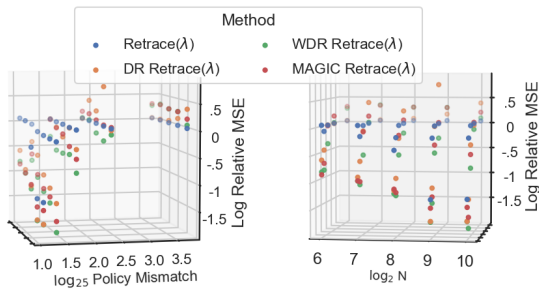


Figure 35: Retrace( $\lambda$ ) Direct vs Hybrid comparison. (Gridworld)

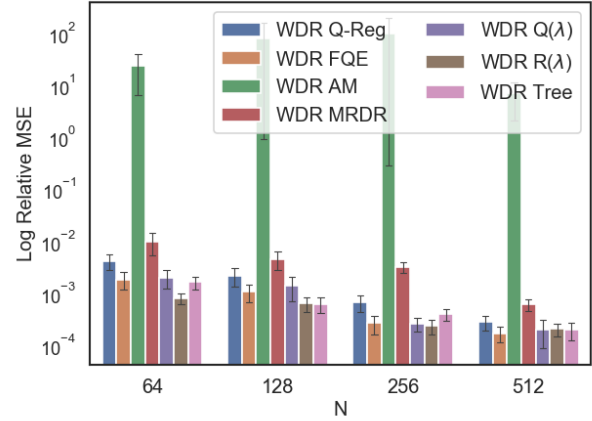


Figure 38: WDR comparison with  $\pi_b = .2$ -Greedy(optimal),  $\pi_e = 1.$ -Greedy(optimal). (Pixel Gridworld)

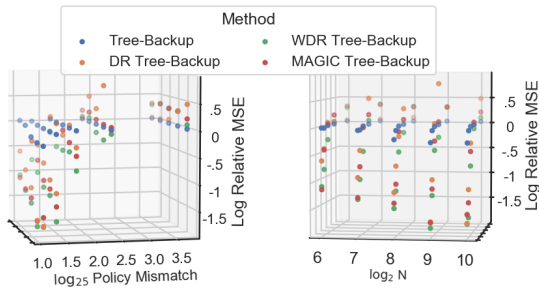


Figure 36: Tree-Backup Direct vs Hybrid comparison. (Gridworld)

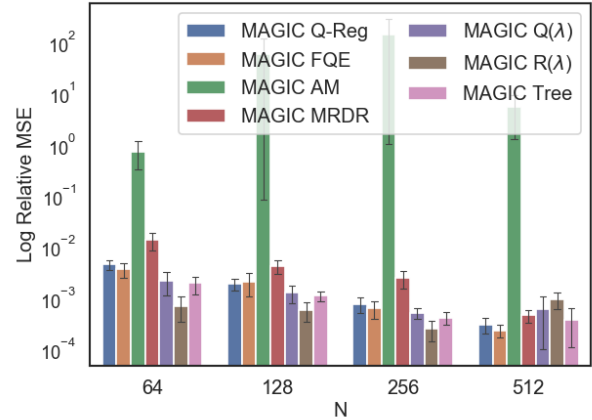


Figure 39: MAGIC comparison with  $\pi_b = .2$ -Greedy(optimal),  $\pi_e = 1.$ -Greedy(optimal). (Pixel Gridworld)

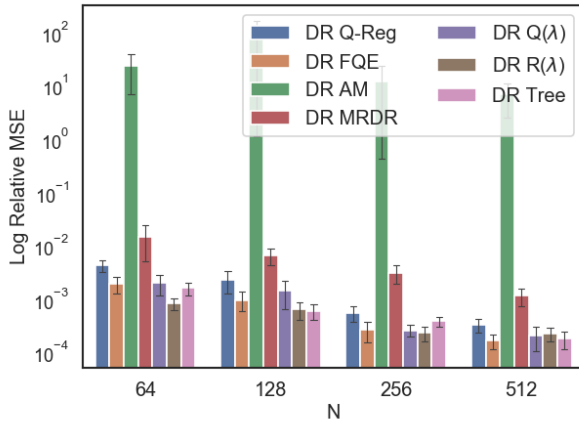


Figure 37: DR comparison with  $\pi_b = .2$ -Greedy(optimal),  $\pi_e = 1.$ -Greedy(optimal). (Pixel Gridworld)

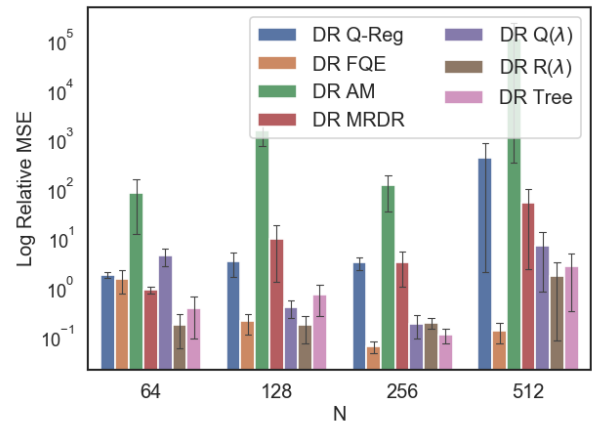


Figure 40: DR comparison with  $\pi_b = .8$ -Greedy(optimal),  $\pi_e = 1.$ -Greedy(optimal). (Pixel Gridworld)

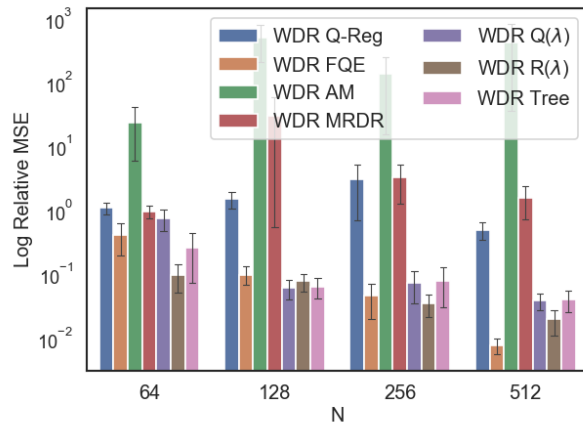


Figure 41: WDR comparison with  $\pi_b = .8$ -Greedy(optimal),  $\pi_e = 1.$ -Greedy(optimal). (Pixel Gridworld)

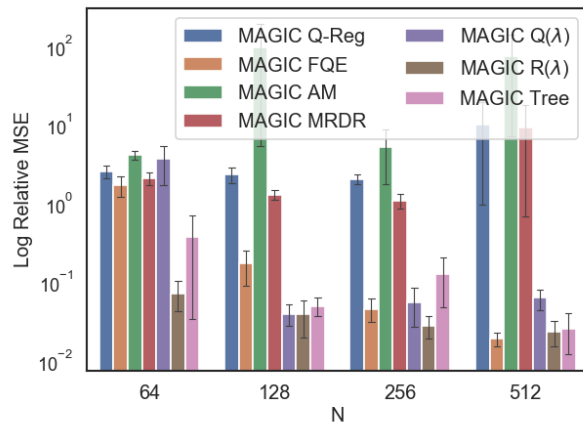


Figure 42: MAGIC comparison with  $\pi_b = .8$ -Greedy(optimal),  $\pi_e = 1.$ -Greedy(optimal). (Pixel Gridworld)



## F Tables of Results, per Environment

## F.1 Detailed Results for Graph

 Table 15: Graph, relative MSE.  $T = 4, N = 8, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.9E-1	5.3E-1	6.4E-1	4.9E-1
Q-REG	2.0E0	5.1E-1	5.7E-1	1.9E0
MRDR	1.7E0	1.7E0	7.0E-1	9.0E-1
FQE	4.8E-1	4.8E-1	4.8E-1	4.8E-1
R( $\lambda$ )	4.8E-1	4.8E-1	4.8E-1	4.8E-1
Q $^\pi$ ( $\lambda$ )	4.8E-1	4.8E-1	4.8E-1	4.8E-1
TREE	4.8E-1	4.8E-1	4.8E-1	<b>4.8E-1</b>
IH	<b>2.9E-1</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	2.4E0	3.3E0
WIS	1.2E0	<b>7.5E-1</b>
NAIVE	3.6E0	-

 Table 16: Graph, relative MSE.  $T = 4, N = 16, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.3E-1	6.7E-1	6.8E-1	4.9E-1
Q-REG	4.3E0	7.0E0	5.9E-1	9.2E-1
MRDR	3.4E0	1.4E1	7.1E-1	2.9E0
FQE	3.9E-1	3.9E-1	3.9E-1	3.9E-1
R( $\lambda$ )	3.9E-1	3.9E-1	3.9E-1	3.9E-1
Q $^\pi$ ( $\lambda$ )	3.9E-1	3.9E-1	3.9E-1	3.9E-1
TREE	<b>3.9E-1</b>	<b>3.9E-1</b>	4.0E-1	4.0E-1
IH	4.8E-1	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	6.7E1	7.9E0
WIS	1.2E0	<b>7.1E-1</b>
NAIVE	3.9E0	-

 Table 17: Graph, relative MSE.  $T = 4, N = 32, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.7E-1	2.8E-1	2.7E-1	1.8E-1
Q-REG	5.0E-1	2.4E-1	3.7E-1	3.6E-1
MRDR	7.6E-1	3.1E-1	5.0E-1	3.1E-1
FQE	1.5E-1	1.5E-1	1.5E-1	1.5E-1
R( $\lambda$ )	1.5E-1	1.5E-1	1.5E-1	1.5E-1
Q $^\pi$ ( $\lambda$ )	1.5E-1	1.5E-1	1.5E-1	1.5E-1
TREE	1.5E-1	1.5E-1	1.5E-1	<b>1.5E-1</b>
IH	<b>3.9E-2</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	5.6E-1	3.8E-1
WIS	5.0E-1	<b>1.9E-1</b>
NAIVE	3.8E0	-

 Table 18: Graph, relative MSE.  $T = 4, N = 64, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.8E-2	1.1E-1	1.2E-1	3.6E-2
Q-REG	4.6E-1	1.2E-1	4.5E-2	2.4E-1
MRDR	3.4E-1	3.2E-1	1.2E-1	3.2E-1
FQE	3.4E-2	3.4E-2	3.4E-2	3.4E-2
R( $\lambda$ )	3.4E-2	3.4E-2	<b>3.4E-2</b>	3.4E-2
Q $^\pi$ ( $\lambda$ )	3.4E-2	3.4E-2	3.4E-2	3.4E-2
TREE	<b>3.4E-2</b>	3.4E-2	3.4E-2	3.4E-2
IH	4.6E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	3.9E0	5.4E-1
WIS	6.8E-1	<b>2.0E-1</b>
NAIVE	4.0E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 19: Graph, relative MSE.  $T = 4, N = 128, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	2.5E-3	6.1E-2	5.2E-2	5.8E-3
Q-REG	4.3E-1	9.7E-2	9.9E-3	1.4E-1
MRDR	3.9E-1	2.5E-1	6.9E-2	1.3E-1
FQE	1.2E-5	1.2E-5	1.2E-5	1.2E-5
R( $\lambda$ )	<b>1.2E-5</b>	1.2E-5	<b>9.0E-6</b>	1.2E-5
Q $^\pi$ ( $\lambda$ )	1.2E-5	1.2E-5	1.2E-5	1.2E-5
TREE	1.4E-5	1.2E-5	1.1E-5	1.4E-5
IH	2.5E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	4.7E0	5.9E-1
WIS	2.2E-1	<b>4.5E-2</b>
NAIVE	3.9E0	-

Table 21: Graph, relative MSE.  $T = 4, N = 512, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	7.8E-4	2.5E-2	2.9E-2	7.5E-3
Q-REG	4.1E-2	1.3E-3	4.6E-4	2.6E-2
MRDR	4.6E-2	2.0E-2	2.4E-2	3.3E-2
FQE	7.0E-6	7.0E-6	7.0E-6	7.0E-6
R( $\lambda$ )	7.0E-6	7.0E-6	8.0E-6	7.0E-6
Q $^\pi$ ( $\lambda$ )	7.0E-6	7.0E-6	7.0E-6	7.0E-6
TREE	<b>5.0E-6</b>	7.0E-6	9.0E-6	<b>5.0E-6</b>
IH	2.2E-3	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	5.0E-1	5.5E-2
WIS	1.5E-1	<b>1.5E-2</b>
NAIVE	4.0E0	-

Table 20: Graph, relative MSE.  $T = 4, N = 256, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	1.3E-3	2.6E-2	2.2E-2	7.4E-3
Q-REG	6.9E-2	7.8E-3	1.4E-3	5.2E-2
MRDR	8.6E-2	6.9E-2	1.1E-1	2.7E-2
FQE	<b>1.4E-8</b>	1.4E-8	1.4E-8	<b>1.4E-8</b>
R( $\lambda$ )	2.5E-8	2.5E-8	6.7E-7	3.3E-8
Q $^\pi$ ( $\lambda$ )	1.4E-8	1.4E-8	1.4E-8	1.5E-8
TREE	2.6E-8	4.9E-8	2.4E-6	1.7E-8
IH	7.5E-3	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	5.5E-1	9.7E-2
WIS	2.4E-1	<b>5.4E-2</b>
NAIVE	4.1E0	-

Table 22: Graph, relative MSE.  $T = 4, N = 1024, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	4.7E-4	6.5E-3	6.2E-3	4.8E-3
Q-REG	4.8E-2	6.2E-4	3.7E-4	3.6E-3
MRDR	2.9E-2	4.6E-3	2.9E-2	3.0E-2
FQE	4.4E-5	4.4E-5	4.4E-5	4.4E-5
R( $\lambda$ )	4.4E-5	4.4E-5	4.3E-5	4.4E-5
Q $^\pi$ ( $\lambda$ )	4.4E-5	4.4E-5	4.4E-5	4.4E-5
TREE	<b>4.0E-5</b>	4.4E-5	4.3E-5	<b>4.1E-5</b>
IH	1.8E-3	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	3.6E-1	3.7E-2
WIS	5.2E-2	<b>1.2E-2</b>
NAIVE	4.0E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 23: Graph, relative MSE.  $T = 4, N = 8, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	6.3E-1	7.9E-1	6.5E-1	6.1E-1
Q-REG	7.4E-1	1.0E0	1.8E0	7.7E-1
MRDR	6.4E-1	1.0E0	8.6E-1	6.8E-1
FQE	5.6E-1	5.8E-1	5.7E-1	5.6E-1
R( $\lambda$ )	<b>5.5E-1</b>	6.0E-1	5.7E-1	5.5E-1
Q $^\pi$ ( $\lambda$ )	5.5E-1	8.8E-1	<b>5.4E-1</b>	5.5E-1
TREE	5.5E-1	5.9E-1	5.7E-1	5.5E-1
IH	1.3E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.1E0	<b>8.0E-1</b>
WIS	1.5E0	1.2E0
NAIVE	3.5E0	-

Table 25: Graph, relative MSE.  $T = 4, N = 32, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.5E-1	4.7E-1	3.6E-1	2.5E-1
Q-REG	3.3E-1	4.5E-1	3.0E-1	3.3E-1
MRDR	4.0E-1	2.4E-1	2.7E-1	4.6E-1
FQE	2.2E-1	2.7E-1	2.6E-1	<b>2.2E-1</b>
R( $\lambda$ )	2.2E-1	2.7E-1	2.9E-1	2.3E-1
Q $^\pi$ ( $\lambda$ )	2.7E-1	2.8E-1	2.6E-1	2.8E-1
TREE	2.2E-1	2.8E-1	2.9E-1	2.3E-1
IH	<b>1.6E-1</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	7.7E-1	<b>2.7E-1</b>
WIS	9.9E-1	3.9E-1
NAIVE	3.7E0	-

Table 24: Graph, relative MSE.  $T = 4, N = 16, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	6.5E-1	7.8E-1	6.9E-1	7.5E-1
Q-REG	6.6E-1	5.1E-1	5.0E-1	6.8E-1
MRDR	6.5E-1	5.7E-1	<b>4.7E-1</b>	7.8E-1
FQE	<b>6.0E-1</b>	8.0E-1	6.4E-1	6.0E-1
R( $\lambda$ )	6.1E-1	7.3E-1	6.4E-1	6.1E-1
Q $^\pi$ ( $\lambda$ )	6.2E-1	7.1E-1	6.5E-1	6.2E-1
TREE	6.1E-1	7.4E-1	6.5E-1	6.1E-1
IH	6.8E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	7.9E-1	<b>6.7E-1</b>
WIS	1.7E0	8.6E-1
NAIVE	4.2E0	-

Table 26: Graph, relative MSE.  $T = 4, N = 64, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.1E-1	3.1E-1	2.0E-1	2.0E-1
Q-REG	7.7E-1	2.5E-1	2.4E-1	3.1E-1
MRDR	6.1E-1	2.2E-1	<b>1.6E-1</b>	3.0E-1
FQE	2.0E-1	2.3E-1	2.5E-1	2.0E-1
R( $\lambda$ )	2.2E-1	2.1E-1	2.4E-1	2.2E-1
Q $^\pi$ ( $\lambda$ )	2.0E-1	1.7E-1	2.0E-1	2.0E-1
TREE	2.2E-1	2.1E-1	2.4E-1	2.2E-1
IH	<b>1.5E-1</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	9.0E0	5.8E-1
WIS	6.6E-1	<b>1.7E-1</b>
NAIVE	3.8E0	-

Table 27: Graph, relative MSE.  $T = 4, N = 128, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	<b>2.5E-2</b>	5.3E-1	2.7E-1	9.8E-2
Q-REG	8.0E-1	1.4E-1	1.3E-1	6.7E-1
MRDR	3.8E-1	1.4E-1	1.4E-1	5.6E-1
FQE	2.8E-2	3.5E-1	1.4E-1	<b>2.9E-2</b>
R( $\lambda$ )	6.3E-2	2.7E-1	1.3E-1	6.3E-2
Q $^\pi$ ( $\lambda$ )	1.0E-1	3.5E-1	1.5E-1	1.0E-1
TREE	5.6E-2	2.7E-1	1.3E-1	5.6E-2
IH	2.7E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	4.0E0	7.0E-1
WIS	5.7E-1	<b>1.4E-1</b>
NAIVE	3.8E0	-

Table 29: Graph, relative MSE.  $T = 4, N = 512, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	9.2E-3	1.4E-1	1.6E-1	2.4E-2
Q-REG	8.6E-2	1.3E-1	1.1E-1	1.2E-1
MRDR	1.0E-1	1.1E-1	1.6E-1	1.7E-1
FQE	<b>8.3E-3</b>	7.1E-2	6.2E-2	<b>8.3E-3</b>
R( $\lambda$ )	1.2E-2	7.2E-2	6.4E-2	1.2E-2
Q $^\pi$ ( $\lambda$ )	1.3E-2	7.7E-2	6.7E-2	1.3E-2
TREE	1.1E-2	7.1E-2	6.4E-2	1.1E-2
IH	1.5E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	3.1E-1	9.1E-2
WIS	2.1E-1	<b>7.5E-2</b>
NAIVE	4.1E0	-

Table 28: Graph, relative MSE.  $T = 4, N = 256, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.8E-2	1.4E-1	7.0E-2	3.6E-2
Q-REG	3.3E-1	6.8E-2	6.8E-2	2.6E-1
MRDR	2.4E-1	5.3E-2	6.0E-2	2.3E-1
FQE	1.7E-2	2.3E-1	7.1E-2	<b>1.8E-2</b>
R( $\lambda$ )	2.6E-2	1.7E-1	6.8E-2	2.6E-2
Q $^\pi$ ( $\lambda$ )	4.0E-2	2.3E-1	7.6E-2	4.1E-2
TREE	2.4E-2	1.7E-1	6.8E-2	2.4E-2
IH	<b>1.1E-2</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	2.9E0	2.9E-1
WIS	2.9E-1	<b>5.4E-2</b>
NAIVE	3.9E0	-

Table 30: Graph, relative MSE.  $T = 4, N = 1024, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	<b>6.6E-3</b>	1.0E-1	8.2E-2	3.5E-2
Q-REG	2.9E-2	2.5E-2	2.3E-2	2.8E-2
MRDR	1.8E-2	2.0E-2	2.3E-2	2.0E-2
FQE	8.4E-3	2.7E-2	2.3E-2	1.1E-2
R( $\lambda$ )	8.0E-3	2.6E-2	2.3E-2	1.1E-2
Q $^\pi$ ( $\lambda$ )	1.0E-2	2.8E-2	2.4E-2	<b>1.0E-2</b>
TREE	7.7E-3	2.6E-2	2.3E-2	1.1E-2
IH	1.3E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.9E-1	2.7E-2
WIS	8.4E-2	<b>2.3E-2</b>
NAIVE	4.0E0	-

Table 31: Graph, relative MSE.  $T = 4, N = 8, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.2E-1	4.5E-1	1.2E0	6.0E-1
Q-REG	3.9E-1	6.9E-1	2.0E0	4.5E-1
MRDR	4.4E-1	9.3E-1	2.1E0	3.9E-1
FQE	<b>2.7E-1</b>	2.8E-1	<b>2.5E-1</b>	2.7E-1
R( $\lambda$ )	2.8E-1	2.9E-1	2.5E-1	2.8E-1
Q $^\pi$ ( $\lambda$ )	3.9E-1	3.3E-1	3.6E-1	3.9E-1
TREE	2.8E-1	2.8E-1	2.5E-1	2.8E-1
IH	6.2E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.4E0	<b>4.7E-1</b>
WIS	2.7E0	1.4E0
NAIVE	4.0E0	-

 Table 33: Graph, relative MSE.  $T = 4, N = 32, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.2E-1	2.2E0	9.8E-1	5.6E-1
Q-REG	6.5E0	1.4E1	4.2E-1	2.7E0
MRDR	3.8E0	2.2E1	4.6E-1	3.9E0
FQE	4.9E-1	5.8E-1	4.1E-1	4.9E-1
R( $\lambda$ )	<b>3.8E-1</b>	4.1E-1	3.7E-1	4.0E-1
Q $^\pi$ ( $\lambda$ )	5.3E-1	8.5E-1	4.7E-1	5.1E-1
TREE	3.9E-1	4.2E-1	<b>3.7E-1</b>	4.0E-1
IH	5.2E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	7.4E1	1.1E1
WIS	1.6E0	<b>5.7E-1</b>
NAIVE	3.5E0	-

 Table 32: Graph, relative MSE.  $T = 4, N = 16, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	8.9E-1	8.7E0	7.6E-1	8.1E-1
Q-REG	3.2E1	1.2E1	<b>5.8E-1</b>	3.2E1
MRDR	1.8E1	4.5E1	1.1E0	1.8E1
FQE	8.1E-1	1.3E0	7.4E-1	8.1E-1
R( $\lambda$ )	<b>7.2E-1</b>	1.8E0	8.3E-1	6.9E-1
Q $^\pi$ ( $\lambda$ )	1.6E0	1.9E0	1.0E0	1.6E0
TREE	7.4E-1	1.8E0	8.3E-1	7.2E-1
IH	1.1E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.8E2	3.0E1
WIS	2.3E0	<b>6.8E-1</b>
NAIVE	3.9E0	-

 Table 34: Graph, relative MSE.  $T = 4, N = 64, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.0E-1	2.0E0	6.1E-1	2.9E-1
Q-REG	1.0E0	6.4E-1	5.6E-1	1.0E0
MRDR	8.2E-1	7.7E-1	7.7E-1	7.8E-1
FQE	2.0E-1	9.7E-1	4.2E-1	<b>2.0E-1</b>
R( $\lambda$ )	2.5E-1	1.0E0	4.7E-1	2.4E-1
Q $^\pi$ ( $\lambda$ )	4.4E-1	8.4E-1	4.5E-1	4.4E-1
TREE	2.4E-1	1.0E0	4.7E-1	2.4E-1
IH	<b>1.4E-1</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.3E0	9.3E-1
WIS	2.0E0	<b>7.8E-1</b>
NAIVE	4.2E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 35: Graph, relative MSE.  $T = 4, N = 128, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.1E-1	8.8E-1	5.4E-1	1.0E-1
Q-REG	1.4E0	6.8E-1	3.5E-1	9.4E-1
MRDR	6.5E-1	4.3E-1	2.3E-1	1.7E0
FQE	8.8E-2	9.1E-1	4.9E-1	8.8E-2
R( $\lambda$ )	9.8E-2	8.1E-1	4.7E-1	1.0E-1
Q $^\pi$ ( $\lambda$ )	<b>7.2E-2</b>	9.6E-1	5.2E-1	<b>5.3E-2</b>
TREE	9.9E-2	8.1E-1	4.7E-1	1.1E-1
IH	1.8E-1	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	5.3E0	1.1E0
WIS	8.2E-1	<b>5.1E-1</b>
NAIVE	4.0E0	-

Table 37: Graph, relative MSE.  $T = 4, N = 512, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.4E-2	1.3E-1	1.2E-1	4.3E-2
Q-REG	5.5E-2	1.2E-1	1.1E-1	4.3E-2
MRDR	4.3E-2	7.5E-2	1.1E-1	1.1E-1
FQE	<b>1.3E-2</b>	8.7E-2	9.8E-2	<b>1.3E-2</b>
R( $\lambda$ )	2.2E-2	9.2E-2	9.9E-2	2.5E-2
Q $^\pi$ ( $\lambda$ )	1.8E-2	8.9E-2	1.0E-1	1.6E-2
TREE	2.2E-2	9.2E-2	9.8E-2	2.4E-2
IH	1.5E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	3.2E-1	6.9E-2
WIS	1.4E-1	<b>6.7E-2</b>
NAIVE	4.0E0	-

Table 36: Graph, relative MSE.  $T = 4, N = 256, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	7.5E-2	2.8E-1	2.7E-1	1.2E-1
Q-REG	2.8E-1	1.8E-1	2.0E-1	8.1E-2
MRDR	2.6E-1	1.0E-1	1.4E-1	3.0E-1
FQE	6.6E-2	1.9E-1	2.2E-1	<b>6.8E-2</b>
R( $\lambda$ )	1.2E-1	2.0E-1	2.1E-1	1.1E-1
Q $^\pi$ ( $\lambda$ )	1.1E-1	1.8E-1	2.1E-1	1.1E-1
TREE	1.1E-1	2.0E-1	2.1E-1	1.1E-1
IH	<b>5.7E-2</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	7.2E-1	2.3E-1
WIS	5.8E-1	<b>2.1E-1</b>
NAIVE	4.3E0	-

Table 38: Graph, relative MSE.  $T = 4, N = 1024, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.1E-2	8.9E-2	9.7E-2	2.4E-2
Q-REG	6.7E-2	6.6E-2	6.1E-2	3.4E-2
MRDR	5.4E-2	7.3E-2	1.0E-1	7.3E-2
FQE	<b>9.2E-3</b>	6.2E-2	5.8E-2	<b>1.0E-2</b>
R( $\lambda$ )	1.7E-2	6.4E-2	5.9E-2	1.5E-2
Q $^\pi$ ( $\lambda$ )	3.0E-2	6.4E-2	5.8E-2	1.8E-2
TREE	1.6E-2	6.4E-2	5.9E-2	1.5E-2
IH	1.8E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	2.6E-1	8.4E-2
WIS	2.0E-1	<b>4.3E-2</b>
NAIVE	3.9E0	-

Table 39: Graph, relative MSE.  $T = 4, N = 8, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.0E0	3.4E0	2.8E0	1.5E0
Q-REG	8.4E1	9.4E0	1.6E0	8.4E1
MRDR	4.1E1	4.3E1	2.7E0	4.1E1
FQE	9.0E-1	2.5E0	9.8E-1	9.0E-1
$R(\lambda)$	<b>8.9E-1</b>	1.2E0	<b>8.2E-1</b>	8.9E-1
$Q^\pi(\lambda)$	1.1E0	2.1E0	9.6E-1	1.1E0
TREE	9.0E-1	1.2E0	8.3E-1	9.0E-1
IH	1.7E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.6E1	5.9E1
WIS	3.9E0	<b>2.5E0</b>
NAIVE	6.3E0	-

 Table 41: Graph, relative MSE.  $T = 4, N = 32, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	7.3E-1	8.9E-1	1.2E0	<b>6.8E-1</b>
Q-REG	3.4E0	1.8E0	2.5E0	3.4E0
MRDR	2.1E0	9.4E-1	1.6E0	3.1E0
FQE	<b>7.1E-1</b>	4.3E0	1.8E0	7.1E-1
$R(\lambda)$	9.7E-1	3.1E0	1.9E0	9.7E-1
$Q^\pi(\lambda)$	2.6E0	5.4E0	2.4E0	2.6E0
TREE	8.8E-1	3.1E0	1.8E0	8.8E-1
IH	1.6E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.5E0</b>	3.4E0
WIS	3.5E0	2.6E0
NAIVE	4.4E0	-

 Table 40: Graph, relative MSE.  $T = 4, N = 16, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.3E0	7.8E0	3.3E0	2.2E0
Q-REG	5.9E0	2.0E0	2.3E0	5.0E0
MRDR	5.3E0	<b>1.5E0</b>	2.4E0	4.6E0
FQE	2.3E0	4.0E0	2.7E0	2.3E0
$R(\lambda)$	<b>2.0E0</b>	2.4E0	2.2E0	2.0E0
$Q^\pi(\lambda)$	2.9E0	2.6E0	1.9E0	2.9E0
TREE	2.0E0	2.5E0	2.3E0	2.0E0
IH	2.3E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.1E1	1.0E1
WIS	5.8E0	<b>4.9E0</b>
NAIVE	5.8E0	-

 Table 42: Graph, relative MSE.  $T = 4, N = 64, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.9E-1	2.6E0	1.9E0	4.8E-1
Q-REG	3.2E0	3.2E0	1.6E0	3.2E0
MRDR	2.9E0	4.3E0	2.5E0	2.0E0
FQE	4.7E-1	8.3E-1	7.4E-1	4.8E-1
$R(\lambda)$	5.3E-1	8.8E-1	7.8E-1	5.4E-1
$Q^\pi(\lambda)$	<b>3.1E-1</b>	8.9E-1	6.2E-1	<b>3.4E-1</b>
TREE	5.2E-1	8.6E-1	7.8E-1	5.3E-1
IH	3.8E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	5.8E1	4.0E0
WIS	4.6E0	<b>1.4E0</b>
NAIVE	3.5E0	-



**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 43: Graph, relative MSE.  $T = 4, N = 128, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.3E-1	3.0E0	1.9E0	1.2E0
Q-REG	2.5E0	1.1E0	7.3E-1	2.3E0
MRDR	2.5E0	1.8E0	9.4E-1	3.1E0
FQE	3.7E-1	1.3E0	9.3E-1	3.8E-1
R( $\lambda$ )	3.8E-1	1.4E0	9.4E-1	3.7E-1
Q $^\pi$ ( $\lambda$ )	4.4E-1	1.2E0	8.4E-1	<b>2.8E-1</b>
TREE	3.7E-1	1.4E0	9.5E-1	3.7E-1
IH	<b>3.5E-1</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.6E1	3.3E0
WIS	1.8E0	<b>1.0E0</b>
NAIVE	3.9E0	-

Table 45: Graph, relative MSE.  $T = 4, N = 512, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	7.3E-2	5.1E-1	4.0E-1	1.3E-1
Q-REG	3.3E-1	1.8E-1	1.7E-1	1.8E-1
MRDR	1.8E-1	7.8E-2	6.6E-2	2.2E-1
FQE	4.1E-2	2.0E-1	1.9E-1	<b>4.3E-2</b>
R( $\lambda$ )	1.1E-1	2.1E-1	2.0E-1	9.9E-2
Q $^\pi$ ( $\lambda$ )	1.1E-1	2.0E-1	1.9E-1	8.1E-2
TREE	1.0E-1	2.1E-1	2.0E-1	9.5E-2
IH	<b>3.3E-2</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	3.7E-1	2.4E-1
WIS	<b>1.5E-1</b>	2.0E-1
NAIVE	3.9E0	-

Table 44: Graph, relative MSE.  $T = 4, N = 256, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.4E-1	1.2E0	9.8E-1	1.9E-1
Q-REG	5.3E-1	2.6E-1	2.6E-1	3.5E-1
MRDR	5.4E-1	3.7E-1	3.3E-1	3.4E-1
FQE	1.6E-1	7.0E-1	4.6E-1	<b>1.5E-1</b>
R( $\lambda$ )	<b>1.5E-1</b>	5.0E-1	3.6E-1	1.6E-1
Q $^\pi$ ( $\lambda$ )	2.0E-1	7.5E-1	4.8E-1	2.2E-1
TREE	1.5E-1	5.0E-1	3.6E-1	1.6E-1
IH	1.7E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.5E0	6.6E-1
WIS	5.2E-1	<b>3.4E-1</b>
NAIVE	4.6E0	-

Table 46: Graph, relative MSE.  $T = 4, N = 1024, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.4E-2	3.0E-1	3.4E-1	8.5E-2
Q-REG	4.9E-1	2.4E-1	2.6E-1	2.3E-1
MRDR	3.5E-1	3.4E-1	2.9E-1	4.1E-1
FQE	<b>2.1E-2</b>	2.1E-1	2.2E-1	<b>2.3E-2</b>
R( $\lambda$ )	3.2E-2	2.0E-1	2.2E-1	3.0E-2
Q $^\pi$ ( $\lambda$ )	4.5E-2	2.1E-1	2.3E-1	3.7E-2
TREE	3.1E-2	2.0E-1	2.2E-1	3.0E-2
IH	2.3E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.5E0	4.5E-1
WIS	9.2E-1	<b>3.2E-1</b>
NAIVE	3.9E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 47: Graph, relative MSE.  $T = 4, N = 8, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.2E0	<b>1.2E0</b>	1.6E0	1.2E0
Q-REG	2.5E0	2.5E0	2.3E0	2.5E0
MRDR	2.3E0	4.2E0	1.7E0	2.1E0
FQE	1.3E0	1.3E0	1.3E0	1.3E0
$R(\lambda)$	1.3E0	1.3E0	1.3E0	1.3E0
$Q^\pi(\lambda)$	1.3E0	1.3E0	1.3E0	1.3E0
TREE	1.2E0	1.3E0	1.3E0	1.3E0
IH	<b>1.1E0</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>3.8E0</b>	3.8E0
WIS	3.9E0	3.9E0
NAIVE	3.9E0	-

Table 49: Graph, relative MSE.  $T = 4, N = 32, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.8E-1	1.1E0	7.2E-1	4.8E-1
Q-REG	1.5E1	2.2E1	2.2E0	1.5E1
MRDR	1.4E1	4.2E1	3.0E0	1.5E1
FQE	4.3E-1	4.3E-1	4.3E-1	4.3E-1
$R(\lambda)$	4.2E-1	4.3E-1	4.3E-1	4.3E-1
$Q^\pi(\lambda)$	4.3E-1	4.3E-1	4.3E-1	4.3E-1
TREE	4.2E-1	4.2E-1	4.3E-1	<b>4.2E-1</b>
IH	<b>4.0E-1</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	3.4E1	3.4E1
WIS	<b>1.1E0</b>	1.1E0
NAIVE	3.5E0	-

Table 48: Graph, relative MSE.  $T = 4, N = 16, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	7.4E-1	1.3E0	8.8E-1	7.5E-1
Q-REG	2.5E0	1.7E0	1.6E0	1.0E0
MRDR	2.7E0	1.8E0	<b>5.7E-1</b>	1.3E0
FQE	6.8E-1	6.8E-1	6.8E-1	6.8E-1
$R(\lambda)$	6.8E-1	6.8E-1	7.0E-1	6.8E-1
$Q^\pi(\lambda)$	6.8E-1	6.8E-1	6.8E-1	6.8E-1
TREE	<b>6.8E-1</b>	6.8E-1	7.2E-1	6.8E-1
IH	7.4E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	3.9E0	3.9E0
WIS	<b>2.2E0</b>	2.2E0
NAIVE	3.8E0	-

Table 50: Graph, relative MSE.  $T = 4, N = 64, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	<b>2.7E-1</b>	4.8E-1	6.0E-1	2.9E-1
Q-REG	7.2E-1	6.4E-1	4.3E-1	4.8E-1
MRDR	9.3E-1	5.6E-1	5.9E-1	5.1E-1
FQE	2.7E-1	2.7E-1	2.7E-1	2.7E-1
$R(\lambda)$	2.7E-1	2.7E-1	2.7E-1	2.7E-1
$Q^\pi(\lambda)$	2.7E-1	2.7E-1	2.7E-1	2.7E-1
TREE	2.7E-1	2.7E-1	2.7E-1	<b>2.7E-1</b>
IH	4.1E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>6.7E-1</b>	6.7E-1
WIS	1.1E0	1.1E0
NAIVE	3.9E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 51: Graph, relative MSE.  $T = 4, N = 128, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.9E-3	1.4E-1	1.6E-1	3.3E-2
Q-REG	5.8E0	7.9E-1	1.1E-1	6.7E0
MRDR	4.2E0	2.2E0	7.5E-1	4.1E0
FQE	9.0E-6	9.0E-6	9.0E-6	9.0E-6
R( $\lambda$ )	<b>7.0E-6</b>	7.0E-6	1.0E-5	7.0E-6
Q $^\pi$ ( $\lambda$ )	9.0E-6	9.0E-6	9.0E-6	9.0E-6
TREE	1.5E-5	<b>7.0E-6</b>	2.3E-5	1.4E-5
IH	2.4E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	4.8E0	4.8E0
WIS	1.0E-1	<b>1.0E-1</b>
NAIVE	4.0E0	-

Table 53: Graph, relative MSE.  $T = 4, N = 512, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.1E-3	5.1E-2	4.0E-2	2.0E-2
Q-REG	4.8E-1	3.9E-2	1.2E-2	1.9E-1
MRDR	3.7E-1	2.7E-1	2.7E-1	3.6E-1
FQE	9.3E-7	9.3E-7	9.3E-7	9.3E-7
R( $\lambda$ )	8.3E-7	8.4E-7	3.0E-6	8.8E-7
Q $^\pi$ ( $\lambda$ )	9.7E-7	9.3E-7	9.3E-7	9.5E-7
TREE	<b>1.8E-7</b>	9.3E-7	8.5E-6	<b>1.5E-7</b>
IH	5.6E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	4.9E-1	4.9E-1
WIS	<b>2.8E-1</b>	2.8E-1
NAIVE	3.9E0	-

Table 52: Graph, relative MSE.  $T = 4, N = 256, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.6E-3	4.3E-2	6.0E-2	4.0E-2
Q-REG	8.3E-1	2.5E0	1.8E0	6.8E0
MRDR	7.9E-1	5.9E-1	6.3E-1	8.0E-1
FQE	3.0E-5	3.0E-5	3.0E-5	3.0E-5
R( $\lambda$ )	<b>3.0E-5</b>	3.0E-5	2.1E-5	3.0E-5
Q $^\pi$ ( $\lambda$ )	3.0E-5	3.0E-5	3.0E-5	3.0E-5
TREE	4.5E-5	3.2E-5	<b>2.1E-5</b>	4.3E-5
IH	1.0E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	7.8E-1	7.8E-1
WIS	3.6E-1	<b>3.6E-1</b>
NAIVE	4.1E0	-

Table 54: Graph, relative MSE.  $T = 4, N = 1024, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.5E-3	1.1E-2	1.1E-2	6.8E-3
Q-REG	5.1E-1	1.8E-2	4.3E-3	3.6E-2
MRDR	5.3E-1	1.8E-1	4.7E-1	7.5E-1
FQE	1.5E-5	1.5E-5	1.5E-5	1.5E-5
R( $\lambda$ )	1.6E-5	1.6E-5	<b>1.5E-5</b>	1.6E-5
Q $^\pi$ ( $\lambda$ )	<b>1.5E-5</b>	1.5E-5	1.5E-5	1.5E-5
TREE	2.7E-5	1.7E-5	1.6E-5	2.6E-5
IH	2.1E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	4.5E-1	4.5E-1
WIS	<b>1.2E-1</b>	1.2E-1
NAIVE	4.1E0	-

Table 55: Graph, relative MSE.  $T = 4, N = 8, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.5E0	1.1E1	5.6E0	5.2E0
Q-REG	1.2E1	1.1E1	6.4E0	1.2E1
MRDR	1.1E1	6.5E0	3.6E0	1.1E1
FQE	5.7E0	5.0E0	3.6E0	5.7E0
R( $\lambda$ )	5.4E0	5.6E0	<b>3.4E0</b>	5.4E0
Q $^\pi$ ( $\lambda$ )	<b>3.9E0</b>	1.9E1	4.0E0	3.9E0
TREE	5.5E0	5.4E0	3.4E0	5.5E0
IH	1.5E1	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.4E1	4.0E1
WIS	1.3E1	<b>1.0E1</b>
NAIVE	8.6E0	-

Table 57: Graph, relative MSE.  $T = 4, N = 32, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.7E0	5.3E0	6.1E0	2.9E0
Q-REG	3.5E0	3.9E0	3.9E0	3.0E0
MRDR	<b>1.3E0</b>	<b>1.8E0</b>	2.2E0	2.2E0
FQE	2.6E0	2.8E0	4.2E0	2.6E0
R( $\lambda$ )	2.6E0	2.7E0	3.6E0	2.6E0
Q $^\pi$ ( $\lambda$ )	3.8E0	2.6E0	2.8E0	3.7E0
TREE	2.6E0	2.7E0	3.6E0	2.6E0
IH	2.3E0	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	2.0E0	<b>1.9E0</b>
WIS	4.7E0	4.9E0
NAIVE	4.7E0	-

Table 56: Graph, relative MSE.  $T = 4, N = 16, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	<b>3.4E0</b>	4.2E0	5.2E0	3.5E0
Q-REG	5.8E0	4.8E0	1.2E1	8.6E0
MRDR	6.9E0	3.6E0	5.3E0	5.4E0
FQE	3.4E0	4.5E0	5.2E0	3.4E0
R( $\lambda$ )	3.6E0	3.9E0	4.7E0	3.6E0
Q $^\pi$ ( $\lambda$ )	5.5E0	<b>2.8E0</b>	4.5E0	5.5E0
TREE	3.6E0	4.0E0	4.8E0	3.6E0
IH	7.3E0	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	<b>3.6E0</b>	5.4E0
WIS	7.3E0	7.1E0
NAIVE	4.0E0	-

Table 58: Graph, relative MSE.  $T = 4, N = 64, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.3E0	9.0E0	7.9E0	3.1E0
Q-REG	5.4E1	7.6E0	5.0E0	5.2E1
MRDR	2.8E1	1.3E1	<b>2.3E0</b>	2.8E1
FQE	2.8E0	6.4E0	3.8E0	2.8E0
R( $\lambda$ )	3.5E0	5.1E0	3.8E0	3.5E0
Q $^\pi$ ( $\lambda$ )	4.8E0	6.8E0	3.9E0	4.8E0
TREE	3.4E0	5.1E0	3.8E0	3.4E0
IH	<b>1.8E0</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.3E1	3.2E1
WIS	5.8E0	<b>3.8E0</b>
NAIVE	4.0E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 59: Graph, relative MSE.  $T = 4, N = 128, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	1.5E0	4.2E0	2.5E0	1.5E0	
Q-REG	3.2E1	1.2E1	2.5E0	1.5E1	
MRDR	1.9E1	2.9E1	3.6E0	1.6E1	
FQE	1.5E0	1.9E0	2.4E0	1.5E0	
R( $\lambda$ )	1.4E0	1.9E0	2.3E0	1.5E0	
Q $^\pi$ ( $\lambda$ )	2.1E0	1.8E0	2.4E0	2.1E0	
TREE	<b>1.4E0</b>	2.0E0	2.3E0	<b>1.4E0</b>	
IH	2.2E0	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	3.7E1	3.0E1
WIS	<b>3.1E0</b>	3.9E0
NAIVE	4.1E0	-

Table 61: Graph, relative MSE.  $T = 4, N = 512, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	<b>2.5E-1</b>	2.4E0	2.1E0	3.1E-1	
Q-REG	1.6E0	1.6E0	1.2E0	1.2E0	
MRDR	1.8E0	1.4E0	1.0E0	1.2E0	
FQE	2.8E-1	1.1E0	1.1E0	<b>2.9E-1</b>	
R( $\lambda$ )	4.2E-1	1.2E0	1.1E0	4.3E-1	
Q $^\pi$ ( $\lambda$ )	4.9E-1	1.1E0	1.1E0	4.9E-1	
TREE	4.0E-1	1.2E0	1.1E0	4.0E-1	
IH	3.0E-1	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	5.5E-1	1.5E0
WIS	<b>3.8E-1</b>	9.3E-1
NAIVE	3.9E0	-

Table 60: Graph, relative MSE.  $T = 4, N = 256, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	<b>1.0E0</b>	1.2E1	6.1E0	1.5E0	
Q-REG	9.5E0	7.2E0	5.6E0	1.7E1	
MRDR	6.7E0	4.2E0	4.8E0	9.7E0	
FQE	1.1E0	8.0E0	5.2E0	<b>1.1E0</b>	
R( $\lambda$ )	2.1E0	7.5E0	5.2E0	2.1E0	
Q $^\pi$ ( $\lambda$ )	1.6E0	8.4E0	5.4E0	1.6E0	
TREE	2.0E0	7.4E0	5.1E0	2.0E0	
IH	1.3E0	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	1.6E1	8.6E0
WIS	7.5E0	<b>5.2E0</b>
NAIVE	3.8E0	-

Table 62: Graph, relative MSE.  $T = 4, N = 1024, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	<b>5.1E-2</b>	8.5E-1	7.2E-1	1.7E-1	
Q-REG	5.2E-1	7.1E-1	5.3E-1	4.8E-1	
MRDR	3.9E-1	5.3E-1	4.9E-1	8.3E-1	
FQE	6.8E-2	4.1E-1	4.2E-1	<b>6.8E-2</b>	
R( $\lambda$ )	8.8E-2	4.3E-1	4.4E-1	8.9E-2	
Q $^\pi$ ( $\lambda$ )	7.8E-2	4.1E-1	4.1E-1	7.8E-2	
TREE	8.5E-2	4.3E-1	4.4E-1	8.6E-2	
IH	5.1E-2	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	8.7E-1	5.2E-1
WIS	1.1E0	<b>4.6E-1</b>
NAIVE	3.9E0	-

Table 63: Graph, relative MSE.  $T = 4, N = 8, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	3.8E0	7.7E0	7.0E0	3.7E0	
Q-REG	6.4E2	1.8E3	1.6E1	6.3E2	
MRDR	5.2E2	3.7E3	2.5E1	5.1E2	
FQE	3.5E0	2.9E0	3.3E0	3.4E0	
R( $\lambda$ )	3.4E0	<b>2.8E0</b>	3.1E0	3.3E0	
Q $^\pi$ ( $\lambda$ )	5.1E0	2.9E0	2.8E0	4.5E0	
TREE	3.4E0	2.8E0	3.1E0	3.4E0	
IH	<b>2.3E0</b>	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	1.2E3	1.2E3
WIS	<b>9.6E0</b>	9.6E0
NAIVE	6.0E0	-

Table 65: Graph, relative MSE.  $T = 4, N = 32, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	3.0E0	2.6E1	1.1E1	3.5E0	
Q-REG	9.1E0	2.2E1	1.4E1	7.4E0	
MRDR	1.5E1	2.4E0	<b>2.2E0</b>	2.9E0	
FQE	2.6E0	2.1E1	5.3E0	2.6E0	
R( $\lambda$ )	3.5E0	1.2E1	4.1E0	3.5E0	
Q $^\pi$ ( $\lambda$ )	5.6E0	2.1E1	5.5E0	5.3E0	
TREE	3.3E0	1.2E1	4.2E0	3.2E0	
IH	<b>5.8E-1</b>	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	8.9E0	8.9E0
WIS	<b>8.0E0</b>	8.0E0
NAIVE	3.9E0	-

Table 64: Graph, relative MSE.  $T = 4, N = 16, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	2.3E0	8.7E0	6.2E0	2.3E0	
Q-REG	1.7E0	4.1E0	2.1E1	2.0E1	
MRDR	3.3E0	5.6E0	6.7E1	5.1E1	
FQE	1.8E0	3.9E0	1.4E0	1.8E0	
R( $\lambda$ )	1.5E0	1.4E0	1.5E0	1.5E0	
Q $^\pi$ ( $\lambda$ )	3.6E0	4.2E0	2.6E0	3.3E0	
TREE	1.5E0	<b>1.4E0</b>	1.5E0	1.5E0	
IH	<b>7.9E-1</b>	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	<b>2.5E0</b>	2.5E0
WIS	5.7E0	5.7E0
NAIVE	4.5E0	-

Table 66: Graph, relative MSE.  $T = 4, N = 64, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	3.2E-1	1.6E1	8.5E0	3.2E-1	
Q-REG	2.2E1	4.3E0	2.3E0	2.3E1	
MRDR	1.7E1	9.4E0	4.6E0	1.9E1	
FQE	<b>2.4E-1</b>	5.3E0	2.8E0	<b>2.4E-1</b>	
R( $\lambda$ )	9.6E-1	5.4E0	2.8E0	9.6E-1	
Q $^\pi$ ( $\lambda$ )	7.6E-1	5.5E0	2.4E0	8.8E-1	
TREE	7.7E-1	5.4E0	2.8E0	7.7E-1	
IH	2.6E-1	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	2.0E1	2.0E1
WIS	<b>3.0E0</b>	3.0E0
NAIVE	4.0E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 67: Graph, relative MSE.  $T = 4, N = 128, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.2E-1	2.1E0	1.5E0	6.0E-1
Q-REG	1.7E1	2.8E0	2.9E0	1.5E1
MRDR	1.4E1	1.1E1	9.8E0	2.1E1
FQE	3.6E-1	2.3E0	1.8E0	<b>3.6E-1</b>
R( $\lambda$ )	6.8E-1	2.1E0	1.8E0	6.8E-1
Q $^\pi$ ( $\lambda$ )	4.5E-1	2.5E0	1.9E0	4.8E-1
TREE	6.5E-1	2.1E0	1.8E0	6.5E-1
IH	<b>3.0E-1</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.0E1	2.0E1
WIS	<b>5.1E0</b>	5.1E0
NAIVE	4.8E0	-

Table 69: Graph, relative MSE.  $T = 4, N = 512, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	<b>5.1E-2</b>	1.4E0	1.3E0	1.5E-1
Q-REG	1.4E0	4.7E-1	3.7E-1	9.8E-1
MRDR	1.8E0	5.1E-1	9.4E-1	1.8E0
FQE	6.1E-2	3.2E-1	3.1E-1	<b>6.4E-2</b>
R( $\lambda$ )	9.8E-2	3.3E-1	3.3E-1	1.0E-1
Q $^\pi$ ( $\lambda$ )	2.2E-1	3.3E-1	3.3E-1	1.9E-1
TREE	9.0E-2	3.3E-1	3.3E-1	9.4E-2
IH	1.2E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.2E0	1.2E0
WIS	<b>1.0E0</b>	1.0E0
NAIVE	4.1E0	-

Table 68: Graph, relative MSE.  $T = 4, N = 256, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.6E-1	1.9E0	2.3E0	4.9E-1
Q-REG	3.4E-1	7.5E-1	5.7E-1	2.7E-1
MRDR	4.8E-1	5.3E-1	2.1E0	1.9E0
FQE	<b>1.4E-1</b>	6.5E-1	5.6E-1	<b>1.3E-1</b>
R( $\lambda$ )	2.7E-1	7.1E-1	5.9E-1	2.8E-1
Q $^\pi$ ( $\lambda$ )	2.5E-1	6.6E-1	5.5E-1	2.2E-1
TREE	2.7E-1	7.1E-1	5.9E-1	2.8E-1
IH	2.0E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>3.3E-1</b>	3.3E-1
WIS	3.4E-1	3.4E-1
NAIVE	3.9E0	-

Table 70: Graph, relative MSE.  $T = 4, N = 1024, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.1E-2	2.7E-1	3.1E-1	4.5E-2
Q-REG	2.6E-1	2.4E-1	2.3E-1	1.8E-1
MRDR	1.1E0	3.1E-1	3.0E-1	6.5E-1
FQE	<b>2.5E-2</b>	2.1E-1	2.0E-1	2.5E-2
R( $\lambda$ )	4.0E-2	2.1E-1	2.0E-1	3.9E-2
Q $^\pi$ ( $\lambda$ )	4.9E-2	2.2E-1	2.1E-1	<b>2.3E-2</b>
TREE	4.0E-2	2.1E-1	2.0E-1	3.9E-2
IH	7.6E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>3.0E-1</b>	3.0E-1
WIS	3.4E-1	3.4E-1
NAIVE	3.9E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 71: Graph, relative MSE.  $T = 4, N = 8, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.6E1	8.6E1	5.5E1	1.6E1
Q-REG	3.7E0	6.2E1	3.3E1	3.7E0
MRDR	<b>3.5E0</b>	9.6E1	2.6E1	<b>3.5E0</b>
FQE	1.1E1	2.3E1	1.7E1	1.1E1
R( $\lambda$ )	9.5E0	9.4E0	1.1E1	9.5E0
$Q^\pi(\lambda)$	1.1E1	1.8E1	1.2E1	1.1E1
TREE	9.7E0	9.6E0	1.2E1	9.7E0
IH	2.0E1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	6.6E1	<b>1.1E1</b>
WIS	2.6E1	1.4E1
NAIVE	1.2E1	-

Table 73: Graph, relative MSE.  $T = 4, N = 32, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	<b>8.8E0</b>	3.2E1	1.5E1	<b>8.4E0</b>
Q-REG	4.0E1	5.6E1	1.8E1	4.0E1
MRDR	3.3E1	7.1E1	2.1E1	2.7E1
FQE	9.6E0	2.0E1	1.1E1	9.6E0
R( $\lambda$ )	1.3E1	2.6E1	1.6E1	1.3E1
$Q^\pi(\lambda)$	1.3E1	2.2E1	1.5E1	1.3E1
TREE	1.3E1	2.6E1	1.6E1	1.3E1
IH	1.5E1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.0E1	3.7E1
WIS	1.6E1	<b>1.2E1</b>
NAIVE	3.3E0	-

Table 72: Graph, relative MSE.  $T = 4, N = 16, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.6E1	7.1E1	2.4E1	1.4E1
Q-REG	8.0E2	1.8E3	8.5E1	7.9E2
MRDR	7.2E2	4.1E3	1.2E2	6.9E2
FQE	<b>1.3E1</b>	1.7E2	1.8E1	<b>1.3E1</b>
R( $\lambda$ )	1.4E1	5.2E1	1.5E1	1.3E1
$Q^\pi(\lambda)$	2.4E1	1.9E2	1.8E1	2.4E1
TREE	1.3E1	5.3E1	1.5E1	1.3E1
IH	2.0E1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	7.2E3	2.0E3
WIS	4.2E1	<b>3.2E1</b>
NAIVE	8.7E0	-

Table 74: Graph, relative MSE.  $T = 4, N = 64, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	6.4E0	3.6E1	3.7E1	7.5E0
Q-REG	6.4E1	2.3E1	1.0E1	6.6E1
MRDR	4.3E1	4.3E1	6.8E0	5.4E1
FQE	6.4E0	8.7E0	8.6E0	<b>6.4E0</b>
R( $\lambda$ )	7.1E0	7.4E0	7.0E0	7.1E0
$Q^\pi(\lambda)$	8.0E0	1.2E1	9.8E0	8.1E0
TREE	7.1E0	7.2E0	6.9E0	7.1E0
IH	<b>5.1E0</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	9.3E1	8.0E1
WIS	1.4E1	<b>9.1E0</b>
NAIVE	6.5E0	-



**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 75: Graph, relative MSE.  $T = 4, N = 128, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	2.8E0	5.3E1	2.0E1	<b>2.8E0</b>	
Q-REG	4.4E1	1.2E1	1.4E1	3.9E1	
MRDR	3.5E1	2.1E1	1.6E1	4.6E1	
FQE	2.8E0	4.4E1	1.4E1	2.8E0	
R( $\lambda$ )	5.2E0	2.9E1	1.4E1	4.4E0	
Q $^\pi$ ( $\lambda$ )	6.4E0	4.2E1	1.5E1	5.3E0	
TREE	4.9E0	3.0E1	1.4E1	4.3E0	
IH	<b>2.6E0</b>	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	1.3E2	5.7E1
WIS	3.1E1	<b>1.5E1</b>
NAIVE	5.4E0	-

Table 77: Graph, relative MSE.  $T = 4, N = 512, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	6.4E-1	7.3E0	6.0E0	1.1E0	
Q-REG	4.0E0	1.9E0	1.8E0	2.6E0	
MRDR	2.8E0	1.8E0	2.5E0	2.5E0	
FQE	<b>5.5E-1</b>	2.0E0	1.5E0	<b>5.4E-1</b>	
R( $\lambda$ )	7.0E-1	1.9E0	1.5E0	6.2E-1	
Q $^\pi$ ( $\lambda$ )	1.1E0	1.9E0	1.4E0	7.6E-1	
TREE	6.7E-1	2.0E0	1.5E0	6.1E-1	
IH	8.0E-1	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	1.3E1	4.1E0
WIS	7.4E0	<b>2.7E0</b>
NAIVE	4.0E0	-

Table 76: Graph, relative MSE.  $T = 4, N = 256, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	9.1E-1	2.6E0	4.8E0	1.5E0	
Q-REG	5.4E0	2.3E0	2.4E0	4.4E0	
MRDR	4.2E0	3.4E0	5.5E0	3.8E0	
FQE	1.1E0	2.8E0	3.0E0	1.1E0	
R( $\lambda$ )	<b>8.0E-1</b>	2.3E0	2.9E0	<b>8.1E-1</b>	
Q $^\pi$ ( $\lambda$ )	9.1E-1	2.7E0	2.9E0	1.1E0	
TREE	8.3E-1	2.2E0	2.9E0	8.3E-1	
IH	1.1E0	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	<b>2.2E0</b>	5.0E0
WIS	5.6E0	6.4E0
NAIVE	3.8E0	-

Table 78: Graph, relative MSE.  $T = 4, N = 1024, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	<b>1.0E-1</b>	6.5E0	4.4E0	3.5E-1	
Q-REG	3.6E0	1.4E0	1.5E0	2.1E0	
MRDR	3.0E0	1.8E0	3.0E0	2.4E0	
FQE	1.2E-1	2.2E0	1.6E0	<b>1.2E-1</b>	
R( $\lambda$ )	2.0E-1	2.1E0	1.5E0	1.5E-1	
Q $^\pi$ ( $\lambda$ )	7.9E-1	2.2E0	1.6E0	4.2E-1	
TREE	1.8E-1	2.1E0	1.6E0	1.4E-1	
IH	1.7E-1	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	1.4E1	4.0E0
WIS	1.2E1	<b>2.9E0</b>
NAIVE	4.6E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 79: Graph, relative MSE.  $T = 16, N = 8, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	8.5E-1	8.7E-1	1.0E0	9.1E-1
Q-REG	6.8E-1	9.0E-1	4.8E0	2.2E0
MRDR	7.2E-1	9.8E-1	6.5E0	6.1E0
FQE	8.5E-1	8.5E-1	8.5E-1	8.5E-1
R( $\lambda$ )	8.5E-1	8.4E-1	1.4E0	1.3E0
Q $^\pi$ ( $\lambda$ )	8.5E-1	8.5E-1	8.5E-1	8.5E-1
TREE	8.5E-1	<b>8.3E-1</b>	1.5E0	1.4E0
IH	<b>7.5E-2</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	<b>6.5E-1</b>
WIS	2.6E0	1.8E0
NAIVE	4.2E0	-

Table 81: Graph, relative MSE.  $T = 16, N = 32, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.7E-1	5.2E-1	<b>4.3E-1</b>	5.6E-1
Q-REG	5.9E-1	5.0E-1	1.2E0	8.5E-1
MRDR	5.9E-1	8.3E-1	5.2E0	5.3E0
FQE	5.4E-1	5.4E-1	5.4E-1	5.4E-1
R( $\lambda$ )	6.1E-1	6.0E-1	6.3E-1	7.2E-1
Q $^\pi$ ( $\lambda$ )	5.5E-1	5.4E-1	5.4E-1	5.4E-1
TREE	6.4E-1	6.1E-1	6.5E-1	7.4E-1
IH	<b>1.7E-2</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	<b>6.1E-1</b>
WIS	1.7E0	7.4E-1
NAIVE	4.1E0	-

Table 80: Graph, relative MSE.  $T = 16, N = 16, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	6.6E-1	7.4E-1	1.0E0	7.1E-1
Q-REG	4.4E-1	<b>5.0E-1</b>	9.4E-1	8.7E-1
MRDR	5.3E-1	8.7E-1	2.3E0	2.1E0
FQE	6.5E-1	6.5E-1	6.5E-1	6.5E-1
R( $\lambda$ )	6.6E-1	6.5E-1	9.4E-1	9.4E-1
Q $^\pi$ ( $\lambda$ )	6.5E-1	6.5E-1	6.5E-1	6.5E-1
TREE	6.7E-1	6.5E-1	1.0E0	1.0E0
IH	<b>1.1E-2</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	<b>4.5E-1</b>
WIS	2.1E0	1.2E0
NAIVE	3.9E0	-

Table 82: Graph, relative MSE.  $T = 16, N = 64, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.7E-1	2.5E-1	2.7E-1	<b>1.7E-1</b>
Q-REG	4.9E-1	9.2E0	1.6E1	4.8E-1
MRDR	4.7E-1	7.4E-1	2.2E0	2.2E0
FQE	1.7E-1	1.7E-1	1.7E-1	1.7E-1
R( $\lambda$ )	3.4E-1	3.3E-1	4.3E-1	4.9E-1
Q $^\pi$ ( $\lambda$ )	1.7E-1	1.7E-1	1.7E-1	1.7E-1
TREE	4.0E-1	3.6E-1	4.6E-1	5.4E-1
IH	<b>5.9E-3</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	<b>5.0E-1</b>
WIS	1.5E0	6.7E-1
NAIVE	4.0E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 83: Graph, relative MSE.  $T = 16, N = 128, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.1E-2	1.5E-1	2.5E-1	4.0E-2
Q-REG	4.0E-1	7.1E0	1.7E0	3.8E-1
MRDR	3.5E-1	5.6E-1	6.0E0	6.0E0
FQE	2.0E-2	2.0E-2	2.0E-2	2.0E-2
R( $\lambda$ )	2.2E-1	1.9E-1	1.0E-1	2.3E-1
Q $^\pi$ ( $\lambda$ )	2.0E-2	<b>2.0E-2</b>	2.0E-2	2.0E-2
TREE	3.0E-1	2.6E-1	1.3E-1	3.0E-1
IH	<b>3.2E-3</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	9.0E-1	3.8E-1
WIS	9.9E-1	<b>2.7E-1</b>
NAIVE	4.0E0	-

Table 85: Graph, relative MSE.  $T = 16, N = 512, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.2E-4	5.9E-1	1.8E-1	6.2E-4
Q-REG	3.1E0	3.4E1	1.8E1	3.3E0
MRDR	1.6E0	1.4E2	4.1E1	1.0E1
FQE	3.6E-7	3.6E-7	3.6E-7	3.6E-7
R( $\lambda$ )	1.0E-1	8.3E-1	8.7E-2	1.3E-1
Q $^\pi$ ( $\lambda$ )	<b>3.6E-7</b>	3.6E-7	3.6E-7	<b>3.6E-7</b>
TREE	1.8E-1	2.2E0	1.3E-1	2.1E-1
IH	5.5E-4	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	4.2E0	2.6E0
WIS	8.6E-1	<b>2.4E-1</b>
NAIVE	4.0E0	-

Table 84: Graph, relative MSE.  $T = 16, N = 256, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.3E-4	1.7E1	2.5E-1	5.5E-4
Q-REG	2.5E1	4.0E2	2.3E1	9.9E0
MRDR	1.9E1	1.2E3	1.9E1	2.2E1
FQE	<b>9.9E-8</b>	9.9E-8	9.9E-8	9.9E-8
R( $\lambda$ )	9.7E-2	2.7E0	7.1E-2	1.2E-1
Q $^\pi$ ( $\lambda$ )	1.0E-7	1.1E-7	<b>9.9E-8</b>	1.0E-7
TREE	1.8E-1	1.2E1	9.7E-2	1.9E-1
IH	6.9E-4	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.2E1	2.8E1
WIS	7.0E-1	<b>2.2E-1</b>
NAIVE	4.0E0	-

Table 86: Graph, relative MSE.  $T = 16, N = 1024, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.5E-5	1.2E-1	5.6E-2	3.8E-4
Q-REG	2.8E-1	1.4E0	3.0E-1	2.4E-1
MRDR	3.7E-1	6.0E-1	3.9E0	3.9E0
FQE	<b>1.0E-6</b>	1.0E-6	1.0E-6	1.0E-6
R( $\lambda$ )	9.3E-2	7.5E-2	5.8E-2	9.3E-2
Q $^\pi$ ( $\lambda$ )	1.0E-6	<b>1.0E-6</b>	1.0E-6	1.0E-6
TREE	1.7E-1	1.2E-1	7.6E-2	1.7E-1
IH	8.5E-4	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	9.3E-1	2.7E-1
WIS	6.9E-1	<b>1.5E-1</b>
NAIVE	4.0E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 87: Graph, relative MSE.  $T = 16, N = 8, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	7.8E-1	2.0E0	1.2E0	7.8E-1
Q-REG	9.5E-1	2.5E1	2.2E1	1.1E0
MRDR	1.1E0	6.5E1	2.5E1	2.6E0
FQE	7.6E-1	7.1E-1	9.5E-1	7.6E-1
$R(\lambda)$	7.7E-1	7.6E-1	1.1E0	8.7E-1
$Q^\pi(\lambda)$	7.6E-1	7.2E-1	7.8E-1	7.6E-1
TREE	7.7E-1	<b>7.0E-1</b>	1.2E0	9.0E-1
IH	<b>2.9E-1</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	<b>9.7E-1</b>	4.8E0
WIS	1.9E0	1.4E0
NAIVE	3.7E0	-

Table 89: Graph, relative MSE.  $T = 16, N = 32, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.7E-1	7.8E-1	1.3E0	<b>5.5E-1</b>
Q-REG	6.9E-1	1.2E0	4.2E0	1.6E0
MRDR	1.5E0	2.5E0	8.6E0	8.1E0
FQE	5.8E-1	7.1E-1	5.7E-1	5.8E-1
$R(\lambda)$	6.6E-1	6.7E-1	1.0E0	1.0E0
$Q^\pi(\lambda)$	6.7E-1	1.3E0	5.8E-1	6.7E-1
TREE	6.8E-1	6.8E-1	1.1E0	1.1E0
IH	<b>6.5E-2</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.0E0	<b>6.9E-1</b>
WIS	2.3E0	1.2E0
NAIVE	4.1E0	-

Table 88: Graph, relative MSE.  $T = 16, N = 16, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	6.8E-1	1.5E1	<b>3.2E-1</b>	6.2E-1
Q-REG	2.1E0	2.0E1	2.0E0	2.8E0
MRDR	7.8E-1	3.9E0	4.1E1	4.2E1
FQE	6.9E-1	8.6E-1	6.8E-1	7.0E-1
$R(\lambda)$	7.4E-1	7.7E-1	9.8E-1	1.0E0
$Q^\pi(\lambda)$	7.1E-1	1.8E0	6.2E-1	7.1E-1
TREE	7.6E-1	7.8E-1	1.0E0	9.9E-1
IH	<b>1.2E-1</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	9.9E-1	<b>6.0E-1</b>
WIS	1.5E0	9.9E-1
NAIVE	3.8E0	-

Table 90: Graph, relative MSE.  $T = 16, N = 64, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.6E-1	2.9E-1	3.3E-1	2.6E-1
Q-REG	4.9E-1	3.6E-1	8.1E-1	7.4E-1
MRDR	4.8E-1	9.1E-1	3.3E0	3.4E0
FQE	2.6E-1	2.5E-1	<b>2.2E-1</b>	2.4E-1
$R(\lambda)$	4.4E-1	3.9E-1	2.9E-1	5.2E-1
$Q^\pi(\lambda)$	2.7E-1	2.8E-1	3.0E-1	2.7E-1
TREE	4.7E-1	4.1E-1	3.3E-1	5.8E-1
IH	<b>5.2E-2</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.0E0	5.4E-1
WIS	1.1E0	<b>3.4E-1</b>
NAIVE	3.9E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 91: Graph, relative MSE.  $T = 16, N = 128, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	6.4E-2	6.1E-1	2.9E-1	6.2E-2
Q-REG	4.3E-1	3.1E-1	2.3E0	3.6E-1
MRDR	3.7E-1	5.8E-1	2.7E0	2.7E0
FQE	6.5E-2	8.3E-2	1.5E-1	6.5E-2
R( $\lambda$ )	1.9E-1	1.8E-1	7.1E-2	1.9E-1
Q $^\pi$ ( $\lambda$ )	5.8E-2	<b>5.1E-2</b>	1.3E-1	5.8E-2
TREE	2.6E-1	2.2E-1	8.9E-2	2.6E-1
IH	<b>3.0E-2</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	9.8E-1	4.1E-1
WIS	1.3E0	<b>2.1E-1</b>
NAIVE	4.0E0	-

Table 93: Graph, relative MSE.  $T = 16, N = 512, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.4E-3	3.2E-1	2.1E-1	7.4E-3
Q-REG	4.1E-1	8.0E-1	2.0E-1	4.0E-1
MRDR	2.7E-1	5.0E-1	3.4E0	3.4E0
FQE	<b>5.0E-3</b>	4.3E-2	3.0E-2	<b>4.8E-3</b>
R( $\lambda$ )	1.1E-1	1.7E-1	1.4E-1	1.3E-1
Q $^\pi$ ( $\lambda$ )	1.4E-2	6.2E-2	3.6E-2	1.3E-2
TREE	1.9E-1	2.5E-1	1.8E-1	2.1E-1
IH	6.2E-3	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.0E0	4.1E-1
WIS	1.0E0	<b>3.0E-1</b>
NAIVE	4.0E0	-

Table 92: Graph, relative MSE.  $T = 16, N = 256, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	9.6E-3	7.0E-1	4.2E-1	8.8E-3
Q-REG	3.5E-1	3.2E0	6.6E-1	3.7E-1
MRDR	3.6E-1	6.4E-1	2.1E0	1.8E0
FQE	<b>8.4E-3</b>	9.4E-2	3.4E-2	<b>8.3E-3</b>
R( $\lambda$ )	1.4E-1	1.9E-1	9.0E-2	1.4E-1
Q $^\pi$ ( $\lambda$ )	5.7E-2	1.4E-1	6.5E-2	5.7E-2
TREE	2.3E-1	2.9E-1	1.4E-1	2.4E-1
IH	1.2E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	9.7E-1	4.0E-1
WIS	1.1E0	<b>3.0E-1</b>
NAIVE	3.9E0	-

Table 94: Graph, relative MSE.  $T = 16, N = 1024, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.3E-3	1.5E1	1.7E-1	2.4E-2
Q-REG	1.5E0	9.5E0	1.7E1	1.4E0
MRDR	9.7E-1	5.0E1	3.0E1	8.0E0
FQE	<b>8.6E-4</b>	3.0E0	2.4E-2	<b>9.8E-4</b>
R( $\lambda$ )	1.1E-1	5.8E-1	8.3E-2	9.1E-2
Q $^\pi$ ( $\lambda$ )	7.7E-3	3.1E0	2.1E-2	7.7E-3
TREE	2.0E-1	1.7E0	1.1E-1	2.0E-1
IH	1.5E-3	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.4E2	1.8E0
WIS	8.0E-1	<b>1.8E-1</b>
NAIVE	4.0E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 95: Graph, relative MSE.  $T = 16, N = 8, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	7.5E-1	1.0E0	1.2E0	7.1E-1
Q-REG	7.6E-1	8.3E-1	2.5E0	7.3E-1
MRDR	7.9E-1	1.6E0	3.0E0	1.1E0
FQE	7.3E-1	6.8E-1	6.6E-1	7.3E-1
R( $\lambda$ )	7.2E-1	6.9E-1	1.1E0	8.7E-1
Q $^\pi$ ( $\lambda$ )	6.2E-1	8.5E-1	8.7E-1	<b>6.2E-1</b>
TREE	7.4E-1	6.9E-1	1.3E0	9.6E-1
IH	<b>2.3E-1</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	<b>1.0E0</b>	1.1E0
WIS	2.7E0	1.9E0
NAIVE	4.2E0	-

Table 97: Graph, relative MSE.  $T = 16, N = 32, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.3E-1	2.3E0	1.3E0	3.3E-1
Q-REG	7.6E-1	5.0E-1	5.9E0	8.5E-1
MRDR	9.9E-1	7.8E-1	2.2E1	1.2E0
FQE	1.9E-1	2.7E-1	2.7E-1	<b>1.9E-1</b>
R( $\lambda$ )	3.6E-1	3.6E-1	5.7E-1	6.0E-1
Q $^\pi$ ( $\lambda$ )	4.4E-1	4.6E-1	3.2E-1	4.3E-1
TREE	4.1E-1	4.5E-1	6.7E-1	6.8E-1
IH	<b>4.8E-2</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.0E0	<b>6.9E-1</b>
WIS	2.4E0	8.1E-1
NAIVE	4.2E0	-

Table 96: Graph, relative MSE.  $T = 16, N = 16, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.3E-1	3.9E-1	1.8E0	5.4E-1
Q-REG	5.4E-1	4.1E-1	7.6E0	8.5E-1
MRDR	5.0E-1	8.0E-1	9.5E0	8.7E0
FQE	5.1E-1	4.7E-1	7.3E-1	5.1E-1
R( $\lambda$ )	5.0E-1	5.1E-1	1.4E0	7.6E-1
Q $^\pi$ ( $\lambda$ )	3.9E-1	3.3E-1	<b>3.1E-1</b>	3.9E-1
TREE	5.4E-1	5.6E-1	1.7E0	1.0E0
IH	<b>1.6E-1</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.0E0	<b>5.3E-1</b>
WIS	2.3E0	2.0E0
NAIVE	4.1E0	-

Table 98: Graph, relative MSE.  $T = 16, N = 64, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.3E-2	2.3E1	1.6E0	6.2E-2
Q-REG	4.0E0	2.3E0	3.7E0	3.9E0
MRDR	2.1E0	2.5E1	1.1E1	1.1E1
FQE	5.1E-2	1.8E0	3.2E-1	<b>5.1E-2</b>
R( $\lambda$ )	2.0E-1	1.7E-1	2.7E-1	3.2E-1
Q $^\pi$ ( $\lambda$ )	3.2E-1	1.9E0	7.7E-1	3.4E-1
TREE	2.5E-1	7.3E-1	2.5E-1	3.2E-1
IH	<b>1.3E-2</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.1E0	3.1E0
WIS	1.9E0	<b>3.1E-1</b>
NAIVE	4.2E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 99: Graph, relative MSE.  $T = 16, N = 128, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.6E-2	4.4E0	1.5E0	2.1E-1
Q-REG	3.3E-1	3.3E1	1.4E1	4.0E-1
MRDR	3.8E0	5.0E0	1.7E1	1.4E1
FQE	<b>1.8E-2</b>	1.4E-1	7.4E-2	<b>1.7E-2</b>
R( $\lambda$ )	2.7E-1	2.7E-1	2.0E-1	2.9E-1
Q $^\pi$ ( $\lambda$ )	1.3E-1	5.3E-1	1.6E-1	1.1E-1
TREE	3.3E-1	2.8E-1	2.3E-1	3.5E-1
IH	2.2E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	9.3E-1	<b>2.6E-1</b>
WIS	8.3E-1	2.9E-1
NAIVE	4.0E0	-

Table 101: Graph, relative MSE.  $T = 16, N = 512, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	8.1E-3	5.5E-1	1.1E0	8.0E-3
Q-REG	3.9E-1	2.7E1	1.3E1	3.6E-1
MRDR	9.0E-1	1.1E0	1.3E1	1.2E1
FQE	<b>4.9E-3</b>	5.5E-2	1.3E-1	<b>4.8E-3</b>
R( $\lambda$ )	1.3E-1	1.5E-1	1.7E-1	1.6E-1
Q $^\pi$ ( $\lambda$ )	9.4E-3	9.5E-2	1.4E-1	8.2E-3
TREE	2.2E-1	2.1E-1	2.1E-1	2.5E-1
IH	5.8E-3	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	9.9E-1	3.9E-1
WIS	1.3E0	<b>3.1E-1</b>
NAIVE	4.0E0	-

Table 100: Graph, relative MSE.  $T = 16, N = 256, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.1E-2	3.9E0	1.1E0	<b>1.0E-2</b>
Q-REG	2.7E-1	1.3E0	5.4E-2	1.8E-1
MRDR	4.3E-1	1.2E0	8.3E0	8.3E0
FQE	7.4E-3	5.5E-2	9.5E-2	1.4E-2
R( $\lambda$ )	1.6E-1	1.3E-1	1.5E-1	1.6E-1
Q $^\pi$ ( $\lambda$ )	1.1E-1	1.3E-1	1.3E-1	1.1E-1
TREE	2.0E-1	1.6E-1	1.9E-1	2.1E-1
IH	<b>6.8E-3</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.0E0	<b>2.8E-1</b>
WIS	1.6E0	2.9E-1
NAIVE	4.0E0	-

Table 102: Graph, relative MSE.  $T = 16, N = 1024, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	<b>1.8E-3</b>	9.5E-1	6.8E-1	<b>3.8E-3</b>
Q-REG	3.9E-1	2.0E-1	3.9E-1	5.5E-1
MRDR	4.0E-1	4.6E-1	2.0E0	2.0E0
FQE	2.7E-3	1.8E-1	8.4E-2	2.6E-2
R( $\lambda$ )	1.5E-1	1.5E-1	4.9E-2	1.5E-1
Q $^\pi$ ( $\lambda$ )	1.3E-2	1.9E-1	7.9E-2	2.0E-2
TREE	2.2E-1	2.2E-1	4.6E-2	2.2E-1
IH	2.0E-3	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	9.3E-1	3.9E-1
WIS	6.4E-1	<b>5.4E-2</b>
NAIVE	4.0E0	-

Table 103: Graph, relative MSE.  $T = 16, N = 8, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	9.2E-1	1.4E0	1.9E0	8.4E-1
Q-REG	1.3E0	1.8E0	3.0E0	2.0E0
MRDR	1.3E0	8.1E-1	2.7E0	2.3E0
FQE	9.3E-1	8.0E-1	1.0E0	9.3E-1
$R(\lambda)$	<b>8.6E-1</b>	<b>7.7E-1</b>	2.0E0	1.3E0
$Q^\pi(\lambda)$	1.4E0	4.7E0	8.7E-1	1.4E0
TREE	8.6E-1	7.8E-1	2.1E0	1.3E0
IH	1.2E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.0E0</b>	1.6E0
WIS	3.4E0	3.2E0
NAIVE	4.5E0	-

Table 105: Graph, relative MSE.  $T = 16, N = 32, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.2E-1	4.2E1	2.1E0	2.3E-1
Q-REG	1.4E0	6.8E1	7.5E1	1.1E0
MRDR	1.6E0	1.3E2	7.4E1	4.8E0
FQE	<b>1.9E-1</b>	2.2E-1	4.7E-1	<b>1.9E-1</b>
$R(\lambda)$	4.0E-1	1.3E0	8.9E-1	8.6E-1
$Q^\pi(\lambda)$	1.8E0	3.1E0	2.8E0	1.8E0
TREE	4.4E-1	3.1E0	9.1E-1	8.6E-1
IH	2.1E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>9.9E-1</b>	3.2E0
WIS	2.8E0	1.2E0
NAIVE	4.1E0	-

Table 104: Graph, relative MSE.  $T = 16, N = 16, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	7.4E-1	1.5E0	2.3E0	1.1E0
Q-REG	8.6E-1	2.4E0	4.5E1	7.9E-1
MRDR	9.6E-1	1.4E0	3.0E1	6.3E0
FQE	7.7E-1	9.1E-1	1.0E0	<b>7.7E-1</b>
$R(\lambda)$	8.8E-1	8.8E-1	1.4E0	9.4E-1
$Q^\pi(\lambda)$	1.4E0	1.4E0	1.2E0	1.4E0
TREE	9.2E-1	9.2E-1	1.5E0	9.6E-1
IH	<b>4.3E-1</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	9.9E-1	<b>8.3E-1</b>
WIS	1.6E0	1.3E0
NAIVE	3.7E0	-

Table 106: Graph, relative MSE.  $T = 16, N = 64, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	6.6E-2	2.6E0	5.2E0	9.4E-2
Q-REG	7.1E-1	4.0E0	6.3E0	2.4E0
MRDR	8.1E-1	1.4E0	5.6E0	5.3E0
FQE	<b>5.5E-2</b>	1.5E-1	3.3E-1	<b>5.5E-2</b>
$R(\lambda)$	1.4E-1	2.5E-1	7.2E-1	1.4E-1
$Q^\pi(\lambda)$	1.2E0	9.1E-1	1.1E0	1.2E0
TREE	2.0E-1	3.5E-1	8.5E-1	2.1E-1
IH	6.2E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.2E0	<b>7.7E-1</b>
WIS	3.9E0	1.3E0
NAIVE	4.1E0	-



**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 107: Graph, relative MSE.  $T = 16, N = 128, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.7E-2	3.1E1	3.4E0	5.4E-2
Q-REG	1.9E0	1.1E1	4.9E0	1.8E0
MRDR	4.0E1	2.3E1	1.6E1	1.8E1
FQE	5.1E-2	7.0E0	4.8E-1	<b>5.1E-2</b>
$R(\lambda)$	2.3E-1	1.3E0	5.9E-1	4.9E-1
$Q^\pi(\lambda)$	2.0E-1	1.7E0	5.0E-1	2.1E-1
TREE	3.0E-1	2.9E0	6.6E-1	6.9E-1
IH	<b>4.0E-2</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.7E0	3.0E0
WIS	2.2E0	<b>7.9E-1</b>
NAIVE	4.1E0	-

Table 109: Graph, relative MSE.  $T = 16, N = 512, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	<b>6.3E-3</b>	2.6E0	1.4E0	1.1E-1
Q-REG	4.0E0	9.8E0	1.5E1	5.1E0
MRDR	1.1E1	3.2E1	3.1E2	3.0E2
FQE	7.9E-3	1.5E0	2.4E-1	<b>7.9E-3</b>
$R(\lambda)$	1.5E-1	7.9E-1	2.8E-1	1.5E-1
$Q^\pi(\lambda)$	1.0E-1	3.1E0	2.9E-1	9.3E-2
TREE	2.4E-1	1.1E0	3.5E-1	2.4E-1
IH	8.1E-3	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.6E0	3.5E0
WIS	1.5E0	<b>4.0E-1</b>
NAIVE	4.0E0	-

Table 108: Graph, relative MSE.  $T = 16, N = 256, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.7E-2	5.3E1	3.4E0	4.9E-2
Q-REG	7.9E0	3.2E1	1.9E1	8.4E0
MRDR	8.3E0	1.9E2	4.9E1	2.0E1
FQE	3.8E-2	1.2E0	3.1E-1	<b>4.0E-2</b>
$R(\lambda)$	2.4E-1	2.4E0	4.5E-1	2.4E-1
$Q^\pi(\lambda)$	1.5E-1	7.5E-1	3.1E-1	1.4E-1
TREE	2.6E-1	3.2E0	5.3E-1	2.7E-1
IH	<b>3.0E-2</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	7.9E1	1.0E1
WIS	1.2E0	<b>6.8E-1</b>
NAIVE	4.1E0	-

Table 110: Graph, relative MSE.  $T = 16, N = 1024, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	<b>5.3E-3</b>	1.3E2	1.3E0	9.0E-3
Q-REG	8.1E0	2.6E0	1.4E1	4.0E0
MRDR	3.7E0	1.2E2	3.2E1	8.8E0
FQE	6.4E-3	1.3E1	2.4E-1	<b>6.5E-3</b>
$R(\lambda)$	2.7E-1	3.0E0	2.5E-1	2.7E-1
$Q^\pi(\lambda)$	3.3E-2	2.0E1	2.4E-1	3.4E-2
TREE	3.2E-1	4.9E0	2.9E-1	4.1E-1
IH	5.8E-3	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	4.8E2	6.8E0
WIS	1.3E0	<b>3.2E-1</b>
NAIVE	4.0E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 111: Graph, relative MSE.  $T = 16, N = 8, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.0E0	9.1E-1	<b>9.1E-1</b>	1.0E0
Q-REG	9.7E-1	1.2E0	1.2E1	1.0E0
MRDR	<b>9.7E-1</b>	1.3E0	1.2E1	1.0E0
FQE	1.0E0	1.0E0	1.0E0	1.0E0
R( $\lambda$ )	1.0E0	9.7E-1	4.3E0	3.4E0
Q $^\pi$ ( $\lambda$ )	1.0E0	1.0E0	1.0E0	1.0E0
TREE	1.0E0	9.7E-1	4.3E0	3.4E0
IH	1.2E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>9.7E-1</b>	9.7E-1
WIS	4.3E0	4.3E0
NAIVE	5.3E0	-

Table 113: Graph, relative MSE.  $T = 16, N = 32, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.0E0	2.1E0	2.6E0	1.2E0
Q-REG	9.9E-1	1.0E0	1.2E1	1.1E0
MRDR	9.8E-1	1.0E0	1.3E1	5.7E0
FQE	9.8E-1	9.8E-1	9.8E-1	9.8E-1
R( $\lambda$ )	1.0E0	9.9E-1	1.1E0	1.5E0
Q $^\pi$ ( $\lambda$ )	9.8E-1	<b>9.8E-1</b>	9.8E-1	9.8E-1
TREE	1.0E0	9.9E-1	1.1E0	1.5E0
IH	<b>9.6E-1</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>9.9E-1</b>	9.9E-1
WIS	1.1E0	1.1E0
NAIVE	4.3E0	-

Table 112: Graph, relative MSE.  $T = 16, N = 16, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.0E0	1.1E0	6.2E0	1.0E0
Q-REG	1.0E0	<b>9.9E-1</b>	9.7E0	9.9E-1
MRDR	1.0E0	1.0E0	6.4E0	3.5E0
FQE	1.0E0	1.0E0	1.0E0	1.0E0
R( $\lambda$ )	1.0E0	1.0E0	1.5E0	1.1E0
Q $^\pi$ ( $\lambda$ )	1.0E0	1.0E0	1.0E0	1.0E0
TREE	1.0E0	1.0E0	1.5E0	1.1E0
IH	<b>8.9E-1</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.0E0</b>	1.0E0
WIS	1.5E0	1.5E0
NAIVE	4.2E0	-

Table 114: Graph, relative MSE.  $T = 16, N = 64, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.3E-1	2.2E0	9.4E0	1.2E0
Q-REG	1.0E0	9.6E-1	7.3E0	1.0E0
MRDR	9.9E-1	9.7E-1	5.2E1	3.5E1
FQE	<b>4.9E-1</b>	4.9E-1	4.9E-1	4.9E-1
R( $\lambda$ )	1.0E0	1.0E0	1.8E0	1.8E0
Q $^\pi$ ( $\lambda$ )	4.9E-1	<b>4.9E-1</b>	4.9E-1	4.9E-1
TREE	1.0E0	1.0E0	1.8E0	1.8E0
IH	8.4E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.0E0</b>	1.0E0
WIS	1.8E0	1.8E0
NAIVE	4.0E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 115: Graph, relative MSE.  $T = 16, N = 128, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	9.3E-3	1.8E1	9.1E0	6.1E-2
Q-REG	1.0E0	6.3E0	1.4E1	1.3E1
MRDR	1.1E0	1.2E0	1.4E1	1.4E1
FQE	2.0E-6	2.0E-6	2.0E-6	2.0E-6
R( $\lambda$ )	1.0E0	1.1E0	3.5E0	2.3E0
Q $^\pi$ ( $\lambda$ )	<b>2.0E-6</b>	2.0E-6	2.0E-6	<b>2.0E-6</b>
TREE	1.0E0	1.1E0	3.5E0	2.3E0
IH	7.3E-1	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	<b>1.1E0</b>	1.1E0
WIS	3.5E0	3.5E0
NAIVE	4.2E0	-

Table 117: Graph, relative MSE.  $T = 16, N = 512, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.6E-3	1.8E1	4.6E0	2.6E-2
Q-REG	1.7E1	9.2E2	6.1E2	2.4E1
MRDR	9.5E0	9.6E2	3.7E2	1.3E2
FQE	<b>5.0E-6</b>	5.0E-6	5.0E-6	<b>5.0E-6</b>
R( $\lambda$ )	1.0E0	1.6E1	1.9E0	1.6E0
Q $^\pi$ ( $\lambda$ )	5.0E-6	1.1E-5	5.0E-6	5.0E-6
TREE	1.0E0	1.6E1	1.9E0	1.6E0
IH	8.4E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.6E1	1.6E1
WIS	1.9E0	<b>1.9E0</b>
NAIVE	3.9E0	-

Table 116: Graph, relative MSE.  $T = 16, N = 256, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.4E-3	2.4E1	7.3E0	3.2E-2
Q-REG	9.3E-1	1.2E1	1.2E1	8.4E-1
MRDR	8.6E-1	4.6E0	1.5E2	1.5E2
FQE	<b>5.1E-5</b>	5.1E-5	5.1E-5	5.1E-5
R( $\lambda$ )	1.0E0	9.3E-1	2.0E0	2.2E0
Q $^\pi$ ( $\lambda$ )	5.1E-5	5.1E-5	5.1E-5	<b>5.0E-5</b>
TREE	1.0E0	9.3E-1	2.0E0	2.2E0
IH	1.5E-1	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	<b>9.3E-1</b>	9.3E-1
WIS	2.0E0	2.0E0
NAIVE	4.0E0	-

Table 118: Graph, relative MSE.  $T = 16, N = 1024, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.8E-3	2.3E3	2.7E0	3.2E-3
Q-REG	1.2E3	2.2E3	2.5E1	1.3E3
MRDR	1.8E4	2.5E4	1.4E2	9.6E2
FQE	<b>2.4E-5</b>	2.4E-5	2.4E-5	2.4E-5
R( $\lambda$ )	1.0E0	1.1E3	2.5E0	1.0E0
Q $^\pi$ ( $\lambda$ )	2.4E-5	<b>2.3E-5</b>	2.4E-5	2.4E-5
TREE	1.0E0	1.1E3	2.5E0	1.0E0
IH	2.6E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.1E3	1.1E3
WIS	2.5E0	<b>2.5E0</b>
NAIVE	3.9E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 119: Graph, relative MSE.  $T = 16, N = 8, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	6.7E0	1.3E1	2.5E1	6.0E0	
Q-REG	7.7E2	6.4E3	1.9E4	7.7E2	
MRDR	8.2E2	2.8E4	1.7E4	3.9E2	
FQE	6.8E0	1.1E1	1.9E1	6.9E0	
R( $\lambda$ )	6.9E0	3.2E2	3.1E1	8.5E0	
Q $^\pi$ ( $\lambda$ )	<b>5.8E0</b>	4.2E1	1.9E1	<b>5.8E0</b>	
TREE	6.9E0	7.1E2	3.8E1	9.5E0	
IH	9.9E1	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	5.6E3	8.5E2
WIS	8.5E1	<b>6.2E1</b>
NAIVE	1.9E1	-

Table 121: Graph, relative MSE.  $T = 16, N = 32, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	5.7E0	9.4E1	1.8E2	1.6E1	
Q-REG	7.4E1	5.8E1	5.4E1	7.4E1	
MRDR	6.2E1	3.2E2	9.0E1	9.9E1	
FQE	<b>5.1E0</b>	1.5E1	2.7E1	<b>5.1E0</b>	
R( $\lambda$ )	5.7E0	8.7E0	2.7E1	5.7E0	
Q $^\pi$ ( $\lambda$ )	1.9E1	3.1E2	2.0E1	1.9E1	
TREE	5.5E0	9.9E0	2.8E1	5.5E0	
IH	2.6E1	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.0E0</b>	1.2E2
WIS	4.7E1	3.7E1
NAIVE	7.3E0	-

Table 120: Graph, relative MSE.  $T = 16, N = 16, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	3.8E0	1.2E1	2.7E1	3.4E0	
Q-REG	4.9E2	1.5E3	3.6E2	4.9E2	
MRDR	3.0E2	2.4E3	1.8E2	3.4E2	
FQE	<b>2.3E0</b>	3.5E2	1.7E1	<b>2.3E0</b>	
R( $\lambda$ )	2.4E0	1.3E1	2.0E1	2.4E0	
Q $^\pi$ ( $\lambda$ )	6.7E0	4.2E2	1.5E1	6.7E0	
TREE	2.4E0	1.5E1	2.2E1	2.4E0	
IH	8.2E1	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.4E0</b>	4.9E2
WIS	4.2E1	3.5E1
NAIVE	1.1E1	-

Table 122: Graph, relative MSE.  $T = 16, N = 64, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	4.6E0	4.4E3	9.8E1	4.6E0	
Q-REG	1.1E2	1.4E3	1.3E3	1.1E2	
MRDR	7.1E1	1.2E3	1.8E3	2.3E2	
FQE	<b>3.4E0</b>	8.8E1	6.7E0	<b>3.4E0</b>	
R( $\lambda$ )	5.1E0	1.9E1	1.5E1	7.2E0	
Q $^\pi$ ( $\lambda$ )	4.5E1	8.2E1	3.8E1	4.5E1	
TREE	4.2E0	2.3E1	1.9E1	7.8E0	
IH	1.3E1	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	5.2E1	1.4E2
WIS	4.1E1	<b>2.4E1</b>
NAIVE	6.8E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 123: Graph, relative MSE.  $T = 16, N = 128, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.7E0	9.8E1	1.8E2	7.4E0
Q-REG	1.4E1	4.0E1	7.3E1	9.4E1
MRDR	1.9E1	<b>4.2E0</b>	2.5E2	2.4E2
FQE	<b>4.6E0</b>	6.3E0	2.8E1	4.6E0
$R(\lambda)$	7.4E0	8.2E0	3.1E1	8.3E0
$Q^\pi(\lambda)$	4.7E1	4.5E1	6.5E1	4.7E1
TREE	5.7E0	6.5E0	3.1E1	5.8E0
IH	5.2E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.0E0</b>	6.9E0
WIS	4.7E1	3.1E1
NAIVE	4.0E0	-

Table 125: Graph, relative MSE.  $T = 16, N = 512, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.5E0	1.3E5	6.1E1	1.2E1
Q-REG	1.4E4	1.8E5	2.1E4	1.4E4
MRDR	8.9E3	3.8E5	3.2E4	8.9E3
FQE	1.1E0	2.4E4	1.2E1	<b>1.2E0</b>
$R(\lambda)$	5.1E0	8.1E0	1.7E1	5.1E0
$Q^\pi(\lambda)$	6.5E0	2.2E4	1.2E1	6.5E0
TREE	4.4E0	8.7E1	1.6E1	4.5E0
IH	<b>8.1E-1</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>5.7E0</b>	1.7E4
WIS	8.1E1	1.7E1
NAIVE	4.7E0	-

Table 124: Graph, relative MSE.  $T = 16, N = 256, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.7E0	1.6E3	1.2E2	<b>1.6E0</b>
Q-REG	5.4E1	3.8E2	1.6E2	8.2E0
MRDR	9.3E1	4.8E1	6.0E2	6.0E2
FQE	1.9E0	1.1E2	1.4E1	2.0E0
$R(\lambda)$	4.2E0	8.9E0	1.1E1	9.0E0
$Q^\pi(\lambda)$	4.3E0	3.8E1	1.5E1	4.3E0
TREE	3.4E0	3.0E1	7.7E0	6.6E0
IH	<b>1.6E0</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.6E0</b>	5.3E1
WIS	6.7E1	9.9E0
NAIVE	4.0E0	-

Table 126: Graph, relative MSE.  $T = 16, N = 1024, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.1E0	4.6E1	6.9E1	5.1E0
Q-REG	9.1E0	1.6E2	6.4E1	1.3E1
MRDR	1.3E1	4.5E1	3.1E1	1.5E1
FQE	9.8E-1	7.9E0	8.9E0	<b>9.6E-1</b>
$R(\lambda)$	2.6E0	8.7E0	1.0E1	2.6E0
$Q^\pi(\lambda)$	2.1E0	4.3E0	9.3E0	2.0E0
TREE	2.2E0	9.8E0	9.9E0	2.2E0
IH	<b>9.8E-1</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>7.9E-1</b>	8.5E0
WIS	1.9E1	9.9E0
NAIVE	3.7E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 127: Graph, relative MSE.  $T = 16, N = 8, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.0E0	9.9E-1	8.5E0	1.0E0
Q-REG	9.9E-1	1.3E0	1.8E3	4.0E1
MRDR	9.9E-1	1.3E0	1.4E3	1.3E1
FQE	1.0E0	1.0E0	2.5E0	1.0E0
R( $\lambda$ )	1.0E0	9.9E-1	6.2E0	1.0E0
Q $^\pi$ ( $\lambda$ )	1.1E0	2.2E0	9.9E-1	<b>9.7E-1</b>
TREE	1.0E0	9.9E-1	6.2E0	1.0E0
IH	<b>8.0E-1</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>9.9E-1</b>	9.9E-1
WIS	6.2E0	6.2E0
NAIVE	5.0E0	-

Table 129: Graph, relative MSE.  $T = 16, N = 32, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.8E0	4.9E0	5.5E1	1.7E0
Q-REG	9.8E-1	<b>9.0E-1</b>	6.9E1	9.8E-1
MRDR	<b>9.6E-1</b>	1.1E0	1.5E2	1.5E2
FQE	1.5E0	1.4E0	5.5E0	1.5E0
R( $\lambda$ )	1.0E0	9.8E-1	6.4E0	1.1E0
Q $^\pi$ ( $\lambda$ )	8.5E0	1.1E1	9.8E0	7.0E0
TREE	1.0E0	9.8E-1	6.4E0	1.1E0
IH	1.0E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>9.8E-1</b>	9.8E-1
WIS	6.4E0	6.4E0
NAIVE	4.5E0	-

Table 128: Graph, relative MSE.  $T = 16, N = 16, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.5E0	1.2E2	4.0E1	1.5E0
Q-REG	1.8E0	9.1E0	8.0E1	1.8E0
MRDR	1.8E0	3.7E1	2.1E2	2.0E2
FQE	1.3E0	1.6E0	6.4E0	<b>1.3E0</b>
R( $\lambda$ )	1.0E0	2.8E0	1.1E1	3.7E0
Q $^\pi$ ( $\lambda$ )	2.8E0	2.6E3	2.6E0	2.7E0
TREE	<b>1.0E0</b>	2.8E0	1.1E1	3.7E0
IH	1.3E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>2.8E0</b>	2.8E0
WIS	1.1E1	1.1E1
NAIVE	5.2E0	-

Table 130: Graph, relative MSE.  $T = 16, N = 64, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.3E0	8.9E2	8.3E1	1.8E0
Q-REG	1.0E0	1.0E1	8.1E4	3.3E0
MRDR	1.0E0	1.3E1	4.2E5	4.2E5
FQE	2.1E0	2.1E0	6.4E0	2.2E0
R( $\lambda$ )	1.0E0	<b>1.0E0</b>	1.1E1	2.5E0
Q $^\pi$ ( $\lambda$ )	1.0E1	1.3E1	6.3E0	9.2E0
TREE	<b>1.0E0</b>	1.0E0	1.1E1	2.5E0
IH	1.0E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.0E0</b>	1.0E0
WIS	1.1E1	1.1E1
NAIVE	4.3E0	-

Table 131: Graph, relative MSE.  $T = 16, N = 128, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.4E-1	3.2E3	4.9E1	2.3E0
Q-REG	5.0E0	4.9E2	1.6E1	4.9E0
MRDR	5.2E1	4.6E1	1.1E3	1.1E3
FQE	<b>4.1E-1</b>	6.4E-1	4.8E0	<b>4.2E-1</b>
R( $\lambda$ )	1.0E0	4.3E0	9.0E0	2.8E0
Q $^\pi$ ( $\lambda$ )	3.8E0	4.8E1	8.9E0	4.0E0
TREE	1.0E0	4.3E0	9.0E0	2.7E0
IH	6.8E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>4.3E0</b>	4.3E0
WIS	9.0E0	9.0E0
NAIVE	3.9E0	-

Table 133: Graph, relative MSE.  $T = 16, N = 512, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	8.1E-2	2.1E1	3.8E1	4.7E0
Q-REG	1.6E0	3.5E1	4.0E1	2.4E0
MRDR	1.4E0	4.0E0	2.3E2	2.3E2
FQE	<b>5.3E-2</b>	9.1E0	6.5E0	<b>5.1E-2</b>
R( $\lambda$ )	1.0E0	1.7E0	6.8E0	1.0E0
Q $^\pi$ ( $\lambda$ )	1.1E0	1.0E1	6.2E0	9.9E-1
TREE	1.0E0	1.7E0	6.8E0	1.0E0
IH	1.9E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.7E0</b>	1.7E0
WIS	6.8E0	6.8E0
NAIVE	3.9E0	-

Table 132: Graph, relative MSE.  $T = 16, N = 256, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.1E-1	1.4E3	2.2E1	1.9E-1
Q-REG	1.2E0	6.1E2	3.5E2	4.9E0
MRDR	1.3E1	5.1E1	4.3E2	4.3E2
FQE	<b>9.2E-2</b>	2.1E0	1.8E0	<b>9.8E-2</b>
R( $\lambda$ )	1.0E0	1.3E0	3.9E0	3.5E0
Q $^\pi$ ( $\lambda$ )	1.3E0	3.1E1	2.7E0	1.5E0
TREE	1.0E0	1.3E0	3.9E0	3.5E0
IH	1.3E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.3E0</b>	1.3E0
WIS	3.9E0	3.9E0
NAIVE	4.1E0	-

Table 134: Graph, relative MSE.  $T = 16, N = 1024, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.6E-2	5.0E1	8.6E0	6.2E-2
Q-REG	1.1E0	2.8E1	1.8E1	4.8E0
MRDR	8.7E-1	1.7E0	1.9E2	1.6E2
FQE	<b>3.4E-2</b>	3.9E-1	2.8E0	<b>3.4E-2</b>
R( $\lambda$ )	1.0E0	1.1E0	4.9E0	1.2E0
Q $^\pi$ ( $\lambda$ )	6.6E-1	2.1E0	3.5E0	6.3E-1
TREE	1.0E0	1.1E0	4.9E0	1.2E0
IH	9.3E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.1E0</b>	1.1E0
WIS	4.9E0	4.9E0
NAIVE	4.0E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 135: Graph, relative MSE.  $T = 16, N = 8, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	1.6E1	8.4E3	3.2E2	1.7E1	
Q-REG	9.9E3	6.2E5	4.0E4	9.9E3	
MRDR	7.8E3	1.4E6	1.3E4	7.9E3	
FQE	<b>1.2E1</b>	2.2E3	7.8E1	<b>1.2E1</b>	
R( $\lambda$ )	2.2E1	9.1E3	9.7E1	2.5E1	
Q $^\pi$ ( $\lambda$ )	3.3E1	2.7E3	7.6E1	3.2E1	
TREE	2.0E1	1.3E4	8.8E1	2.0E1	
IH	3.2E2	-	-	-	

IPS		
	STANDARD	PER-DECISION
IS	8.9E2	1.8E4
WIS	1.3E2	<b>6.6E1</b>
NAIVE	1.8E1	-

Table 137: Graph, relative MSE.  $T = 16, N = 32, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	5.5E1	3.4E4	4.4E2	4.8E1	
Q-REG	1.4E4	9.1E4	8.4E4	1.4E4	
MRDR	1.0E5	3.1E5	4.0E5	1.5E5	
FQE	5.2E1	2.0E4	1.5E2	5.2E1	
R( $\lambda$ )	4.3E1	7.0E1	7.9E1	4.3E1	
Q $^\pi$ ( $\lambda$ )	1.5E2	1.5E4	1.6E2	1.5E2	
TREE	<b>3.6E1</b>	2.7E2	9.3E1	<b>3.7E1</b>	
IH	3.8E1	-	-	-	

IPS		
	STANDARD	PER-DECISION
IS	<b>1.0E0</b>	2.1E4
WIS	9.3E1	1.3E2
NAIVE	1.8E1	-

Table 136: Graph, relative MSE.  $T = 16, N = 16, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	<b>2.2E1</b>	8.1E2	3.3E2	<b>2.1E1</b>	
Q-REG	3.0E3	6.1E3	5.2E3	5.3E2	
MRDR	1.9E3	3.3E4	1.6E4	8.5E3	
FQE	2.2E1	4.5E1	3.8E1	2.2E1	
R( $\lambda$ )	2.5E1	3.6E1	1.7E2	2.5E1	
Q $^\pi$ ( $\lambda$ )	2.3E2	3.3E2	2.1E2	2.3E2	
TREE	2.6E1	2.9E1	1.7E2	2.6E1	
IH	1.0E2	-	-	-	

IPS		
	STANDARD	PER-DECISION
IS	<b>1.1E0</b>	3.0E3
WIS	2.4E2	1.9E2
NAIVE	2.5E1	-

Table 138: Graph, relative MSE.  $T = 16, N = 64, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	3.7E1	2.8E3	6.0E2	2.8E1	
Q-REG	3.1E2	6.9E1	4.1E2	3.1E2	
MRDR	3.5E2	7.6E2	3.1E3	3.4E2	
FQE	<b>2.8E1</b>	3.5E2	1.4E2	<b>2.8E1</b>	
R( $\lambda$ )	4.0E1	1.4E2	1.1E2	3.9E1	
Q $^\pi$ ( $\lambda$ )	1.7E2	7.0E2	2.4E2	1.7E2	
TREE	3.3E1	1.5E2	1.2E2	3.2E1	
IH	3.2E1	-	-	-	

IPS		
	STANDARD	PER-DECISION
IS	<b>1.0E0</b>	3.2E2
WIS	2.2E2	1.3E2
NAIVE	9.5E0	-



**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 139: Graph, relative MSE.  $T = 16, N = 128, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.3E1	8.9E2	8.5E2	1.4E1
Q-REG	3.6E1	3.8E4	9.3E3	2.4E1
MRDR	5.7E1	4.3E2	7.7E2	5.6E2
FQE	1.1E1	6.0E1	1.1E2	1.1E1
$R(\lambda)$	9.9E0	3.2E1	9.8E1	1.2E1
$Q^\pi(\lambda)$	4.2E1	2.4E2	1.5E2	4.4E1
TREE	<b>7.8E0</b>	3.0E1	9.8E1	<b>1.0E1</b>
IH	1.6E1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.0E0</b>	3.1E1
WIS	8.3E1	1.1E2
NAIVE	7.8E0	-

Table 141: Graph, relative MSE.  $T = 16, N = 512, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.0E0	2.3E2	4.1E2	4.9E0
Q-REG	1.6E1	5.4E1	1.7E2	8.8E0
MRDR	2.7E1	4.2E1	3.0E2	1.4E2
FQE	2.4E0	1.6E1	1.0E1	5.3E0
$R(\lambda)$	3.4E0	2.0E1	1.7E1	<b>3.2E0</b>
$Q^\pi(\lambda)$	1.3E1	2.9E1	6.4E0	1.2E1
TREE	4.3E0	2.0E1	1.7E1	4.0E0
IH	<b>2.3E0</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.2E0</b>	1.7E1
WIS	4.2E1	1.6E1
NAIVE	4.5E0	-

Table 140: Graph, relative MSE.  $T = 16, N = 256, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	8.9E0	5.1E2	5.2E2	8.4E0
Q-REG	2.5E1	1.8E3	1.9E4	3.2E1
MRDR	1.2E2	5.4E1	2.5E3	2.4E3
FQE	<b>8.3E0</b>	3.4E1	8.0E1	<b>8.3E0</b>
$R(\lambda)$	1.6E1	3.6E1	8.5E1	1.7E1
$Q^\pi(\lambda)$	3.7E1	7.5E1	1.0E2	4.2E1
TREE	1.1E1	3.1E1	8.8E1	1.1E1
IH	8.7E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	9.9E1	<b>2.4E1</b>
WIS	2.1E2	9.5E1
NAIVE	5.2E0	-

Table 142: Graph, relative MSE.  $T = 16, N = 1024, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.5E0	1.0E3	1.6E2	4.7E0
Q-REG	3.2E2	9.7E3	2.3E3	3.2E2
MRDR	6.6E3	2.3E3	1.4E4	8.8E3
FQE	<b>1.5E0</b>	5.0E2	1.6E1	<b>3.6E0</b>
$R(\lambda)$	3.4E0	3.1E2	2.3E1	8.6E0
$Q^\pi(\lambda)$	9.1E0	5.0E2	1.5E1	7.6E0
TREE	2.1E0	3.7E2	2.1E1	6.5E0
IH	1.5E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>4.1E0</b>	3.2E2
WIS	4.5E1	2.0E1
NAIVE	3.3E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 143: Graph, relative MSE.  $T = 4, N = 8, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	7.7E-2	4.9E-2	4.5E-2	4.1E-2
Q-REG	2.2E-1	9.6E-2	9.4E-2	1.8E-1
MRDR	3.6E-1	1.8E-1	1.6E-1	3.6E-1
FQE	3.3E-2	3.3E-2	3.3E-2	3.3E-2
R( $\lambda$ )	3.3E-2	3.3E-2	3.3E-2	3.3E-2
Q $^\pi$ ( $\lambda$ )	3.3E-2	3.3E-2	3.3E-2	3.3E-2
TREE	<b>3.3E-2</b>	3.4E-2	3.4E-2	<b>3.3E-2</b>
IH	1.2E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	5.0E-1	2.7E-1
WIS	3.6E-1	<b>2.4E-1</b>
NAIVE	8.6E-1	-

Table 145: Graph, relative MSE.  $T = 4, N = 32, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	7.0E-3	1.5E-2	1.4E-2	2.7E-3
Q-REG	1.5E-2	1.2E-3	7.2E-4	4.2E-3
MRDR	6.7E-2	3.7E-3	2.0E-3	1.8E-2
FQE	4.2E-7	4.2E-7	4.2E-7	4.2E-7
R( $\lambda$ )	4.2E-7	4.2E-7	4.2E-7	4.2E-7
Q $^\pi$ ( $\lambda$ )	<b>4.2E-7</b>	4.2E-7	4.2E-7	<b>4.2E-7</b>
TREE	1.6E-6	5.3E-7	5.4E-7	1.6E-6
IH	1.1E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	8.4E-2	3.8E-2
WIS	2.1E-2	<b>1.2E-2</b>
NAIVE	5.1E-1	-

Table 144: Graph, relative MSE.  $T = 4, N = 16, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.0E-2	4.0E-2	3.7E-2	2.7E-2
Q-REG	4.4E-2	5.5E-3	4.3E-3	7.1E-3
MRDR	7.9E-2	1.1E-2	7.6E-3	3.6E-2
FQE	4.2E-3	4.2E-3	4.2E-3	4.2E-3
R( $\lambda$ )	4.2E-3	4.2E-3	4.2E-3	4.2E-3
Q $^\pi$ ( $\lambda$ )	4.2E-3	4.2E-3	4.2E-3	4.2E-3
TREE	<b>4.2E-3</b>	4.2E-3	4.2E-3	<b>4.2E-3</b>
IH	2.5E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.0E-1	7.1E-2
WIS	4.4E-2	<b>2.6E-2</b>
NAIVE	5.6E-1	-

Table 146: Graph, relative MSE.  $T = 4, N = 64, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.1E-3	9.5E-3	9.6E-3	2.3E-3
Q-REG	2.0E-2	4.2E-4	1.6E-4	1.3E-3
MRDR	3.2E-2	1.6E-3	8.6E-4	2.1E-3
FQE	1.8E-5	1.8E-5	1.8E-5	1.8E-5
R( $\lambda$ )	1.8E-5	1.8E-5	1.8E-5	1.8E-5
Q $^\pi$ ( $\lambda$ )	1.8E-5	1.8E-5	1.8E-5	1.8E-5
TREE	<b>1.3E-5</b>	1.6E-5	1.6E-5	<b>1.3E-5</b>
IH	9.4E-3	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	4.8E-2	2.5E-2
WIS	1.4E-2	<b>8.5E-3</b>
NAIVE	3.9E-1	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 147: Graph, relative MSE.  $T = 4, N = 128, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.3E-3	3.6E-3	3.8E-3	3.0E-3
Q-REG	6.4E-3	3.1E-4	2.3E-4	1.4E-3
MRDR	3.3E-2	2.0E-3	1.2E-3	1.5E-3
FQE	7.4E-8	7.4E-8	7.4E-8	7.4E-8
R( $\lambda$ )	<b>7.3E-8</b>	7.3E-8	7.2E-8	7.2E-8
Q $^\pi$ ( $\lambda$ )	7.6E-8	7.5E-8	7.5E-8	7.5E-8
TREE	1.0E-7	<b>6.7E-8</b>	6.8E-8	9.1E-8
IH	3.7E-3	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.2E-2	1.3E-2
WIS	7.3E-3	<b>4.5E-3</b>
NAIVE	4.7E-1	-

Table 149: Graph, relative MSE.  $T = 4, N = 512, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.7E-4	2.5E-3	2.4E-3	1.6E-3
Q-REG	3.3E-3	2.9E-5	2.0E-5	1.1E-4
MRDR	1.7E-2	1.2E-4	8.9E-5	1.4E-4
FQE	1.4E-5	1.4E-5	1.4E-5	1.4E-5
R( $\lambda$ )	1.4E-5	1.4E-5	1.4E-5	1.4E-5
Q $^\pi$ ( $\lambda$ )	1.4E-5	1.4E-5	1.4E-5	1.4E-5
TREE	<b>1.0E-5</b>	1.5E-5	1.5E-5	<b>1.1E-5</b>
IH	1.7E-3	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E-2	5.8E-3
WIS	2.6E-3	<b>1.7E-3</b>
NAIVE	4.2E-1	-

Table 148: Graph, relative MSE.  $T = 4, N = 256, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	7.4E-4	3.0E-3	3.0E-3	1.8E-3
Q-REG	3.5E-3	7.6E-6	6.5E-6	2.5E-5
MRDR	2.7E-2	1.7E-4	1.0E-4	2.7E-4
FQE	8.4E-7	8.4E-7	8.4E-7	8.4E-7
R( $\lambda$ )	8.4E-7	8.4E-7	8.4E-7	8.4E-7
Q $^\pi$ ( $\lambda$ )	8.4E-7	8.4E-7	8.4E-7	8.5E-7
TREE	<b>1.2E-7</b>	7.9E-7	8.0E-7	<b>1.3E-7</b>
IH	4.6E-3	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.2E-2	9.2E-3
WIS	<b>2.9E-3</b>	3.3E-3
NAIVE	4.6E-1	-

Table 150: Graph, relative MSE.  $T = 4, N = 1024, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.0E-4	5.5E-4	5.5E-4	2.2E-4
Q-REG	9.3E-4	1.6E-5	1.6E-5	1.7E-5
MRDR	1.6E-2	1.8E-4	2.0E-4	2.4E-4
FQE	1.6E-5	1.6E-5	1.6E-5	1.6E-5
R( $\lambda$ )	1.6E-5	1.6E-5	1.6E-5	1.6E-5
Q $^\pi$ ( $\lambda$ )	1.6E-5	1.6E-5	1.6E-5	1.6E-5
TREE	<b>1.2E-5</b>	1.6E-5	1.6E-5	<b>1.2E-5</b>
IH	5.1E-4	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.4E-3	1.8E-3
WIS	8.6E-4	<b>6.1E-4</b>
NAIVE	4.3E-1	-

Table 151: Graph, relative MSE.  $T = 4, N = 8, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.9E-1	1.7E-1	<b>1.7E-1</b>	1.8E-1
Q-REG	3.1E-1	2.7E-1	2.6E-1	2.8E-1
MRDR	3.2E-1	2.1E-1	2.3E-1	3.2E-1
FQE	<b>1.9E-1</b>	2.5E-1	2.4E-1	1.9E-1
R( $\lambda$ )	2.3E-1	2.5E-1	2.4E-1	2.3E-1
Q $^\pi$ ( $\lambda$ )	2.2E-1	2.2E-1	2.2E-1	2.2E-1
TREE	2.4E-1	2.6E-1	2.6E-1	2.4E-1
IH	3.0E-1	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.0E0	5.8E-1
WIS	5.1E-1	<b>3.7E-1</b>
NAIVE	5.6E-1	-

Table 153: Graph, relative MSE.  $T = 4, N = 32, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.5E-2	4.9E-2	4.7E-2	4.3E-2
Q-REG	3.1E-2	3.0E-2	3.0E-2	3.1E-2
MRDR	5.6E-2	3.1E-2	3.1E-2	4.9E-2
FQE	2.3E-2	2.6E-2	2.6E-2	<b>2.3E-2</b>
R( $\lambda$ )	2.7E-2	2.9E-2	2.8E-2	2.7E-2
Q $^\pi$ ( $\lambda$ )	2.5E-2	3.0E-2	3.0E-2	2.5E-2
TREE	2.8E-2	2.7E-2	2.7E-2	2.8E-2
IH	<b>2.0E-2</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	2.7E-2	2.1E-2
WIS	2.6E-2	<b>2.1E-2</b>
NAIVE	4.6E-1	-

Table 152: Graph, relative MSE.  $T = 4, N = 16, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.0E-1	2.0E-1	1.9E-1	1.0E-1
Q-REG	2.2E-1	1.3E-1	1.2E-1	2.2E-1
MRDR	2.3E-1	1.5E-1	1.3E-1	2.3E-1
FQE	<b>8.1E-2</b>	9.0E-2	8.8E-2	<b>8.1E-2</b>
R( $\lambda$ )	9.6E-2	9.9E-2	9.8E-2	9.6E-2
Q $^\pi$ ( $\lambda$ )	1.0E-1	1.2E-1	1.1E-1	1.0E-1
TREE	9.6E-2	9.4E-2	9.4E-2	9.6E-2
IH	1.3E-1	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	4.3E-1	2.4E-1
WIS	1.8E-1	<b>1.5E-1</b>
NAIVE	5.9E-1	-

Table 154: Graph, relative MSE.  $T = 4, N = 64, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.6E-2	3.4E-2	3.2E-2	1.5E-2
Q-REG	2.9E-2	7.3E-3	7.3E-3	7.3E-3
MRDR	2.6E-2	<b>6.4E-3</b>	6.6E-3	2.3E-2
FQE	8.0E-3	8.8E-3	8.6E-3	7.9E-3
R( $\lambda$ )	<b>7.9E-3</b>	7.8E-3	7.9E-3	7.9E-3
Q $^\pi$ ( $\lambda$ )	8.1E-3	8.4E-3	8.3E-3	8.2E-3
TREE	8.1E-3	7.5E-3	7.7E-3	8.0E-3
IH	1.6E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	8.7E-2	3.3E-2
WIS	1.9E-2	<b>1.3E-2</b>
NAIVE	3.6E-1	-

Table 155: Graph, relative MSE.  $T = 4, N = 128, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.8E-2	4.8E-2	5.0E-2	2.4E-2
Q-REG	1.8E-2	1.0E-2	1.0E-2	<b>7.2E-3</b>
MRDR	4.5E-2	8.6E-3	1.1E-2	2.1E-2
FQE	<b>7.8E-3</b>	1.0E-2	1.0E-2	7.9E-3
$R(\lambda)$	1.0E-2	1.1E-2	1.1E-2	1.0E-2
$Q^\pi(\lambda)$	1.0E-2	1.1E-2	1.1E-2	1.0E-2
TREE	9.7E-3	1.1E-2	1.1E-2	9.8E-3
IH	9.8E-3	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	5.4E-2	1.6E-2
WIS	1.9E-2	<b>1.0E-2</b>
NAIVE	4.7E-1	-

 Table 157: Graph, relative MSE.  $T = 4, N = 512, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.9E-3	3.4E-3	3.3E-3	1.3E-3
Q-REG	3.4E-3	9.3E-4	9.5E-4	2.5E-3
MRDR	2.5E-2	1.4E-3	1.4E-3	1.5E-3
FQE	<b>6.7E-4</b>	1.0E-3	1.0E-3	<b>6.8E-4</b>
$R(\lambda)$	8.5E-4	1.0E-3	1.0E-3	8.5E-4
$Q^\pi(\lambda)$	6.8E-4	1.0E-3	1.0E-3	6.9E-4
TREE	9.0E-4	1.0E-3	1.0E-3	9.1E-4
IH	3.1E-3	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	8.3E-3	5.4E-3
WIS	2.9E-3	<b>2.6E-3</b>
NAIVE	4.7E-1	-

 Table 156: Graph, relative MSE.  $T = 4, N = 256, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	<b>5.2E-3</b>	1.9E-2	1.9E-2	8.0E-3
Q-REG	1.2E-2	8.0E-3	7.8E-3	8.8E-3
MRDR	2.7E-2	7.7E-3	7.5E-3	1.0E-2
FQE	5.9E-3	7.3E-3	7.3E-3	<b>5.9E-3</b>
$R(\lambda)$	7.0E-3	7.3E-3	7.3E-3	7.0E-3
$Q^\pi(\lambda)$	7.2E-3	7.3E-3	7.3E-3	7.2E-3
TREE	7.0E-3	7.3E-3	7.3E-3	7.0E-3
IH	7.7E-3	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.3E-2	1.5E-2
WIS	<b>9.3E-3</b>	9.6E-3
NAIVE	4.2E-1	-

 Table 158: Graph, relative MSE.  $T = 4, N = 1024, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.1E-3	1.8E-3	1.8E-3	1.4E-3
Q-REG	<b>5.3E-4</b>	1.0E-3	1.0E-3	<b>4.2E-4</b>
MRDR	1.9E-2	1.4E-3	1.5E-3	1.5E-3
FQE	9.5E-4	1.0E-3	1.0E-3	9.5E-4
$R(\lambda)$	1.0E-3	1.1E-3	1.1E-3	1.0E-3
$Q^\pi(\lambda)$	1.1E-3	1.0E-3	1.0E-3	1.1E-3
TREE	1.0E-3	1.1E-3	1.1E-3	1.0E-3
IH	8.7E-4	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.6E-3	1.6E-3
WIS	1.1E-3	<b>7.7E-4</b>
NAIVE	4.6E-1	-

Table 159: Graph, relative MSE.  $T = 4, N = 8, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	7.7E-1	8.5E-1	9.0E-1	6.9E-1
Q-REG	9.0E-1	1.2E0	1.1E0	1.0E0
MRDR	6.6E-1	6.9E-1	7.3E-1	6.4E-1
FQE	6.3E-1	9.1E-1	9.0E-1	<b>6.3E-1</b>
$R(\lambda)$	9.5E-1	1.0E0	1.0E0	9.5E-1
$Q^\pi(\lambda)$	8.6E-1	9.3E-1	9.3E-1	8.6E-1
TREE	9.4E-1	9.9E-1	9.9E-1	9.4E-1
IH	<b>4.1E-1</b>	-	-	-

IPS			
	STANDARD		PER-DECISION
IS	1.2E0		9.5E-1
WIS	9.4E-1		<b>8.5E-1</b>
NAIVE	1.2E0		-

 Table 161: Graph, relative MSE.  $T = 4, N = 32, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.1E-1	2.0E-1	2.0E-1	1.9E-1
Q-REG	1.6E-1	<b>1.0E-1</b>	1.1E-1	1.5E-1
MRDR	1.6E-1	1.1E-1	1.1E-1	1.6E-1
FQE	1.3E-1	1.5E-1	1.4E-1	1.3E-1
$R(\lambda)$	1.2E-1	1.3E-1	1.2E-1	1.2E-1
$Q^\pi(\lambda)$	<b>1.1E-1</b>	1.3E-1	1.2E-1	1.1E-1
TREE	1.4E-1	1.3E-1	1.3E-1	1.3E-1
IH	1.5E-1	-	-	-

IPS			
	STANDARD		PER-DECISION
IS	2.1E-1		1.6E-1
WIS	1.8E-1		<b>1.4E-1</b>
NAIVE	4.8E-1		-

 Table 160: Graph, relative MSE.  $T = 4, N = 16, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.1E-1	1.3E-1	1.6E-1	1.8E-1
Q-REG	2.6E-1	2.3E-1	2.2E-1	2.8E-1
MRDR	<b>1.2E-1</b>	1.7E-1	1.6E-1	<b>1.3E-1</b>
FQE	2.0E-1	1.4E-1	1.5E-1	2.0E-1
$R(\lambda)$	1.7E-1	1.8E-1	1.8E-1	1.7E-1
$Q^\pi(\lambda)$	2.1E-1	1.8E-1	1.8E-1	2.1E-1
TREE	1.5E-1	1.7E-1	1.7E-1	1.5E-1
IH	1.9E-1	-	-	-

IPS			
	STANDARD		PER-DECISION
IS	5.8E-1		3.0E-1
WIS	2.9E-1		<b>1.8E-1</b>
NAIVE	6.2E-1		-

 Table 162: Graph, relative MSE.  $T = 4, N = 64, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.3E-1	2.4E-1	2.4E-1	1.5E-1
Q-REG	1.4E-1	9.1E-2	9.4E-2	1.0E-1
MRDR	1.2E-1	7.7E-2	8.0E-2	7.7E-2
FQE	<b>6.2E-2</b>	9.8E-2	9.7E-2	<b>6.3E-2</b>
$R(\lambda)$	8.8E-2	1.0E-1	1.0E-1	7.9E-2
$Q^\pi(\lambda)$	9.6E-2	1.0E-1	1.0E-1	8.7E-2
TREE	7.9E-2	9.6E-2	9.6E-2	7.5E-2
IH	1.3E-1	-	-	-

IPS			
	STANDARD		PER-DECISION
IS	3.9E-1		2.1E-1
WIS	1.9E-1		<b>1.4E-1</b>
NAIVE	4.8E-1		-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 163: Graph, relative MSE.  $T = 4, N = 128, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	9.7E-2	1.1E-1	1.2E-1	8.9E-2
Q-REG	9.1E-2	9.0E-2	9.1E-2	8.0E-2
MRDR	1.3E-1	8.8E-2	9.3E-2	1.1E-1
FQE	<b>7.8E-2</b>	8.4E-2	8.6E-2	<b>7.8E-2</b>
R( $\lambda$ )	9.3E-2	8.9E-2	9.0E-2	9.5E-2
Q $^\pi$ ( $\lambda$ )	1.0E-1	8.8E-2	8.9E-2	9.6E-2
TREE	1.0E-1	9.2E-2	9.2E-2	1.0E-1
IH	1.2E-1	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.1E-1	9.7E-2
WIS	9.8E-2	<b>8.8E-2</b>
NAIVE	5.3E-1	-

Table 165: Graph, relative MSE.  $T = 4, N = 512, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.2E-2	2.0E-2	2.0E-2	1.2E-2
Q-REG	1.4E-2	1.1E-2	1.1E-2	9.6E-3
MRDR	6.6E-3	1.1E-2	1.1E-2	8.2E-3
FQE	<b>5.3E-3</b>	1.1E-2	1.1E-2	<b>5.4E-3</b>
R( $\lambda$ )	9.0E-3	1.1E-2	1.1E-2	7.5E-3
Q $^\pi$ ( $\lambda$ )	9.3E-3	1.1E-2	1.1E-2	7.2E-3
TREE	8.7E-3	1.1E-2	1.1E-2	7.9E-3
IH	8.6E-3	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.4E-2	1.6E-2
WIS	<b>1.2E-2</b>	1.4E-2
NAIVE	3.9E-1	-

Table 164: Graph, relative MSE.  $T = 4, N = 256, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.0E-2	2.9E-2	3.0E-2	<b>1.1E-2</b>
Q-REG	4.1E-2	4.1E-2	4.0E-2	3.7E-2
MRDR	8.6E-2	4.3E-2	4.3E-2	4.8E-2
FQE	<b>2.4E-2</b>	3.8E-2	3.8E-2	2.5E-2
R( $\lambda$ )	3.7E-2	3.9E-2	3.9E-2	3.6E-2
Q $^\pi$ ( $\lambda$ )	3.9E-2	3.9E-2	4.0E-2	3.6E-2
TREE	3.6E-2	3.9E-2	3.9E-2	3.7E-2
IH	3.4E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	5.3E-2	4.6E-2
WIS	4.3E-2	<b>4.0E-2</b>
NAIVE	5.7E-1	-

Table 166: Graph, relative MSE.  $T = 4, N = 1024, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.2E-3	6.8E-3	6.9E-3	3.1E-3
Q-REG	3.6E-3	1.8E-3	1.8E-3	3.3E-3
MRDR	2.1E-2	1.4E-3	<b>1.2E-3</b>	1.0E-2
FQE	<b>1.7E-3</b>	1.8E-3	1.8E-3	1.7E-3
R( $\lambda$ )	1.8E-3	1.8E-3	1.8E-3	2.1E-3
Q $^\pi$ ( $\lambda$ )	2.5E-3	1.8E-3	1.8E-3	2.6E-3
TREE	1.8E-3	1.8E-3	1.8E-3	2.1E-3
IH	2.2E-3	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	4.8E-3	3.3E-3
WIS	3.2E-3	<b>2.7E-3</b>
NAIVE	4.6E-1	-

Table 167: Graph, relative MSE.  $T = 4, N = 8, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	4.3E-1	5.9E-1	7.1E-1	4.3E-1
Q-REG	3.9E-1	3.7E-1	3.6E-1	3.9E-1
MRDR	3.8E-1	<b>2.8E-1</b>	3.4E-1	3.8E-1
FQE	4.1E-1	3.0E-1	3.3E-1	4.1E-1
R( $\lambda$ )	3.2E-1	3.4E-1	3.6E-1	3.2E-1
Q $^\pi$ ( $\lambda$ )	3.7E-1	3.1E-1	3.4E-1	3.7E-1
TREE	<b>3.2E-1</b>	3.5E-1	3.6E-1	3.2E-1
IH	7.9E-1	-	-	-

IPS	
STANDARD	PER-DECISION
IS	1.4E0
WIS	6.9E-1
NAIVE	6.6E-1

Table 169: Graph, relative MSE.  $T = 4, N = 32, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	2.5E-1	4.6E-1	4.6E-1	2.6E-1
Q-REG	3.4E-1	2.9E-1	3.0E-1	3.4E-1
MRDR	3.0E-1	<b>2.4E-1</b>	2.5E-1	3.0E-1
FQE	<b>2.4E-1</b>	3.0E-1	3.1E-1	2.4E-1
R( $\lambda$ )	3.0E-1	3.1E-1	3.1E-1	3.0E-1
Q $^\pi$ ( $\lambda$ )	3.4E-1	3.3E-1	3.3E-1	3.4E-1
TREE	2.9E-1	3.0E-1	2.9E-1	2.9E-1
IH	2.7E-1	-	-	-

IPS	
STANDARD	PER-DECISION
IS	3.8E-1
WIS	3.3E-1
NAIVE	9.0E-1

Table 168: Graph, relative MSE.  $T = 4, N = 16, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	1.2E0	1.7E0	1.9E0	1.4E0
Q-REG	7.7E-1	9.8E-1	9.7E-1	7.7E-1
MRDR	<b>6.6E-1</b>	6.9E-1	7.4E-1	<b>6.5E-1</b>
FQE	9.8E-1	9.3E-1	1.0E0	9.8E-1
R( $\lambda$ )	9.9E-1	1.0E0	1.0E0	9.9E-1
Q $^\pi$ ( $\lambda$ )	9.6E-1	9.6E-1	1.0E0	9.6E-1
TREE	1.0E0	1.0E0	1.1E0	1.0E0
IH	7.1E-1	-	-	-

IPS	
STANDARD	PER-DECISION
IS	<b>7.6E-1</b>
WIS	7.7E-1
NAIVE	1.6E0

Table 170: Graph, relative MSE.  $T = 4, N = 64, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	2.0E-1	1.1E-1	1.2E-1	9.8E-2
Q-REG	<b>6.8E-2</b>	1.0E-1	1.1E-1	8.4E-2
MRDR	1.1E-1	<b>8.4E-2</b>	8.8E-2	1.0E-1
FQE	8.8E-2	9.1E-2	9.3E-2	8.8E-2
R( $\lambda$ )	8.9E-2	1.1E-1	1.0E-1	8.7E-2
Q $^\pi$ ( $\lambda$ )	8.8E-2	1.1E-1	1.1E-1	8.5E-2
TREE	8.3E-2	1.0E-1	9.7E-2	8.4E-2
IH	7.5E-2	-	-	-

IPS	
STANDARD	PER-DECISION
IS	<b>5.9E-2</b>
WIS	8.1E-2
NAIVE	5.3E-1



Table 171: Graph, relative MSE.  $T = 4, N = 128, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.9E-2	8.8E-2	9.3E-2	6.4E-2
Q-REG	9.3E-2	7.4E-2	7.4E-2	7.0E-2
MRDR	1.3E-1	8.1E-2	7.9E-2	1.3E-1
FQE	4.0E-2	6.3E-2	6.8E-2	<b>4.0E-2</b>
$R(\lambda)$	6.1E-2	7.1E-2	7.2E-2	4.7E-2
$Q^\pi(\lambda)$	6.8E-2	7.2E-2	7.4E-2	4.3E-2
TREE	5.9E-2	7.0E-2	7.0E-2	5.3E-2
IH	<b>3.3E-2</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.8E-1	9.7E-2
WIS	1.6E-1	<b>8.9E-2</b>
NAIVE	7.0E-1	-

Table 173: Graph, relative MSE.  $T = 4, N = 512, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.7E-2	2.3E-2	<b>2.3E-2</b>	2.8E-2
Q-REG	4.1E-2	3.5E-2	3.5E-2	3.4E-2
MRDR	5.6E-2	3.9E-2	3.9E-2	5.3E-2
FQE	<b>2.4E-2</b>	3.5E-2	3.5E-2	2.5E-2
$R(\lambda)$	3.3E-2	3.5E-2	3.5E-2	2.8E-2
$Q^\pi(\lambda)$	3.3E-2	3.4E-2	3.5E-2	2.8E-2
TREE	3.2E-2	3.5E-2	3.5E-2	2.9E-2
IH	2.8E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	4.9E-2	4.2E-2
WIS	4.1E-2	<b>3.8E-2</b>
NAIVE	5.1E-1	-

Table 172: Graph, relative MSE.  $T = 4, N = 256, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	8.9E-2	1.7E-1	1.6E-1	9.5E-2
Q-REG	1.3E-1	7.6E-2	7.7E-2	1.1E-1
MRDR	1.3E-1	7.5E-2	7.0E-2	9.6E-2
FQE	<b>4.9E-2</b>	8.0E-2	7.9E-2	<b>5.0E-2</b>
$R(\lambda)$	7.0E-2	7.8E-2	7.8E-2	6.1E-2
$Q^\pi(\lambda)$	7.1E-2	7.9E-2	7.8E-2	6.2E-2
TREE	7.0E-2	7.8E-2	7.8E-2	6.3E-2
IH	7.2E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	2.0E-1	1.5E-1
WIS	1.4E-1	<b>1.2E-1</b>
NAIVE	4.9E-1	-

Table 174: Graph, relative MSE.  $T = 4, N = 1024, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.0E-2	2.2E-2	2.3E-2	1.7E-2
Q-REG	1.5E-2	1.5E-2	1.5E-2	1.5E-2
MRDR	4.1E-2	1.6E-2	1.7E-2	2.6E-2
FQE	<b>1.3E-2</b>	1.5E-2	1.5E-2	<b>1.3E-2</b>
$R(\lambda)$	1.5E-2	1.5E-2	1.5E-2	1.4E-2
$Q^\pi(\lambda)$	1.5E-2	1.5E-2	1.5E-2	1.4E-2
TREE	1.5E-2	1.5E-2	1.5E-2	1.5E-2
IH	1.4E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	2.5E-2	1.6E-2
WIS	2.2E-2	<b>1.5E-2</b>
NAIVE	5.0E-1	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 175: Graph, relative MSE.  $T = 4, N = 8, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	8.6E-2	3.4E-1	2.6E-1	3.2E-1	
Q-REG	2.6E-1	9.6E-2	6.5E-2	3.9E-1	
MRDR	2.7E-1	1.8E-1	1.4E-1	2.9E-1	
FQE	1.9E-2	1.9E-2	1.9E-2	1.2E-1	
R( $\lambda$ )	1.9E-2	1.9E-2	1.9E-2	1.2E-1	
Q $^\pi$ ( $\lambda$ )	1.9E-2	1.9E-2	1.9E-2	1.2E-1	
TREE	<b>1.9E-2</b>	<b>1.9E-2</b>	1.9E-2	1.2E-1	
IH	1.7E-1	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	6.0E-1	6.0E-1
WIS	<b>2.4E-1</b>	2.4E-1
NAIVE	3.6E-1	-

Table 177: Graph, relative MSE.  $T = 4, N = 32, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	1.1E-2	4.8E-2	4.7E-2	2.7E-2	
Q-REG	1.2E-1	2.1E-2	9.0E-3	1.5E-2	
MRDR	9.8E-2	2.2E-2	1.4E-2	4.7E-2	
FQE	4.1E-4	4.1E-4	4.1E-4	4.1E-4	
R( $\lambda$ )	4.2E-4	4.1E-4	4.2E-4	4.2E-4	
Q $^\pi$ ( $\lambda$ )	<b>4.1E-4</b>	4.1E-4	4.1E-4	<b>4.1E-4</b>	
TREE	4.3E-4	4.2E-4	4.2E-4	4.3E-4	
IH	1.0E-1	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	1.1E-1	1.1E-1
WIS	<b>7.3E-2</b>	7.3E-2
NAIVE	5.4E-1	-

Table 176: Graph, relative MSE.  $T = 4, N = 16, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	7.6E-2	1.1E-1	1.1E-1	2.7E-1	
Q-REG	8.3E-2	1.3E-2	1.3E-2	2.3E-1	
MRDR	1.4E-1	3.1E-2	2.4E-2	3.1E-1	
FQE	5.6E-4	5.6E-4	5.6E-4	2.0E-1	
R( $\lambda$ )	<b>5.6E-4</b>	5.6E-4	5.6E-4	2.0E-1	
Q $^\pi$ ( $\lambda$ )	5.6E-4	5.6E-4	5.6E-4	2.0E-1	
TREE	5.9E-4	5.3E-4	<b>5.3E-4</b>	2.0E-1	
IH	1.4E-1	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	6.6E-2	6.6E-2
WIS	5.1E-2	<b>5.1E-2</b>
NAIVE	6.2E-1	-

Table 178: Graph, relative MSE.  $T = 4, N = 64, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	3.3E-2	3.9E-2	3.8E-2	2.5E-1	
Q-REG	8.1E-2	2.0E-3	7.1E-4	2.0E-1	
MRDR	8.5E-2	1.5E-2	1.0E-2	2.1E-1	
FQE	1.1E-9	1.1E-9	1.1E-9	2.0E-1	
R( $\lambda$ )	8.2E-9	5.3E-9	6.6E-9	2.0E-1	
Q $^\pi$ ( $\lambda$ )	<b>7.8E-10</b>	<b>1.1E-9</b>	1.1E-9	2.0E-1	
TREE	6.0E-6	5.9E-7	1.1E-6	2.0E-1	
IH	1.2E-1	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	7.4E-2	7.4E-2
WIS	<b>5.1E-2</b>	5.1E-2
NAIVE	6.0E-1	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 179: Graph, relative MSE.  $T = 4, N = 128, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.3E-2	1.2E-2	1.2E-2	2.2E-2
Q-REG	3.0E-2	3.4E-4	2.1E-4	5.9E-4
MRDR	2.9E-2	1.1E-3	2.4E-3	9.5E-3
FQE	6.9E-7	6.9E-7	6.9E-7	6.9E-7
R( $\lambda$ )	<b>6.7E-7</b>	6.8E-7	6.7E-7	<b>6.7E-7</b>
Q $^\pi$ ( $\lambda$ )	7.0E-7	6.9E-7	6.9E-7	7.0E-7
TREE	2.6E-6	9.0E-7	9.8E-7	2.4E-6
IH	1.2E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.9E-2	2.9E-2
WIS	1.9E-2	<b>1.9E-2</b>
NAIVE	5.4E-1	-

Table 181: Graph, relative MSE.  $T = 4, N = 512, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.0E-3	5.5E-3	5.8E-3	4.0E-3
Q-REG	2.5E-3	1.9E-5	2.0E-5	5.9E-4
MRDR	6.1E-3	5.4E-4	7.9E-4	2.0E-3
FQE	3.2E-5	3.2E-5	3.2E-5	3.2E-5
R( $\lambda$ )	3.2E-5	3.2E-5	3.2E-5	3.2E-5
Q $^\pi$ ( $\lambda$ )	3.2E-5	3.2E-5	3.2E-5	3.2E-5
TREE	<b>1.1E-5</b>	3.2E-5	3.2E-5	<b>1.2E-5</b>
IH	3.4E-3	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.8E-3	2.8E-3
WIS	<b>1.2E-3</b>	1.2E-3
NAIVE	4.6E-1	-

Table 180: Graph, relative MSE.  $T = 4, N = 256, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.7E-3	6.8E-3	6.9E-3	7.2E-3
Q-REG	2.2E-2	1.0E-4	7.4E-5	9.7E-5
MRDR	2.3E-2	4.8E-4	5.5E-4	3.3E-3
FQE	4.9E-7	4.9E-7	4.9E-7	4.9E-7
R( $\lambda$ )	5.1E-7	5.1E-7	5.1E-7	5.2E-7
Q $^\pi$ ( $\lambda$ )	<b>4.9E-7</b>	4.9E-7	4.9E-7	<b>4.9E-7</b>
TREE	9.3E-6	7.4E-7	7.6E-7	8.8E-6
IH	1.4E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.4E-2	2.4E-2
WIS	1.2E-2	<b>1.2E-2</b>
NAIVE	4.6E-1	-

Table 182: Graph, relative MSE.  $T = 4, N = 1024, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.1E-3	1.7E-3	1.6E-3	1.4E-3
Q-REG	3.0E-3	6.3E-5	6.4E-5	5.8E-4
MRDR	4.6E-3	6.0E-4	8.6E-4	2.0E-3
FQE	6.3E-5	6.3E-5	6.3E-5	6.3E-5
R( $\lambda$ )	<b>6.3E-5</b>	6.3E-5	6.3E-5	6.3E-5
Q $^\pi$ ( $\lambda$ )	6.3E-5	6.3E-5	6.3E-5	6.3E-5
TREE	1.1E-4	<b>6.2E-5</b>	6.2E-5	1.0E-4
IH	2.2E-3	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	3.6E-3	3.6E-3
WIS	1.5E-3	<b>1.5E-3</b>
NAIVE	4.5E-1	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 183: Graph, relative MSE.  $T = 4, N = 8, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.3E0	3.9E0	3.1E0	2.8E0
Q-REG	3.0E0	3.8E0	4.1E0	3.0E0
MRDR	<b>1.5E0</b>	2.5E0	3.0E0	<b>1.5E0</b>
FQE	2.8E0	3.8E0	3.8E0	2.8E0
R( $\lambda$ )	4.3E0	4.6E0	4.6E0	4.3E0
Q $^\pi$ ( $\lambda$ )	4.6E0	4.8E0	4.8E0	4.6E0
TREE	3.7E0	4.1E0	4.1E0	3.7E0
IH	3.4E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	3.8E0	<b>2.0E0</b>
WIS	3.5E0	2.4E0
NAIVE	2.7E0	-

Table 185: Graph, relative MSE.  $T = 4, N = 32, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	6.1E-1	6.6E-1	5.9E-1	4.2E-1
Q-REG	5.2E-1	4.2E-1	3.8E-1	5.2E-1
MRDR	3.5E-1	4.1E-1	3.8E-1	3.4E-1
FQE	3.8E-1	3.2E-1	<b>2.9E-1</b>	3.8E-1
R( $\lambda$ )	3.3E-1	3.2E-1	3.1E-1	3.3E-1
Q $^\pi$ ( $\lambda$ )	<b>3.2E-1</b>	3.1E-1	3.0E-1	3.2E-1
TREE	3.7E-1	3.3E-1	3.2E-1	3.7E-1
IH	7.4E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	4.8E-1	<b>4.2E-1</b>
WIS	4.7E-1	4.8E-1
NAIVE	8.0E-1	-

Table 184: Graph, relative MSE.  $T = 4, N = 16, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.9E-1	2.2E0	1.9E0	4.7E-1
Q-REG	3.8E-1	4.0E-1	3.7E-1	3.8E-1
MRDR	2.0E-1	1.8E-1	2.0E-1	2.0E-1
FQE	<b>1.7E-1</b>	1.9E-1	1.9E-1	<b>1.7E-1</b>
R( $\lambda$ )	2.3E-1	2.3E-1	2.3E-1	2.3E-1
Q $^\pi$ ( $\lambda$ )	3.1E-1	2.6E-1	2.7E-1	3.1E-1
TREE	2.3E-1	2.2E-1	2.0E-1	2.3E-1
IH	4.1E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	5.3E-1	4.1E-1
WIS	3.4E-1	<b>2.5E-1</b>
NAIVE	5.1E-1	-

Table 186: Graph, relative MSE.  $T = 4, N = 64, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.7E-1	6.6E-1	6.9E-1	6.3E-1
Q-REG	3.0E-1	2.5E-1	2.5E-1	3.1E-1
MRDR	<b>1.3E-1</b>	1.5E-1	<b>1.5E-1</b>	1.7E-1
FQE	3.5E-1	2.6E-1	2.5E-1	3.4E-1
R( $\lambda$ )	2.6E-1	2.5E-1	2.4E-1	2.6E-1
Q $^\pi$ ( $\lambda$ )	2.8E-1	2.5E-1	2.4E-1	2.8E-1
TREE	2.7E-1	2.5E-1	2.5E-1	2.7E-1
IH	2.5E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.4E-1	<b>1.8E-1</b>
WIS	3.4E-1	2.4E-1
NAIVE	4.1E-1	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 187: Graph, relative MSE.  $T = 4, N = 128, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.4E-1	5.8E-1	5.6E-1	4.1E-1
Q-REG	1.3E-1	9.3E-2	9.1E-2	1.1E-1
MRDR	9.6E-2	1.0E-1	1.0E-1	9.6E-2
FQE	<b>5.4E-2</b>	8.3E-2	8.6E-2	<b>5.5E-2</b>
$R(\lambda)$	7.1E-2	8.3E-2	8.5E-2	7.2E-2
$Q^\pi(\lambda)$	7.2E-2	8.2E-2	8.5E-2	7.3E-2
TREE	6.7E-2	8.2E-2	8.4E-2	6.7E-2
IH	2.1E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.9E-1	1.2E-1
WIS	1.8E-1	<b>1.1E-1</b>
NAIVE	5.7E-1	-

Table 189: Graph, relative MSE.  $T = 4, N = 512, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.0E-2	9.7E-2	9.6E-2	8.1E-2
Q-REG	1.8E-2	1.9E-2	1.9E-2	1.9E-2
MRDR	<b>1.4E-2</b>	1.7E-2	1.7E-2	2.5E-2
FQE	1.6E-2	1.9E-2	1.9E-2	<b>1.6E-2</b>
$R(\lambda)$	1.8E-2	1.8E-2	1.9E-2	1.8E-2
$Q^\pi(\lambda)$	1.8E-2	1.9E-2	1.9E-2	1.8E-2
TREE	1.8E-2	1.8E-2	1.9E-2	1.8E-2
IH	1.7E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	4.3E-2	<b>1.8E-2</b>
WIS	5.0E-2	2.0E-2
NAIVE	4.7E-1	-

Table 188: Graph, relative MSE.  $T = 4, N = 256, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.5E-1	2.8E-1	2.8E-1	2.2E-1
Q-REG	1.3E-1	1.4E-1	1.4E-1	1.4E-1
MRDR	1.6E-1	1.4E-1	1.4E-1	<b>1.0E-1</b>
FQE	1.2E-1	1.4E-1	1.4E-1	1.2E-1
$R(\lambda)$	1.4E-1	1.4E-1	1.4E-1	1.4E-1
$Q^\pi(\lambda)$	1.5E-1	1.4E-1	1.4E-1	1.5E-1
TREE	1.4E-1	1.4E-1	1.4E-1	1.4E-1
IH	<b>1.1E-1</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	8.3E-2	1.4E-1
WIS	<b>7.7E-2</b>	1.3E-1
NAIVE	3.8E-1	-

Table 190: Graph, relative MSE.  $T = 4, N = 1024, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.7E-2	4.1E-2	4.1E-2	3.3E-2
Q-REG	<b>8.7E-3</b>	1.0E-2	1.0E-2	9.4E-3
MRDR	1.0E-2	1.1E-2	1.2E-2	<b>9.2E-3</b>
FQE	9.3E-3	9.9E-3	1.0E-2	9.2E-3
$R(\lambda)$	9.8E-3	1.0E-2	1.0E-2	9.8E-3
$Q^\pi(\lambda)$	1.0E-2	1.0E-2	1.0E-2	1.0E-2
TREE	9.7E-3	9.9E-3	1.0E-2	9.7E-3
IH	9.7E-3	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.2E-2	<b>7.8E-3</b>
WIS	1.3E-2	8.1E-3
NAIVE	4.7E-1	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 191: Graph, relative MSE.  $T = 4, N = 8, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	1.5E0	3.5E0	3.2E0	2.3E0	
Q-REG	3.3E0	2.1E0	2.2E0	2.3E0	
MRDR	1.7E0	<b>1.6E0</b>	1.8E0	1.7E0	
FQE	1.6E0	2.1E0	2.2E0	1.7E0	
R( $\lambda$ )	2.3E0	2.6E0	2.6E0	2.4E0	
Q $^\pi$ ( $\lambda$ )	2.5E0	2.7E0	2.7E0	2.5E0	
TREE	2.0E0	2.5E0	2.5E0	2.2E0	
IH	<b>6.0E-1</b>	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	4.2E0	4.2E0
WIS	2.6E0	<b>2.6E0</b>
NAIVE	2.1E0	-

Table 193: Graph, relative MSE.  $T = 4, N = 32, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	6.5E-1	1.3E0	1.5E0	5.7E-1	
Q-REG	7.7E-1	5.3E-1	5.2E-1	4.8E-1	
MRDR	4.7E-1	5.4E-1	5.4E-1	3.3E-1	
FQE	3.4E-1	3.9E-1	4.1E-1	3.3E-1	
R( $\lambda$ )	3.8E-1	4.4E-1	4.5E-1	3.5E-1	
Q $^\pi$ ( $\lambda$ )	4.9E-1	4.8E-1	5.0E-1	4.2E-1	
TREE	3.4E-1	4.1E-1	4.2E-1	<b>3.2E-1</b>	
IH	<b>2.0E-1</b>	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	7.8E-1	7.8E-1
WIS	<b>5.8E-1</b>	5.8E-1
NAIVE	2.5E-1	-

Table 192: Graph, relative MSE.  $T = 4, N = 16, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	2.7E0	2.4E0	2.3E0	2.2E0	
Q-REG	1.9E0	1.8E0	1.9E0	2.0E0	
MRDR	1.4E0	<b>9.2E-1</b>	9.8E-1	1.3E0	
FQE	1.8E0	1.9E0	1.8E0	2.0E0	
R( $\lambda$ )	2.1E0	2.2E0	2.1E0	2.2E0	
Q $^\pi$ ( $\lambda$ )	2.1E0	2.3E0	2.2E0	2.2E0	
TREE	2.0E0	2.0E0	2.0E0	2.2E0	
IH	<b>1.0E0</b>	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	2.2E0	2.2E0
WIS	<b>1.6E0</b>	1.6E0
NAIVE	1.5E0	-

Table 194: Graph, relative MSE.  $T = 4, N = 64, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	3.7E-1	3.2E-1	3.3E-1	2.7E-1	
Q-REG	1.4E-1	2.1E-1	2.1E-1	2.1E-1	
MRDR	9.0E-2	1.8E-1	2.1E-1	2.0E-1	
FQE	<b>4.2E-2</b>	1.7E-1	1.8E-1	<b>1.4E-1</b>	
R( $\lambda$ )	1.2E-1	1.8E-1	1.9E-1	1.7E-1	
Q $^\pi$ ( $\lambda$ )	1.4E-1	1.8E-1	1.9E-1	1.8E-1	
TREE	1.4E-1	1.8E-1	1.9E-1	1.9E-1	
IH	6.5E-2	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	1.7E-1	1.7E-1
WIS	1.6E-1	<b>1.6E-1</b>
NAIVE	6.1E-1	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 195: Graph, relative MSE.  $T = 4, N = 128, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.0E-1	2.3E-1	2.3E-1	1.7E-1
Q-REG	1.1E-1	9.2E-2	<b>9.1E-2</b>	1.1E-1
MRDR	1.5E-1	1.1E-1	1.2E-1	1.2E-1
FQE	9.6E-2	9.6E-2	9.6E-2	9.5E-2
R( $\lambda$ )	1.0E-1	9.6E-2	9.6E-2	1.2E-1
Q $^\pi$ ( $\lambda$ )	1.1E-1	9.9E-2	9.8E-2	1.2E-1
TREE	1.2E-1	9.5E-2	9.5E-2	1.3E-1
IH	<b>7.5E-2</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.1E-1	1.1E-1
WIS	1.0E-1	<b>1.0E-1</b>
NAIVE	5.2E-1	-

Table 197: Graph, relative MSE.  $T = 4, N = 512, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.4E-2	3.2E-2	3.2E-2	5.4E-2
Q-REG	2.8E-2	1.7E-2	1.7E-2	2.6E-2
MRDR	2.5E-2	2.7E-2	2.9E-2	2.3E-2
FQE	<b>1.4E-2</b>	1.6E-2	1.7E-2	<b>1.4E-2</b>
R( $\lambda$ )	1.6E-2	1.7E-2	1.7E-2	1.6E-2
Q $^\pi$ ( $\lambda$ )	1.6E-2	1.7E-2	1.7E-2	1.7E-2
TREE	1.6E-2	1.7E-2	1.7E-2	1.5E-2
IH	2.3E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.7E-2	2.7E-2
WIS	<b>2.4E-2</b>	2.4E-2
NAIVE	5.2E-1	-

Table 196: Graph, relative MSE.  $T = 4, N = 256, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	8.3E-2	5.4E-2	5.3E-2	8.5E-2
Q-REG	4.7E-2	3.1E-2	3.2E-2	5.4E-2
MRDR	5.4E-2	3.2E-2	3.3E-2	6.3E-2
FQE	3.5E-2	3.0E-2	3.1E-2	3.4E-2
R( $\lambda$ )	3.4E-2	3.1E-2	3.0E-2	3.8E-2
Q $^\pi$ ( $\lambda$ )	<b>3.2E-2</b>	3.0E-2	3.1E-2	3.4E-2
TREE	3.7E-2	3.0E-2	<b>3.0E-2</b>	4.0E-2
IH	6.7E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>4.5E-2</b>	4.5E-2
WIS	4.6E-2	4.6E-2
NAIVE	6.2E-1	-

Table 198: Graph, relative MSE.  $T = 4, N = 1024, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	8.0E-3	2.8E-2	2.8E-2	2.1E-2
Q-REG	8.4E-3	8.3E-3	8.5E-3	7.9E-3
MRDR	1.3E-2	1.1E-2	1.2E-2	<b>5.9E-3</b>
FQE	9.3E-3	8.5E-3	8.6E-3	9.2E-3
R( $\lambda$ )	7.9E-3	8.5E-3	8.6E-3	6.6E-3
Q $^\pi$ ( $\lambda$ )	8.5E-3	8.5E-3	8.6E-3	6.1E-3
TREE	<b>7.7E-3</b>	8.5E-3	8.6E-3	7.2E-3
IH	7.8E-3	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>9.0E-3</b>	9.0E-3
WIS	9.1E-3	9.1E-3
NAIVE	5.1E-1	-

Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning

---

Table 199: Graph, relative MSE.  $T = 4, N = 8, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	2.4E1	2.5E1	2.4E1	1.7E1	
Q-REG	1.7E1	2.1E1	2.0E1	1.7E1	
MRDR	<b>9.7E0</b>	1.8E1	1.8E1	<b>9.7E0</b>	
FQE	1.7E1	1.8E1	1.8E1	1.7E1	
R( $\lambda$ )	1.8E1	1.9E1	1.8E1	1.8E1	
Q $^\pi$ ( $\lambda$ )	1.7E1	1.8E1	1.8E1	1.7E1	
TREE	1.9E1	1.9E1	1.8E1	1.9E1	
IH	1.8E1	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	<b>6.4E0</b>	1.9E1
WIS	6.9E0	1.9E1
NAIVE	1.3E1	-

Table 201: Graph, relative MSE.  $T = 4, N = 32, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	3.4E0	7.3E0	6.8E0	5.6E0	
Q-REG	1.9E0	1.5E0	1.5E0	1.9E0	
MRDR	1.2E0	1.5E0	1.5E0	1.2E0	
FQE	1.1E0	1.3E0	1.3E0	<b>1.2E0</b>	
R( $\lambda$ )	1.3E0	1.4E0	1.4E0	1.3E0	
Q $^\pi$ ( $\lambda$ )	1.4E0	1.4E0	1.4E0	1.4E0	
TREE	1.4E0	1.4E0	1.4E0	1.4E0	
IH	<b>7.4E-1</b>	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	1.7E0	1.7E0
WIS	1.5E0	<b>1.3E0</b>
NAIVE	1.2E0	-

Table 200: Graph, relative MSE.  $T = 4, N = 16, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	5.6E0	9.8E0	1.1E1	5.8E0	
Q-REG	4.7E0	3.3E0	3.3E0	4.7E0	
MRDR	3.1E0	3.0E0	3.0E0	3.1E0	
FQE	<b>2.1E0</b>	3.7E0	3.6E0	<b>2.1E0</b>	
R( $\lambda$ )	2.7E0	3.4E0	3.3E0	2.7E0	
Q $^\pi$ ( $\lambda$ )	3.6E0	4.0E0	3.8E0	3.6E0	
TREE	2.4E0	3.2E0	3.0E0	2.4E0	
IH	3.1E0	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	5.8E0	5.3E0
WIS	<b>4.8E0</b>	4.9E0
NAIVE	2.6E0	-

Table 202: Graph, relative MSE.  $T = 4, N = 64, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	9.7E-1	1.8E0	1.7E0	1.1E0	
Q-REG	1.6E0	1.9E0	1.9E0	1.1E0	
MRDR	8.0E-1	1.8E0	1.9E0	7.7E-1	
FQE	5.3E-1	1.9E0	1.9E0	<b>5.6E-1</b>	
R( $\lambda$ )	1.5E0	1.9E0	2.0E0	1.0E0	
Q $^\pi$ ( $\lambda$ )	1.4E0	1.9E0	1.9E0	1.0E0	
TREE	1.4E0	2.0E0	2.0E0	1.2E0	
IH	<b>4.9E-1</b>	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.6E0</b>	1.7E0
WIS	1.8E0	1.8E0
NAIVE	2.0E-1	-



**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 203: Graph, relative MSE.  $T = 4, N = 128, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	7.1E-1	1.2E0	1.2E0	7.7E-1
Q-REG	3.5E-1	4.5E-1	4.4E-1	2.9E-1
MRDR	2.5E-1	3.9E-1	4.0E-1	2.7E-1
FQE	2.1E-1	4.8E-1	4.6E-1	<b>2.2E-1</b>
R( $\lambda$ )	3.8E-1	4.7E-1	4.5E-1	3.6E-1
Q $^\pi$ ( $\lambda$ )	2.8E-1	4.6E-1	4.4E-1	3.2E-1
TREE	4.5E-1	4.7E-1	4.6E-1	4.1E-1
IH	<b>1.3E-1</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	4.8E-1	<b>3.3E-1</b>
WIS	5.1E-1	3.6E-1
NAIVE	5.5E-1	-

Table 205: Graph, relative MSE.  $T = 4, N = 512, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.7E-1	3.9E-1	4.0E-1	4.5E-1
Q-REG	2.5E-1	2.2E-1	2.2E-1	2.2E-1
MRDR	2.1E-1	2.3E-1	2.3E-1	1.9E-1
FQE	<b>1.4E-1</b>	2.1E-1	2.1E-1	<b>1.4E-1</b>
R( $\lambda$ )	1.9E-1	2.1E-1	2.1E-1	1.8E-1
Q $^\pi$ ( $\lambda$ )	2.1E-1	2.1E-1	2.1E-1	1.7E-1
TREE	2.0E-1	2.1E-1	2.1E-1	1.9E-1
IH	1.6E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.8E-1	2.3E-1
WIS	2.7E-1	<b>2.3E-1</b>
NAIVE	3.8E-1	-

Table 204: Graph, relative MSE.  $T = 4, N = 256, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.9E-1	5.4E-1	4.9E-1	2.2E-1
Q-REG	2.8E-1	2.7E-1	2.6E-1	2.2E-1
MRDR	2.0E-1	2.8E-1	2.8E-1	<b>1.9E-1</b>
FQE	2.0E-1	3.0E-1	2.8E-1	2.0E-1
R( $\lambda$ )	2.3E-1	2.8E-1	2.8E-1	1.9E-1
Q $^\pi$ ( $\lambda$ )	2.6E-1	2.9E-1	2.8E-1	2.0E-1
TREE	2.2E-1	2.8E-1	2.8E-1	2.0E-1
IH	<b>1.8E-1</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	3.3E-1	2.9E-1
WIS	3.2E-1	<b>2.8E-1</b>
NAIVE	6.2E-1	-

Table 206: Graph, relative MSE.  $T = 4, N = 1024, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.2E-1	1.1E-1	1.2E-1	7.1E-2
Q-REG	6.4E-2	6.6E-2	6.6E-2	6.0E-2
MRDR	5.1E-2	6.6E-2	6.7E-2	4.2E-2
FQE	3.6E-2	6.5E-2	6.6E-2	<b>3.6E-2</b>
R( $\lambda$ )	6.1E-2	6.6E-2	6.6E-2	6.2E-2
Q $^\pi$ ( $\lambda$ )	6.3E-2	6.7E-2	6.7E-2	6.0E-2
TREE	6.6E-2	6.5E-2	6.5E-2	6.8E-2
IH	<b>2.6E-2</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	9.3E-2	6.6E-2
WIS	8.9E-2	<b>6.2E-2</b>
NAIVE	4.8E-1	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 207: Graph, relative MSE.  $T = 16, N = 8, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.1E-1	2.2E-1	1.8E-1	<b>1.0E-1</b>
Q-REG	4.6E-1	2.0E-1	2.3E-1	2.8E-1
MRDR	3.4E-1	1.3E0	8.6E-1	8.1E-1
FQE	1.1E-1	1.1E-1	1.1E-1	1.1E-1
R( $\lambda$ )	1.1E-1	1.1E-1	1.1E-1	1.1E-1
Q $^\pi$ ( $\lambda$ )	1.1E-1	1.1E-1	1.1E-1	1.1E-1
TREE	1.8E-1	1.4E-1	1.2E-1	1.7E-1
IH	<b>3.1E-2</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.1E0	3.0E-1
WIS	1.3E-1	<b>9.6E-2</b>
NAIVE	6.2E-1	-

Table 209: Graph, relative MSE.  $T = 16, N = 32, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.0E-3	4.1E-2	2.7E-2	1.1E-3
Q-REG	7.7E-2	8.8E-3	2.6E-3	4.8E-2
MRDR	7.4E-2	2.4E-1	1.4E-1	1.1E-1
FQE	<b>5.0E-6</b>	5.0E-6	5.0E-6	5.0E-6
R( $\lambda$ )	7.0E-6	7.0E-6	7.0E-6	7.0E-6
Q $^\pi$ ( $\lambda$ )	5.0E-6	5.0E-6	5.0E-6	<b>5.0E-6</b>
TREE	5.6E-2	9.1E-3	6.5E-3	1.4E-2
IH	4.1E-3	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	4.2E-1	1.1E-1
WIS	3.1E-2	<b>6.7E-3</b>
NAIVE	4.1E-1	-

Table 208: Graph, relative MSE.  $T = 16, N = 16, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.3E-2	5.4E-2	5.8E-2	9.8E-3
Q-REG	1.7E-1	1.7E-1	1.3E-1	1.0E-1
MRDR	1.7E-1	8.9E-1	4.9E-1	2.9E-1
FQE	6.1E-3	6.1E-3	6.1E-3	6.1E-3
R( $\lambda$ )	<b>6.0E-3</b>	6.0E-3	<b>6.0E-3</b>	6.0E-3
Q $^\pi$ ( $\lambda$ )	6.1E-3	6.1E-3	6.1E-3	6.1E-3
TREE	7.9E-2	2.8E-2	1.9E-2	4.6E-2
IH	8.8E-3	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	4.7E-1	1.9E-1
WIS	7.5E-2	<b>2.4E-2</b>
NAIVE	4.6E-1	-

Table 210: Graph, relative MSE.  $T = 16, N = 64, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.3E-3	2.1E-2	2.3E-2	6.9E-4
Q-REG	3.9E-2	5.4E-3	5.3E-3	1.9E-2
MRDR	5.3E-2	1.5E-1	1.2E-1	5.3E-2
FQE	<b>5.5E-8</b>	5.5E-8	5.5E-8	5.5E-8
R( $\lambda$ )	1.9E-7	1.6E-7	1.5E-7	1.9E-7
Q $^\pi$ ( $\lambda$ )	5.6E-8	<b>5.5E-8</b>	5.5E-8	5.6E-8
TREE	5.9E-2	7.3E-3	3.6E-3	3.6E-3
IH	1.7E-3	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.4E-1	4.7E-2
WIS	1.3E-2	<b>4.3E-3</b>
NAIVE	4.7E-1	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 211: Graph, relative MSE.  $T = 16, N = 128, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.6E-4	7.1E-3	6.6E-3	2.6E-4
Q-REG	1.6E-2	3.6E-3	2.3E-3	2.7E-3
MRDR	2.0E-2	1.8E-2	1.5E-2	2.8E-2
FQE	3.0E-6	3.0E-6	3.0E-6	3.0E-6
R( $\lambda$ )	3.0E-6	3.0E-6	<b>3.0E-6</b>	3.0E-6
Q $^\pi$ ( $\lambda$ )	<b>3.0E-6</b>	3.0E-6	3.0E-6	3.0E-6
TREE	4.8E-2	2.6E-3	1.2E-3	1.2E-3
IH	6.7E-4	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	5.7E-2	2.0E-2
WIS	6.0E-3	<b>1.8E-3</b>
NAIVE	4.6E-1	-

Table 212: Graph, relative MSE.  $T = 16, N = 256, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.8E-4	1.5E-3	1.6E-3	6.5E-4
Q-REG	1.8E-2	2.3E-3	9.4E-4	9.9E-4
MRDR	1.4E-2	3.7E-2	2.5E-2	1.7E-2
FQE	1.0E-6	1.0E-6	1.0E-6	1.0E-6
R( $\lambda$ )	<b>1.0E-6</b>	1.0E-6	<b>1.0E-6</b>	1.0E-6
Q $^\pi$ ( $\lambda$ )	1.0E-6	1.0E-6	1.0E-6	1.0E-6
TREE	3.0E-3	2.6E-4	2.4E-4	2.4E-4
IH	3.4E-4	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.6E-1	2.0E-2
WIS	1.6E-2	<b>1.8E-3</b>
NAIVE	4.5E-1	-

Table 213: Graph, relative MSE.  $T = 16, N = 512, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.0E-5	3.3E-3	3.3E-3	1.5E-3
Q-REG	4.5E-3	7.7E-5	3.3E-5	4.5E-5
MRDR	5.0E-3	2.6E-3	2.7E-3	5.7E-3
FQE	7.7E-7	7.7E-7	7.7E-7	7.7E-7
R( $\lambda$ )	8.1E-7	8.0E-7	7.9E-7	8.1E-7
Q $^\pi$ ( $\lambda$ )	<b>7.6E-7</b>	7.7E-7	7.7E-7	<b>7.6E-7</b>
TREE	2.2E-3	1.2E-4	1.1E-4	1.1E-4
IH	1.7E-4	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	4.0E-2	4.7E-3
WIS	4.0E-3	<b>6.0E-4</b>
NAIVE	4.5E-1	-

Table 214: Graph, relative MSE.  $T = 16, N = 1024, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.8E-5	4.7E-4	4.8E-4	7.2E-5
Q-REG	2.1E-3	1.4E-5	1.0E-5	2.9E-5
MRDR	2.9E-3	5.0E-4	4.6E-4	9.6E-4
FQE	2.7E-7	2.7E-7	2.7E-7	2.7E-7
R( $\lambda$ )	2.9E-7	2.8E-7	2.8E-7	2.9E-7
Q $^\pi$ ( $\lambda$ )	<b>2.7E-7</b>	2.7E-7	2.7E-7	<b>2.7E-7</b>
TREE	1.8E-3	7.7E-5	5.8E-5	5.8E-5
IH	9.6E-5	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.7E-2	2.1E-3
WIS	1.8E-3	<b>2.9E-4</b>
NAIVE	4.4E-1	-

Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning

---

Table 215: Graph, relative MSE.  $T = 16, N = 8, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.5E-1	5.0E-1	4.3E-1	2.1E-1
Q-REG	4.3E-1	3.1E-1	2.6E-1	4.5E-1
MRDR	2.8E-1	1.1E0	4.5E-1	3.0E-1
FQE	1.3E-1	1.4E-1	1.3E-1	1.3E-1
R( $\lambda$ )	1.4E-1	1.5E-1	1.4E-1	1.4E-1
Q $^\pi$ ( $\lambda$ )	1.5E-1	1.6E-1	1.5E-1	1.5E-1
TREE	2.1E-1	1.2E-1	<b>7.0E-2</b>	2.1E-1
IH	<b>4.2E-2</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	4.9E-1	3.5E-1
WIS	1.1E-1	<b>6.0E-2</b>
NAIVE	4.4E-1	-

Table 217: Graph, relative MSE.  $T = 16, N = 32, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.0E-2	8.5E-2	7.5E-2	1.1E-2
Q-REG	7.3E-2	4.8E-2	4.6E-2	6.4E-2
MRDR	1.1E-1	2.0E-1	1.8E-1	1.3E-1
FQE	9.8E-3	2.6E-2	2.8E-2	<b>9.8E-3</b>
R( $\lambda$ )	2.9E-2	3.5E-2	3.6E-2	2.9E-2
Q $^\pi$ ( $\lambda$ )	2.9E-2	4.4E-2	4.1E-2	2.9E-2
TREE	1.1E-1	4.7E-2	3.3E-2	4.4E-2
IH	<b>8.4E-3</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	3.0E-1	6.2E-2
WIS	5.4E-2	<b>3.6E-2</b>
NAIVE	4.4E-1	-

Table 216: Graph, relative MSE.  $T = 16, N = 16, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.8E-2	3.2E-1	1.6E-1	7.6E-2
Q-REG	3.4E-1	4.3E-1	1.1E-1	3.4E-1
MRDR	2.4E-1	2.3E0	6.4E-1	2.4E-1
FQE	<b>1.4E-2</b>	3.6E-2	<b>1.0E-2</b>	1.4E-2
R( $\lambda$ )	1.6E-2	1.4E-2	1.6E-2	1.6E-2
Q $^\pi$ ( $\lambda$ )	2.8E-2	2.3E-2	1.9E-2	2.8E-2
TREE	6.7E-2	3.2E-2	1.7E-2	3.9E-2
IH	1.5E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.6E1	4.5E-1
WIS	1.5E-1	<b>3.2E-2</b>
NAIVE	4.3E-1	-

Table 218: Graph, relative MSE.  $T = 16, N = 64, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.1E-3	3.3E-2	2.1E-2	4.3E-3
Q-REG	4.7E-2	1.4E-2	1.1E-2	3.3E-2
MRDR	5.8E-2	4.7E-2	4.1E-2	7.2E-2
FQE	<b>4.1E-3</b>	1.3E-2	1.2E-2	<b>4.2E-3</b>
R( $\lambda$ )	5.8E-3	1.1E-2	1.1E-2	5.9E-3
Q $^\pi$ ( $\lambda$ )	6.7E-3	1.3E-2	1.2E-2	6.7E-3
TREE	5.5E-2	2.2E-2	1.9E-2	3.6E-2
IH	7.8E-3	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.7E-1	6.4E-2
WIS	6.9E-2	<b>2.4E-2</b>
NAIVE	4.1E-1	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 219: Graph, relative MSE.  $T = 16, N = 128, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	6.6E-3	2.6E-2	2.7E-2	6.3E-3
Q-REG	1.5E-2	5.1E-3	5.4E-3	6.8E-3
MRDR	2.0E-2	2.8E-2	2.5E-2	2.5E-2
FQE	<b>2.1E-3</b>	4.9E-3	4.9E-3	<b>2.1E-3</b>
R( $\lambda$ )	3.5E-3	4.3E-3	4.5E-3	3.5E-3
Q $^\pi$ ( $\lambda$ )	4.7E-3	5.0E-3	4.8E-3	4.6E-3
TREE	4.7E-2	4.9E-3	5.1E-3	8.9E-3
IH	3.3E-3	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	7.6E-2	2.3E-2
WIS	9.2E-3	<b>4.5E-3</b>
NAIVE	4.5E-1	-

Table 221: Graph, relative MSE.  $T = 16, N = 512, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	9.6E-4	5.6E-3	5.2E-3	3.6E-3
Q-REG	7.0E-3	8.2E-4	8.1E-4	9.0E-4
MRDR	4.1E-3	4.5E-3	4.3E-3	6.7E-3
FQE	<b>5.5E-4</b>	9.0E-4	8.6E-4	<b>5.6E-4</b>
R( $\lambda$ )	6.5E-4	8.1E-4	8.0E-4	6.6E-4
Q $^\pi$ ( $\lambda$ )	7.3E-4	8.8E-4	8.3E-4	7.1E-4
TREE	3.0E-3	7.8E-4	8.4E-4	1.8E-3
IH	7.5E-4	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	3.0E-2	8.1E-3
WIS	5.7E-3	<b>1.3E-3</b>
NAIVE	4.5E-1	-

Table 220: Graph, relative MSE.  $T = 16, N = 256, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.9E-3	1.2E-2	1.3E-2	4.5E-3
Q-REG	3.0E-2	3.1E-3	3.4E-3	6.3E-3
MRDR	2.7E-2	1.4E-2	1.2E-2	1.4E-2
FQE	<b>1.1E-3</b>	4.3E-3	4.1E-3	<b>1.1E-3</b>
R( $\lambda$ )	2.2E-3	3.7E-3	3.8E-3	2.3E-3
Q $^\pi$ ( $\lambda$ )	1.2E-3	3.8E-3	3.9E-3	1.3E-3
TREE	5.5E-3	4.2E-3	4.0E-3	5.4E-3
IH	2.1E-3	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.3E-1	3.5E-2
WIS	1.8E-2	<b>7.3E-3</b>
NAIVE	4.3E-1	-

Table 222: Graph, relative MSE.  $T = 16, N = 1024, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.9E-4	3.4E-3	3.7E-3	4.2E-4
Q-REG	1.8E-3	9.9E-4	1.0E-3	1.6E-3
MRDR	1.8E-3	1.1E-3	1.2E-3	1.4E-3
FQE	2.9E-4	9.4E-4	9.8E-4	<b>3.1E-4</b>
R( $\lambda$ )	6.8E-4	9.7E-4	9.9E-4	6.9E-4
Q $^\pi$ ( $\lambda$ )	5.4E-4	9.7E-4	1.0E-3	5.5E-4
TREE	1.9E-3	8.3E-4	8.9E-4	3.1E-4
IH	<b>1.7E-4</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.0E-2	1.1E-3
WIS	3.7E-3	<b>6.1E-4</b>
NAIVE	4.4E-1	-

Table 223: Graph, relative MSE.  $T = 16, N = 8, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.5E-1	7.2E-1	4.7E-1	<b>1.5E-1</b>
Q-REG	5.4E-1	5.1E-1	4.9E-1	5.4E-1
MRDR	4.1E-1	1.2E0	6.6E-1	4.2E-1
FQE	1.7E-1	3.8E-1	2.2E-1	1.7E-1
R( $\lambda$ )	1.7E-1	1.7E-1	1.7E-1	1.7E-1
Q $^\pi$ ( $\lambda$ )	1.9E-1	2.6E-1	2.0E-1	1.9E-1
TREE	2.4E-1	2.1E-1	2.1E-1	2.3E-1
IH	<b>1.5E-1</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	8.9E-1	4.6E-1
WIS	5.1E-1	<b>2.9E-1</b>
NAIVE	5.8E-1	-

Table 225: Graph, relative MSE.  $T = 16, N = 32, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.1E-2	9.7E-2	1.4E-1	3.5E-2
Q-REG	1.4E-1	1.0E-1	1.0E-1	1.4E-1
MRDR	1.4E-1	4.5E-1	3.6E-1	1.7E-1
FQE	2.0E-2	7.2E-2	5.7E-2	<b>2.0E-2</b>
R( $\lambda$ )	4.0E-2	3.8E-2	3.9E-2	4.1E-2
Q $^\pi$ ( $\lambda$ )	3.7E-2	6.4E-2	4.3E-2	3.7E-2
TREE	5.9E-2	5.1E-2	5.2E-2	6.2E-2
IH	<b>1.2E-2</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	6.5E-1	1.4E-1
WIS	7.6E-2	<b>7.5E-2</b>
NAIVE	4.2E-1	-

Table 224: Graph, relative MSE.  $T = 16, N = 16, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.2E-2	6.0E-1	4.5E-1	3.4E-2
Q-REG	1.8E-1	1.5E-1	2.2E-1	1.8E-1
MRDR	1.3E-1	1.2E0	1.1E0	2.3E-1
FQE	<b>2.3E-2</b>	8.7E-2	7.4E-2	<b>2.3E-2</b>
R( $\lambda$ )	3.2E-2	3.9E-2	4.3E-2	3.2E-2
Q $^\pi$ ( $\lambda$ )	6.0E-2	5.8E-2	6.2E-2	5.9E-2
TREE	1.0E-1	9.5E-2	8.1E-2	8.7E-2
IH	3.2E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	4.8E-1	1.6E-1
WIS	1.4E-1	<b>8.3E-2</b>
NAIVE	3.2E-1	-

Table 226: Graph, relative MSE.  $T = 16, N = 64, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.5E-2	9.3E-2	1.2E-1	2.1E-2
Q-REG	6.1E-2	4.5E-2	4.5E-2	6.4E-2
MRDR	7.1E-2	7.7E-2	6.9E-2	9.1E-2
FQE	1.3E-2	3.3E-2	4.1E-2	<b>1.3E-2</b>
R( $\lambda$ )	2.9E-2	3.4E-2	3.9E-2	3.1E-2
Q $^\pi$ ( $\lambda$ )	3.0E-2	2.9E-2	3.8E-2	3.0E-2
TREE	6.5E-2	3.6E-2	5.1E-2	6.3E-2
IH	<b>1.0E-2</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.5E-1	5.7E-2
WIS	7.1E-2	<b>4.5E-2</b>
NAIVE	4.5E-1	-

Table 227: Graph, relative MSE.  $T = 16, N = 128, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.3E-2	4.5E-2	3.7E-2	1.0E-2
Q-REG	2.9E-2	1.1E-2	1.2E-2	1.9E-2
MRDR	4.1E-2	2.9E-2	2.5E-2	5.2E-2
FQE	<b>6.5E-3</b>	9.9E-3	9.6E-3	6.7E-3
R( $\lambda$ )	6.7E-3	9.3E-3	9.6E-3	7.3E-3
Q $^\pi$ ( $\lambda$ )	8.7E-3	8.5E-3	9.2E-3	<b>5.9E-3</b>
TREE	5.4E-2	1.2E-2	1.1E-2	3.8E-2
IH	8.9E-3	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.6E-1	2.8E-2
WIS	6.2E-2	<b>1.2E-2</b>
NAIVE	4.5E-1	-

 Table 229: Graph, relative MSE.  $T = 16, N = 512, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	<b>8.8E-4</b>	1.2E-2	1.2E-2	1.7E-3
Q-REG	4.5E-3	5.9E-3	5.9E-3	5.7E-3
MRDR	4.1E-3	6.1E-3	6.2E-3	8.0E-3
FQE	1.1E-3	6.8E-3	6.4E-3	<b>1.2E-3</b>
R( $\lambda$ )	3.3E-3	5.9E-3	5.7E-3	3.3E-3
Q $^\pi$ ( $\lambda$ )	3.7E-3	7.1E-3	6.8E-3	3.8E-3
TREE	1.2E-2	5.7E-3	5.7E-3	1.2E-2
IH	9.3E-4	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	2.6E-2	<b>5.3E-3</b>
WIS	1.3E-2	6.3E-3
NAIVE	4.3E-1	-

 Table 228: Graph, relative MSE.  $T = 16, N = 256, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.0E-3	4.6E-2	5.3E-2	3.6E-2
Q-REG	1.3E-2	6.0E-3	6.2E-3	8.9E-3
MRDR	2.8E-2	1.7E-2	1.8E-2	3.0E-2
FQE	<b>2.5E-3</b>	7.1E-3	7.4E-3	<b>2.5E-3</b>
R( $\lambda$ )	7.0E-3	7.1E-3	7.0E-3	7.0E-3
Q $^\pi$ ( $\lambda$ )	3.8E-3	7.7E-3	7.9E-3	3.8E-3
TREE	3.4E-2	1.1E-2	1.1E-2	1.9E-2
IH	3.7E-3	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	5.2E-2	1.5E-2
WIS	1.7E-2	<b>1.2E-2</b>
NAIVE	4.8E-1	-

 Table 230: Graph, relative MSE.  $T = 16, N = 1024, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.1E-3	8.3E-3	8.7E-3	2.1E-3
Q-REG	3.8E-3	1.6E-3	1.6E-3	2.1E-3
MRDR	3.1E-3	1.6E-3	1.5E-3	3.0E-3
FQE	<b>7.7E-4</b>	1.7E-3	1.7E-3	<b>7.5E-4</b>
R( $\lambda$ )	1.6E-3	1.8E-3	1.8E-3	1.6E-3
Q $^\pi$ ( $\lambda$ )	9.0E-4	1.4E-3	1.5E-3	1.0E-3
TREE	2.8E-3	1.9E-3	2.0E-3	9.0E-4
IH	9.4E-4	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.9E-2	5.0E-3
WIS	5.4E-3	<b>1.7E-3</b>
NAIVE	4.3E-1	-

Table 231: Graph, relative MSE.  $T = 16, N = 8, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.4E-1	1.3E0	1.2E0	4.2E-1
Q-REG	6.3E-1	4.7E-1	<b>3.6E-1</b>	6.3E-1
MRDR	4.6E-1	2.6E0	7.8E-1	4.6E-1
FQE	3.6E-1	4.3E-1	4.6E-1	3.6E-1
R( $\lambda$ )	4.5E-1	4.7E-1	4.5E-1	4.5E-1
Q $^\pi$ ( $\lambda$ )	5.8E-1	6.4E-1	6.1E-1	5.8E-1
TREE	4.7E-1	3.9E-1	3.6E-1	4.7E-1
IH	<b>2.3E-1</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.0E0	7.5E-1
WIS	6.5E-1	<b>4.4E-1</b>
NAIVE	5.2E-1	-

Table 233: Graph, relative MSE.  $T = 16, N = 32, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	6.8E-2	7.0E-1	3.6E-1	4.9E-2
Q-REG	3.3E-1	8.7E-2	9.3E-2	2.5E-1
MRDR	2.5E-1	3.5E-1	3.2E-1	3.6E-1
FQE	<b>3.0E-2</b>	2.5E-1	1.6E-1	<b>3.0E-2</b>
R( $\lambda$ )	7.0E-2	1.1E-1	9.1E-2	7.0E-2
Q $^\pi$ ( $\lambda$ )	7.4E-2	3.0E-1	2.0E-1	7.4E-2
TREE	1.3E-1	8.5E-2	7.5E-2	1.3E-1
IH	5.5E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.7E0	3.5E-1
WIS	5.4E-1	<b>1.4E-1</b>
NAIVE	4.2E-1	-

Table 232: Graph, relative MSE.  $T = 16, N = 16, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.3E-1	8.6E-1	6.6E-1	1.7E-1
Q-REG	1.7E-1	4.6E-1	5.4E-1	2.8E-1
MRDR	2.3E-1	1.1E0	1.3E0	2.3E-1
FQE	1.5E-1	5.2E-1	3.2E-1	<b>1.5E-1</b>
R( $\lambda$ )	2.5E-1	2.8E-1	2.6E-1	2.5E-1
Q $^\pi$ ( $\lambda$ )	3.0E-1	5.6E-1	4.7E-1	3.0E-1
TREE	3.1E-1	2.3E-1	2.1E-1	3.1E-1
IH	<b>9.8E-2</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.5E0	1.8E-1
WIS	5.1E-1	<b>1.5E-1</b>
NAIVE	6.2E-1	-

Table 234: Graph, relative MSE.  $T = 16, N = 64, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.6E-2	2.9E-1	2.8E-1	3.4E-2
Q-REG	6.5E-2	3.7E-2	3.3E-2	4.4E-2
MRDR	7.7E-2	1.6E-1	1.2E-1	8.8E-2
FQE	1.6E-2	2.7E-2	2.1E-2	<b>1.6E-2</b>
R( $\lambda$ )	2.0E-2	2.6E-2	2.4E-2	2.0E-2
Q $^\pi$ ( $\lambda$ )	2.3E-2	3.8E-2	2.9E-2	2.0E-2
TREE	8.8E-2	2.4E-2	2.5E-2	8.6E-2
IH	<b>1.4E-2</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	6.6E-1	6.9E-2
WIS	1.3E-1	<b>2.2E-2</b>
NAIVE	4.8E-1	-



**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 235: Graph, relative MSE.  $T = 16, N = 128, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	6.0E-3	1.2E-1	1.2E-1	5.4E-3
Q-REG	1.2E-1	1.2E-2	1.0E-2	2.6E-2
MRDR	5.7E-2	2.0E-1	1.2E-1	5.5E-2
FQE	<b>2.6E-3</b>	5.2E-2	3.1E-2	<b>2.8E-3</b>
R( $\lambda$ )	1.2E-2	2.6E-2	2.0E-2	1.2E-2
Q $^\pi$ ( $\lambda$ )	1.1E-2	4.4E-2	2.6E-2	1.1E-2
TREE	4.5E-2	3.6E-2	2.8E-2	3.6E-2
IH	4.1E-3	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	3.0E-1	1.1E-1
WIS	4.4E-2	<b>4.1E-2</b>
NAIVE	4.5E-1	-

Table 237: Graph, relative MSE.  $T = 16, N = 512, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.3E-3	3.0E-2	3.2E-2	9.4E-3
Q-REG	1.4E-2	1.4E-2	1.3E-2	1.1E-2
MRDR	8.3E-3	6.6E-3	6.3E-3	9.8E-3
FQE	5.0E-3	1.2E-2	1.2E-2	<b>5.0E-3</b>
R( $\lambda$ )	8.8E-3	1.3E-2	1.3E-2	8.1E-3
Q $^\pi$ ( $\lambda$ )	1.3E-2	1.3E-2	1.3E-2	1.0E-2
TREE	9.4E-3	1.3E-2	1.3E-2	1.0E-2
IH	<b>4.2E-3</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.0E-1	1.5E-2
WIS	2.1E-2	<b>1.2E-2</b>
NAIVE	4.2E-1	-

Table 236: Graph, relative MSE.  $T = 16, N = 256, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.6E-3	4.6E-2	4.5E-2	6.2E-3
Q-REG	8.3E-3	1.4E-2	1.5E-2	1.1E-2
MRDR	1.3E-2	2.5E-2	2.3E-2	9.2E-3
FQE	3.8E-3	1.1E-2	1.3E-2	<b>3.5E-3</b>
R( $\lambda$ )	9.0E-3	1.3E-2	1.3E-2	9.5E-3
Q $^\pi$ ( $\lambda$ )	6.3E-3	1.4E-2	1.6E-2	6.4E-3
TREE	3.1E-2	1.5E-2	1.5E-2	1.9E-2
IH	<b>3.5E-3</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.4E-1	<b>1.1E-2</b>
WIS	6.6E-2	1.2E-2
NAIVE	4.4E-1	-

Table 238: Graph, relative MSE.  $T = 16, N = 1024, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.3E-3	1.1E-2	1.1E-2	3.8E-3
Q-REG	6.0E-3	5.0E-3	5.0E-3	5.6E-3
MRDR	5.2E-3	3.4E-3	3.8E-3	4.9E-3
FQE	1.1E-3	5.3E-3	5.6E-3	<b>1.1E-3</b>
R( $\lambda$ )	2.7E-3	5.1E-3	5.2E-3	2.4E-3
Q $^\pi$ ( $\lambda$ )	2.0E-3	5.0E-3	5.3E-3	1.5E-3
TREE	1.3E-2	5.0E-3	5.4E-3	8.8E-3
IH	<b>9.8E-4</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	9.2E-3	5.4E-3
WIS	7.3E-3	<b>5.4E-3</b>
NAIVE	4.6E-1	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 239: Graph, relative MSE.  $T = 16, N = 8, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	6.5E-1	1.3E0	3.6E0	5.7E-1
Q-REG	1.7E0	3.2E0	5.5E0	1.7E0
MRDR	8.7E-1	6.3E0	5.8E0	9.5E-1
FQE	4.3E-1	4.3E-1	4.3E-1	4.8E-1
R( $\lambda$ )	4.6E-1	4.5E-1	4.4E-1	5.1E-1
Q $^\pi$ ( $\lambda$ )	<b>4.3E-1</b>	<b>4.3E-1</b>	4.3E-1	4.8E-1
TREE	9.8E-1	7.1E-1	6.6E-1	8.5E-1
IH	6.9E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	1.0E0
WIS	<b>7.0E-1</b>	7.0E-1
NAIVE	2.6E-1	-

Table 241: Graph, relative MSE.  $T = 16, N = 32, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.4E-2	5.0E-1	3.4E-1	1.2E-1
Q-REG	1.6E0	5.3E-1	1.7E-1	6.3E-1
MRDR	1.1E0	8.7E0	3.6E0	1.8E0
FQE	3.6E-3	3.6E-3	<b>3.6E-3</b>	1.0E-1
R( $\lambda$ )	4.6E-3	4.5E-3	4.5E-3	1.0E-1
Q $^\pi$ ( $\lambda$ )	<b>3.6E-3</b>	3.6E-3	3.6E-3	1.0E-1
TREE	9.1E-1	8.5E-1	1.0E-1	5.3E-1
IH	6.1E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.9E0	1.9E0
WIS	<b>1.6E-1</b>	1.6E-1
NAIVE	3.3E-1	-

Table 240: Graph, relative MSE.  $T = 16, N = 16, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.2E-1	2.0E0	1.3E0	2.2E-1
Q-REG	5.6E-1	1.8E0	4.8E-1	7.8E-1
MRDR	4.4E-1	2.2E0	1.9E0	4.2E-1
FQE	<b>4.6E-2</b>	4.6E-2	4.6E-2	1.4E-1
R( $\lambda$ )	4.6E-2	4.6E-2	<b>4.5E-2</b>	1.4E-1
Q $^\pi$ ( $\lambda$ )	4.6E-2	4.6E-2	4.6E-2	1.4E-1
TREE	9.9E-1	5.7E-1	1.6E-1	8.7E-1
IH	6.2E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	6.2E-1	6.2E-1
WIS	<b>1.7E-1</b>	1.7E-1
NAIVE	3.5E-1	-

Table 242: Graph, relative MSE.  $T = 16, N = 64, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.3E-2	3.2E-1	3.4E-1	1.3E-1
Q-REG	1.8E-1	1.1E-1	9.6E-2	2.1E-1
MRDR	1.8E-1	7.2E-1	8.0E-1	4.9E-1
FQE	2.0E-6	2.0E-6	2.0E-6	1.0E-1
R( $\lambda$ )	4.4E-5	3.2E-5	4.2E-5	1.0E-1
Q $^\pi$ ( $\lambda$ )	<b>2.0E-6</b>	2.0E-6	<b>2.0E-6</b>	1.0E-1
TREE	9.7E-1	1.8E-1	9.3E-2	6.1E-1
IH	6.2E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.8E-1	1.8E-1
WIS	<b>9.1E-2</b>	9.1E-2
NAIVE	4.7E-1	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 243: Graph, relative MSE.  $T = 16, N = 128, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	8.8E-3	6.9E-1	6.1E-1	3.1E-1
Q-REG	1.7E-1	4.7E-2	2.4E-2	1.1E-1
MRDR	1.5E-1	4.0E-1	2.9E-1	3.1E-1
FQE	<b>2.2E-7</b>	2.2E-7	2.2E-7	2.2E-7
R( $\lambda$ )	1.2E-5	9.7E-6	1.0E-5	1.2E-5
Q $^\pi$ ( $\lambda$ )	2.3E-7	2.2E-7	<b>2.2E-7</b>	2.3E-7
TREE	9.4E-1	1.6E-1	7.0E-2	1.9E-1
IH	6.1E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.2E-1	2.2E-1
WIS	<b>7.6E-2</b>	7.6E-2
NAIVE	5.7E-1	-

Table 245: Graph, relative MSE.  $T = 16, N = 512, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.2E-3	6.4E-2	5.8E-2	3.4E-3
Q-REG	4.8E-2	2.4E-3	1.7E-3	4.4E-3
MRDR	3.3E-2	5.0E-2	6.0E-2	6.9E-2
FQE	2.2E-5	2.2E-5	2.2E-5	2.2E-5
R( $\lambda$ )	2.3E-5	2.0E-5	<b>2.0E-5</b>	2.3E-5
Q $^\pi$ ( $\lambda$ )	<b>2.2E-5</b>	2.2E-5	2.2E-5	2.2E-5
TREE	9.6E-1	4.5E-2	2.2E-2	2.2E-2
IH	3.8E-3	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	5.3E-2	5.3E-2
WIS	2.5E-2	<b>2.5E-2</b>
NAIVE	4.3E-1	-

Table 244: Graph, relative MSE.  $T = 16, N = 256, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.5E-3	1.7E-1	1.6E-1	9.2E-2
Q-REG	2.2E-1	4.6E-3	2.1E-3	2.1E-2
MRDR	1.7E-1	1.5E-1	1.2E-1	2.2E-1
FQE	9.0E-6	9.0E-6	9.0E-6	9.0E-6
R( $\lambda$ )	2.3E-5	2.3E-5	2.6E-5	2.3E-5
Q $^\pi$ ( $\lambda$ )	<b>9.0E-6</b>	9.0E-6	9.0E-6	<b>9.0E-6</b>
TREE	9.5E-1	2.0E-1	1.3E-1	1.6E-1
IH	1.3E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.3E-1	2.3E-1
WIS	1.3E-1	<b>1.3E-1</b>
NAIVE	4.9E-1	-

Table 246: Graph, relative MSE.  $T = 16, N = 1024, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.3E-3	3.0E-2	2.6E-2	8.3E-4
Q-REG	2.4E-2	5.7E-4	3.2E-4	1.1E-3
MRDR	2.0E-2	1.6E-2	1.1E-2	1.1E-2
FQE	<b>2.5E-5</b>	2.5E-5	2.5E-5	2.5E-5
R( $\lambda$ )	2.5E-5	2.7E-5	2.8E-5	2.5E-5
Q $^\pi$ ( $\lambda$ )	2.5E-5	<b>2.5E-5</b>	2.5E-5	2.5E-5
TREE	1.0E0	2.5E-2	1.4E-2	1.4E-2
IH	2.0E-3	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.5E-2	2.5E-2
WIS	<b>1.4E-2</b>	1.4E-2
NAIVE	4.5E-1	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 247: Graph, relative MSE.  $T = 16, N = 8, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	8.4E0	9.8E1	4.9E1	8.9E0	
Q-REG	1.5E1	2.1E1	1.5E1	1.5E1	
MRDR	6.6E0	1.2E1	8.7E0	6.6E0	
FQE	<b>5.8E0</b>	1.1E1	1.1E1	<b>5.8E0</b>	
R( $\lambda$ )	8.3E0	8.3E0	8.6E0	8.3E0	
Q $^\pi$ ( $\lambda$ )	1.0E1	1.1E1	9.3E0	1.0E1	
TREE	7.8E0	9.3E0	1.0E1	7.8E0	
IH	7.8E0	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	1.1E1	1.5E1
WIS	2.0E1	<b>1.0E1</b>
NAIVE	6.0E0	-

Table 249: Graph, relative MSE.  $T = 16, N = 32, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	5.9E0	2.1E1	2.1E1	6.5E0	
Q-REG	1.4E0	7.3E0	4.9E0	1.4E0	
MRDR	<b>1.0E0</b>	8.8E0	4.6E0	<b>1.1E0</b>	
FQE	5.2E0	7.4E0	3.7E0	5.2E0	
R( $\lambda$ )	3.3E0	4.1E0	3.3E0	3.3E0	
Q $^\pi$ ( $\lambda$ )	3.8E0	9.6E0	3.6E0	3.8E0	
TREE	4.0E0	2.1E0	2.4E0	4.0E0	
IH	4.6E0	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	6.8E0	<b>1.5E0</b>
WIS	9.5E0	1.9E0
NAIVE	2.3E0	-

Table 248: Graph, relative MSE.  $T = 16, N = 16, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	6.4E0	5.9E1	4.9E1	1.3E1	
Q-REG	2.8E0	3.8E0	3.2E0	2.8E0	
MRDR	<b>1.9E0</b>	1.9E0	<b>1.1E0</b>	1.9E0	
FQE	5.9E0	1.1E1	9.5E0	5.9E0	
R( $\lambda$ )	6.3E0	5.5E0	6.1E0	6.3E0	
Q $^\pi$ ( $\lambda$ )	7.1E0	8.0E0	7.9E0	7.1E0	
TREE	3.8E0	3.7E0	4.7E0	4.1E0	
IH	2.9E0	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	4.1E0	<b>3.2E0</b>
WIS	9.0E0	3.2E0
NAIVE	4.5E0	-

Table 250: Graph, relative MSE.  $T = 16, N = 64, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	1.8E0	1.3E1	1.2E1	3.7E0	
Q-REG	2.3E0	2.6E0	2.5E0	2.2E0	
MRDR	1.7E0	2.9E0	2.8E0	1.6E0	
FQE	<b>1.1E0</b>	2.8E0	2.5E0	<b>1.1E0</b>	
R( $\lambda$ )	1.2E0	2.0E0	2.0E0	1.2E0	
Q $^\pi$ ( $\lambda$ )	1.7E0	2.6E0	2.6E0	1.7E0	
TREE	1.4E0	2.1E0	2.0E0	1.4E0	
IH	1.9E0	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	7.0E0	2.3E0
WIS	7.7E0	<b>2.0E0</b>
NAIVE	2.0E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 251: Graph, relative MSE.  $T = 16, N = 128, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	9.5E-1	1.9E0	1.5E0	1.2E0
Q-REG	9.8E-1	1.4E0	1.3E0	9.2E-1
MRDR	1.1E0	2.8E0	1.5E0	<b>7.3E-1</b>
FQE	9.1E-1	1.5E0	1.1E0	8.8E-1
R( $\lambda$ )	9.9E-1	1.2E0	1.0E0	9.9E-1
Q $^\pi$ ( $\lambda$ )	<b>7.5E-1</b>	1.3E0	1.1E0	7.4E-1
TREE	1.2E0	1.1E0	1.0E0	9.4E-1
IH	1.5E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	3.7E0	1.1E0
WIS	4.0E0	<b>8.9E-1</b>
NAIVE	1.2E0	-

Table 253: Graph, relative MSE.  $T = 16, N = 512, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.7E-1	1.2E0	1.2E0	2.2E-1
Q-REG	3.2E-1	2.5E-1	2.4E-1	3.4E-1
MRDR	3.9E-1	2.8E-1	2.7E-1	3.8E-1
FQE	<b>1.0E-1</b>	2.2E-1	2.3E-1	<b>9.9E-2</b>
R( $\lambda$ )	2.1E-1	2.3E-1	2.4E-1	2.1E-1
Q $^\pi$ ( $\lambda$ )	1.9E-1	2.2E-1	2.3E-1	1.8E-1
TREE	9.8E-1	3.0E-1	2.7E-1	4.3E-1
IH	1.1E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.4E0	3.3E-1
WIS	1.2E0	<b>2.9E-1</b>
NAIVE	6.3E-1	-

Table 252: Graph, relative MSE.  $T = 16, N = 256, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.6E-1	4.4E0	4.2E0	2.1E0
Q-REG	1.1E0	1.4E0	1.3E0	1.3E0
MRDR	9.4E-1	7.8E-1	7.8E-1	8.2E-1
FQE	3.8E-1	1.3E0	1.1E0	<b>3.8E-1</b>
R( $\lambda$ )	6.6E-1	1.1E0	1.1E0	6.7E-1
Q $^\pi$ ( $\lambda$ )	8.0E-1	1.4E0	1.2E0	7.9E-1
TREE	1.0E0	9.8E-1	9.4E-1	9.4E-1
IH	<b>3.6E-1</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	3.1E0	1.1E0
WIS	2.5E0	<b>1.0E0</b>
NAIVE	7.1E-1	-

Table 254: Graph, relative MSE.  $T = 16, N = 1024, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.2E-1	9.3E-1	9.2E-1	<b>5.7E-2</b>
Q-REG	4.0E-1	4.2E-1	4.1E-1	4.3E-1
MRDR	2.9E-1	3.2E-1	3.3E-1	3.6E-1
FQE	<b>1.0E-1</b>	4.0E-1	4.0E-1	1.6E-1
R( $\lambda$ )	1.7E-1	4.0E-1	4.0E-1	1.8E-1
Q $^\pi$ ( $\lambda$ )	1.5E-1	4.0E-1	4.0E-1	1.5E-1
TREE	7.1E-1	3.7E-1	4.1E-1	2.5E-1
IH	1.0E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	8.7E-1	<b>3.8E-1</b>
WIS	9.6E-1	4.2E-1
NAIVE	3.5E-1	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 255: Graph, relative MSE.  $T = 16, N = 8, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.0E0	1.4E1	1.5E1	1.5E0
Q-REG	1.9E1	5.1E1	4.1E1	3.0E0
MRDR	3.0E0	2.7E1	2.0E1	<b>8.9E-1</b>
FQE	1.1E0	6.4E0	4.8E0	1.3E0
R( $\lambda$ )	1.2E0	1.2E0	1.3E0	1.2E0
Q $^\pi$ ( $\lambda$ )	2.0E0	4.3E0	3.0E0	1.4E0
TREE	1.0E0	4.0E0	3.8E0	1.3E0
IH	<b>6.5E-1</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	8.7E0	8.7E0
WIS	<b>5.2E0</b>	5.2E0
NAIVE	1.7E0	-

Table 257: Graph, relative MSE.  $T = 16, N = 32, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.8E-1	1.5E1	1.0E1	<b>3.8E-1</b>
Q-REG	2.1E0	3.5E0	2.7E0	2.8E0
MRDR	1.6E0	4.5E0	3.1E0	3.4E0
FQE	<b>4.4E-1</b>	1.9E0	2.0E0	4.5E-1
R( $\lambda$ )	1.4E0	1.7E0	1.7E0	1.4E0
Q $^\pi$ ( $\lambda$ )	9.7E-1	1.9E0	1.6E0	1.0E0
TREE	9.4E-1	1.9E0	2.7E0	9.4E-1
IH	7.4E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>2.1E0</b>	2.1E0
WIS	2.8E0	2.8E0
NAIVE	1.7E0	-

Table 256: Graph, relative MSE.  $T = 16, N = 16, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	8.3E-1	5.5E0	3.7E0	1.3E0
Q-REG	8.0E0	3.6E1	1.9E1	7.8E0
MRDR	6.4E0	1.4E2	5.2E1	6.4E0
FQE	<b>3.8E-1</b>	1.9E0	1.8E0	<b>3.6E-1</b>
R( $\lambda$ )	1.1E0	1.4E0	1.4E0	1.0E0
Q $^\pi$ ( $\lambda$ )	1.4E0	1.1E0	1.4E0	1.1E0
TREE	8.9E-1	4.9E0	2.4E0	8.5E-1
IH	6.8E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.2E1	1.2E1
WIS	2.9E0	<b>2.9E0</b>
NAIVE	1.3E0	-

Table 258: Graph, relative MSE.  $T = 16, N = 64, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.2E-1	3.3E0	2.5E0	2.3E-1
Q-REG	1.0E0	3.8E-1	4.0E-1	9.3E-1
MRDR	7.3E-1	2.2E0	2.0E0	8.4E-1
FQE	<b>7.8E-2</b>	5.6E-1	5.6E-1	<b>7.9E-2</b>
R( $\lambda$ )	2.2E-1	3.7E-1	3.6E-1	2.3E-1
Q $^\pi$ ( $\lambda$ )	3.8E-1	3.6E-1	3.4E-1	2.9E-1
TREE	1.0E0	9.9E-1	8.7E-1	7.8E-1
IH	5.1E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	1.0E0
WIS	<b>8.7E-1</b>	8.7E-1
NAIVE	5.9E-1	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 259: Graph, relative MSE.  $T = 16, N = 128, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.3E-1	4.4E0	4.4E0	1.4E0
Q-REG	9.4E-1	7.4E-1	6.7E-1	9.4E-1
MRDR	7.9E-1	3.9E0	1.6E0	6.7E-1
FQE	<b>6.3E-2</b>	1.2E0	8.2E-1	<b>6.7E-2</b>
$R(\lambda)$	3.6E-1	7.2E-1	6.3E-1	3.7E-1
$Q^\pi(\lambda)$	2.5E-1	1.1E0	8.5E-1	2.2E-1
TREE	9.6E-1	8.0E-1	6.8E-1	8.1E-1
IH	5.3E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	9.0E-1	9.0E-1
WIS	<b>7.3E-1</b>	7.3E-1
NAIVE	6.2E-1	-

Table 261: Graph, relative MSE.  $T = 16, N = 512, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	6.2E-2	4.7E-1	3.9E-1	5.3E-2
Q-REG	2.4E-1	1.4E-1	1.3E-1	1.5E-1
MRDR	1.8E-1	4.8E-1	4.2E-1	2.5E-1
FQE	3.5E-2	1.8E-1	1.5E-1	<b>3.6E-2</b>
$R(\lambda)$	8.3E-2	1.5E-1	1.4E-1	8.4E-2
$Q^\pi(\lambda)$	1.0E-1	1.8E-1	1.5E-1	8.0E-2
TREE	9.9E-1	2.5E-1	2.2E-1	3.7E-1
IH	<b>3.4E-2</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.5E-1	2.5E-1
WIS	<b>2.2E-1</b>	2.2E-1
NAIVE	4.8E-1	-

Table 260: Graph, relative MSE.  $T = 16, N = 256, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	<b>3.3E-2</b>	9.9E-1	9.3E-1	1.1E-1
Q-REG	4.5E-1	3.4E-1	3.2E-1	2.9E-1
MRDR	5.2E-1	5.6E-1	5.1E-1	6.8E-1
FQE	3.6E-2	3.9E-1	3.2E-1	<b>3.3E-2</b>
$R(\lambda)$	1.6E-1	3.0E-1	2.9E-1	1.7E-1
$Q^\pi(\lambda)$	1.3E-1	3.7E-1	3.2E-1	1.4E-1
TREE	1.0E0	4.5E-1	4.0E-1	7.1E-1
IH	3.5E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	4.5E-1	4.5E-1
WIS	<b>4.0E-1</b>	4.0E-1
NAIVE	4.0E-1	-

Table 262: Graph, relative MSE.  $T = 16, N = 1024, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	<b>6.8E-3</b>	1.3E-1	1.2E-1	1.3E-2
Q-REG	1.2E-1	7.4E-2	7.1E-2	1.1E-1
MRDR	1.2E-1	6.5E-2	6.6E-2	8.8E-2
FQE	1.1E-2	6.6E-2	6.4E-2	<b>1.2E-2</b>
$R(\lambda)$	3.2E-2	6.8E-2	6.8E-2	3.3E-2
$Q^\pi(\lambda)$	2.9E-2	7.1E-2	6.8E-2	2.6E-2
TREE	1.0E0	1.2E-1	8.1E-2	8.1E-2
IH	1.7E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.2E-1	1.2E-1
WIS	<b>8.1E-2</b>	8.1E-2
NAIVE	4.4E-1	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 263: Graph, relative MSE.  $T = 16, N = 8, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	4.8E1	2.2E2	1.2E2	5.4E1	
Q-REG	1.3E2	3.2E2	2.2E2	1.3E2	
MRDR	6.7E1	4.4E2	2.0E2	6.2E1	
FQE	<b>2.6E1</b>	1.1E2	5.2E1	<b>2.6E1</b>	
R( $\lambda$ )	4.2E1	5.2E1	4.0E1	4.2E1	
Q $^\pi$ ( $\lambda$ )	3.8E1	8.1E1	3.6E1	3.8E1	
TREE	3.3E1	7.7E1	6.3E1	4.0E1	
IH	5.8E1	-	-	-	

IPS		
	STANDARD	PER-DECISION
IS	1.3E2	1.3E2
WIS	1.5E2	<b>8.6E1</b>
NAIVE	2.8E1	-

Table 265: Graph, relative MSE.  $T = 16, N = 32, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	5.2E0	1.5E2	1.1E2	4.8E0	
Q-REG	6.5E1	5.4E1	3.7E1	6.5E1	
MRDR	4.3E1	3.2E2	1.2E2	4.3E1	
FQE	3.9E0	3.8E1	2.2E1	<b>3.9E0</b>	
R( $\lambda$ )	9.8E0	1.5E1	1.3E1	9.8E0	
Q $^\pi$ ( $\lambda$ )	1.4E1	3.1E1	1.8E1	1.4E1	
TREE	2.3E1	2.6E1	1.8E1	2.2E1	
IH	<b>3.8E0</b>	-	-	-	

IPS		
	STANDARD	PER-DECISION
IS	5.5E1	6.3E1
WIS	3.6E1	<b>2.7E1</b>
NAIVE	5.0E0	-

Table 264: Graph, relative MSE.  $T = 16, N = 16, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	2.2E1	1.0E2	1.0E2	2.2E1	
Q-REG	5.3E1	4.9E1	4.8E1	5.6E1	
MRDR	2.6E1	7.5E1	4.5E1	2.0E1	
FQE	2.2E1	4.0E1	4.9E1	2.1E1	
R( $\lambda$ )	3.0E1	3.5E1	3.6E1	3.0E1	
Q $^\pi$ ( $\lambda$ )	6.0E1	7.3E1	7.4E1	6.0E1	
TREE	<b>1.4E1</b>	1.5E1	2.5E1	<b>1.4E1</b>	
IH	1.7E1	-	-	-	

IPS		
	STANDARD	PER-DECISION
IS	1.1E2	<b>4.3E1</b>
WIS	9.0E1	4.3E1
NAIVE	1.5E1	-

Table 266: Graph, relative MSE.  $T = 16, N = 64, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	6.6E0	3.5E1	3.5E1	6.5E0	
Q-REG	5.7E0	1.4E1	1.5E1	6.0E0	
MRDR	5.9E0	5.2E0	5.1E0	4.8E0	
FQE	5.3E0	8.5E0	9.7E0	5.2E0	
R( $\lambda$ )	7.9E0	8.3E0	8.5E0	7.8E0	
Q $^\pi$ ( $\lambda$ )	7.1E0	8.7E0	9.4E0	6.5E0	
TREE	<b>3.0E0</b>	5.8E0	7.5E0	<b>3.0E0</b>	
IH	3.8E0	-	-	-	

IPS		
	STANDARD	PER-DECISION
IS	1.3E1	<b>6.6E0</b>
WIS	2.0E1	8.6E0
NAIVE	5.9E0	-



**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 267: Graph, relative MSE.  $T = 16, N = 128, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	6.1E0	2.7E1	3.0E1	3.8E0	
Q-REG	<b>2.5E0</b>	6.2E0	5.2E0	2.4E0	
MRDR	3.9E0	6.3E0	5.8E0	<b>1.6E0</b>	
FQE	2.7E0	3.5E0	3.3E0	2.6E0	
R( $\lambda$ )	3.1E0	3.7E0	3.5E0	2.9E0	
Q $^\pi$ ( $\lambda$ )	5.1E0	5.0E0	4.4E0	5.1E0	
TREE	3.1E0	2.4E0	2.3E0	2.9E0	
IH	3.2E0	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	3.7E0	2.7E0
WIS	4.7E0	<b>2.6E0</b>
NAIVE	1.9E0	-

Table 269: Graph, relative MSE.  $T = 16, N = 512, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	8.5E-1	1.0E1	1.0E1	5.1E0	
Q-REG	1.4E0	1.2E0	1.3E0	1.3E0	
MRDR	7.9E-1	1.5E0	1.4E0	7.6E-1	
FQE	4.4E-1	1.1E0	1.1E0	<b>4.3E-1</b>	
R( $\lambda$ )	4.3E-1	1.1E0	1.1E0	4.4E-1	
Q $^\pi$ ( $\lambda$ )	5.0E-1	1.1E0	1.1E0	4.6E-1	
TREE	1.2E0	1.0E0	1.1E0	7.9E-1	
IH	<b>4.2E-1</b>	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	6.1E0	<b>1.3E0</b>
WIS	5.5E0	1.4E0
NAIVE	9.9E-1	-

Table 268: Graph, relative MSE.  $T = 16, N = 256, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	1.5E0	7.6E0	8.0E0	1.6E0	
Q-REG	2.1E0	3.0E0	3.0E0	1.7E0	
MRDR	1.9E0	1.7E0	1.8E0	1.7E0	
FQE	<b>4.5E-1</b>	2.2E0	2.5E0	<b>4.8E-1</b>	
R( $\lambda$ )	1.8E0	2.7E0	2.8E0	1.8E0	
Q $^\pi$ ( $\lambda$ )	1.4E0	2.3E0	2.5E0	1.3E0	
TREE	1.2E0	3.0E0	3.1E0	1.1E0	
IH	4.5E-1	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	3.4E0	<b>2.2E0</b>
WIS	4.4E0	2.6E0
NAIVE	2.4E-1	-

Table 270: Graph, relative MSE.  $T = 16, N = 1024, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	5.5E-1	9.0E-1	8.9E-1	6.2E-1	
Q-REG	5.9E-1	5.3E-1	5.6E-1	5.7E-1	
MRDR	6.7E-1	5.7E-1	6.2E-1	6.5E-1	
FQE	3.6E-1	5.9E-1	5.9E-1	<b>3.4E-1</b>	
R( $\lambda$ )	5.9E-1	6.0E-1	5.8E-1	5.9E-1	
Q $^\pi$ ( $\lambda$ )	4.2E-1	5.8E-1	5.6E-1	4.3E-1	
TREE	1.5E0	6.0E-1	5.8E-1	1.4E0	
IH	<b>3.2E-1</b>	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	5.8E0	5.9E-1
WIS	6.2E0	<b>5.6E-1</b>
NAIVE	8.9E-1	-

## F.2 Detailed Results for Graph-POMDP

 Table 271: Graph-POMDP, relative MSE.  $T = 2, N = 256, H = 2, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	9.4E-1	1.2E-1	3.7E-2	3.7E-2
Q-REG	1.5E-1	1.1E-2	3.5E-3	1.4E-2
MRDR	1.4E0	1.2E-2	3.3E-2	3.3E-2
FQE	8.7E-1	2.3E-2	2.7E-3	5.7E-2
R( $\lambda$ )	5.1E-1	1.0E-2	<b>2.4E-3</b>	3.2E-2
Q $^\pi$ ( $\lambda$ )	<b>5.3E-2</b>	9.2E-3	2.8E-3	2.9E-2
TREE	3.8E-1	8.2E-3	2.4E-3	2.2E-2
IH	8.4E-1	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.3E-1	4.5E-2
WIS	1.5E-2	<b>6.0E-3</b>
NAIVE	3.8E0	-

 Table 272: Graph-POMDP, relative MSE.  $T = 2, N = 512, H = 2, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	9.6E-1	3.8E-2	1.3E-2	1.3E-2
Q-REG	1.6E-1	1.4E-3	1.4E-3	8.7E-3
MRDR	1.5E0	1.4E-3	2.8E-2	2.8E-2
FQE	9.5E-1	9.7E-3	1.3E-3	1.3E-3
R( $\lambda$ )	5.6E-1	3.3E-3	9.9E-4	6.1E-3
Q $^\pi$ ( $\lambda$ )	<b>5.5E-2</b>	5.9E-3	1.7E-3	<b>9.7E-4</b>
TREE	4.2E-1	2.3E-3	1.0E-3	2.4E-2
IH	9.4E-1	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	6.3E-2	2.0E-2
WIS	1.5E-2	<b>4.3E-3</b>
NAIVE	4.0E0	-

 Table 273: Graph-POMDP, relative MSE.  $T = 2, N = 1024, H = 2, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	9.7E-1	6.7E-3	2.4E-3	2.4E-3
Q-REG	1.7E-1	3.1E-3	<b>1.2E-3</b>	4.3E-3
MRDR	1.5E0	3.0E-3	5.5E-3	5.5E-3
FQE	1.0E0	7.8E-3	1.6E-3	1.6E-3
R( $\lambda$ )	6.3E-1	5.0E-3	1.4E-3	1.4E-3
Q $^\pi$ ( $\lambda$ )	<b>9.0E-2</b>	3.3E-3	1.3E-3	1.3E-3
TREE	4.9E-1	4.5E-3	1.3E-3	1.4E-3
IH	1.0E0	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.6E-2	1.0E-2
WIS	3.3E-3	<b>2.5E-3</b>
NAIVE	4.0E0	-

 Table 274: Graph-POMDP, relative MSE.  $T = 2, N = 256, H = 2, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.1E0	1.9E-1	1.5E-1	4.5E-1
Q-REG	2.8E-1	5.7E-2	4.4E-2	<b>3.4E-2</b>
MRDR	1.7E0	5.4E-2	1.2E-1	1.2E-1
FQE	1.0E0	8.1E-2	4.8E-2	2.9E-1
R( $\lambda$ )	6.7E-1	6.4E-2	4.4E-2	1.5E-1
Q $^\pi$ ( $\lambda$ )	<b>1.0E-1</b>	5.7E-2	4.3E-2	7.4E-2
TREE	5.3E-1	6.2E-2	4.4E-2	1.1E-1
IH	1.0E0	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	9.8E-2	6.0E-2
WIS	9.7E-2	<b>5.3E-2</b>
NAIVE	4.0E0	-

Table 275: Graph-POMDP, relative MSE.  $T = 2, N = 512, H = 2, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.1E0	4.7E-2	3.0E-2	6.2E-2
Q-REG	2.0E-1	1.4E-2	1.1E-2	<b>6.8E-3</b>
MRDR	1.5E0	1.6E-2	3.9E-2	3.9E-2
FQE	1.2E0	1.4E-2	8.1E-3	8.1E-3
$R(\lambda)$	7.3E-1	9.8E-3	8.5E-3	6.3E-2
$Q^\pi(\lambda)$	<b>1.3E-1</b>	1.2E-2	9.1E-3	2.1E-2
TREE	5.9E-1	9.5E-3	8.5E-3	6.1E-2
IH	1.2E0	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	5.5E-2	1.7E-2
WIS	2.7E-2	<b>8.6E-3</b>
NAIVE	4.1E0	-

 Table 277: Graph-POMDP, relative MSE.  $T = 2, N = 256, H = 2, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.6E0	6.5E-1	5.5E-1	1.5E0
Q-REG	5.7E-1	3.1E-1	2.9E-1	3.5E-1
MRDR	2.3E0	3.0E-1	2.1E-1	<b>2.1E-1</b>
FQE	1.7E0	4.2E-1	3.3E-1	8.5E-1
$R(\lambda)$	1.2E0	3.7E-1	3.1E-1	6.8E-1
$Q^\pi(\lambda)$	<b>4.0E-1</b>	2.9E-1	3.0E-1	3.1E-1
TREE	1.1E0	3.5E-1	3.1E-1	6.3E-1
IH	1.8E0	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	8.2E-1	4.1E-1
WIS	7.3E-1	<b>3.6E-1</b>
NAIVE	4.6E0	-

 Table 276: Graph-POMDP, relative MSE.  $T = 2, N = 1024, H = 2, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	9.6E-1	4.2E-2	2.6E-2	2.6E-2
Q-REG	1.7E-1	1.1E-2	1.1E-2	2.1E-2
MRDR	1.5E0	<b>1.0E-2</b>	1.4E-2	1.4E-2
FQE	1.0E0	2.1E-2	1.2E-2	1.2E-2
$R(\lambda)$	6.1E-1	1.6E-2	1.1E-2	1.3E-2
$Q^\pi(\lambda)$	<b>6.4E-2</b>	1.0E-2	1.1E-2	2.4E-2
TREE	4.7E-1	1.4E-2	1.1E-2	2.1E-2
IH	1.0E0	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	5.0E-2	2.3E-2
WIS	3.3E-2	<b>1.4E-2</b>
NAIVE	4.0E0	-

 Table 278: Graph-POMDP, relative MSE.  $T = 2, N = 512, H = 2, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.1E0	1.2E-1	1.1E-1	4.8E-1
Q-REG	2.7E-1	5.5E-2	5.7E-2	6.1E-2
MRDR	1.8E0	5.6E-2	6.8E-2	6.8E-2
FQE	1.2E0	5.7E-2	5.6E-2	3.6E-1
$R(\lambda)$	7.8E-1	5.4E-2	5.6E-2	1.8E-1
$Q^\pi(\lambda)$	<b>1.2E-1</b>	5.4E-2	5.6E-2	<b>3.6E-2</b>
TREE	6.3E-1	5.3E-2	5.6E-2	1.2E-1
IH	1.2E0	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.5E-1	5.9E-2
WIS	1.4E-1	<b>5.6E-2</b>
NAIVE	4.3E0	-

Table 279: Graph-POMDP, relative MSE.  $T = 2, N = 1024, H = 2, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.0E0	1.4E-1	6.8E-2	6.8E-2
Q-REG	2.8E-1	6.7E-2	6.4E-2	7.1E-2
MRDR	1.7E0	6.6E-2	7.6E-2	7.6E-2
FQE	1.1E0	1.1E-1	7.3E-2	2.3E-1
R( $\lambda$ )	6.9E-1	8.6E-2	6.7E-2	1.4E-1
Q $^\pi$ ( $\lambda$ )	<b>9.0E-2</b>	<b>6.1E-2</b>	6.4E-2	9.5E-2
TREE	5.4E-1	8.1E-2	6.7E-2	9.0E-2
IH	1.1E0	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	3.0E-1	1.2E-1
WIS	2.3E-1	<b>8.3E-2</b>
NAIVE	4.0E0	-

 Table 280: Graph-POMDP, relative MSE.  $T = 2, N = 256, H = 2, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	7.4E-1	2.8E-1	2.6E-1	8.4E-1
Q-REG	1.8E-1	1.7E-1	1.6E-1	1.5E-1
MRDR	1.4E0	1.7E-1	2.6E-1	2.6E-1
FQE	6.5E-1	1.8E-1	1.5E-1	1.2E-1
R( $\lambda$ )	3.5E-1	1.8E-1	1.5E-1	<b>9.7E-2</b>
Q $^\pi$ ( $\lambda$ )	<b>1.2E-1</b>	2.3E-1	1.7E-1	4.9E-1
TREE	2.4E-1	1.8E-1	1.5E-1	1.3E-1
IH	6.8E-1	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.9E-1	1.8E-1
WIS	2.4E-1	<b>1.4E-1</b>
NAIVE	3.7E0	-

 Table 281: Graph-POMDP, relative MSE.  $T = 2, N = 512, H = 2, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.2E0	4.5E-1	4.7E-1	6.5E-1
Q-REG	2.2E-1	7.6E-2	7.0E-2	8.5E-2
MRDR	1.7E0	7.9E-2	1.0E-1	1.0E-1
FQE	1.1E0	<b>6.3E-2</b>	6.4E-2	2.8E-1
R( $\lambda$ )	6.8E-1	6.7E-2	6.7E-2	2.1E-1
Q $^\pi$ ( $\lambda$ )	<b>1.1E-1</b>	7.8E-2	6.8E-2	1.8E-1
TREE	5.5E-1	6.8E-2	6.7E-2	1.8E-1
IH	1.1E0	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.0E-1	6.6E-2
WIS	9.4E-2	<b>6.0E-2</b>
NAIVE	4.3E0	-

 Table 282: Graph-POMDP, relative MSE.  $T = 2, N = 1024, H = 2, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.0E0	1.3E-1	1.5E-1	4.7E-1
Q-REG	1.9E-1	7.4E-2	6.9E-2	7.2E-2
MRDR	1.5E0	7.2E-2	9.2E-2	9.2E-2
FQE	1.0E0	6.5E-2	<b>6.3E-2</b>	2.6E-1
R( $\lambda$ )	6.3E-1	6.7E-2	6.5E-2	1.8E-1
Q $^\pi$ ( $\lambda$ )	<b>1.4E-1</b>	8.0E-2	6.8E-2	1.0E-1
TREE	5.0E-1	6.9E-2	6.5E-2	1.5E-1
IH	1.0E0	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.0E-1	<b>5.4E-2</b>
WIS	9.0E-2	5.9E-2
NAIVE	4.0E0	-

Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning

---

Table 283: Graph-POMDP, relative MSE.  $T = 2, N = 256, H = 2, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.9E0	3.0E-1	7.8E-2	7.8E-2
Q-REG	<b>1.2E-1</b>	1.5E-2	5.8E-3	3.1E-2
MRDR	1.0E0	<b>5.3E-3</b>	3.2E-2	3.2E-2
FQE	3.8E0	2.7E-1	2.6E-2	2.6E-2
$R(\lambda)$	1.8E0	1.4E-1	1.8E-2	1.8E-2
$Q^\pi(\lambda)$	2.6E-1	2.5E-2	8.4E-3	1.1E-1
TREE	1.8E0	1.4E-1	1.9E-2	1.9E-2
IH	3.8E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.2E-1	1.2E-1
WIS	<b>2.6E-2</b>	2.6E-2
NAIVE	3.9E0	-

Table 285: Graph-POMDP, relative MSE.  $T = 2, N = 1024, H = 2, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.0E0	5.5E-2	1.7E-2	1.7E-2
Q-REG	<b>1.0E-1</b>	2.4E-3	<b>9.6E-4</b>	1.2E-2
MRDR	1.1E0	1.3E-3	1.2E-2	1.2E-2
FQE	3.9E0	7.6E-2	5.7E-3	5.7E-3
$R(\lambda)$	1.8E0	3.7E-2	3.8E-3	3.8E-3
$Q^\pi(\lambda)$	2.7E-1	7.3E-3	1.6E-3	1.6E-3
TREE	1.8E0	3.8E-2	4.1E-3	4.1E-3
IH	3.9E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	3.3E-2	3.3E-2
WIS	<b>5.8E-3</b>	5.8E-3
NAIVE	3.9E0	-

Table 284: Graph-POMDP, relative MSE.  $T = 2, N = 512, H = 2, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.0E0	5.6E-2	4.2E-2	4.2E-2
Q-REG	<b>1.1E-1</b>	6.1E-3	<b>5.0E-3</b>	3.7E-2
MRDR	1.1E0	5.1E-3	8.5E-3	8.5E-3
FQE	4.0E0	3.7E-2	7.4E-3	7.4E-3
$R(\lambda)$	1.9E0	2.1E-2	6.5E-3	6.5E-3
$Q^\pi(\lambda)$	3.8E-1	8.4E-3	5.5E-3	2.6E-2
TREE	1.9E0	2.2E-2	6.6E-3	6.6E-3
IH	4.1E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.7E-2	1.7E-2
WIS	<b>7.4E-3</b>	7.4E-3
NAIVE	4.1E0	-

Table 286: Graph-POMDP, relative MSE.  $T = 2, N = 256, H = 2, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.0E0	4.9E-1	3.5E-1	7.6E-1
Q-REG	3.4E-1	9.8E-2	8.9E-2	2.6E-1
MRDR	1.6E0	9.5E-2	1.6E-1	1.6E-1
FQE	4.2E0	3.0E-1	9.9E-2	3.3E-1
$R(\lambda)$	2.1E0	1.8E-1	9.1E-2	6.2E-1
$Q^\pi(\lambda)$	<b>3.1E-1</b>	8.9E-2	<b>8.8E-2</b>	2.2E-1
TREE	2.1E0	1.8E-1	9.1E-2	6.3E-1
IH	4.2E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	8.2E-2	1.3E-1
WIS	<b>8.0E-2</b>	9.5E-2
NAIVE	3.8E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 287: Graph-POMDP, relative MSE.  $T = 2, N = 512, H = 2, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.7E0	3.3E-1	2.2E-1	2.2E-1
Q-REG	<b>2.3E-1</b>	9.5E-2	9.9E-2	3.2E-1
MRDR	1.5E0	1.0E-1	1.4E-1	1.4E-1
FQE	4.7E0	1.0E-1	9.0E-2	9.0E-2
R( $\lambda$ )	2.2E0	8.7E-2	9.3E-2	9.3E-2
Q $^\pi$ ( $\lambda$ )	2.7E-1	9.2E-2	9.7E-2	2.4E-1
TREE	2.2E0	<b>8.7E-2</b>	9.2E-2	9.2E-2
IH	4.5E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	9.1E-2	<b>8.7E-2</b>
WIS	1.1E-1	9.0E-2
NAIVE	4.5E0	-

Table 289: Graph-POMDP, relative MSE.  $T = 2, N = 256, H = 2, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.8E0	1.2E0	6.4E-1	2.0E0
Q-REG	3.7E-1	2.7E-1	2.6E-1	3.4E-1
MRDR	1.1E0	<b>2.5E-1</b>	2.8E-1	2.8E-1
FQE	3.8E0	8.6E-1	3.3E-1	3.3E-1
R( $\lambda$ )	1.7E0	6.0E-1	3.1E-1	3.4E-1
Q $^\pi$ ( $\lambda$ )	<b>2.5E-1</b>	3.9E-1	2.9E-1	2.5E-1
TREE	1.7E0	6.1E-1	3.2E-1	3.4E-1
IH	3.7E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	6.6E-1	6.6E-1
WIS	<b>3.4E-1</b>	3.4E-1
NAIVE	3.9E0	-

Table 288: Graph-POMDP, relative MSE.  $T = 2, N = 1024, H = 2, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.0E0	2.4E-1	1.8E-1	1.8E-1
Q-REG	<b>1.8E-1</b>	5.5E-2	5.2E-2	1.3E-1
MRDR	1.2E0	<b>5.1E-2</b>	5.4E-2	5.4E-2
FQE	4.0E0	1.2E-1	5.9E-2	5.9E-2
R( $\lambda$ )	1.9E0	8.8E-2	5.6E-2	5.6E-2
Q $^\pi$ ( $\lambda$ )	2.7E-1	6.2E-2	5.3E-2	1.4E-1
TREE	1.9E0	8.9E-2	5.7E-2	5.7E-2
IH	4.0E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E-1	8.4E-2
WIS	7.7E-2	<b>5.8E-2</b>
NAIVE	3.9E0	-

Table 290: Graph-POMDP, relative MSE.  $T = 2, N = 512, H = 2, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.1E0	4.3E-1	3.5E-1	1.3E0
Q-REG	<b>2.2E-1</b>	8.7E-2	9.7E-2	2.1E-1
MRDR	1.3E0	1.1E-1	2.0E-1	2.0E-1
FQE	4.1E0	1.6E-1	8.0E-2	8.0E-2
R( $\lambda$ )	1.9E0	1.0E-1	8.2E-2	3.8E-1
Q $^\pi$ ( $\lambda$ )	3.1E-1	<b>7.9E-2</b>	9.1E-2	2.9E-1
TREE	1.9E0	1.0E-1	8.2E-2	5.3E-1
IH	4.1E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	9.1E-2	9.1E-2
WIS	7.9E-2	<b>7.9E-2</b>
NAIVE	3.9E0	-

Table 291: Graph-POMDP, relative MSE.  $T = 2, N = 1024, H = 2, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.2E0	2.6E-1	2.4E-1	2.4E-1
Q-REG	<b>1.5E-1</b>	1.3E-1	1.3E-1	1.5E-1
MRDR	1.0E0	1.3E-1	1.5E-1	1.5E-1
FQE	4.0E0	1.6E-1	1.3E-1	1.3E-1
R( $\lambda$ )	1.9E0	1.4E-1	1.3E-1	1.3E-1
Q $^\pi$ ( $\lambda$ )	3.1E-1	<b>1.2E-1</b>	1.3E-1	2.8E-1
TREE	1.9E0	1.4E-1	1.3E-1	1.3E-1
IH	4.1E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.3E-1	1.3E-1
WIS	<b>1.3E-1</b>	1.3E-1
NAIVE	3.9E0	-

 Table 292: Graph-POMDP, relative MSE.  $T = 2, N = 256, H = 2, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.1E0	1.9E0	1.7E0	2.9E0
Q-REG	8.9E-1	1.2E0	1.4E0	1.5E0
MRDR	1.1E0	1.4E0	1.7E0	1.6E0
FQE	3.5E0	1.3E0	1.3E0	2.4E0
R( $\lambda$ )	1.7E0	1.2E0	1.3E0	2.3E0
Q $^\pi$ ( $\lambda$ )	<b>8.2E-1</b>	<b>1.2E0</b>	1.4E0	1.5E0
TREE	1.8E0	1.2E0	1.3E0	1.5E0
IH	3.6E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.6E0	1.3E0
WIS	2.8E0	<b>1.3E0</b>
NAIVE	4.0E0	-

 Table 293: Graph-POMDP, relative MSE.  $T = 2, N = 512, H = 2, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.5E0	1.3E0	1.1E0	3.6E0
Q-REG	<b>1.1E-1</b>	1.7E-1	1.9E-1	1.9E-1
MRDR	9.0E-1	1.9E-1	2.8E-1	4.1E-1
FQE	4.3E0	2.6E-1	1.7E-1	1.7E-1
R( $\lambda$ )	1.8E0	1.9E-1	1.7E-1	8.7E-1
Q $^\pi$ ( $\lambda$ )	2.6E-1	<b>1.6E-1</b>	1.8E-1	3.4E-1
TREE	1.9E0	1.9E-1	1.7E-1	8.7E-1
IH	4.3E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	5.0E-1	2.4E-1
WIS	4.0E-1	<b>1.7E-1</b>
NAIVE	4.4E0	-

 Table 294: Graph-POMDP, relative MSE.  $T = 2, N = 1024, H = 2, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.5E0	5.7E-1	5.1E-1	1.7E0
Q-REG	<b>4.2E-1</b>	2.7E-1	2.7E-1	4.3E-1
MRDR	1.7E0	2.6E-1	<b>2.4E-1</b>	3.1E-1
FQE	4.4E0	3.4E-1	2.8E-1	2.8E-1
R( $\lambda$ )	2.2E0	3.1E-1	2.8E-1	1.0E0
Q $^\pi$ ( $\lambda$ )	4.7E-1	2.8E-1	2.7E-1	5.2E-1
TREE	2.2E0	3.1E-1	2.8E-1	1.0E0
IH	4.5E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	5.6E-1	3.2E-1
WIS	5.1E-1	<b>2.8E-1</b>
NAIVE	4.0E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 295: Graph-POMDP, relative MSE.  $T = 16, N = 256, H = 6, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.0E-1	2.4E1	1.0E-1	9.6E-2
Q-REG	1.7E0	1.1E1	1.6E-1	1.8E0
MRDR	1.6E0	1.3E1	5.6E1	5.6E1
FQE	1.4E-2	1.9E0	6.3E-3	4.9E-3
R( $\lambda$ )	2.7E-2	4.9E0	1.3E-1	1.3E-1
Q $^\pi$ ( $\lambda$ )	<b>7.1E-3</b>	3.2E0	7.5E-3	<b>2.5E-3</b>
TREE	8.4E-3	1.0E1	1.9E-1	1.9E-1
IH	1.0E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	9.1E-1	4.3E0
WIS	8.8E-1	<b>2.8E-1</b>
NAIVE	4.0E0	-

Table 297: Graph-POMDP, relative MSE.  $T = 16, N = 1024, H = 6, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.1E-1	1.6E-1	8.1E-2	9.2E-2
Q-REG	2.6E-1	3.7E-1	6.6E-2	1.5E-1
MRDR	2.0E-1	2.6E-1	1.2E0	1.2E0
FQE	1.4E-2	2.9E-2	9.0E-3	4.1E-3
R( $\lambda$ )	2.0E-2	1.3E-1	9.4E-2	9.8E-2
Q $^\pi$ ( $\lambda$ )	5.8E-3	2.6E-2	7.4E-3	<b>6.2E-4</b>
TREE	<b>2.0E-3</b>	1.6E-1	1.0E-1	1.0E-1
IH	1.1E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.1E0	2.1E-1
WIS	7.0E-1	<b>1.6E-1</b>
NAIVE	4.0E0	-

Table 296: Graph-POMDP, relative MSE.  $T = 16, N = 512, H = 6, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.1E-1	1.3E0	1.1E-1	1.1E-1
Q-REG	1.5E0	3.7E-1	9.0E-1	1.4E0
MRDR	5.4E-1	1.1E0	5.3E0	5.3E0
FQE	1.9E-2	1.1E0	2.0E-2	8.7E-3
R( $\lambda$ )	3.1E-2	1.8E0	1.3E-1	1.3E-1
Q $^\pi$ ( $\lambda$ )	9.1E-3	8.0E-1	1.2E-2	<b>1.7E-3</b>
TREE	<b>6.3E-3</b>	1.8E0	1.4E-1	1.4E-1
IH	1.4E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	3.9E0	2.0E0
WIS	9.4E-1	<b>2.4E-1</b>
NAIVE	4.0E0	-

Table 298: Graph-POMDP, relative MSE.  $T = 16, N = 256, H = 6, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.3E-1	2.2E-1	1.6E-1	1.0E-1
Q-REG	4.5E-1	1.4E-1	1.3E-1	3.9E-1
MRDR	4.2E-1	3.5E-1	2.9E0	2.9E0
FQE	2.1E-2	2.8E-2	6.1E-2	<b>1.1E-2</b>
R( $\lambda$ )	5.6E-2	1.5E-1	2.0E-1	2.0E-1
Q $^\pi$ ( $\lambda$ )	2.0E-2	1.8E-2	5.1E-2	1.2E-2
TREE	2.4E-2	2.0E-1	2.2E-1	2.2E-1
IH	<b>1.7E-2</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	9.2E-1	3.8E-1
WIS	1.2E0	<b>3.4E-1</b>
NAIVE	4.0E0	-



Table 299: Graph-POMDP, relative MSE.  $T = 16, N = 512, H = 6, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.1E-1	1.9E-1	2.0E-1	1.5E-1
Q-REG	4.2E-1	1.2E-1	2.0E-1	3.3E-1
MRDR	3.6E-1	3.3E-1	2.1E0	2.0E0
FQE	1.8E-2	3.5E-2	7.0E-2	7.4E-3
R( $\lambda$ )	2.8E-2	1.5E-1	1.9E-1	2.0E-1
Q $^\pi$ ( $\lambda$ )	<b>8.5E-3</b>	2.7E-2	6.7E-2	<b>3.0E-3</b>
TREE	1.3E-2	2.2E-1	2.0E-1	2.0E-1
IH	1.5E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	9.9E-1	3.4E-1
WIS	1.1E0	<b>2.9E-1</b>
NAIVE	4.0E0	-

 Table 301: Graph-POMDP, relative MSE.  $T = 16, N = 256, H = 6, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.5E-1	1.9E0	4.0E-1	1.9E-1
Q-REG	3.3E-1	2.1E-1	5.7E-1	2.4E-1
MRDR	3.1E-1	4.6E-1	7.9E0	7.9E0
FQE	4.1E-2	2.9E-1	1.6E-1	<b>2.7E-2</b>
R( $\lambda$ )	4.9E-2	4.7E-1	3.5E-1	9.8E-2
Q $^\pi$ ( $\lambda$ )	6.0E-2	2.2E-1	1.6E-1	4.8E-2
TREE	<b>2.7E-2</b>	7.1E-1	3.7E-1	9.0E-2
IH	3.8E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.0E0	5.6E-1
WIS	2.1E0	<b>4.5E-1</b>
NAIVE	4.1E0	-

 Table 300: Graph-POMDP, relative MSE.  $T = 16, N = 1024, H = 6, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.2E-1	6.2E0	3.8E-1	1.7E-1
Q-REG	5.5E-1	3.2E0	1.3E-1	2.4E-1
MRDR	8.5E-1	3.5E-1	3.1E0	3.1E0
FQE	1.5E-2	9.2E-1	5.0E-2	7.0E-3
R( $\lambda$ )	5.6E-2	1.8E0	2.1E-1	2.1E-1
Q $^\pi$ ( $\lambda$ )	1.1E-2	7.0E-1	4.3E-2	<b>4.2E-3</b>
TREE	1.5E-2	2.3E0	2.7E-1	2.7E-1
IH	<b>1.1E-2</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	9.6E-1	5.7E-1
WIS	1.0E0	<b>3.1E-1</b>
NAIVE	4.0E0	-

 Table 302: Graph-POMDP, relative MSE.  $T = 16, N = 512, H = 6, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.3E-1	1.5E-1	6.2E-1	3.5E-1
Q-REG	3.1E-1	1.8E-1	1.1E-1	2.5E-1
MRDR	3.7E-1	4.8E-1	2.3E0	2.4E0
FQE	2.7E-2	1.4E-1	1.7E-1	<b>2.3E-2</b>
R( $\lambda$ )	6.0E-2	1.4E-1	4.1E-1	1.8E-1
Q $^\pi$ ( $\lambda$ )	4.1E-2	2.0E-1	1.7E-1	3.0E-2
TREE	3.7E-2	1.8E-1	4.2E-1	1.9E-1
IH	<b>2.1E-2</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.0E0	<b>2.7E-1</b>
WIS	1.5E0	5.2E-1
NAIVE	4.0E0	-

Table 303: Graph-POMDP, relative MSE.  $T = 16, N = 1024, H = 6, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.2E-1	4.7E-1	3.4E-1	1.8E-1
Q-REG	4.6E-1	1.1E-1	1.5E-1	3.9E-1
MRDR	7.4E-1	3.3E-1	5.8E0	5.7E0
FQE	1.9E-2	5.2E-2	1.1E-1	<b>8.6E-3</b>
$R(\lambda)$	2.7E-2	1.2E-1	1.9E-1	2.0E-1
$Q^\pi(\lambda)$	2.2E-2	6.0E-2	9.6E-2	1.0E-2
TREE	<b>9.0E-3</b>	2.3E-1	1.9E-1	1.9E-1
IH	1.3E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.0E0	3.3E-1
WIS	1.0E0	<b>2.8E-1</b>
NAIVE	4.0E0	-

 Table 304: Graph-POMDP, relative MSE.  $T = 16, N = 256, H = 6, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.9E-1	1.6E0	4.9E-1	3.4E-1
Q-REG	6.5E-1	2.7E-1	3.4E-1	4.8E-1
MRDR	9.1E-1	7.5E-1	2.3E0	2.3E0
FQE	<b>1.3E-2</b>	2.7E-1	2.1E-1	<b>9.4E-3</b>
$R(\lambda)$	3.6E-2	2.9E-1	2.8E-1	1.8E-1
$Q^\pi(\lambda)$	3.9E-2	2.6E-1	2.2E-1	3.8E-2
TREE	3.2E-2	3.3E-1	3.2E-1	2.5E-1
IH	1.3E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	9.8E-1	8.4E-1
WIS	1.1E0	<b>4.2E-1</b>
NAIVE	4.0E0	-

 Table 305: Graph-POMDP, relative MSE.  $T = 16, N = 512, H = 6, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.1E-1	1.5E0	5.9E-1	3.1E-1
Q-REG	4.6E-1	2.4E-1	6.2E-1	7.0E-1
MRDR	3.6E-1	7.8E-1	8.9E0	9.0E0
FQE	3.7E-2	3.0E-1	2.5E-1	5.8E-2
$R(\lambda)$	4.4E-2	4.4E-1	4.3E-1	1.8E-1
$Q^\pi(\lambda)$	<b>3.0E-2</b>	2.5E-1	2.4E-1	<b>4.8E-2</b>
TREE	3.9E-2	5.1E-1	4.5E-1	2.3E-1
IH	3.2E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	9.0E-1	5.9E-1
WIS	1.2E0	<b>5.6E-1</b>
NAIVE	4.0E0	-

 Table 306: Graph-POMDP, relative MSE.  $T = 16, N = 1024, H = 6, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.2E-1	3.1E-1	1.2E0	3.1E-1
Q-REG	4.1E-1	3.4E-1	4.6E-1	3.9E-1
MRDR	7.7E-1	3.9E-1	2.6E0	2.7E0
FQE	2.0E-2	4.8E-2	3.1E-1	3.4E-2
$R(\lambda)$	4.7E-2	1.2E-1	4.4E-1	8.2E-2
$Q^\pi(\lambda)$	3.6E-2	<b>3.3E-2</b>	2.8E-1	5.0E-2
TREE	2.2E-2	1.4E-1	4.4E-1	2.1E-1
IH	<b>1.6E-2</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.0E0	<b>3.2E-1</b>
WIS	1.7E0	5.3E-1
NAIVE	4.0E0	-

Table 307: Graph-POMDP, relative MSE.  $T = 16, N = 256, H = 6, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	1.5E0	5.6E0	5.8E0	1.6E0	
Q-REG	8.9E-1	6.1E-1	1.7E0	1.2E0	
MRDR	7.5E-1	1.5E0	1.2E2	1.2E2	
FQE	4.3E0	3.9E0	1.7E0	4.1E0	
R( $\lambda$ )	1.0E0	9.2E-1	1.6E0	1.4E0	
Q $^\pi$ ( $\lambda$ )	<b>1.6E-1</b>	<b>9.3E-2</b>	1.2E0	1.6E-1	
TREE	1.0E0	9.2E-1	1.6E0	1.4E0	
IH	4.3E0	-	-	-	

IPS		
	STANDARD	PER-DECISION
IS	<b>9.2E-1</b>	9.2E-1
WIS	1.6E0	1.6E0
NAIVE	4.1E0	-

Table 309: Graph-POMDP, relative MSE.  $T = 16, N = 1024, H = 6, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	1.5E0	3.2E0	8.7E0	1.7E0	
Q-REG	9.8E-1	7.0E-1	2.0E0	9.2E-1	
MRDR	1.0E0	1.2E0	1.7E1	1.6E1	
FQE	4.1E0	3.8E0	2.6E0	3.2E0	
R( $\lambda$ )	1.0E0	9.6E-1	2.6E0	1.8E0	
Q $^\pi$ ( $\lambda$ )	<b>3.8E-2</b>	3.8E-2	1.2E0	<b>3.2E-2</b>	
TREE	1.0E0	9.6E-1	2.6E0	1.8E0	
IH	4.0E0	-	-	-	

IPS		
	STANDARD	PER-DECISION
IS	<b>9.6E-1</b>	9.6E-1
WIS	2.6E0	2.6E0
NAIVE	4.1E0	-

Table 308: Graph-POMDP, relative MSE.  $T = 16, N = 512, H = 6, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	1.5E0	1.7E0	5.4E0	1.6E0	
Q-REG	9.9E-1	4.1E0	1.1E0	9.1E-1	
MRDR	9.6E-1	1.2E0	4.9E1	4.9E1	
FQE	4.1E0	3.6E0	1.8E0	3.8E0	
R( $\lambda$ )	1.0E0	1.1E0	1.8E0	1.4E0	
Q $^\pi$ ( $\lambda$ )	<b>9.6E-2</b>	2.3E-1	9.7E-1	<b>8.7E-2</b>	
TREE	1.0E0	1.1E0	1.8E0	1.4E0	
IH	4.0E0	-	-	-	

IPS		
	STANDARD	PER-DECISION
IS	<b>1.1E0</b>	1.1E0
WIS	1.8E0	1.8E0
NAIVE	4.0E0	-

Table 310: Graph-POMDP, relative MSE.  $T = 16, N = 256, H = 6, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	<b>9.3E-1</b>	8.7E1	3.8E1	<b>1.7E0</b>	
Q-REG	8.0E1	4.4E1	2.2E1	3.6E1	
MRDR	4.0E2	5.3E0	1.8E2	2.1E2	
FQE	2.4E0	4.8E1	1.1E1	2.4E0	
R( $\lambda$ )	2.0E0	3.3E1	1.2E1	2.1E0	
Q $^\pi$ ( $\lambda$ )	2.9E0	3.7E1	1.1E1	2.5E0	
TREE	2.0E0	4.1E1	1.1E1	2.2E0	
IH	2.2E0	-	-	-	

IPS		
	STANDARD	PER-DECISION
IS	<b>8.2E-1</b>	3.5E1
WIS	4.5E1	1.1E1
NAIVE	3.3E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 311: Graph-POMDP, relative MSE.  $T = 16, N = 512, H = 6, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	<b>1.2E0</b>	1.2E2	6.4E1	2.0E0	
Q-REG	6.2E1	1.6E1	6.4E1	6.2E1	
MRDR	2.6E1	1.5E1	1.2E2	1.1E2	
FQE	4.8E0	1.1E2	5.0E1	3.8E0	
R( $\lambda$ )	2.8E0	1.0E2	5.4E1	2.4E0	
Q $^\pi$ ( $\lambda$ )	3.2E0	9.6E1	4.9E1	3.4E0	
TREE	2.2E0	1.0E2	5.0E1	<b>1.6E0</b>	
IH	5.1E0	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	<b>2.0E1</b>	1.1E2
WIS	4.5E1	5.0E1
NAIVE	4.0E0	-

Table 313: Graph-POMDP, relative MSE.  $T = 16, N = 256, H = 6, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	1.5E0	7.2E2	2.1E1	6.9E0	
Q-REG	1.0E0	<b>7.9E-1</b>	4.6E0	2.6E0	
MRDR	2.3E0	2.0E0	3.1E2	3.1E2	
FQE	3.8E0	3.5E0	6.4E0	3.7E0	
R( $\lambda$ )	1.0E0	1.0E0	6.1E0	2.0E0	
Q $^\pi$ ( $\lambda$ )	2.0E0	2.7E0	3.2E0	1.9E0	
TREE	<b>1.0E0</b>	1.0E0	6.1E0	2.0E0	
IH	3.6E0	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.0E0</b>	1.0E0
WIS	6.1E0	6.1E0
NAIVE	4.0E0	-

Table 312: Graph-POMDP, relative MSE.  $T = 16, N = 1024, H = 6, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	1.9E0	5.9E2	2.2E1	1.8E0	
Q-REG	5.9E1	3.0E1	4.1E1	8.0E1	
MRDR	4.5E1	2.1E1	2.7E1	2.9E1	
FQE	4.9E0	9.5E1	3.5E1	4.8E0	
R( $\lambda$ )	2.3E0	7.9E1	3.5E1	2.6E0	
Q $^\pi$ ( $\lambda$ )	<b>1.4E0</b>	7.0E1	3.8E1	<b>1.4E0</b>	
TREE	2.7E0	8.3E1	3.5E1	2.8E0	
IH	4.9E0	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	1.2E2	8.1E1
WIS	<b>3.1E1</b>	3.5E1
NAIVE	3.9E0	-

Table 314: Graph-POMDP, relative MSE.  $T = 16, N = 512, H = 6, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	1.5E0	1.9E0	1.4E1	2.4E0	
Q-REG	1.1E0	1.3E0	6.8E0	1.1E0	
MRDR	1.4E0	1.1E0	2.3E1	1.7E1	
FQE	4.4E0	4.2E0	7.5E0	4.3E0	
R( $\lambda$ )	1.0E0	1.0E0	7.5E0	1.0E0	
Q $^\pi$ ( $\lambda$ )	<b>9.0E-1</b>	<b>5.1E-1</b>	6.4E0	8.7E-1	
TREE	1.0E0	1.0E0	7.5E0	1.0E0	
IH	4.3E0	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.0E0</b>	1.0E0
WIS	7.5E0	7.5E0
NAIVE	3.9E0	-

Table 315: Graph-POMDP, relative MSE.  $T = 16, N = 1024, H = 6, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	1.4E0	2.4E1	1.4E1	1.5E0	
Q-REG	1.2E0	1.4E0	4.9E0	4.5E0	
MRDR	1.3E0	2.0E0	6.9E1	6.8E1	
FQE	4.1E0	4.1E0	4.7E0	3.9E0	
$R(\lambda)$	1.0E0	1.2E0	4.8E0	1.1E0	
$Q^\pi(\lambda)$	<b>3.0E-1</b>	3.1E-1	3.4E0	<b>2.9E-1</b>	
TREE	1.0E0	1.2E0	4.8E0	1.1E0	
IH	3.9E0	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.2E0</b>	1.2E0
WIS	4.8E0	4.8E0
NAIVE	3.9E0	-

 Table 316: Graph-POMDP, relative MSE.  $T = 16, N = 256, H = 6, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	<b>3.1E0</b>	2.2E3	2.5E2	<b>5.5E0</b>	
Q-REG	4.7E1	6.3E1	5.0E1	5.9E1	
MRDR	5.4E1	3.7E1	1.5E2	1.4E2	
FQE	1.2E1	9.4E1	7.6E1	1.0E1	
$R(\lambda)$	2.6E1	1.2E2	8.2E1	2.5E1	
$Q^\pi(\lambda)$	1.2E1	6.8E1	7.0E1	1.1E1	
TREE	2.3E1	1.1E2	8.4E1	1.8E1	
IH	1.2E1	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	<b>9.9E-1</b>	6.0E1
WIS	6.5E1	7.3E1
NAIVE	7.5E0	-

 Table 317: Graph-POMDP, relative MSE.  $T = 16, N = 512, H = 6, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	<b>2.3E0</b>	8.8E2	2.6E2	<b>3.6E0</b>	
Q-REG	1.8E2	9.7E1	6.5E1	1.4E2	
MRDR	5.9E2	7.4E1	2.2E2	4.0E2	
FQE	9.0E0	1.7E2	6.7E1	1.1E1	
$R(\lambda)$	1.2E1	1.6E2	7.1E1	1.4E1	
$Q^\pi(\lambda)$	1.4E1	2.2E2	7.1E1	1.5E1	
TREE	1.3E1	1.7E2	6.8E1	1.4E1	
IH	8.8E0	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.7E0</b>	2.0E2
WIS	5.0E1	6.6E1
NAIVE	5.9E0	-

 Table 318: Graph-POMDP, relative MSE.  $T = 16, N = 1024, H = 6, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	<b>1.9E0</b>	1.7E2	1.9E2	4.1E0	
Q-REG	1.9E2	3.1E2	2.0E2	1.8E2	
MRDR	1.6E3	7.5E0	9.4E1	9.9E1	
FQE	4.8E0	5.6E1	5.6E1	5.0E0	
$R(\lambda)$	2.8E0	4.6E1	5.4E1	<b>2.8E0</b>	
$Q^\pi(\lambda)$	2.7E0	4.4E1	4.7E1	3.0E0	
TREE	2.6E0	4.6E1	5.4E1	2.8E0	
IH	4.9E0	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.2E0</b>	4.7E1
WIS	5.0E1	5.4E1
NAIVE	5.0E0	-

Table 319: Graph-POMDP, relative MSE.  $T = 2, N = 256, H = 2, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.1E-1	3.6E-3	2.8E-3	2.8E-3
Q-REG	1.1E-1	2.9E-3	2.0E-3	3.2E-3
MRDR	4.7E-1	2.5E-3	<b>1.4E-3</b>	1.9E-3
FQE	1.2E-1	5.4E-3	4.1E-3	3.2E-3
R( $\lambda$ )	1.7E-1	3.0E-3	1.8E-3	7.2E-3
Q $^\pi$ ( $\lambda$ )	<b>5.7E-2</b>	2.7E-3	2.0E-3	2.0E-3
TREE	2.1E-1	2.9E-3	1.7E-3	6.9E-3
IH	1.2E-1	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.7E-2	1.4E-2
WIS	9.0E-3	<b>7.5E-3</b>
NAIVE	4.6E-1	-

 Table 321: Graph-POMDP, relative MSE.  $T = 2, N = 1024, H = 2, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.2E-1	2.2E-3	1.8E-3	1.8E-3
Q-REG	1.1E-1	4.8E-4	3.7E-4	3.7E-4
MRDR	5.1E-1	5.1E-4	6.4E-4	6.4E-4
FQE	1.3E-1	7.5E-4	5.7E-4	5.7E-4
R( $\lambda$ )	1.8E-1	4.7E-4	<b>3.6E-4</b>	1.3E-3
Q $^\pi$ ( $\lambda$ )	<b>6.4E-2</b>	4.8E-4	3.6E-4	3.6E-4
TREE	2.2E-1	4.7E-4	3.6E-4	1.6E-3
IH	1.3E-1	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	3.4E-3	2.5E-3
WIS	1.6E-3	<b>1.2E-3</b>
NAIVE	4.9E-1	-

 Table 320: Graph-POMDP, relative MSE.  $T = 2, N = 512, H = 2, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.1E-1	2.9E-3	2.0E-3	2.0E-3
Q-REG	1.0E-1	1.7E-3	<b>1.3E-3</b>	1.3E-3
MRDR	4.8E-1	1.9E-3	2.0E-3	2.0E-3
FQE	1.1E-1	1.8E-3	1.3E-3	1.3E-3
R( $\lambda$ )	1.6E-1	1.8E-3	1.4E-3	6.9E-3
Q $^\pi$ ( $\lambda$ )	<b>5.0E-2</b>	1.9E-3	1.4E-3	1.4E-3
TREE	2.0E-1	1.8E-3	1.4E-3	4.4E-3
IH	1.1E-1	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	6.5E-3	4.5E-3
WIS	2.9E-3	<b>2.2E-3</b>
NAIVE	4.5E-1	-

 Table 322: Graph-POMDP, relative MSE.  $T = 2, N = 256, H = 2, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	7.4E-2	2.5E-2	2.3E-2	2.6E-2
Q-REG	8.4E-2	8.2E-3	7.2E-3	3.2E-2
MRDR	4.0E-1	8.0E-3	7.3E-3	8.7E-3
FQE	7.8E-2	9.9E-3	8.7E-3	2.2E-2
R( $\lambda$ )	1.3E-1	8.3E-3	7.2E-3	3.1E-2
Q $^\pi$ ( $\lambda$ )	<b>3.9E-2</b>	8.3E-3	7.3E-3	7.3E-3
TREE	1.6E-1	8.2E-3	<b>7.0E-3</b>	2.1E-2
IH	7.9E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.8E-2	1.9E-2
WIS	<b>1.1E-2</b>	1.2E-2
NAIVE	3.6E-1	-

Table 323: Graph-POMDP, relative MSE.  $T = 2, N = 512, H = 2, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.5E-1	1.8E-2	1.7E-2	2.0E-2
Q-REG	1.2E-1	5.7E-3	5.5E-3	<b>4.3E-3</b>
MRDR	4.9E-1	5.6E-3	5.4E-3	1.0E-2
FQE	1.3E-1	6.4E-3	6.2E-3	4.4E-3
R( $\lambda$ )	1.8E-1	5.6E-3	5.4E-3	1.1E-2
Q $^\pi$ ( $\lambda$ )	<b>6.5E-2</b>	5.5E-3	5.4E-3	5.4E-3
TREE	2.2E-1	5.6E-3	5.3E-3	1.2E-2
IH	1.3E-1	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.1E-2	9.9E-3
WIS	<b>7.6E-3</b>	7.7E-3
NAIVE	4.7E-1	-

 Table 325: Graph-POMDP, relative MSE.  $T = 2, N = 256, H = 2, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.9E-1	5.8E-2	5.8E-2	2.4E-1
Q-REG	1.7E-1	4.9E-2	4.9E-2	1.5E-1
MRDR	5.6E-1	4.8E-2	4.9E-2	1.7E-1
FQE	2.2E-1	4.8E-2	4.8E-2	1.6E-1
R( $\lambda$ )	2.6E-1	4.8E-2	4.8E-2	1.5E-1
Q $^\pi$ ( $\lambda$ )	<b>1.3E-1</b>	<b>4.8E-2</b>	4.8E-2	1.6E-1
TREE	3.1E-1	4.8E-2	4.8E-2	1.5E-1
IH	2.1E-1	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	5.5E-2	<b>4.9E-2</b>
WIS	5.4E-2	4.9E-2
NAIVE	5.8E-1	-

 Table 324: Graph-POMDP, relative MSE.  $T = 2, N = 1024, H = 2, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.0E-1	4.7E-3	4.4E-3	4.4E-3
Q-REG	1.1E-1	2.1E-3	1.9E-3	1.9E-3
MRDR	5.0E-1	2.1E-3	2.2E-3	2.2E-3
FQE	1.2E-1	2.2E-3	1.9E-3	<b>1.9E-3</b>
R( $\lambda$ )	1.8E-1	2.1E-3	1.9E-3	4.7E-3
Q $^\pi$ ( $\lambda$ )	<b>6.5E-2</b>	2.2E-3	1.9E-3	1.9E-3
TREE	2.2E-1	2.1E-3	1.9E-3	2.3E-3
IH	1.2E-1	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	2.7E-3	2.7E-3
WIS	2.1E-3	<b>2.1E-3</b>
NAIVE	4.8E-1	-

 Table 326: Graph-POMDP, relative MSE.  $T = 2, N = 512, H = 2, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.4E-1	2.7E-2	2.5E-2	5.6E-2
Q-REG	1.2E-1	2.5E-2	2.3E-2	6.1E-2
MRDR	4.8E-1	2.4E-2	<b>2.1E-2</b>	2.9E-2
FQE	1.3E-1	2.8E-2	2.6E-2	4.8E-2
R( $\lambda$ )	1.8E-1	2.5E-2	2.3E-2	4.2E-2
Q $^\pi$ ( $\lambda$ )	<b>6.7E-2</b>	2.4E-2	2.3E-2	5.9E-2
TREE	2.3E-1	2.5E-2	2.2E-2	4.4E-2
IH	1.3E-1	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	4.5E-2	3.6E-2
WIS	3.8E-2	<b>3.0E-2</b>
NAIVE	4.7E-1	-

Table 327: Graph-POMDP, relative MSE.  $T = 2, N = 1024, H = 2, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.6E-1	2.3E-2	2.1E-2	4.0E-2
Q-REG	1.4E-1	2.0E-2	1.9E-2	2.1E-2
MRDR	5.1E-1	2.0E-2	1.8E-2	3.4E-2
FQE	1.6E-1	2.2E-2	2.1E-2	2.1E-2
R( $\lambda$ )	2.1E-1	2.0E-2	1.9E-2	3.2E-2
Q $^\pi$ ( $\lambda$ )	<b>8.7E-2</b>	2.0E-2	1.9E-2	<b>1.6E-2</b>
TREE	2.6E-1	2.0E-2	1.9E-2	3.5E-2
IH	1.5E-1	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	3.5E-2	2.8E-2
WIS	2.9E-2	<b>2.3E-2</b>
NAIVE	4.7E-1	-

 Table 328: Graph-POMDP, relative MSE.  $T = 2, N = 256, H = 2, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.7E-1	7.7E-2	7.4E-2	2.4E-1
Q-REG	1.9E-1	6.1E-2	6.0E-2	7.9E-2
MRDR	5.6E-1	5.9E-2	<b>5.5E-2</b>	1.6E-1
FQE	2.5E-1	6.8E-2	6.6E-2	1.3E-1
R( $\lambda$ )	2.6E-1	6.1E-2	6.0E-2	9.2E-2
Q $^\pi$ ( $\lambda$ )	<b>1.3E-1</b>	6.0E-2	6.0E-2	9.5E-2
TREE	3.2E-1	6.2E-2	5.9E-2	1.1E-1
IH	2.5E-1	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	7.7E-2	7.9E-2
WIS	<b>6.8E-2</b>	7.3E-2
NAIVE	5.8E-1	-

 Table 329: Graph-POMDP, relative MSE.  $T = 2, N = 512, H = 2, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.8E-1	5.5E-2	5.6E-2	1.4E-1
Q-REG	1.4E-1	3.5E-2	3.6E-2	5.7E-2
MRDR	5.5E-1	3.6E-2	3.8E-2	9.5E-2
FQE	1.8E-1	<b>3.3E-2</b>	3.3E-2	5.8E-2
R( $\lambda$ )	2.3E-1	3.5E-2	3.6E-2	4.8E-2
Q $^\pi$ ( $\lambda$ )	<b>1.0E-1</b>	3.6E-2	3.6E-2	5.3E-2
TREE	2.8E-1	3.5E-2	3.6E-2	5.2E-2
IH	1.8E-1	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	3.2E-2	<b>3.0E-2</b>
WIS	3.4E-2	3.2E-2
NAIVE	5.6E-1	-

 Table 330: Graph-POMDP, relative MSE.  $T = 2, N = 1024, H = 2, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.3E-1	4.4E-2	4.4E-2	1.0E-1
Q-REG	1.5E-1	1.8E-2	1.8E-2	2.8E-2
MRDR	5.3E-1	1.8E-2	<b>1.7E-2</b>	7.9E-2
FQE	1.7E-1	1.9E-2	1.9E-2	2.6E-2
R( $\lambda$ )	2.2E-1	1.8E-2	1.8E-2	1.9E-2
Q $^\pi$ ( $\lambda$ )	<b>9.6E-2</b>	1.8E-2	1.8E-2	2.5E-2
TREE	2.6E-1	1.8E-2	1.8E-2	2.2E-2
IH	1.6E-1	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	2.2E-2	2.2E-2
WIS	<b>2.0E-2</b>	2.1E-2
NAIVE	5.2E-1	-



**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 331: Graph-POMDP, relative MSE.  $T = 2, N = 256, H = 2, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.4E-1	2.4E-2	1.9E-2	1.9E-2
Q-REG	<b>2.9E-2</b>	2.9E-3	3.0E-3	5.6E-3
MRDR	3.0E-1	3.3E-3	4.3E-3	4.3E-3
FQE	4.5E-1	8.0E-3	4.7E-3	4.7E-3
R( $\lambda$ )	1.6E-1	2.7E-3	2.8E-3	2.8E-3
Q $^\pi$ ( $\lambda$ )	3.4E-2	<b>2.5E-3</b>	2.7E-3	3.3E-3
TREE	2.7E-1	3.1E-3	2.8E-3	2.8E-3
IH	4.5E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E-2	1.0E-2
WIS	4.6E-3	<b>4.6E-3</b>
NAIVE	4.3E-1	-

Table 333: Graph-POMDP, relative MSE.  $T = 2, N = 1024, H = 2, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.4E-1	2.2E-3	1.9E-3	1.9E-3
Q-REG	<b>2.4E-2</b>	4.4E-4	5.1E-4	5.1E-4
MRDR	3.1E-1	5.0E-4	8.5E-4	8.5E-4
FQE	4.4E-1	7.9E-4	4.7E-4	4.7E-4
R( $\lambda$ )	1.4E-1	3.8E-4	4.7E-4	4.7E-4
Q $^\pi$ ( $\lambda$ )	2.4E-2	4.6E-4	5.2E-4	5.2E-4
TREE	2.5E-1	<b>3.6E-4</b>	4.1E-4	4.1E-4
IH	4.4E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.1E-3	1.1E-3
WIS	4.6E-4	<b>4.6E-4</b>
NAIVE	4.4E-1	-

Table 332: Graph-POMDP, relative MSE.  $T = 2, N = 512, H = 2, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.9E-1	7.9E-3	7.8E-3	7.8E-3
Q-REG	<b>3.4E-2</b>	9.8E-4	8.9E-4	1.2E-3
MRDR	3.7E-1	<b>8.9E-4</b>	1.1E-3	1.1E-3
FQE	5.1E-1	4.9E-3	2.8E-3	2.8E-3
R( $\lambda$ )	1.8E-1	1.3E-3	9.2E-4	9.2E-4
Q $^\pi$ ( $\lambda$ )	4.1E-2	1.0E-3	9.0E-4	9.0E-4
TREE	2.8E-1	1.7E-3	1.0E-3	1.0E-3
IH	5.1E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	5.7E-3	5.7E-3
WIS	<b>2.8E-3</b>	2.8E-3
NAIVE	5.2E-1	-

Table 334: Graph-POMDP, relative MSE.  $T = 2, N = 256, H = 2, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	6.0E-1	6.3E-2	6.2E-2	1.3E-1
Q-REG	<b>4.4E-2</b>	1.6E-2	1.6E-2	5.0E-2
MRDR	3.6E-1	1.6E-2	1.7E-2	2.4E-2
FQE	5.3E-1	1.9E-2	1.7E-2	1.7E-2
R( $\lambda$ )	1.9E-1	1.5E-2	1.6E-2	5.1E-2
Q $^\pi$ ( $\lambda$ )	5.3E-2	1.5E-2	1.6E-2	4.2E-2
TREE	2.9E-1	1.5E-2	<b>1.5E-2</b>	2.7E-2
IH	5.3E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.8E-2	1.8E-2
WIS	1.6E-2	<b>1.6E-2</b>
NAIVE	5.1E-1	-

Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning

---

Table 335: Graph-POMDP, relative MSE.  $T = 2, N = 512, H = 2, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.7E-1	2.7E-2	2.6E-2	6.4E-2
Q-REG	<b>4.7E-2</b>	1.7E-2	1.6E-2	5.7E-2
MRDR	3.5E-1	1.6E-2	1.6E-2	1.6E-2
FQE	5.0E-1	1.9E-2	1.8E-2	1.8E-2
R( $\lambda$ )	2.0E-1	1.7E-2	1.6E-2	2.7E-2
Q $^\pi$ ( $\lambda$ )	5.0E-2	1.6E-2	<b>1.6E-2</b>	5.2E-2
TREE	3.2E-1	1.7E-2	1.6E-2	1.6E-2
IH	5.0E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.9E-2	2.0E-2
WIS	<b>1.8E-2</b>	1.8E-2
NAIVE	4.8E-1	-

Table 337: Graph-POMDP, relative MSE.  $T = 2, N = 256, H = 2, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.0E-1	1.4E-1	1.4E-1	3.8E-1
Q-REG	<b>6.8E-2</b>	3.7E-2	3.7E-2	6.7E-2
MRDR	3.6E-1	3.7E-2	3.8E-2	1.0E-1
FQE	5.3E-1	<b>3.5E-2</b>	3.5E-2	7.1E-2
R( $\lambda$ )	2.2E-1	3.6E-2	3.7E-2	1.7E-1
Q $^\pi$ ( $\lambda$ )	7.1E-2	3.6E-2	3.7E-2	6.8E-2
TREE	3.5E-1	3.6E-2	3.7E-2	1.1E-1
IH	5.3E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	3.7E-2	3.7E-2
WIS	3.6E-2	<b>3.6E-2</b>
NAIVE	4.8E-1	-

Table 336: Graph-POMDP, relative MSE.  $T = 2, N = 1024, H = 2, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.0E-1	2.2E-2	1.9E-2	1.9E-2
Q-REG	<b>3.8E-2</b>	1.1E-2	1.0E-2	3.5E-2
MRDR	3.5E-1	1.0E-2	<b>9.1E-3</b>	9.1E-3
FQE	4.9E-1	1.7E-2	1.4E-2	1.4E-2
R( $\lambda$ )	1.7E-1	1.2E-2	1.1E-2	1.1E-2
Q $^\pi$ ( $\lambda$ )	4.0E-2	1.1E-2	1.1E-2	3.3E-2
TREE	2.8E-1	1.3E-2	1.1E-2	1.1E-2
IH	4.9E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.8E-2	1.8E-2
WIS	1.5E-2	<b>1.4E-2</b>
NAIVE	5.0E-1	-

Table 338: Graph-POMDP, relative MSE.  $T = 2, N = 512, H = 2, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.8E-1	9.3E-2	8.9E-2	2.1E-1
Q-REG	6.2E-2	2.0E-2	2.0E-2	6.6E-2
MRDR	3.9E-1	<b>2.0E-2</b>	2.0E-2	2.0E-2
FQE	5.5E-1	2.1E-2	2.0E-2	2.0E-2
R( $\lambda$ )	2.2E-1	2.0E-2	2.0E-2	3.3E-2
Q $^\pi$ ( $\lambda$ )	<b>5.9E-2</b>	2.0E-2	2.0E-2	6.3E-2
TREE	3.5E-1	2.0E-2	2.0E-2	2.0E-2
IH	5.5E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.1E-2	2.1E-2
WIS	<b>2.0E-2</b>	2.0E-2
NAIVE	5.1E-1	-

Table 339: Graph-POMDP, relative MSE.  $T = 2, N = 1024, H = 2, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.1E-1	2.2E-2	2.1E-2	4.2E-2
Q-REG	<b>1.7E-2</b>	7.4E-3	7.3E-3	1.8E-2
MRDR	2.7E-1	7.2E-3	<b>6.8E-3</b>	6.8E-3
FQE	4.0E-1	9.5E-3	8.6E-3	8.6E-3
R( $\lambda$ )	1.2E-1	7.7E-3	7.4E-3	7.4E-3
Q $^\pi$ ( $\lambda$ )	2.1E-2	7.5E-3	7.4E-3	2.2E-2
TREE	2.2E-1	8.0E-3	7.5E-3	7.5E-3
IH	4.1E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	9.9E-3	9.9E-3
WIS	8.6E-3	<b>8.6E-3</b>
NAIVE	3.9E-1	-

 Table 340: Graph-POMDP, relative MSE.  $T = 2, N = 256, H = 2, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.8E-1	1.2E-1	1.1E-1	2.4E-1
Q-REG	9.1E-2	1.1E-1	1.1E-1	1.3E-1
MRDR	3.1E-1	1.1E-1	<b>1.0E-1</b>	2.1E-1
FQE	3.8E-1	1.3E-1	1.2E-1	2.9E-1
R( $\lambda$ )	1.8E-1	1.1E-1	1.1E-1	1.8E-1
Q $^\pi$ ( $\lambda$ )	<b>8.4E-2</b>	1.1E-1	1.1E-1	1.3E-1
TREE	2.6E-1	1.2E-1	1.1E-1	2.5E-1
IH	3.8E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.6E-1	1.4E-1
WIS	1.4E-1	<b>1.2E-1</b>
NAIVE	4.1E-1	-

 Table 341: Graph-POMDP, relative MSE.  $T = 2, N = 512, H = 2, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.1E-1	3.1E-2	3.0E-2	3.5E-1
Q-REG	3.7E-2	1.5E-2	1.5E-2	9.5E-2
MRDR	3.3E-1	1.5E-2	1.6E-2	1.2E-1
FQE	5.7E-1	1.7E-2	1.5E-2	9.4E-2
R( $\lambda$ )	1.7E-1	1.5E-2	1.5E-2	1.5E-1
Q $^\pi$ ( $\lambda$ )	<b>3.5E-2</b>	1.5E-2	1.5E-2	8.7E-2
TREE	3.1E-1	1.5E-2	<b>1.5E-2</b>	1.6E-1
IH	5.7E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.2E-2	1.7E-2
WIS	1.8E-2	<b>1.5E-2</b>
NAIVE	5.2E-1	-

 Table 342: Graph-POMDP, relative MSE.  $T = 2, N = 1024, H = 2, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	7.0E-1	1.0E-1	1.0E-1	2.1E-1
Q-REG	7.0E-2	5.7E-2	5.7E-2	1.2E-1
MRDR	3.5E-1	5.8E-2	5.9E-2	1.2E-1
FQE	5.5E-1	5.6E-2	<b>5.6E-2</b>	5.6E-2
R( $\lambda$ )	1.9E-1	5.6E-2	5.7E-2	2.0E-1
Q $^\pi$ ( $\lambda$ )	<b>7.0E-2</b>	5.7E-2	5.7E-2	1.2E-1
TREE	3.2E-1	5.6E-2	5.7E-2	1.1E-1
IH	5.5E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	5.3E-2	5.7E-2
WIS	<b>5.3E-2</b>	5.6E-2
NAIVE	5.2E-1	-

Table 343: Graph-POMDP, relative MSE.  $T = 16, N = 256, H = 6, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.9E-1	5.4E-3	2.1E-3	2.1E-3
Q-REG	1.5E-2	4.4E-4	<b>1.2E-4</b>	1.4E-4
MRDR	1.5E-2	5.2E-3	3.6E-3	9.4E-3
FQE	3.0E-3	2.2E-4	1.7E-4	1.2E-4
R( $\lambda$ )	<b>7.3E-4</b>	2.4E-4	1.7E-4	4.9E-4
Q $^\pi$ ( $\lambda$ )	6.5E-3	2.9E-4	1.6E-4	2.6E-4
TREE	4.8E-3	3.2E-3	4.8E-4	5.2E-4
IH	8.7E-4	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	2.2E-1	2.2E-2
WIS	1.3E-2	<b>2.0E-3</b>
NAIVE	4.3E-1	-

 Table 345: Graph-POMDP, relative MSE.  $T = 16, N = 1024, H = 6, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.9E-1	9.8E-4	8.9E-4	8.9E-4
Q-REG	1.2E-2	1.2E-4	1.1E-4	1.4E-4
MRDR	1.2E-2	7.0E-4	7.1E-4	3.0E-3
FQE	2.8E-3	1.4E-4	1.3E-4	<b>6.3E-5</b>
R( $\lambda$ )	<b>6.5E-4</b>	1.7E-4	1.1E-4	1.6E-4
Q $^\pi$ ( $\lambda$ )	6.3E-3	1.3E-4	1.1E-4	1.7E-4
TREE	6.1E-3	4.3E-4	2.3E-4	5.9E-4
IH	1.5E-3	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	2.0E-2	3.3E-3
WIS	2.1E-3	<b>4.9E-4</b>
NAIVE	4.5E-1	-

 Table 344: Graph-POMDP, relative MSE.  $T = 16, N = 512, H = 6, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.8E-1	1.4E-3	7.6E-4	7.6E-4
Q-REG	9.3E-3	1.2E-4	3.5E-5	8.2E-5
MRDR	9.5E-3	6.3E-4	4.5E-4	3.1E-3
FQE	2.6E-3	1.1E-4	5.7E-5	3.6E-5
R( $\lambda$ )	<b>6.3E-4</b>	1.2E-4	<b>3.0E-5</b>	7.5E-5
Q $^\pi$ ( $\lambda$ )	6.0E-3	9.9E-5	3.0E-5	1.0E-4
TREE	5.7E-3	8.0E-4	2.0E-4	7.0E-4
IH	1.3E-3	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	4.7E-2	2.9E-3
WIS	4.8E-3	<b>4.1E-4</b>
NAIVE	4.4E-1	-

 Table 346: Graph-POMDP, relative MSE.  $T = 16, N = 256, H = 6, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.9E-1	1.8E-2	9.5E-3	9.5E-3
Q-REG	2.1E-2	4.2E-3	3.1E-3	5.6E-3
MRDR	1.6E-2	6.6E-3	6.3E-3	1.2E-2
FQE	3.0E-3	2.7E-3	2.4E-3	<b>1.0E-3</b>
R( $\lambda$ )	2.8E-3	2.7E-3	2.5E-3	2.3E-3
Q $^\pi$ ( $\lambda$ )	7.0E-3	2.8E-3	2.6E-3	2.1E-3
TREE	6.5E-3	1.8E-3	2.1E-3	2.4E-3
IH	<b>1.8E-3</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.1E-1	3.1E-2
WIS	7.3E-3	<b>2.5E-3</b>
NAIVE	4.4E-1	-

Table 347: Graph-POMDP, relative MSE.  $T = 16, N = 512, H = 6, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.9E-1	8.8E-3	5.4E-3	5.4E-3
Q-REG	4.7E-3	2.0E-3	2.1E-3	1.5E-3
MRDR	4.9E-3	1.7E-3	1.8E-3	2.9E-3
FQE	2.1E-3	2.5E-3	2.3E-3	1.6E-3
$R(\lambda)$	1.1E-3	2.4E-3	2.3E-3	2.3E-3
$Q^\pi(\lambda)$	5.4E-3	2.3E-3	2.1E-3	1.4E-3
TREE	2.5E-3	3.7E-3	2.6E-3	<b>8.8E-4</b>
IH	<b>5.9E-4</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	9.5E-2	1.3E-2
WIS	6.9E-3	<b>3.0E-3</b>
NAIVE	4.2E-1	-

 Table 349: Graph-POMDP, relative MSE.  $T = 16, N = 256, H = 6, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.7E-1	4.2E-3	6.1E-3	6.1E-3
Q-REG	1.9E-2	6.6E-3	6.1E-3	<b>3.0E-3</b>
MRDR	2.0E-2	1.3E-2	9.6E-3	8.7E-3
FQE	<b>3.6E-3</b>	7.4E-3	6.7E-3	3.2E-3
$R(\lambda)$	4.5E-3	7.2E-3	6.4E-3	6.1E-3
$Q^\pi(\lambda)$	1.2E-2	6.9E-3	6.5E-3	6.1E-3
TREE	6.1E-3	5.0E-3	6.7E-3	4.9E-3
IH	3.8E-3	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	6.3E-2	9.8E-3
WIS	1.7E-2	<b>4.4E-3</b>
NAIVE	4.5E-1	-

 Table 348: Graph-POMDP, relative MSE.  $T = 16, N = 1024, H = 6, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.9E-1	4.1E-3	3.8E-3	3.8E-3
Q-REG	5.7E-3	7.9E-4	8.6E-4	1.2E-3
MRDR	5.7E-3	1.4E-3	1.5E-3	1.1E-3
FQE	3.4E-3	7.7E-4	8.4E-4	<b>3.0E-4</b>
$R(\lambda)$	<b>1.1E-3</b>	7.7E-4	8.4E-4	6.2E-4
$Q^\pi(\lambda)$	7.8E-3	7.8E-4	8.5E-4	1.7E-3
TREE	6.3E-3	9.5E-4	9.5E-4	1.3E-3
IH	1.5E-3	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.1E-2	2.3E-3
WIS	2.9E-3	<b>8.8E-4</b>
NAIVE	4.4E-1	-

 Table 350: Graph-POMDP, relative MSE.  $T = 16, N = 512, H = 6, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.9E-1	1.2E-2	9.3E-3	9.3E-3
Q-REG	1.0E-2	2.7E-3	2.2E-3	6.5E-3
MRDR	1.0E-2	3.6E-3	3.3E-3	6.3E-3
FQE	4.0E-3	2.9E-3	2.3E-3	<b>1.1E-3</b>
$R(\lambda)$	2.7E-3	2.8E-3	2.3E-3	2.4E-3
$Q^\pi(\lambda)$	6.9E-3	2.9E-3	2.3E-3	1.9E-3
TREE	7.7E-3	2.6E-3	1.8E-3	3.0E-3
IH	<b>1.6E-3</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	3.4E-2	1.1E-2
WIS	8.3E-3	<b>2.7E-3</b>
NAIVE	4.4E-1	-

Table 351: Graph-POMDP, relative MSE.  $T = 16, N = 1024, H = 6, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.0E-1	4.9E-3	5.2E-3	5.2E-3
Q-REG	7.4E-3	1.8E-3	2.0E-3	3.2E-3
MRDR	9.1E-3	2.0E-3	2.1E-3	4.2E-3
FQE	4.9E-3	2.3E-3	2.1E-3	<b>6.1E-4</b>
$R(\lambda)$	<b>1.5E-3</b>	2.0E-3	2.0E-3	1.6E-3
$Q^\pi(\lambda)$	6.9E-3	2.0E-3	2.0E-3	1.3E-3
TREE	5.5E-3	2.8E-3	2.1E-3	1.9E-3
IH	2.4E-3	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	2.8E-2	7.7E-3
WIS	7.4E-3	<b>2.4E-3</b>
NAIVE	4.6E-1	-

 Table 352: Graph-POMDP, relative MSE.  $T = 16, N = 256, H = 6, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.2E-1	5.4E-2	4.4E-2	4.4E-2
Q-REG	3.2E-2	1.8E-2	1.8E-2	2.8E-2
MRDR	3.0E-2	1.9E-2	1.7E-2	2.5E-2
FQE	7.2E-3	1.8E-2	1.9E-2	<b>4.1E-3</b>
$R(\lambda)$	1.2E-2	1.9E-2	1.9E-2	1.2E-2
$Q^\pi(\lambda)$	1.7E-2	1.8E-2	1.9E-2	1.1E-2
TREE	1.2E-2	1.9E-2	1.8E-2	8.5E-3
IH	<b>6.9E-3</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.5E-1	3.7E-2
WIS	8.6E-2	<b>1.9E-2</b>
NAIVE	4.7E-1	-

 Table 353: Graph-POMDP, relative MSE.  $T = 16, N = 512, H = 6, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.7E-1	2.3E-2	1.5E-2	1.5E-2
Q-REG	2.5E-2	8.4E-3	8.5E-3	1.4E-2
MRDR	2.2E-2	6.6E-3	7.7E-3	1.5E-2
FQE	4.5E-3	9.8E-3	8.5E-3	<b>3.0E-3</b>
$R(\lambda)$	6.2E-3	9.0E-3	8.7E-3	6.0E-3
$Q^\pi(\lambda)$	9.4E-3	9.6E-3	8.6E-3	4.6E-3
TREE	8.5E-3	1.0E-2	8.0E-3	6.0E-3
IH	<b>2.8E-3</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	9.1E-2	1.9E-2
WIS	3.2E-2	<b>8.8E-3</b>
NAIVE	4.5E-1	-

 Table 354: Graph-POMDP, relative MSE.  $T = 16, N = 1024, H = 6, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Dense rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.3E-1	1.2E-2	1.1E-2	1.1E-2
Q-REG	1.8E-2	1.2E-2	1.2E-2	1.3E-2
MRDR	1.7E-2	1.2E-2	1.2E-2	1.1E-2
FQE	7.5E-3	1.3E-2	1.2E-2	7.0E-3
$R(\lambda)$	7.6E-3	1.2E-2	1.2E-2	<b>6.2E-3</b>
$Q^\pi(\lambda)$	1.2E-2	1.2E-2	1.2E-2	6.5E-3
TREE	1.5E-2	1.3E-2	1.2E-2	9.3E-3
IH	<b>4.9E-3</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.4E-2	1.4E-2
WIS	1.4E-2	<b>1.2E-2</b>
NAIVE	4.6E-1	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 355: Graph-POMDP, relative MSE.  $T = 16, N = 256, H = 6, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	8.6E-1	6.2E-2	3.7E-2	2.5E-1
Q-REG	1.2E-1	3.4E-2	3.2E-2	6.4E-2
MRDR	1.1E-1	1.9E-1	1.2E-1	1.2E-1
FQE	5.2E-1	1.3E-1	5.1E-2	1.7E-1
R( $\lambda$ )	2.9E-2	3.5E-2	3.3E-2	3.1E-2
Q $^\pi$ ( $\lambda$ )	<b>1.8E-2</b>	3.5E-2	3.3E-2	<b>1.9E-2</b>
TREE	1.0E0	1.6E-1	5.0E-2	5.0E-2
IH	5.1E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.6E-1	1.6E-1
WIS	<b>5.0E-2</b>	5.0E-2
NAIVE	4.9E-1	-

Table 357: Graph-POMDP, relative MSE.  $T = 16, N = 1024, H = 6, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	8.5E-1	4.5E-2	2.2E-2	2.2E-2
Q-REG	3.7E-2	3.6E-3	3.8E-3	1.1E-2
MRDR	4.4E-2	1.7E-2	1.3E-2	2.6E-2
FQE	4.4E-1	2.4E-2	8.4E-3	8.4E-3
R( $\lambda$ )	<b>4.1E-3</b>	5.1E-3	4.7E-3	5.1E-3
Q $^\pi$ ( $\lambda$ )	4.3E-3	5.6E-3	4.3E-3	<b>2.9E-3</b>
TREE	1.0E0	3.6E-2	8.3E-3	8.3E-3
IH	4.5E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	3.6E-2	3.6E-2
WIS	8.3E-3	<b>8.3E-3</b>
NAIVE	4.3E-1	-

Table 356: Graph-POMDP, relative MSE.  $T = 16, N = 512, H = 6, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	8.4E-1	7.0E-2	6.6E-2	1.6E-1
Q-REG	1.1E-2	9.5E-3	9.4E-3	1.1E-2
MRDR	1.0E-2	3.6E-2	2.2E-2	6.7E-3
FQE	3.9E-1	8.8E-3	3.9E-3	3.8E-2
R( $\lambda$ )	<b>7.6E-3</b>	5.5E-3	6.6E-3	9.2E-3
Q $^\pi$ ( $\lambda$ )	1.3E-2	7.5E-3	8.5E-3	1.1E-2
TREE	1.0E0	2.1E-2	<b>3.9E-3</b>	3.9E-3
IH	4.0E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.1E-2	2.1E-2
WIS	3.9E-3	<b>3.9E-3</b>
NAIVE	4.2E-1	-

Table 358: Graph-POMDP, relative MSE.  $T = 16, N = 256, H = 6, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.1E0	3.2E0	3.0E0	7.4E-1
Q-REG	6.0E-1	4.6E-1	4.4E-1	6.4E-1
MRDR	4.0E-1	4.8E-1	4.2E-1	4.8E-1
FQE	4.1E-1	7.2E-1	5.5E-1	4.5E-1
R( $\lambda$ )	2.3E-1	6.4E-1	5.4E-1	3.1E-1
Q $^\pi$ ( $\lambda$ )	<b>1.6E-1</b>	6.6E-1	5.0E-1	<b>1.6E-1</b>
TREE	1.5E0	7.3E-1	5.5E-1	1.0E0
IH	4.0E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.6E0	7.9E-1
WIS	2.6E0	<b>5.6E-1</b>
NAIVE	5.5E-1	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 359: Graph-POMDP, relative MSE.  $T = 16, N = 512, H = 6, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.3E0	6.4E-1	5.9E-1	1.1E0
Q-REG	2.6E-1	2.6E-1	2.6E-1	2.4E-1
MRDR	2.2E-1	2.6E-1	2.8E-1	2.0E-1
FQE	6.7E-1	2.9E-1	2.8E-1	6.1E-1
R( $\lambda$ )	1.5E-1	2.6E-1	2.6E-1	1.5E-1
Q $^\pi$ ( $\lambda$ )	<b>1.4E-1</b>	2.6E-1	2.6E-1	<b>1.5E-1</b>
TREE	1.1E0	3.1E-1	2.8E-1	1.0E0
IH	6.8E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	6.6E-1	2.9E-1
WIS	7.7E-1	<b>2.8E-1</b>
NAIVE	5.6E-1	-

Table 361: Graph-POMDP, relative MSE.  $T = 16, N = 256, H = 6, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	9.5E-1	8.5E-1	8.5E-1	8.1E-1
Q-REG	1.5E-1	2.0E-1	2.5E-1	1.5E-1
MRDR	1.5E-1	2.0E-1	2.8E-1	2.9E-1
FQE	5.4E-1	1.7E-1	2.1E-1	4.7E-1
R( $\lambda$ )	1.4E-1	2.3E-1	2.5E-1	1.4E-1
Q $^\pi$ ( $\lambda$ )	<b>8.4E-2</b>	2.0E-1	2.5E-1	<b>8.6E-2</b>
TREE	1.0E0	1.7E-1	2.0E-1	4.9E-1
IH	5.5E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.7E-1</b>	1.7E-1
WIS	2.0E-1	2.0E-1
NAIVE	5.7E-1	-

Table 360: Graph-POMDP, relative MSE.  $T = 16, N = 1024, H = 6, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	8.3E-1	5.5E-1	5.8E-1	6.2E-1
Q-REG	3.4E-1	3.6E-1	3.6E-1	3.2E-1
MRDR	2.9E-1	3.6E-1	3.5E-1	2.7E-1
FQE	4.8E-1	3.9E-1	3.8E-1	4.1E-1
R( $\lambda$ )	2.1E-1	3.6E-1	3.6E-1	2.1E-1
Q $^\pi$ ( $\lambda$ )	<b>1.4E-1</b>	3.8E-1	3.6E-1	<b>1.3E-1</b>
TREE	1.1E0	3.8E-1	3.8E-1	6.3E-1
IH	4.8E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.3E0	3.8E-1
WIS	1.3E0	<b>3.8E-1</b>
NAIVE	4.5E-1	-

Table 362: Graph-POMDP, relative MSE.  $T = 16, N = 512, H = 6, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	8.5E-1	4.8E-1	4.6E-1	5.6E-1
Q-REG	1.8E-1	1.3E-1	1.3E-1	1.6E-1
MRDR	1.6E-1	1.4E-1	1.4E-1	1.7E-1
FQE	4.2E-1	1.9E-1	1.5E-1	2.6E-1
R( $\lambda$ )	3.6E-2	1.4E-1	1.3E-1	3.4E-2
Q $^\pi$ ( $\lambda$ )	<b>3.2E-2</b>	1.5E-1	1.3E-1	<b>2.7E-2</b>
TREE	1.0E0	2.1E-1	1.5E-1	1.5E-1
IH	4.3E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.1E-1	2.1E-1
WIS	<b>1.5E-1</b>	1.5E-1
NAIVE	4.6E-1	-



Table 363: Graph-POMDP, relative MSE.  $T = 16, N = 1024, H = 6, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	8.6E-1	3.1E-1	3.1E-1	5.2E-1
Q-REG	1.2E-1	7.7E-2	7.5E-2	6.4E-2
MRDR	9.3E-2	6.4E-2	7.0E-2	8.1E-2
FQE	4.6E-1	1.3E-1	9.2E-2	2.2E-1
$R(\lambda)$	3.0E-2	8.2E-2	7.6E-2	3.0E-2
$Q^\pi(\lambda)$	<b>2.6E-2</b>	8.5E-2	7.5E-2	<b>2.1E-2</b>
TREE	1.0E0	1.4E-1	9.3E-2	9.3E-2
IH	4.7E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.4E-1	1.4E-1
WIS	<b>9.3E-2</b>	9.3E-2
NAIVE	4.9E-1	-

 Table 364: Graph-POMDP, relative MSE.  $T = 16, N = 256, H = 6, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	<b>9.1E-1</b>	6.1E0	6.5E0	1.0E0
Q-REG	4.1E0	4.2E0	4.1E0	3.5E0
MRDR	3.7E0	3.4E0	3.4E0	3.0E0
FQE	9.8E-1	3.7E0	3.9E0	<b>6.2E-1</b>
$R(\lambda)$	3.3E0	4.1E0	4.1E0	2.7E0
$Q^\pi(\lambda)$	2.8E0	4.4E0	4.3E0	2.3E0
TREE	3.6E0	3.8E0	4.0E0	2.9E0
IH	1.1E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.2E1	<b>3.7E0</b>
WIS	1.0E1	4.0E0
NAIVE	1.1E0	-

 Table 365: Graph-POMDP, relative MSE.  $T = 16, N = 512, H = 6, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.2E0	4.7E0	4.5E0	2.0E0
Q-REG	9.4E-1	1.3E0	1.3E0	1.0E0
MRDR	1.0E0	1.6E0	1.6E0	1.1E0
FQE	1.3E0	1.2E0	1.2E0	1.3E0
$R(\lambda)$	1.0E0	1.2E0	1.3E0	1.1E0
$Q^\pi(\lambda)$	<b>9.0E-1</b>	1.2E0	1.2E0	<b>9.0E-1</b>
TREE	2.7E0	1.1E0	1.2E0	2.7E0
IH	1.3E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.2E0	1.2E0
WIS	2.3E0	<b>1.2E0</b>
NAIVE	1.1E0	-

 Table 366: Graph-POMDP, relative MSE.  $T = 16, N = 1024, H = 6, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$ . Stochastic environment. Stochastic rewards. Sparse rewards.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	8.8E-1	1.2E0	1.1E0	1.1E0
Q-REG	5.6E-1	5.1E-1	5.0E-1	4.7E-1
MRDR	5.9E-1	5.9E-1	5.8E-1	4.8E-1
FQE	5.7E-1	5.6E-1	5.3E-1	5.0E-1
$R(\lambda)$	2.7E-1	5.1E-1	5.0E-1	2.2E-1
$Q^\pi(\lambda)$	<b>1.3E-1</b>	4.6E-1	4.6E-1	<b>1.5E-1</b>
TREE	9.9E-1	5.9E-1	5.4E-1	9.6E-1
IH	5.7E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	7.3E-1	5.8E-1
WIS	7.8E-1	<b>5.4E-1</b>
NAIVE	6.7E-1	-

**F.3 Detailed Results for Graph Mountain Car (Graph-MC)**

 Table 367: Graph-MC, relative MSE.  $T = 250, N = 128, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ .

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.4E-1	5.1E-1	5.0E0	4.2E0
Q-REG	1.5E2	1.3E1	3.5E3	1.5E2
MRDR	9.7E2	1.3E1	7.6E4	2.7E4
FQE	<b>4.0E-1</b>	4.0E-1	1.8E-1	<b>1.5E-1</b>
R( $\lambda$ )	4.4E-1	9.4E0	1.7E1	1.7E1
Q $^\pi$ ( $\lambda$ )	1.9E130	1.9E129	1.8E131	1.0E0
TREE	4.4E-1	9.4E0	1.7E1	1.7E1
IH	2.0E1	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	<b>1.0E0</b>	9.4E0
WIS	2.0E1	2.0E1
NAIVE	2.0E1	-

 Table 368: Graph-MC, relative MSE.  $T = 250, N = 256, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ .

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.8E-1	5.1E-1	5.4E0	4.8E0
Q-REG	3.3E-1	3.7E-1	4.3E0	1.1E0
MRDR	<b>1.8E-1</b>	4.3E-1	1.1E4	1.1E4
FQE	3.7E-1	3.7E-1	1.8E-1	<b>1.3E-1</b>
R( $\lambda$ )	3.7E-1	3.7E-1	1.6E1	1.6E1
Q $^\pi$ ( $\lambda$ )	9.1E118	7.5E117	1.5E119	1.0E0
TREE	3.9E-1	3.7E-1	1.6E1	1.5E1
IH	2.1E1	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.0E0	<b>3.3E-1</b>
WIS	2.1E1	2.1E1
NAIVE	2.1E1	-

 Table 369: Graph-MC, relative MSE.  $T = 250, N = 512, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ .

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.9E-1	3.9E-1	5.8E0	5.3E0
Q-REG	2.7E-1	1.1E0	2.7E0	8.8E-1
MRDR	<b>1.4E-1</b>	1.6E0	9.1E3	9.1E3
FQE	2.9E-1	2.9E-1	1.3E-1	<b>8.8E-2</b>
R( $\lambda$ )	3.0E-1	3.0E-1	1.7E1	1.7E1
Q $^\pi$ ( $\lambda$ )	4.9E146	9.2E145	1.0E147	1.0E0
TREE	3.1E-1	2.9E-1	1.8E1	1.7E1
IH	2.1E1	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.0E0	<b>4.4E-1</b>
WIS	2.1E1	2.1E1
NAIVE	2.1E1	-

 Table 370: Graph-MC, relative MSE.  $T = 250, N = 1024, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$ .

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.6E-1	3.9E-1	5.7E0	5.2E0
Q-REG	2.8E-1	2.4E-1	3.8E0	2.0E0
MRDR	<b>1.7E-1</b>	2.8E-1	2.2E4	2.2E4
FQE	2.6E-1	2.6E-1	1.1E-1	<b>7.7E-2</b>
R( $\lambda$ )	2.6E-1	2.6E-1	1.6E1	1.5E1
Q $^\pi$ ( $\lambda$ )	9.6E121	1.5E122	7.2E124	1.0E0
TREE	2.8E-1	2.6E-1	1.6E1	1.6E1
IH	2.0E1	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.0E0	<b>2.7E-1</b>
WIS	2.0E1	2.0E1
NAIVE	2.0E1	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 371: Graph-MC, relative MSE.  $T = 250, N = 128, \pi_b(a = 0) = 0.5, \pi_e(a = 0) = 0.5$ .

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.5E-1	1.7E-3	1.7E-3	3.7E-3
Q-REG	1.7E-2	1.5E-4	1.5E-4	1.3E-4
MRDR	3.0E-2	5.0E-3	5.0E-3	6.1E-3
FQE	3.5E-2	<b>6.0E-5</b>	6.0E-5	6.5E-5
R( $\lambda$ )	1.7E-1	7.0E-4	7.0E-4	1.5E-3
Q $^\pi$ ( $\lambda$ )	1.6E-1	7.1E-4	7.1E-4	1.6E-3
TREE	7.9E-1	3.0E-4	3.0E-4	5.4E-4
IH	<b>2.2E-4</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>2.2E-4</b>	2.2E-4
WIS	2.2E-4	2.2E-4
NAIVE	2.2E-4	-

Table 373: Graph-MC, relative MSE.  $T = 250, N = 512, \pi_b(a = 0) = 0.5, \pi_e(a = 0) = 0.5$ .

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.4E-1	1.9E-4	1.9E-4	2.7E-4
Q-REG	1.7E-2	<b>3.0E-6</b>	3.0E-6	6.0E-6
MRDR	2.9E-2	2.7E-4	2.7E-4	2.6E-4
FQE	3.4E-2	4.0E-6	4.0E-6	9.0E-6
R( $\lambda$ )	1.6E-1	7.5E-5	7.5E-5	6.6E-5
Q $^\pi$ ( $\lambda$ )	1.5E-1	7.4E-5	7.4E-5	7.4E-5
TREE	7.9E-1	6.6E-5	6.6E-5	8.7E-5
IH	<b>5.9E-5</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>5.9E-5</b>	5.9E-5
WIS	5.9E-5	5.9E-5
NAIVE	5.9E-5	-

Table 372: Graph-MC, relative MSE.  $T = 250, N = 256, \pi_b(a = 0) = 0.5, \pi_e(a = 0) = 0.5$ .

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.4E-1	2.8E-4	2.8E-4	6.7E-4
Q-REG	1.5E-2	2.6E-5	2.6E-5	2.3E-5
MRDR	2.6E-2	1.5E-3	1.5E-3	1.4E-3
FQE	3.4E-2	<b>9.0E-6</b>	9.0E-6	1.1E-5
R( $\lambda$ )	1.7E-1	2.2E-4	2.2E-4	3.0E-4
Q $^\pi$ ( $\lambda$ )	1.6E-1	2.2E-4	2.2E-4	3.0E-4
TREE	7.9E-1	1.1E-4	1.1E-4	1.5E-4
IH	<b>1.0E-4</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.0E-4</b>	1.0E-4
WIS	1.0E-4	1.0E-4
NAIVE	1.0E-4	-

Table 374: Graph-MC, relative MSE.  $T = 250, N = 1024, \pi_b(a = 0) = 0.5, \pi_e(a = 0) = 0.5$ .

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.4E-1	4.7E-5	4.7E-5	5.3E-5
Q-REG	1.7E-2	5.0E-6	5.0E-6	6.0E-6
MRDR	2.9E-2	5.5E-5	5.5E-5	6.5E-5
FQE	3.6E-2	3.0E-6	3.0E-6	<b>2.0E-6</b>
R( $\lambda$ )	1.6E-1	3.5E-5	3.5E-5	4.1E-5
Q $^\pi$ ( $\lambda$ )	1.5E-1	3.4E-5	3.4E-5	4.1E-5
TREE	7.9E-1	2.4E-5	2.4E-5	2.4E-5
IH	<b>3.2E-5</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>3.2E-5</b>	3.2E-5
WIS	3.2E-5	3.2E-5
NAIVE	3.2E-5	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 375: Graph-MC, relative MSE.  $T = 250, N = 128, \pi_b(a = 0) = 0.5, \pi_e(a = 0) = 0.6$ .

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	7.6E-2	9.9E-3	7.1E-3	7.1E-3
Q-REG	7.3E-3	4.1E-4	7.1E-4	1.1E-3
MRDR	5.2E-3	3.4E-2	4.8E-2	5.6E-2
FQE	4.0E-3	<b>4.7E-5</b>	2.7E-4	2.7E-4
R( $\lambda$ )	1.1E-1	3.0E-3	2.4E-3	2.4E-3
Q $^\pi$ ( $\lambda$ )	<b>1.4E-3</b>	3.1E-4	9.8E-5	9.8E-5
TREE	5.8E-1	9.5E-3	1.5E-2	1.5E-2
IH	4.0E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	4.6E-2	<b>1.1E-2</b>
WIS	1.4E-2	1.8E-2
NAIVE	1.6E0	-

Table 377: Graph-MC, relative MSE.  $T = 250, N = 512, \pi_b(a = 0) = 0.5, \pi_e(a = 0) = 0.6$ .

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.2E-2	4.3E-3	3.6E-3	3.5E-3
Q-REG	5.9E-3	3.4E-5	2.4E-5	2.5E-5
MRDR	4.5E-3	4.4E-3	4.8E-3	7.4E-3
FQE	4.3E-3	8.0E-6	1.4E-5	1.4E-5
R( $\lambda$ )	7.9E-2	4.9E-4	1.5E-4	1.5E-4
Q $^\pi$ ( $\lambda$ )	<b>2.1E-4</b>	6.0E-6	<b>6.0E-6</b>	6.0E-6
TREE	5.7E-1	2.4E-3	1.1E-3	1.1E-3
IH	3.4E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.7E-2	3.0E-3
WIS	2.9E-3	<b>1.8E-3</b>
NAIVE	1.5E0	-

Table 376: Graph-MC, relative MSE.  $T = 250, N = 256, \pi_b(a = 0) = 0.5, \pi_e(a = 0) = 0.6$ .

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.4E-2	4.8E-3	5.5E-3	5.5E-3
Q-REG	3.3E-3	<b>1.8E-4</b>	2.6E-4	4.1E-4
MRDR	2.6E-3	4.3E-3	4.6E-3	6.7E-3
FQE	2.8E-3	2.3E-4	2.4E-4	2.4E-4
R( $\lambda$ )	8.7E-2	6.2E-4	4.9E-4	4.9E-4
Q $^\pi$ ( $\lambda$ )	<b>7.0E-6</b>	2.0E-4	2.0E-4	2.0E-4
TREE	5.7E-1	2.6E-3	1.5E-3	1.5E-3
IH	2.9E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.8E-2	3.6E-3
WIS	2.7E-3	<b>2.2E-3</b>
NAIVE	1.6E0	-

Table 378: Graph-MC, relative MSE.  $T = 250, N = 1024, \pi_b(a = 0) = 0.5, \pi_e(a = 0) = 0.6$ .

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.6E-2	7.3E-4	5.2E-4	5.5E-4
Q-REG	3.3E-3	5.3E-6	7.0E-6	8.0E-6
MRDR	1.8E-3	1.3E-3	1.4E-3	2.1E-3
FQE	4.3E-3	2.6E-6	6.0E-6	6.0E-6
R( $\lambda$ )	7.7E-2	6.6E-5	2.3E-4	2.4E-4
Q $^\pi$ ( $\lambda$ )	<b>1.6E-4</b>	<b>9.4E-7</b>	2.0E-6	2.0E-6
TREE	5.7E-1	4.0E-4	7.1E-4	7.2E-4
IH	2.9E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.2E-2	<b>6.8E-4</b>
WIS	1.2E-3	9.2E-4
NAIVE	1.5E0	-

Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning

---

Table 379: Graph-MC, relative MSE.  $T = 250, N = 128, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.5$ .

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.2E-1	4.3E-1	1.7E-1	3.1E-1
Q-REG	2.0E0	1.2E3	5.0E0	1.4E-1
MRDR	2.2E0	4.6E0	1.4E1	1.5E1
FQE	<b>2.9E-2</b>	1.4E0	<b>2.9E-2</b>	2.9E-2
R( $\lambda$ )	3.2E-1	6.0E-1	6.7E-2	1.2E-1
Q $^\pi$ ( $\lambda$ )	1.0E-1	5.3E-1	6.4E-2	9.8E-2
TREE	7.2E-1	1.5E0	3.4E-2	1.6E-1
IH	3.6E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.5E2	1.5E0
WIS	1.5E-1	<b>3.1E-2</b>
NAIVE	3.1E-1	-

Table 381: Graph-MC, relative MSE.  $T = 250, N = 512, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.5$ .

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.6E-1	1.0E-1	6.6E-2	1.2E-1
Q-REG	6.1E-2	2.2E0	2.1E-1	6.4E-2
MRDR	4.6E-2	4.8E-1	1.2E0	1.2E0
FQE	<b>5.6E-3</b>	4.8E-3	<b>3.8E-3</b>	5.8E-3
R( $\lambda$ )	2.7E-1	6.4E-2	2.0E-2	6.8E-2
Q $^\pi$ ( $\lambda$ )	2.1E-2	2.2E-2	7.3E-3	1.5E-2
TREE	7.2E-1	5.7E-2	1.1E-2	7.7E-2
IH	1.4E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	6.9E-1	5.7E-2
WIS	9.1E-2	<b>8.8E-3</b>
NAIVE	3.0E-1	-

Table 380: Graph-MC, relative MSE.  $T = 250, N = 256, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.5$ .

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.8E-1	1.1E-1	9.0E-2	1.8E-1
Q-REG	8.4E-2	2.4E-1	9.5E-2	5.0E-2
MRDR	8.3E-2	6.2E-1	1.0E0	9.9E-1
FQE	<b>6.2E-3</b>	4.8E-3	<b>3.2E-3</b>	6.5E-3
R( $\lambda$ )	3.0E-1	1.1E-1	2.6E-2	8.6E-2
Q $^\pi$ ( $\lambda$ )	1.4E-2	2.5E-2	5.5E-3	9.3E-3
TREE	7.1E-1	8.2E-2	1.3E-2	1.1E-1
IH	3.3E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	6.1E-1	9.1E-2
WIS	7.4E-2	<b>9.5E-3</b>
NAIVE	3.0E-1	-

Table 382: Graph-MC, relative MSE.  $T = 250, N = 1024, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.5$ .

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.4E-1	7.4E-2	3.7E-2	7.3E-2
Q-REG	4.2E-2	2.0E-2	3.0E-2	3.0E-2
MRDR	3.2E-2	2.6E-1	5.1E-1	3.4E-1
FQE	5.4E-3	2.3E-2	<b>2.7E-3</b>	5.7E-3
R( $\lambda$ )	2.7E-1	2.4E-2	1.2E-2	6.3E-2
Q $^\pi$ ( $\lambda$ )	1.7E-2	1.5E-2	2.7E-3	1.1E-2
TREE	7.2E-1	4.0E-2	6.1E-3	5.4E-2
IH	<b>1.6E-3</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	3.3E0	4.0E-2
WIS	5.7E-2	<b>3.1E-3</b>
NAIVE	3.1E-1	-

Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning

---

Table 383: Graph-MC, relative MSE.  $T = 250, N = 128, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.6$ .

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.8E-2	7.7E-3	7.7E-3	1.8E-2
Q-REG	2.2E-3	3.5E-4	3.5E-4	1.7E-4
MRDR	2.3E-3	1.8E-2	1.8E-2	8.0E-3
FQE	<b>6.3E-5</b>	3.7E-5	3.7E-5	<b>3.6E-5</b>
R( $\lambda$ )	3.9E-4	1.4E-4	1.4E-4	2.9E-4
Q $^\pi$ ( $\lambda$ )	3.6E-4	1.3E-4	1.3E-4	2.7E-4
TREE	4.7E-1	1.4E-3	1.4E-3	3.5E-3
IH	1.8E-3	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.8E-3</b>	1.8E-3
WIS	1.8E-3	1.8E-3
NAIVE	1.8E-3	-

Table 385: Graph-MC, relative MSE.  $T = 250, N = 512, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.6$ .

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.0E-2	1.2E-3	1.2E-3	1.3E-3
Q-REG	6.6E-4	1.7E-5	1.7E-5	1.7E-5
MRDR	7.6E-4	6.6E-4	6.6E-4	6.1E-4
FQE	1.3E-5	1.8E-5	1.8E-5	1.8E-5
R( $\lambda$ )	<b>1.0E-5</b>	1.5E-5	1.5E-5	<b>1.3E-5</b>
Q $^\pi$ ( $\lambda$ )	1.1E-5	1.5E-5	1.5E-5	1.3E-5
TREE	4.6E-1	3.4E-4	3.4E-4	3.8E-4
IH	3.4E-4	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>3.4E-4</b>	3.4E-4
WIS	3.4E-4	3.4E-4
NAIVE	3.4E-4	-

Table 384: Graph-MC, relative MSE.  $T = 250, N = 256, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.6$ .

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.3E-2	1.1E-3	1.1E-3	3.7E-3
Q-REG	9.7E-4	8.9E-5	8.9E-5	8.2E-5
MRDR	1.0E-3	3.1E-3	3.1E-3	2.1E-3
FQE	<b>4.8E-5</b>	3.7E-5	3.7E-5	<b>3.6E-5</b>
R( $\lambda$ )	1.4E-4	8.4E-5	8.4E-5	9.7E-5
Q $^\pi$ ( $\lambda$ )	1.3E-4	8.3E-5	8.3E-5	9.4E-5
TREE	4.7E-1	2.8E-4	2.8E-4	9.3E-4
IH	3.6E-4	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>3.6E-4</b>	3.6E-4
WIS	3.6E-4	3.6E-4
NAIVE	3.6E-4	-

Table 386: Graph-MC, relative MSE.  $T = 250, N = 1024, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.6$ .

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.0E-2	6.0E-4	6.0E-4	7.1E-4
Q-REG	4.3E-4	<b>2.0E-6</b>	2.0E-6	2.0E-6
MRDR	4.6E-4	2.1E-4	2.1E-4	2.0E-4
FQE	<b>2.0E-6</b>	2.0E-6	2.0E-6	2.0E-6
R( $\lambda$ )	6.0E-6	3.0E-6	3.0E-6	3.0E-6
Q $^\pi$ ( $\lambda$ )	5.0E-6	3.0E-6	3.0E-6	3.0E-6
TREE	4.5E-1	4.9E-4	4.9E-4	4.7E-4
IH	4.0E-4	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>4.0E-4</b>	4.0E-4
WIS	4.0E-4	4.0E-4
NAIVE	4.0E-4	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

Table 387: Graph-MC, relative MSE.  $T = 250, N = 128, \pi_b(a = 0) = 0.8, \pi_e(a = 0) = 0.2$ .

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	9.0E-1	9.0E-1	9.1E-1	9.0E-1
Q-REG	8.2E-1	9.0E-1	8.5E-1	8.1E-1
MRDR	7.4E-1	9.2E-1	8.8E0	8.7E0
FQE	8.7E-1	8.7E-1	8.7E-1	8.7E-1
R( $\lambda$ )	8.7E-1	8.7E-1	5.8E-1	6.1E-1
Q $^\pi$ ( $\lambda$ )	1.1E125	1.0E123	9.3E124	1.0E0
TREE	8.8E-1	8.7E-1	<b>5.8E-1</b>	6.1E-1
IH	<b>4.5E-1</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	8.2E-1
WIS	6.7E-1	<b>4.9E-1</b>
NAIVE	6.7E-1	-

Table 389: Graph-MC, relative MSE.  $T = 250, N = 512, \pi_b(a = 0) = 0.8, \pi_e(a = 0) = 0.2$ .

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	8.7E-1	8.8E-1	9.0E-1	8.7E-1
Q-REG	8.4E-1	8.3E-1	7.1E-1	7.8E-1
MRDR	7.5E-1	8.4E-1	1.5E1	1.5E1
FQE	8.5E-1	8.5E-1	8.5E-1	8.5E-1
R( $\lambda$ )	8.5E-1	8.5E-1	5.1E-1	5.4E-1
Q $^\pi$ ( $\lambda$ )	3.4E114	3.3E112	3.9E114	1.0E0
TREE	8.5E-1	8.5E-1	<b>5.1E-1</b>	5.5E-1
IH	<b>4.3E-1</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	8.4E-1
WIS	6.7E-1	<b>4.3E-1</b>
NAIVE	6.7E-1	-

Table 388: Graph-MC, relative MSE.  $T = 250, N = 256, \pi_b(a = 0) = 0.8, \pi_e(a = 0) = 0.2$ .

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	8.8E-1	8.9E-1	9.0E-1	8.8E-1
Q-REG	7.2E-1	1.0E0	1.2E0	6.0E-1
MRDR	7.0E-1	1.2E0	9.2E0	8.5E0
FQE	8.3E-1	8.3E-1	8.3E-1	8.3E-1
R( $\lambda$ )	8.4E-1	8.0E-1	5.5E-1	5.8E-1
Q $^\pi$ ( $\lambda$ )	5.4E107	1.6E110	2.9E108	1.0E0
TREE	8.5E-1	7.9E-1	<b>5.4E-1</b>	5.7E-1
IH	<b>3.3E-1</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	6.9E-1
WIS	6.7E-1	<b>4.6E-1</b>
NAIVE	6.7E-1	-

Table 390: Graph-MC, relative MSE.  $T = 250, N = 1024, \pi_b(a = 0) = 0.8, \pi_e(a = 0) = 0.2$ .

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	8.5E-1	8.1E-1	8.2E-1	8.5E-1
Q-REG	7.5E-1	8.3E-1	9.6E-1	9.5E-1
MRDR	6.9E-1	1.1E0	7.6E0	7.6E0
FQE	8.2E-1	8.2E-1	8.2E-1	8.2E-1
R( $\lambda$ )	8.2E-1	8.1E-1	5.4E-1	5.6E-1
Q $^\pi$ ( $\lambda$ )	2.4E112	1.5E110	6.8E112	1.0E0
TREE	8.3E-1	7.9E-1	<b>5.2E-1</b>	5.4E-1
IH	<b>3.5E-1</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	7.3E-1
WIS	6.7E-1	<b>4.1E-1</b>
NAIVE	6.7E-1	-

## F.4 Detailed Results for Mountain Car (MC)

 Table 391: MC, relative MSE. Model Type: linear.  $T = 250, N = 128, \pi_b = 0.10$ -Greedy(DDQN),  $\pi_e = 0.00$ -Greedy(DDQN).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	7.0E-2	1.3E-2	7.6E-3	1.0E-2
Q-REG	2.2E-1	2.6E-3	7.9E-4	2.4E-3
MRDR	9.2E-1	8.2E-3	5.9E-4	5.9E-4
FQE	5.7E-1	7.5E-3	1.2E-3	1.1E-3
R( $\lambda$ )	1.7E-1	1.8E-3	3.8E-4	3.9E-3
Q $^\pi$ ( $\lambda$ )	1.5E-1	1.9E-3	3.9E-4	5.1E-3
TREE	1.7E-1	1.8E-3	<b>3.7E-4</b>	4.3E-3
IH	<b>3.2E-2</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	4.4E-2	9.4E-3
WIS	<b>5.2E-4</b>	6.0E-4
NAIVE	3.2E-2	-

 Table 392: MC, relative MSE. Model Type: linear.  $T = 250, N = 256, \pi_b = 0.10$ -Greedy(DDQN),  $\pi_e = 0.00$ -Greedy(DDQN).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	7.5E-2	7.0E-3	4.7E-4	1.0E-3
Q-REG	2.0E-1	2.5E-3	8.6E-4	6.5E-4
MRDR	9.2E-1	2.9E-3	2.6E-4	<b>1.4E-4</b>
FQE	5.8E-1	5.3E-3	1.7E-3	1.7E-3
R( $\lambda$ )	1.7E-1	2.2E-3	5.6E-4	2.2E-3
Q $^\pi$ ( $\lambda$ )	1.5E-1	2.3E-3	6.1E-4	1.5E-3
TREE	1.7E-1	2.2E-3	5.8E-4	1.6E-3
IH	<b>3.2E-2</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.7E-2	3.3E-3
WIS	4.7E-4	<b>3.4E-4</b>
NAIVE	3.2E-2	-

 Table 393: MC, relative MSE. Model Type: NN.  $T = 250, N = 128, \pi_b = 0.10$ -Greedy(DDQN),  $\pi_e = 0.00$ -Greedy(DDQN).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.0E-1	9.7E-3	9.5E-3	1.5E-2
Q-REG	2.0E-1	2.7E-3	8.9E-4	3.2E-3
MRDR	8.8E-1	6.7E-3	8.4E-4	<b>5.2E-4</b>
FQE	1.2E-2	9.5E-4	7.6E-4	7.6E-4
R( $\lambda$ )	<b>8.9E-3</b>	6.4E-3	1.7E-3	5.2E-3
Q $^\pi$ ( $\lambda$ )	1.4E-1	5.7E-3	1.6E-3	2.8E-3
TREE	8.7E-2	5.7E-3	1.5E-3	6.6E-3
IH	3.3E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	4.0E-2	7.4E-3
WIS	8.2E-4	<b>8.1E-4</b>
NAIVE	3.3E-2	-

 Table 394: MC, relative MSE. Model Type: NN.  $T = 250, N = 256, \pi_b = 0.10$ -Greedy(DDQN),  $\pi_e = 0.00$ -Greedy(DDQN).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.4E-2	2.1E-3	2.0E-3	1.1E-2
Q-REG	1.1E-1	1.1E-3	4.2E-4	4.5E-4
MRDR	5.9E-1	2.5E-3	6.1E-4	1.2E-3
FQE	<b>8.6E-3</b>	2.2E-4	<b>1.8E-4</b>	1.9E-4
R( $\lambda$ )	1.6E-1	1.5E-3	1.2E-3	2.5E-3
Q $^\pi$ ( $\lambda$ )	3.4E-2	1.2E-3	6.9E-4	2.6E-3
TREE	1.4E-2	1.9E-3	1.1E-3	1.5E-3
IH	3.1E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.6E-2	5.2E-3
WIS	8.1E-4	<b>5.0E-4</b>
NAIVE	3.1E-2	-



**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 395: MC, relative MSE. Model Type: linear.  $T = 250, N = 128, \pi_b = 0.10$ -Greedy(DDQN),  $\pi_e = 1.00$ -Greedy(DDQN).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.2E-1	4.1E-1	2.4E0	6.5E-1
Q-REG	9.2E-1	8.6E-1	2.0E0	5.5E-1
MRDR	9.9E-1	8.8E-1	4.3E-1	4.4E-1
FQE	6.0E-1	5.2E-1	<b>2.3E-1</b>	2.7E-1
$R(\lambda)$	7.3E-1	6.4E-1	4.6E-1	4.8E-1
$Q^\pi(\lambda)$	6.4E-1	5.9E-1	4.6E-1	4.5E-1
TREE	7.3E-1	6.4E-1	4.7E-1	4.8E-1
IH	<b>4.1E-1</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	8.9E-1
WIS	4.9E-1	<b>4.3E-1</b>
NAIVE	4.1E-1	-

Table 397: MC, relative MSE. Model Type: NN.  $T = 250, N = 128, \pi_b = 0.10$ -Greedy(DDQN),  $\pi_e = 1.00$ -Greedy(DDQN).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	9.5E-1	8.5E-1	7.1E-1	7.7E-1
Q-REG	9.2E-1	7.5E-1	4.3E-1	4.8E-1
MRDR	1.0E0	8.5E-1	4.2E-1	4.5E-1
FQE	<b>6.1E-2</b>	4.9E-2	6.8E-2	<b>4.8E-2</b>
$R(\lambda)$	3.5E-1	3.0E-1	2.0E-1	1.6E-1
$Q^\pi(\lambda)$	1.8E0	1.4E0	3.1E0	3.8E0
TREE	2.4E-1	2.4E-1	2.1E-1	1.5E-1
IH	4.1E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	8.4E-1
WIS	4.7E-1	<b>4.3E-1</b>
NAIVE	4.1E-1	-

Table 396: MC, relative MSE. Model Type: linear.  $T = 250, N = 256, \pi_b = 0.10$ -Greedy(DDQN),  $\pi_e = 1.00$ -Greedy(DDQN).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.2E-1	6.3E-1	2.2E0	<b>2.1E-1</b>
Q-REG	8.8E-1	7.6E-1	5.8E-1	6.5E-1
MRDR	9.9E-1	8.6E-1	4.0E-1	4.2E-1
FQE	6.0E-1	4.9E-1	2.4E-1	3.0E-1
$R(\lambda)$	7.3E-1	6.2E-1	4.0E-1	4.5E-1
$Q^\pi(\lambda)$	6.4E-1	5.4E-1	3.5E-1	4.4E-1
TREE	7.3E-1	6.2E-1	4.1E-1	4.5E-1
IH	<b>4.0E-1</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	8.6E-1
WIS	4.7E-1	<b>4.1E-1</b>
NAIVE	4.1E-1	-

Table 398: MC, relative MSE. Model Type: NN.  $T = 250, N = 256, \pi_b = 0.10$ -Greedy(DDQN),  $\pi_e = 1.00$ -Greedy(DDQN).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	9.5E-1	8.3E-1	5.7E-1	7.4E-1
Q-REG	9.2E-1	7.8E-1	4.4E-1	4.9E-1
MRDR	1.0E0	8.7E-1	4.2E-1	4.4E-1
FQE	<b>4.3E-2</b>	3.3E-2	<b>1.6E-2</b>	3.2E-2
$R(\lambda)$	5.7E-1	4.6E-1	2.0E-1	2.2E-1
$Q^\pi(\lambda)$	2.1E1	2.0E1	1.8E1	2.7E1
TREE	4.5E-1	3.4E-1	1.3E-1	1.7E-1
IH	4.0E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	8.7E-1
WIS	4.8E-1	<b>4.2E-1</b>
NAIVE	4.1E-1	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 399: MC, relative MSE. Model Type: linear.  $T = 250, N = 128, \pi_b = 1.00$ -Greedy(DDQN),  $\pi_e = 0.00$ -Greedy(DDQN).

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	1.4E-1	<b>3.2E-1</b>	3.0E0	2.7E0
Q-REG	6.6E-1	5.1E2	7.9E26	6.7E1
MRDR	9.4E-1	9.7E0	5.0E0	4.9E0
FQE	2.2E-1	1.8E0	2.9E0	2.7E0
R( $\lambda$ )	1.6E-2	3.4E0	1.9E0	1.6E0
Q $^\pi$ ( $\lambda$ )	9.1E-2	3.6E0	2.0E0	1.8E0
TREE	<b>1.5E-2</b>	3.4E0	1.9E0	1.7E0
IH	5.2E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.0E0</b>	1.1E0
WIS	-	-
NAIVE	5.2E0	-

Table 401: MC, relative MSE. Model Type: NN.  $T = 250, N = 128, \pi_b = 1.00$ -Greedy(DDQN),  $\pi_e = 0.00$ -Greedy(DDQN).

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	8.6E-1	7.3E-1	1.7E1	4.4E0
Q-REG	8.4E-1	5.5E-1	1.1E1	1.1E1
MRDR	9.4E-1	5.0E-1	7.5E0	7.3E0
FQE	<b>1.2E-1</b>	<b>3.9E-1</b>	2.3E0	2.2E0
R( $\lambda$ )	1.5E0	1.7E0	4.9E0	4.7E0
Q $^\pi$ ( $\lambda$ )	5.4E0	5.6E0	4.5E0	4.4E0
TREE	2.0E0	2.3E0	4.8E0	4.6E0
IH	5.2E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	<b>5.4E-1</b>
WIS	-	-
NAIVE	5.2E0	-

Table 400: MC, relative MSE. Model Type: linear.  $T = 250, N = 256, \pi_b = 1.00$ -Greedy(DDQN),  $\pi_e = 0.00$ -Greedy(DDQN).

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	1.2E-1	1.7E-1	3.6E0	3.4E0
Q-REG	7.0E-1	2.5E23	1.5E27	1.9E1
MRDR	9.6E-1	6.7E-1	5.1E0	5.0E0
FQE	2.1E-1	2.3E-1	3.0E0	2.8E0
R( $\lambda$ )	1.6E-2	<b>8.0E-3</b>	1.8E0	1.7E0
Q $^\pi$ ( $\lambda$ )	8.2E-2	2.2E-1	1.6E0	1.6E0
TREE	<b>1.5E-2</b>	8.4E-3	1.7E0	1.7E0
IH	5.1E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	<b>6.7E-1</b>
WIS	-	-
NAIVE	5.1E0	-

Table 402: MC, relative MSE. Model Type: NN.  $T = 250, N = 256, \pi_b = 1.00$ -Greedy(DDQN),  $\pi_e = 0.00$ -Greedy(DDQN).

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	8.6E-1	4.7E-1	5.1E0	4.8E0
Q-REG	7.7E-1	4.8E-1	1.1E1	1.1E1
MRDR	9.0E-1	4.7E-1	7.0E0	6.9E0
FQE	<b>2.7E-2</b>	<b>1.3E-1</b>	3.2E-1	2.9E-1
R( $\lambda$ )	1.0E0	1.2E0	4.3E0	4.3E0
Q $^\pi$ ( $\lambda$ )	5.6E0	5.6E0	4.4E0	4.5E0
TREE	2.1E0	2.5E0	4.7E0	4.7E0
IH	5.1E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	<b>5.3E-1</b>
WIS	-	-
NAIVE	5.1E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 403: MC, relative MSE. Model Type: linear.  $T = 250, N = 128, \pi_b = 1.00$ -Greedy(DDQN),  $\pi_e = 0.10$ -Greedy(DDQN).

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	9.3E-2	3.2E-1	2.9E0	1.0E0
Q-REG	6.3E-1	4.6E-1	4.7E0	1.6E0
MRDR	8.5E-1	6.0E-1	3.5E-1	3.5E-1
FQE	2.7E-1	1.4E-1	7.1E-1	6.1E-1
R( $\lambda$ )	6.5E-2	1.8E-2	3.5E-1	2.9E-1
Q $^\pi$ ( $\lambda$ )	<b>9.4E-3</b>	1.1E-1	4.9E-1	3.9E-1
TREE	6.3E-2	<b>1.8E-2</b>	3.7E-1	3.1E-1
IH	3.0E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	5.4E-1
WIS	<b>2.8E-1</b>	3.6E-1
NAIVE	3.0E0	-

Table 405: MC, relative MSE. Model Type: NN.  $T = 250, N = 128, \pi_b = 1.00$ -Greedy(DDQN),  $\pi_e = 0.10$ -Greedy(DDQN).

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	8.7E-1	6.1E-1	7.7E-1	7.5E-1
Q-REG	8.1E-1	5.7E-1	3.7E-1	4.1E-1
MRDR	9.6E-1	6.9E-1	6.5E-1	6.6E-1
FQE	<b>1.7E-2</b>	<b>3.9E-3</b>	3.9E-1	2.0E-1
R( $\lambda$ )	8.1E-1	9.0E-1	1.5E0	1.4E0
Q $^\pi$ ( $\lambda$ )	3.7E0	3.8E0	3.2E0	3.2E0
TREE	8.9E-1	9.8E-1	1.6E0	1.4E0
IH	3.1E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	7.1E-1
WIS	<b>2.0E-1</b>	3.9E-1
NAIVE	3.1E0	-

Table 404: MC, relative MSE. Model Type: linear.  $T = 250, N = 256, \pi_b = 1.00$ -Greedy(DDQN),  $\pi_e = 0.10$ -Greedy(DDQN).

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	7.4E-2	1.1E0	4.1E0	7.6E-1
Q-REG	6.6E-1	4.2E-1	7.1E-1	7.4E-1
MRDR	7.4E-1	4.6E-1	5.2E-1	3.5E-1
FQE	2.6E-1	1.7E-1	7.0E-1	3.9E-1
R( $\lambda$ )	6.2E-2	<b>1.6E-2</b>	2.9E-1	2.4E-1
Q $^\pi$ ( $\lambda$ )	<b>1.4E-2</b>	1.0E-1	4.8E-1	4.7E-1
TREE	5.9E-2	1.6E-2	3.3E-1	2.6E-1
IH	3.1E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	5.7E-1
WIS	<b>2.3E-2</b>	2.2E-1
NAIVE	3.1E0	-

Table 406: MC, relative MSE. Model Type: NN.  $T = 250, N = 256, \pi_b = 1.00$ -Greedy(DDQN),  $\pi_e = 0.10$ -Greedy(DDQN).

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	8.8E-1	4.0E-1	5.8E-1	4.0E-1
Q-REG	7.6E-1	4.1E-1	1.9E-1	2.1E-1
MRDR	9.7E-1	5.3E-1	3.1E-1	3.2E-1
FQE	<b>5.0E-3</b>	1.4E-1	1.3E-1	<b>1.5E-2</b>
R( $\lambda$ )	1.1E0	1.2E0	2.1E0	1.6E0
Q $^\pi$ ( $\lambda$ )	2.7E0	2.8E0	2.4E0	2.4E0
TREE	8.8E-1	1.0E0	1.6E0	1.3E0
IH	3.1E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	5.4E-1
WIS	<b>8.7E-2</b>	2.5E-1
NAIVE	3.1E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 407: MC, relative MSE. Model Type: linear.  $T = 250, N = 128, \pi_b = 0.10$ -Greedy(DDQN),  $\pi_e = 0.00$ -Greedy(DDQN).

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	7.0E-2	1.3E-2	7.6E-3	1.0E-2
Q-REG	2.2E-1	2.6E-3	7.9E-4	2.4E-3
MRDR	9.2E-1	8.2E-3	5.9E-4	5.9E-4
FQE	5.7E-1	7.5E-3	1.2E-3	1.1E-3
R( $\lambda$ )	1.7E-1	1.8E-3	3.8E-4	3.9E-3
Q $^\pi$ ( $\lambda$ )	1.5E-1	1.9E-3	3.9E-4	5.1E-3
TREE	1.7E-1	1.8E-3	<b>3.7E-4</b>	4.3E-3
IH	<b>3.2E-2</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	4.4E-2	9.4E-3
WIS	<b>5.2E-4</b>	6.0E-4
NAIVE	3.2E-2	-

Table 409: MC, relative MSE. Model Type: NN.  $T = 250, N = 128, \pi_b = 0.10$ -Greedy(DDQN),  $\pi_e = 0.00$ -Greedy(DDQN).

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	2.0E-1	9.7E-3	9.5E-3	1.5E-2
Q-REG	2.0E-1	2.7E-3	8.9E-4	3.2E-3
MRDR	8.8E-1	6.7E-3	8.4E-4	<b>5.2E-4</b>
FQE	1.2E-2	9.5E-4	7.6E-4	7.6E-4
R( $\lambda$ )	<b>8.9E-3</b>	6.4E-3	1.7E-3	5.2E-3
Q $^\pi$ ( $\lambda$ )	1.4E-1	5.7E-3	1.6E-3	2.8E-3
TREE	8.7E-2	5.7E-3	1.5E-3	6.6E-3
IH	3.3E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	4.0E-2	7.4E-3
WIS	8.2E-4	<b>8.1E-4</b>
NAIVE	3.3E-2	-

Table 408: MC, relative MSE. Model Type: linear.  $T = 250, N = 256, \pi_b = 0.10$ -Greedy(DDQN),  $\pi_e = 0.00$ -Greedy(DDQN).

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	7.5E-2	7.0E-3	4.7E-4	1.0E-3
Q-REG	2.0E-1	2.5E-3	8.6E-4	6.5E-4
MRDR	9.2E-1	2.9E-3	2.6E-4	<b>1.4E-4</b>
FQE	5.8E-1	5.3E-3	1.7E-3	1.7E-3
R( $\lambda$ )	1.7E-1	2.2E-3	5.6E-4	2.2E-3
Q $^\pi$ ( $\lambda$ )	1.5E-1	2.3E-3	6.1E-4	1.5E-3
TREE	1.7E-1	2.2E-3	5.8E-4	1.6E-3
IH	<b>3.2E-2</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.7E-2	3.3E-3
WIS	4.7E-4	<b>3.4E-4</b>
NAIVE	3.2E-2	-

Table 410: MC, relative MSE. Model Type: NN.  $T = 250, N = 256, \pi_b = 0.10$ -Greedy(DDQN),  $\pi_e = 0.00$ -Greedy(DDQN).

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	5.4E-2	2.1E-3	2.0E-3	1.1E-2
Q-REG	1.1E-1	1.1E-3	4.2E-4	4.5E-4
MRDR	5.9E-1	2.5E-3	6.1E-4	1.2E-3
FQE	<b>8.6E-3</b>	2.2E-4	<b>1.8E-4</b>	1.9E-4
R( $\lambda$ )	1.6E-1	1.5E-3	1.2E-3	2.5E-3
Q $^\pi$ ( $\lambda$ )	3.4E-2	1.2E-3	6.9E-4	2.6E-3
TREE	1.4E-2	1.9E-3	1.1E-3	1.5E-3
IH	3.1E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.6E-2	5.2E-3
WIS	8.1E-4	<b>5.0E-4</b>
NAIVE	3.1E-2	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 411: MC, relative MSE. Model Type: linear.  $T = 250, N = 128, \pi_b = 0.10$ -Greedy(DDQN),  $\pi_e = 1.00$ -Greedy(DDQN).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.2E-1	4.1E-1	2.4E0	6.5E-1
Q-REG	9.2E-1	8.6E-1	2.0E0	5.5E-1
MRDR	9.9E-1	8.8E-1	4.3E-1	4.4E-1
FQE	6.0E-1	5.2E-1	<b>2.3E-1</b>	2.7E-1
R( $\lambda$ )	7.3E-1	6.4E-1	4.6E-1	4.8E-1
Q $^\pi$ ( $\lambda$ )	6.4E-1	5.9E-1	4.6E-1	4.5E-1
TREE	7.3E-1	6.4E-1	4.7E-1	4.8E-1
IH	<b>4.1E-1</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	8.9E-1
WIS	4.9E-1	<b>4.3E-1</b>
NAIVE	4.1E-1	-

Table 413: MC, relative MSE. Model Type: NN.  $T = 250, N = 128, \pi_b = 0.10$ -Greedy(DDQN),  $\pi_e = 1.00$ -Greedy(DDQN).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	9.5E-1	8.5E-1	7.1E-1	7.7E-1
Q-REG	9.2E-1	7.5E-1	4.3E-1	4.8E-1
MRDR	1.0E0	8.5E-1	4.2E-1	4.5E-1
FQE	<b>6.1E-2</b>	4.9E-2	6.8E-2	<b>4.8E-2</b>
R( $\lambda$ )	3.5E-1	3.0E-1	2.0E-1	1.6E-1
Q $^\pi$ ( $\lambda$ )	1.8E0	1.4E0	3.1E0	3.8E0
TREE	2.4E-1	2.4E-1	2.1E-1	1.5E-1
IH	4.1E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	8.4E-1
WIS	4.7E-1	<b>4.3E-1</b>
NAIVE	4.1E-1	-

Table 412: MC, relative MSE. Model Type: linear.  $T = 250, N = 256, \pi_b = 0.10$ -Greedy(DDQN),  $\pi_e = 1.00$ -Greedy(DDQN).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.2E-1	6.3E-1	2.2E0	<b>2.1E-1</b>
Q-REG	8.8E-1	7.6E-1	5.8E-1	6.5E-1
MRDR	9.9E-1	8.6E-1	4.0E-1	4.2E-1
FQE	6.0E-1	4.9E-1	2.4E-1	3.0E-1
R( $\lambda$ )	7.3E-1	6.2E-1	4.0E-1	4.5E-1
Q $^\pi$ ( $\lambda$ )	6.4E-1	5.4E-1	3.5E-1	4.4E-1
TREE	7.3E-1	6.2E-1	4.1E-1	4.5E-1
IH	<b>4.0E-1</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	8.6E-1
WIS	4.7E-1	<b>4.1E-1</b>
NAIVE	4.1E-1	-

Table 414: MC, relative MSE. Model Type: NN.  $T = 250, N = 256, \pi_b = 0.10$ -Greedy(DDQN),  $\pi_e = 1.00$ -Greedy(DDQN).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	9.5E-1	8.3E-1	5.7E-1	7.4E-1
Q-REG	9.2E-1	7.8E-1	4.4E-1	4.9E-1
MRDR	1.0E0	8.7E-1	4.2E-1	4.4E-1
FQE	<b>4.3E-2</b>	3.3E-2	<b>1.6E-2</b>	3.2E-2
R( $\lambda$ )	5.7E-1	4.6E-1	2.0E-1	2.2E-1
Q $^\pi$ ( $\lambda$ )	2.1E1	2.0E1	1.8E1	2.7E1
TREE	4.5E-1	3.4E-1	1.3E-1	1.7E-1
IH	4.0E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	8.7E-1
WIS	4.8E-1	<b>4.2E-1</b>
NAIVE	4.1E-1	-

Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning

---

Table 415: MC, relative MSE. Model Type: linear.  $T = 250, N = 128, \pi_b = 1.00$ -Greedy(DDQN),  $\pi_e = 0.00$ -Greedy(DDQN).

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	1.4E-1	<b>3.2E-1</b>	3.0E0	2.7E0
Q-REG	6.6E-1	5.1E2	7.9E26	6.7E1
MRDR	9.4E-1	9.7E0	5.0E0	4.9E0
FQE	2.2E-1	1.8E0	2.9E0	2.7E0
R( $\lambda$ )	1.6E-2	3.4E0	1.9E0	1.6E0
Q $^\pi$ ( $\lambda$ )	9.1E-2	3.6E0	2.0E0	1.8E0
TREE	<b>1.5E-2</b>	3.4E0	1.9E0	1.7E0
IH	5.2E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.0E0</b>	1.1E0
WIS	-	-
NAIVE	5.2E0	-

Table 417: MC, relative MSE. Model Type: NN.  $T = 250, N = 128, \pi_b = 1.00$ -Greedy(DDQN),  $\pi_e = 0.00$ -Greedy(DDQN).

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	8.6E-1	7.3E-1	1.7E1	4.4E0
Q-REG	8.4E-1	5.5E-1	1.1E1	1.1E1
MRDR	9.4E-1	5.0E-1	7.5E0	7.3E0
FQE	<b>1.2E-1</b>	<b>3.9E-1</b>	2.3E0	2.2E0
R( $\lambda$ )	1.5E0	1.7E0	4.9E0	4.7E0
Q $^\pi$ ( $\lambda$ )	5.4E0	5.6E0	4.5E0	4.4E0
TREE	2.0E0	2.3E0	4.8E0	4.6E0
IH	5.2E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	<b>5.4E-1</b>
WIS	-	-
NAIVE	5.2E0	-

Table 416: MC, relative MSE. Model Type: linear.  $T = 250, N = 256, \pi_b = 1.00$ -Greedy(DDQN),  $\pi_e = 0.00$ -Greedy(DDQN).

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	1.2E-1	1.7E-1	3.6E0	3.4E0
Q-REG	7.0E-1	2.5E23	1.5E27	1.9E1
MRDR	9.6E-1	6.7E-1	5.1E0	5.0E0
FQE	2.1E-1	2.3E-1	3.0E0	2.8E0
R( $\lambda$ )	1.6E-2	<b>8.0E-3</b>	1.8E0	1.7E0
Q $^\pi$ ( $\lambda$ )	8.2E-2	2.2E-1	1.6E0	1.6E0
TREE	<b>1.5E-2</b>	8.4E-3	1.7E0	1.7E0
IH	5.1E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	<b>6.7E-1</b>
WIS	-	-
NAIVE	5.1E0	-

Table 418: MC, relative MSE. Model Type: NN.  $T = 250, N = 256, \pi_b = 1.00$ -Greedy(DDQN),  $\pi_e = 0.00$ -Greedy(DDQN).

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	8.6E-1	4.7E-1	5.1E0	4.8E0
Q-REG	7.7E-1	4.8E-1	1.1E1	1.1E1
MRDR	9.0E-1	4.7E-1	7.0E0	6.9E0
FQE	<b>2.7E-2</b>	<b>1.3E-1</b>	3.2E-1	2.9E-1
R( $\lambda$ )	1.0E0	1.2E0	4.3E0	4.3E0
Q $^\pi$ ( $\lambda$ )	5.6E0	5.6E0	4.4E0	4.5E0
TREE	2.1E0	2.5E0	4.7E0	4.7E0
IH	5.1E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	<b>5.3E-1</b>
WIS	-	-
NAIVE	5.1E0	-

Table 419: MC, relative MSE. Model Type: linear.  $T = 250, N = 128, \pi_b = 1.00$ -Greedy(DDQN),  $\pi_e = 0.10$ -Greedy(DDQN).

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	9.3E-2	3.2E-1	2.9E0	1.0E0
Q-REG	6.3E-1	4.6E-1	4.7E0	1.6E0
MRDR	8.5E-1	6.0E-1	3.5E-1	3.5E-1
FQE	2.7E-1	1.4E-1	7.1E-1	6.1E-1
R( $\lambda$ )	6.5E-2	1.8E-2	3.5E-1	2.9E-1
Q $^\pi$ ( $\lambda$ )	<b>9.4E-3</b>	1.1E-1	4.9E-1	3.9E-1
TREE	6.3E-2	<b>1.8E-2</b>	3.7E-1	3.1E-1
IH	3.0E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	5.4E-1
WIS	<b>2.8E-1</b>	3.6E-1
NAIVE	3.0E0	-

 Table 421: MC, relative MSE. Model Type: NN.  $T = 250, N = 128, \pi_b = 1.00$ -Greedy(DDQN),  $\pi_e = 0.10$ -Greedy(DDQN).

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	8.7E-1	6.1E-1	7.7E-1	7.5E-1
Q-REG	8.1E-1	5.7E-1	3.7E-1	4.1E-1
MRDR	9.6E-1	6.9E-1	6.5E-1	6.6E-1
FQE	<b>1.7E-2</b>	<b>3.9E-3</b>	3.9E-1	2.0E-1
R( $\lambda$ )	8.1E-1	9.0E-1	1.5E0	1.4E0
Q $^\pi$ ( $\lambda$ )	3.7E0	3.8E0	3.2E0	3.2E0
TREE	8.9E-1	9.8E-1	1.6E0	1.4E0
IH	3.1E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	7.1E-1
WIS	<b>2.0E-1</b>	3.9E-1
NAIVE	3.1E0	-

 Table 420: MC, relative MSE. Model Type: linear.  $T = 250, N = 256, \pi_b = 1.00$ -Greedy(DDQN),  $\pi_e = 0.10$ -Greedy(DDQN).

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	7.4E-2	1.1E0	4.1E0	7.6E-1
Q-REG	6.6E-1	4.2E-1	7.1E-1	7.4E-1
MRDR	7.4E-1	4.6E-1	5.2E-1	3.5E-1
FQE	2.6E-1	1.7E-1	7.0E-1	3.9E-1
R( $\lambda$ )	6.2E-2	<b>1.6E-2</b>	2.9E-1	2.4E-1
Q $^\pi$ ( $\lambda$ )	<b>1.4E-2</b>	1.0E-1	4.8E-1	4.7E-1
TREE	5.9E-2	1.6E-2	3.3E-1	2.6E-1
IH	3.1E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	5.7E-1
WIS	<b>2.3E-2</b>	2.2E-1
NAIVE	3.1E0	-

 Table 422: MC, relative MSE. Model Type: NN.  $T = 250, N = 256, \pi_b = 1.00$ -Greedy(DDQN),  $\pi_e = 0.10$ -Greedy(DDQN).

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	8.8E-1	4.0E-1	5.8E-1	4.0E-1
Q-REG	7.6E-1	4.1E-1	1.9E-1	2.1E-1
MRDR	9.7E-1	5.3E-1	3.1E-1	3.2E-1
FQE	<b>5.0E-3</b>	1.4E-1	1.3E-1	<b>1.5E-2</b>
R( $\lambda$ )	1.1E0	1.2E0	2.1E0	1.6E0
Q $^\pi$ ( $\lambda$ )	2.7E0	2.8E0	2.4E0	2.4E0
TREE	8.8E-1	1.0E0	1.6E0	1.3E0
IH	3.1E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	5.4E-1
WIS	<b>8.7E-2</b>	2.5E-1
NAIVE	3.1E0	-

**F.5 Detailed Results for Pixel-Based Mountain Car (Pix-MC)**

 Table 423: Pixel MC, relative MSE. Model Type: conv.  $T = 150, N = 512, \pi_b = 0.10$ -Greedy(DDQN),  $\pi_e = 0.00$ -Greedy(DDQN).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.6E5	2.1E5	2.0E5	1.1E5
Q-REG	6.8E-3	8.8E-3	9.1E-3	9.5E-3
MRDR	4.7E-3	3.0E-2	4.1E-2	1.8E-2
FQE	<b>3.2E-3</b>	1.1E-3	1.8E-3	<b>9.8E-4</b>
R( $\lambda$ )	-	-	-	-
Q $^\pi$ ( $\lambda$ )	-	-	-	-
TREE	-	-	-	-
IH	4.4E-3	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	4.4E-1	5.5E-3
WIS	5.2E-3	<b>3.8E-4</b>
NAIVE	1.0E-5	-

 Table 424: Pixel MC, relative MSE. Model Type: conv.  $T = 150, N = 512, \pi_b = 0.25$ -Greedy(DDQN),  $\pi_e = 0.00$ -Greedy(DDQN).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.4E5	8.2E4	7.7E5	5.7E5
Q-REG	1.0E-1	3.6E-3	6.9E-3	1.0E-2
MRDR	1.3E-1	9.3E-3	8.6E-3	4.4E-3
FQE	<b>2.6E-3</b>	7.1E-4	6.4E-4	<b>1.7E-4</b>
R( $\lambda$ )	-	-	-	-
Q $^\pi$ ( $\lambda$ )	-	-	-	-
TREE	-	-	-	-
IH	1.3E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	<b>1.6E-2</b>
WIS	-	-
NAIVE	7.5E-5	-

 Table 425: Pixel MC, relative MSE. Model Type: conv.  $T = 150, N = 512, \pi_b = 0.25$ -Greedy(DDQN),  $\pi_e = 0.10$ -Greedy(DDQN).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.1E2	1.2E3	2.6E3	9.5E2
Q-REG	3.8E-3	3.8E-2	3.1E-2	2.2E-2
MRDR	3.6E-2	4.5E-3	4.2E-3	2.6E-3
FQE	<b>1.5E-3</b>	8.0E-4	8.9E-4	<b>7.3E-4</b>
R( $\lambda$ )	-	-	-	-
Q $^\pi$ ( $\lambda$ )	-	-	-	-
TREE	-	-	-	-
IH	3.8E-3	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.3E-1	4.1E-3
WIS	3.4E-4	<b>7.8E-5</b>
NAIVE	3.1E-5	-



## F.6 Detailed Results for Gridworld

 Table 426: Gridworld, relative MSE.  $T = 25, N = 64, \pi_b = 0.20$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.2E-2	4.3E-2	4.4E-2	4.0E-2
Q-REG	1.0E-1	4.7E-2	4.2E-2	4.7E-2
MRDR	1.6E-1	4.5E-2	3.1E-2	2.9E-2
FQE	3.7E-2	3.6E-2	3.6E-2	3.7E-2
$R(\lambda)$	1.4E0	7.1E-2	2.6E-2	2.0E-2
$Q^\pi(\lambda)$	2.3E0	7.5E-2	6.4E-2	2.9E-2
TREE	1.1E0	6.6E-2	7.4E-3	<b>5.7E-3</b>
IH	<b>2.1E-2</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	4.7E-2	6.4E-2
WIS	1.6E-2	<b>6.6E-3</b>
NAIVE	9.6E-2	-

 Table 427: Gridworld, relative MSE.  $T = 25, N = 128, \pi_b = 0.20$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	<b>3.1E-3</b>	1.9E-3	2.1E-3	<b>1.7E-3</b>
Q-REG	3.9E-2	1.5E-2	1.3E-2	1.7E-2
MRDR	7.8E-2	1.1E-2	6.9E-3	8.7E-3
FQE	1.2E-2	1.1E-2	1.0E-2	1.1E-2
$R(\lambda)$	1.4E0	1.7E-2	1.2E-2	1.3E-2
$Q^\pi(\lambda)$	1.6E0	1.5E-2	8.8E-3	7.1E-3
TREE	9.1E-1	1.7E-2	2.6E-3	3.0E-3
IH	1.1E-2	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.3E-2	1.9E-2
WIS	4.1E-3	<b>1.1E-3</b>
NAIVE	9.6E-2	-

 Table 428: Gridworld, relative MSE.  $T = 25, N = 256, \pi_b = 0.20$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	<b>1.0E-3</b>	1.5E-3	1.6E-3	<b>1.2E-3</b>
Q-REG	6.6E-3	3.4E-3	3.3E-3	4.1E-3
MRDR	2.9E-2	3.0E-3	2.4E-3	3.5E-3
FQE	3.5E-3	2.3E-3	2.2E-3	2.6E-3
$R(\lambda)$	1.7E-1	5.7E-3	4.4E-3	6.4E-3
$Q^\pi(\lambda)$	2.3E-1	4.0E-3	2.9E-3	2.7E-3
TREE	4.3E-1	4.8E-3	1.2E-3	1.9E-3
IH	4.8E-3	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	4.8E-3	6.3E-3
WIS	1.0E-3	<b>5.1E-4</b>
NAIVE	1.1E-1	-

 Table 429: Gridworld, relative MSE.  $T = 25, N = 512, \pi_b = 0.20$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.9E-4	1.2E-3	1.2E-3	1.0E-3
Q-REG	5.0E-3	4.1E-4	4.0E-4	5.9E-4
MRDR	2.8E-2	4.1E-4	3.8E-4	6.1E-4
FQE	5.6E-4	2.4E-4	<b>2.3E-4</b>	3.1E-4
$R(\lambda)$	2.8E-3	4.4E-4	4.1E-4	1.3E-3
$Q^\pi(\lambda)$	<b>2.4E-4</b>	2.4E-4	2.4E-4	2.5E-4
TREE	3.4E-1	4.2E-4	2.4E-4	3.9E-4
IH	4.1E-3	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	9.9E-4	1.3E-3
WIS	4.0E-4	<b>3.7E-4</b>
NAIVE	9.0E-2	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 430: Gridworld, relative MSE.  $T = 25, N = 1024, \pi_b = 0.20$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.).

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	3.6E-4	1.7E-4	1.7E-4	<b>1.1E-4</b>	
Q-REG	2.0E-3	3.2E-4	3.2E-4	5.4E-4	
MRDR	2.3E-2	4.0E-4	3.9E-4	7.9E-4	
FQE	8.2E-4	3.3E-4	3.3E-4	4.7E-4	
R( $\lambda$ )	3.0E-3	4.6E-4	4.6E-4	1.5E-3	
Q $^\pi$ ( $\lambda$ )	<b>3.6E-4</b>	3.4E-4	3.4E-4	3.6E-4	
TREE	3.4E-1	3.9E-4	3.6E-4	8.2E-4	
IH	1.8E-3	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	7.4E-4	8.4E-4
WIS	<b>3.1E-4</b>	3.4E-4
NAIVE	9.1E-2	-

Table 432: Gridworld, relative MSE.  $T = 25, N = 128, \pi_b = 0.40$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.).

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	3.4E-2	1.2E-2	1.7E-2	1.2E-2	
Q-REG	2.0E-1	3.4E-2	2.7E-2	3.2E-2	
MRDR	2.1E-1	5.8E-2	2.0E-2	2.1E-2	
FQE	2.8E-2	2.1E-3	<b>9.8E-4</b>	1.5E-3	
R( $\lambda$ )	5.1E-1	6.1E-2	8.3E-3	1.1E-2	
Q $^\pi$ ( $\lambda$ )	<b>1.4E-3</b>	1.3E-3	1.1E-3	1.5E-3	
TREE	8.9E-1	1.5E-1	1.3E-2	3.8E-2	
IH	3.5E-2	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	2.6E-1	2.9E-1
WIS	<b>1.8E-2</b>	3.7E-2
NAIVE	1.2E0	-

Table 431: Gridworld, relative MSE.  $T = 25, N = 64, \pi_b = 0.40$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.).

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	2.0E-1	1.4E-1	1.2E-1	1.8E-1	
Q-REG	2.4E-1	1.1E-1	1.5E-1	3.8E-2	
MRDR	2.9E-1	8.6E-2	6.5E-2	3.1E-2	
FQE	<b>3.7E-2</b>	7.8E-3	<b>2.5E-3</b>	6.6E-3	
R( $\lambda$ )	7.7E-1	1.5E-1	2.8E-2	5.0E-2	
Q $^\pi$ ( $\lambda$ )	5.7E-2	8.2E-3	5.2E-3	5.9E-3	
TREE	1.0E0	1.7E-1	2.9E-2	1.6E-1	
IH	4.8E-2	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	1.2E-1	1.8E-1
WIS	<b>1.1E-2</b>	3.4E-2
NAIVE	1.2E0	-

Table 433: Gridworld, relative MSE.  $T = 25, N = 256, \pi_b = 0.40$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.).

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	1.8E-2	2.1E-2	2.0E-2	1.8E-2	
Q-REG	7.9E-2	3.5E-3	2.9E-3	4.6E-3	
MRDR	8.1E-2	1.1E-2	6.7E-3	4.1E-3	
FQE	2.1E-2	4.8E-4	<b>2.7E-4</b>	4.9E-4	
R( $\lambda$ )	4.7E-1	1.5E-2	3.4E-3	4.4E-3	
Q $^\pi$ ( $\lambda$ )	<b>4.2E-4</b>	3.1E-4	2.9E-4	4.9E-4	
TREE	8.7E-1	4.2E-2	4.9E-3	8.9E-3	
IH	2.8E-2	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	6.6E-2	8.5E-2
WIS	<b>3.8E-3</b>	1.1E-2
NAIVE	1.3E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 434: Gridworld, relative MSE.  $T = 25, N = 512, \pi_b = 0.40$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.).

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	1.1E-2	7.2E-3	6.9E-3	5.6E-3	
Q-REG	2.0E-2	6.2E-4	6.2E-4	9.3E-4	
MRDR	9.3E-2	1.0E-3	3.5E-4	6.8E-4	
FQE	1.8E-2	1.1E-4	4.6E-5	1.4E-4	
$R(\lambda)$	4.6E-1	6.7E-3	1.8E-3	3.1E-3	
$Q^\pi(\lambda)$	<b>5.7E-5</b>	3.6E-5	<b>2.9E-5</b>	9.1E-5	
TREE	8.8E-1	1.3E-2	3.5E-3	6.1E-3	
IH	2.7E-2	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	1.2E-2	1.7E-2
WIS	<b>1.1E-3</b>	4.2E-3
NAIVE	1.2E0	-

Table 436: Gridworld, relative MSE.  $T = 25, N = 64, \pi_b = 0.60$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.).

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	6.0E-1	3.6E-1	3.2E-1	5.3E-1	
Q-REG	2.7E0	2.7E0	2.4E0	1.3E0	
MRDR	1.3E0	8.1E0	2.7E0	1.1E0	
FQE	1.2E-1	1.1E-1	1.9E-2	1.4E-2	
$R(\lambda)$	1.2E0	1.2E0	2.4E-1	1.1E0	
$Q^\pi(\lambda)$	<b>1.8E-2</b>	2.3E-2	<b>1.1E-2</b>	1.5E-2	
TREE	1.2E0	1.5E0	3.4E-1	1.3E0	
IH	1.0E-1	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	1.1E0	1.6E0
WIS	<b>2.9E-1</b>	4.5E-1
NAIVE	3.9E0	-

Table 435: Gridworld, relative MSE.  $T = 25, N = 1024, \pi_b = 0.40$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.).

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	1.4E-2	3.3E-3	3.3E-3	2.7E-3	
Q-REG	6.6E-3	1.5E-4	1.5E-4	2.3E-4	
MRDR	6.0E-2	6.6E-4	2.7E-4	4.9E-4	
FQE	1.8E-2	5.9E-5	5.4E-5	9.0E-5	
$R(\lambda)$	4.8E-1	2.6E-3	2.2E-4	5.8E-4	
$Q^\pi(\lambda)$	<b>2.7E-5</b>	<b>3.3E-5</b>	3.5E-5	5.7E-5	
TREE	9.0E-1	6.1E-3	5.6E-4	2.0E-3	
IH	2.8E-2	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	7.8E-3	9.6E-3
WIS	<b>5.3E-4</b>	1.1E-3
NAIVE	1.2E0	-

Table 437: Gridworld, relative MSE.  $T = 25, N = 128, \pi_b = 0.60$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.).

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	3.9E-1	2.9E-1	2.5E-1	2.6E-1	
Q-REG	5.6E-1	1.8E0	1.0E0	3.1E-1	
MRDR	5.0E-1	3.7E0	1.7E0	4.6E-1	
FQE	1.1E-1	2.0E-2	1.3E-2	2.9E-3	
$R(\lambda)$	1.2E0	7.4E-1	9.3E-2	6.9E-1	
$Q^\pi(\lambda)$	<b>2.9E-3</b>	3.4E-3	<b>2.7E-3</b>	2.9E-3	
TREE	1.1E0	9.9E-1	1.3E-1	1.0E0	
IH	8.7E-2	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	5.5E-1	1.4E0
WIS	2.9E-1	<b>1.6E-1</b>
NAIVE	4.0E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 438: Gridworld, relative MSE.  $T = 25, N = 256, \pi_b = 0.60$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.5E-1	2.1E-1	1.9E-1	1.3E-1
Q-REG	3.0E-1	1.6E-1	2.5E-1	1.3E-1
MRDR	4.8E-1	3.9E-1	2.6E-1	3.6E-1
FQE	1.2E-1	5.1E-3	3.0E-3	2.4E-3
R( $\lambda$ )	1.2E0	4.0E-1	9.6E-2	5.2E-1
Q $^\pi$ ( $\lambda$ )	<b>1.4E-3</b>	1.1E-3	<b>7.4E-4</b>	1.4E-3
TREE	1.2E0	4.7E-1	1.2E-1	7.2E-1
IH	8.0E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	3.0E-1	5.2E-1
WIS	<b>5.1E-2</b>	1.3E-1
NAIVE	4.3E0	-

Table 440: Gridworld, relative MSE.  $T = 25, N = 1024, \pi_b = 0.60$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.5E-1	6.4E-2	5.2E-2	4.3E-2
Q-REG	6.2E-2	1.5E-2	2.1E-2	5.7E-2
MRDR	1.7E-1	7.2E-2	3.9E-2	3.1E-1
FQE	1.3E-1	2.8E-3	8.5E-4	6.1E-4
R( $\lambda$ )	1.3E0	2.4E-1	3.3E-2	1.1E-1
Q $^\pi$ ( $\lambda$ )	<b>3.0E-4</b>	3.0E-4	<b>9.8E-5</b>	1.2E-4
TREE	1.2E0	2.8E-1	4.0E-2	1.4E-1
IH	6.7E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.9E-1	3.0E-1
WIS	<b>8.5E-3</b>	4.2E-2
NAIVE	4.2E0	-

Table 439: Gridworld, relative MSE.  $T = 25, N = 512, \pi_b = 0.60$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.2E-1	1.1E-1	9.3E-2	7.2E-2
Q-REG	9.5E-1	9.2E-1	2.8E-1	3.6E-1
MRDR	6.0E-1	4.9E0	1.5E0	2.3E0
FQE	1.3E-1	1.3E-2	2.5E-3	2.2E-3
R( $\lambda$ )	1.2E0	1.2E0	1.2E-1	2.3E-1
Q $^\pi$ ( $\lambda$ )	<b>1.8E-3</b>	1.6E-3	<b>4.6E-4</b>	6.2E-4
TREE	1.1E0	1.5E0	1.5E-1	3.3E-1
IH	7.0E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.3E0	1.7E0
WIS	<b>3.9E-2</b>	1.6E-1
NAIVE	4.3E0	-

Table 441: Gridworld, relative MSE.  $T = 25, N = 64, \pi_b = 0.80$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.1E0	1.4E0	1.3E0	1.1E0
Q-REG	1.1E0	3.2E0	2.1E1	8.9E-1
MRDR	1.1E0	8.4E-1	1.1E0	1.6E0
FQE	3.6E-1	2.0E-1	<b>1.2E-1</b>	2.4E-1
R( $\lambda$ )	1.2E0	1.2E0	1.3E0	2.3E0
Q $^\pi$ ( $\lambda$ )	<b>2.1E-1</b>	2.2E-1	1.6E-1	2.2E-1
TREE	1.2E0	1.3E0	1.4E0	2.6E0
IH	2.1E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>9.4E-1</b>	1.4E0
WIS	1.8E0	1.6E0
NAIVE	8.3E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 442: Gridworld, relative MSE.  $T = 25, N = 128, \pi_b = 0.80$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.0E0	2.3E0	1.2E0	7.2E-1
Q-REG	4.3E0	3.9E1	2.5E1	4.0E0
MRDR	1.2E0	7.2E1	1.6E1	4.3E0
FQE	3.1E-1	9.5E-2	8.6E-2	<b>5.9E-2</b>
$R(\lambda)$	1.3E0	7.8E0	1.4E0	2.4E0
$Q^\pi(\lambda)$	8.0E-2	6.9E-2	7.3E-2	7.5E-2
TREE	1.2E0	1.1E1	1.6E0	2.7E0
IH	<b>7.9E-2</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	8.7E0	1.2E1
WIS	<b>1.2E0</b>	1.7E0
NAIVE	7.6E0	-

Table 444: Gridworld, relative MSE.  $T = 25, N = 512, \pi_b = 0.80$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	7.0E-1	3.5E0	5.8E-1	3.7E-1
Q-REG	9.8E0	1.3E1	2.7E0	3.9E1
MRDR	4.3E0	1.4E2	1.5E1	2.2E1
FQE	2.9E-1	5.0E-1	1.5E-2	1.5E-2
$R(\lambda)$	1.3E0	3.8E1	3.6E-1	1.4E0
$Q^\pi(\lambda)$	<b>1.9E-2</b>	1.7E-1	<b>8.8E-3</b>	1.6E-2
TREE	1.2E0	4.2E1	3.8E-1	1.5E0
IH	9.5E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	3.0E1	4.3E1
WIS	<b>1.4E-1</b>	4.0E-1
NAIVE	7.9E0	-

Table 443: Gridworld, relative MSE.  $T = 25, N = 256, \pi_b = 0.80$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	9.4E-1	1.1E0	1.9E0	8.7E-1
Q-REG	2.9E0	1.3E1	1.2E1	1.1E0
MRDR	8.0E0	3.9E1	2.9E1	5.3E0
FQE	2.7E-1	1.8E-1	3.8E-2	<b>2.1E-2</b>
$R(\lambda)$	1.3E0	3.8E0	5.4E-1	1.7E0
$Q^\pi(\lambda)$	<b>3.0E-2</b>	4.1E-2	2.7E-2	2.8E-2
TREE	1.2E0	4.6E0	5.7E-1	1.8E0
IH	1.4E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	3.8E0	4.8E0
WIS	<b>4.8E-1</b>	5.7E-1
NAIVE	8.1E0	-

Table 445: Gridworld, relative MSE.  $T = 25, N = 1024, \pi_b = 0.80$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.7E-1	4.1E-1	1.3E-1	1.9E-1
Q-REG	3.1E0	1.2E0	1.0E0	4.6E-1
MRDR	1.0E0	1.3E0	1.5E0	1.1E0
FQE	2.8E-1	4.7E-2	1.7E-2	9.6E-3
$R(\lambda)$	1.3E0	1.1E0	7.8E-1	1.5E0
$Q^\pi(\lambda)$	<b>8.1E-3</b>	8.7E-3	3.7E-3	<b>3.2E-3</b>
TREE	1.2E0	1.2E0	8.3E-1	1.5E0
IH	1.1E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	7.0E-1	1.2E0
WIS	<b>2.1E-1</b>	8.5E-1
NAIVE	7.9E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 446: Gridworld, relative MSE.  $T = 25, N = 64, \pi_b = 1.00$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.1E0	<b>1.1E0</b>	1.6E0	1.2E0
Q-REG	1.5E0	2.4E0	3.8E0	2.0E0
MRDR	1.2E0	2.3E0	3.7E0	4.0E0
FQE	1.2E0	1.2E0	1.2E0	1.2E0
R( $\lambda$ )	1.1E0	1.2E0	3.6E0	4.3E0
Q $^\pi$ ( $\lambda$ )	9.9E0	1.3E1	1.0E1	8.6E0
TREE	<b>1.1E0</b>	1.2E0	3.8E0	4.4E0
IH	1.3E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.0E0</b>	1.7E0
WIS	8.4E0	3.9E0
NAIVE	1.1E1	-

Table 448: Gridworld, relative MSE.  $T = 25, N = 256, \pi_b = 1.00$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.2E0	1.2E0	2.4E0	1.2E0
Q-REG	1.7E0	3.2E0	1.1E1	1.8E0
MRDR	1.7E0	2.8E0	1.2E0	2.5E0
FQE	4.4E-1	3.4E-1	<b>2.1E-1</b>	2.9E-1
R( $\lambda$ )	1.2E0	1.5E0	4.1E0	4.1E0
Q $^\pi$ ( $\lambda$ )	<b>3.7E-1</b>	2.9E0	3.5E-1	3.3E-1
TREE	1.1E0	1.6E0	4.4E0	4.2E0
IH	7.2E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.0E0</b>	2.0E0
WIS	4.6E0	4.6E0
NAIVE	1.1E1	-

Table 447: Gridworld, relative MSE.  $T = 25, N = 128, \pi_b = 1.00$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.2E0	1.1E0	1.2E0	1.2E0
Q-REG	1.3E0	2.0E0	9.4E0	1.8E0
MRDR	1.2E0	1.1E0	4.1E0	3.7E0
FQE	8.8E-1	8.6E-1	7.4E-1	8.3E-1
R( $\lambda$ )	1.2E0	1.3E0	5.1E0	4.8E0
Q $^\pi$ ( $\lambda$ )	8.9E-1	7.6E-1	<b>5.6E-1</b>	7.3E-1
TREE	1.1E0	1.3E0	5.4E0	5.1E0
IH	<b>6.9E-1</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.0E0</b>	1.4E0
WIS	5.6E0	5.6E0
NAIVE	1.1E1	-

Table 449: Gridworld, relative MSE.  $T = 25, N = 512, \pi_b = 1.00$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.2E0	4.3E0	1.4E0	1.3E0
Q-REG	3.8E0	5.7E2	3.4E1	3.0E0
MRDR	1.6E0	1.5E3	1.0E2	2.8E0
FQE	3.9E-1	8.3E0	9.1E-2	1.4E-1
R( $\lambda$ )	1.2E0	2.2E0	4.1E0	4.0E0
Q $^\pi$ ( $\lambda$ )	<b>9.6E-2</b>	4.3E-1	<b>7.1E-2</b>	8.0E-2
TREE	1.2E0	4.2E0	4.3E0	4.2E0
IH	2.8E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>9.7E-1</b>	5.5E0
WIS	2.3E0	4.4E0
NAIVE	1.1E1	-

Table 450: Gridworld, relative MSE.  $T = 25, N = 1024, \pi_b = 1.00$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.).

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	1.1E0	1.1E0	2.1E0	9.2E-1	
Q-REG	1.5E0	8.0E1	8.4E1	2.2E1	
MRDR	1.0E0	8.5E0	3.1E1	1.9E1	
FQE	3.5E-1	1.6E-1	5.1E-2	7.5E-2	
R( $\lambda$ )	1.2E0	6.2E0	1.6E0	2.7E0	
$Q^\pi(\lambda)$	<b>3.6E-2</b>	3.0E-2	<b>1.9E-2</b>	3.6E-2	
TREE	1.1E0	9.4E0	1.6E0	2.7E0	
IH	2.4E-1	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	8.9E0	1.0E1
WIS	1.7E0	<b>1.6E0</b>
NAIVE	1.1E1	-

## F.7 Detailed Results for Pixel Gridworld

 Table 451: Pixel-Gridworld, relative MSE.  $T = 25, N = 64, \pi_b = 0.20$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  known. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.3E0	2.3E1	2.3E1	7.3E-1
Q-REG	1.1E-1	4.3E-3	4.1E-3	4.5E-3
MRDR	1.5E-1	1.5E-2	9.7E-3	1.4E-2
FQE	1.8E-2	1.9E-3	1.8E-3	3.6E-3
$R(\lambda)$	<b>1.3E-3</b>	8.3E-4	8.1E-4	<b>6.9E-4</b>
$Q^\pi(\lambda)$	2.0E-3	2.0E-3	2.0E-3	2.1E-3
TREE	2.9E-3	1.6E-3	1.6E-3	2.0E-3
IH	2.1E-3	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	1.1E-2	1.4E-2
WIS	<b>1.3E-3</b>	3.9E-3
NAIVE	9.6E-2	-

 Table 452: Pixel-Gridworld, relative MSE.  $T = 25, N = 128, \pi_b = 0.20$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  known. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.4E2	8.0E1	7.7E1	6.0E1
Q-REG	4.7E-2	2.3E-3	2.1E-3	1.9E-3
MRDR	1.9E-1	6.4E-3	4.6E-3	4.1E-3
FQE	8.9E-3	9.6E-4	1.1E-3	2.0E-3
$R(\lambda)$	1.2E-3	6.5E-4	6.3E-4	<b>5.8E-4</b>
$Q^\pi(\lambda)$	3.0E-3	1.4E-3	1.4E-3	1.3E-3
TREE	3.1E-3	6.0E-4	6.2E-4	1.1E-3
IH	<b>1.1E-3</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	5.0E-3	6.3E-3
WIS	<b>1.6E-3</b>	2.2E-3
NAIVE	9.6E-2	-

 Table 453: Pixel-Gridworld, relative MSE.  $T = 25, N = 256, \pi_b = 0.20$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  known. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.0E3	1.2E1	9.5E1	1.4E2
Q-REG	8.7E-3	5.5E-4	6.7E-4	7.5E-4
MRDR	2.2E-1	3.0E-3	3.2E-3	2.4E-3
FQE	2.9E-3	2.7E-4	2.7E-4	6.3E-4
$R(\lambda)$	<b>6.4E-4</b>	<b>2.3E-4</b>	2.3E-4	2.4E-4
$Q^\pi(\lambda)$	3.5E-3	2.6E-4	2.6E-4	5.0E-4
TREE	7.5E-4	3.8E-4	3.9E-4	4.0E-4
IH	1.6E-3	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	6.2E-3	8.9E-3
WIS	<b>7.3E-4</b>	2.3E-3
NAIVE	1.1E-1	-

 Table 454: Pixel-Gridworld, relative MSE.  $T = 25, N = 512, \pi_b = 0.20$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  known. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.4E2	6.8E0	6.8E0	5.4E0
Q-REG	5.2E-3	3.2E-4	2.8E-4	2.9E-4
MRDR	1.2E-1	1.2E-3	6.1E-4	4.5E-4
FQE	9.7E-4	<b>1.7E-4</b>	1.7E-4	2.3E-4
$R(\lambda)$	2.0E-3	2.2E-4	2.0E-4	9.3E-4
$Q^\pi(\lambda)$	9.5E-4	2.0E-4	2.0E-4	5.9E-4
TREE	1.9E-3	1.8E-4	2.0E-4	3.7E-4
IH	<b>2.6E-4</b>	-	-	-

IPS		
	STANDARD	PER-DECISION
IS	2.3E-3	3.2E-3
WIS	<b>1.9E-4</b>	4.2E-4
NAIVE	9.0E-2	-



Table 455: Pixel-Gridworld, relative MSE.  $T = 25, N = 64, \pi_b = 0.20$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.3E0	7.7E2	1.4E0	2.7E-1
Q-REG	3.5E0	7.3E2	1.4E0	1.7E-1
MRDR	4.0E0	7.9E2	1.5E0	8.3E-1
FQE	1.6E-2	7.7E0	2.3E-2	4.7E-3
R( $\lambda$ )	3.2E-3	8.4E0	4.8E-3	<b>1.9E-3</b>
Q $^\pi$ ( $\lambda$ )	2.5E-3	1.6E1	3.0E-3	3.9E-3
TREE	<b>1.3E-3</b>	9.8E0	7.8E-3	2.0E-3
IH	3.2E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.3E2	3.9E2
WIS	9.2E-2	<b>2.0E-2</b>
NAIVE	7.4E-2	-

Table 457: Pixel-Gridworld, relative MSE.  $T = 25, N = 256, \pi_b = 0.20$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.8E1	8.4E0	4.9E-1	7.1E-1
Q-REG	1.6E-1	2.3E-2	1.6E-2	9.0E-3
MRDR	7.6E-1	1.0E-1	2.1E-2	3.4E-2
FQE	1.9E-3	<b>2.5E-4</b>	2.9E-4	5.0E-4
R( $\lambda$ )	2.2E-3	1.6E-3	1.4E-3	2.4E-3
Q $^\pi$ ( $\lambda$ )	2.3E-3	4.8E-3	2.3E-3	2.1E-3
TREE	<b>4.8E-4</b>	2.3E-3	1.3E-3	1.2E-3
IH	4.2E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	6.0E-2	1.5E-1
WIS	<b>9.2E-3</b>	1.2E-2
NAIVE	9.1E-2	-

Table 456: Pixel-Gridworld, relative MSE.  $T = 25, N = 128, \pi_b = 0.20$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.4E0	2.5E1	1.0E0	6.0E-1
Q-REG	6.9E-1	1.6E1	8.4E-2	5.7E-2
MRDR	1.8E0	2.0E1	1.1E-1	8.7E-2
FQE	1.9E-2	3.3E-1	1.6E-2	4.0E-3
R( $\lambda$ )	1.2E-3	2.7E-2	3.1E-3	<b>1.0E-3</b>
Q $^\pi$ ( $\lambda$ )	2.3E-3	1.8E-1	6.5E-3	2.5E-3
TREE	<b>8.0E-4</b>	5.1E-2	4.5E-3	1.1E-3
IH	2.0E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.7E1	2.6E1
WIS	6.7E-2	<b>3.1E-2</b>
NAIVE	9.2E-2	-

Table 458: Pixel-Gridworld, relative MSE.  $T = 25, N = 64, \pi_b = 0.20$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.5E4	1.3E16	2.8E4	1.3E4
Q-REG	1.6E2	3.8E12	1.4E3	8.8E1
MRDR	1.2E1	1.2E14	7.2E2	3.7E1
FQE	1.8E-1	6.1E12	1.6E0	5.0E-2
R( $\lambda$ )	<b>2.4E-2</b>	7.2E12	6.8E-1	<b>2.7E-2</b>
Q $^\pi$ ( $\lambda$ )	3.7E-2	8.4E12	4.4E-1	3.0E-2
TREE	3.8E-2	7.9E12	5.3E-1	4.2E-2
IH	4.4E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	5.6E12	1.0E12
WIS	5.9E0	<b>5.8E0</b>
NAIVE	4.6E-1	-

Table 459: Pixel-Gridworld, relative MSE.  $T = 25, N = 128, \pi_b = 0.20$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.2E2	7.4E17	3.7E3	2.1E2
Q-REG	1.4E2	1.1E14	9.9E2	1.3E2
MRDR	5.2E0	6.9E7	3.3E1	3.4E0
FQE	1.4E-1	2.4E13	2.5E0	1.7E-1
R( $\lambda$ )	<b>4.6E-2</b>	6.7E12	8.0E-1	1.2E-1
Q $^\pi$ ( $\lambda$ )	5.4E-2	3.6E12	5.8E-1	9.0E-2
TREE	4.7E-2	6.2E12	1.2E0	<b>8.6E-2</b>
IH	9.1E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.2E12	1.3E14
WIS	<b>2.5E0</b>	3.9E0
NAIVE	4.4E-1	-

Table 461: Pixel-Gridworld, relative MSE.  $T = 25, N = 64, \pi_b = 0.40$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  known. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.5E3	3.5E3	3.4E3	2.9E3
Q-REG	1.6E-1	6.3E-2	1.7E-2	9.5E-2
MRDR	4.4E-1	1.3E-1	6.1E-2	3.0E-1
FQE	1.5E-1	6.2E-3	2.7E-3	1.3E-2
R( $\lambda$ )	5.9E-3	3.1E-3	3.2E-3	4.7E-3
Q $^\pi$ ( $\lambda$ )	8.3E-3	<b>1.2E-3</b>	1.4E-3	2.2E-3
TREE	<b>4.2E-3</b>	4.9E-3	5.2E-3	7.1E-3
IH	5.3E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.2E-1	1.7E-1
WIS	<b>8.2E-3</b>	2.8E-2
NAIVE	1.2E0	-

Table 460: Pixel-Gridworld, relative MSE.  $T = 25, N = 256, \pi_b = 0.20$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.7E1	8.7E4	7.9E1	1.2E1
Q-REG	1.5E1	4.6E6	5.8E0	5.0E0
MRDR	4.4E1	1.0E6	1.4E1	1.0E0
FQE	8.2E-2	9.0E6	9.3E-1	7.6E-2
R( $\lambda$ )	3.4E-2	9.6E6	8.0E-1	1.1E-1
Q $^\pi$ ( $\lambda$ )	1.7E-2	7.7E6	1.1E0	<b>5.8E-2</b>
TREE	<b>1.7E-2</b>	1.4E7	5.8E-1	6.4E-2
IH	7.9E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.0E8	1.2E6
WIS	3.9E0	<b>2.3E0</b>
NAIVE	3.6E-1	-

Table 462: Pixel-Gridworld, relative MSE.  $T = 25, N = 128, \pi_b = 0.40$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  known. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.3E0	1.7E0	6.1E-1	1.7E0
Q-REG	1.6E-1	1.1E-2	9.1E-3	5.6E-2
MRDR	5.8E-1	1.6E-1	5.8E-2	1.6E-1
FQE	7.3E-2	5.5E-3	1.4E-3	1.5E-3
R( $\lambda$ )	<b>2.9E-3</b>	6.6E-4	6.3E-4	9.7E-4
Q $^\pi$ ( $\lambda$ )	9.2E-3	1.2E-3	5.3E-4	1.1E-3
TREE	4.1E-3	1.2E-3	<b>4.7E-4</b>	1.1E-3
IH	5.0E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.7E-1	3.2E-1
WIS	<b>1.2E-2</b>	1.9E-2
NAIVE	1.2E0	-

Table 463: Pixel-Gridworld, relative MSE.  $T = 25, N = 256, \pi_b = 0.40$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  known. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.8E0	1.2E0	1.1E0	5.1E-1
Q-REG	2.1E-1	6.0E-3	5.9E-3	1.1E-2
MRDR	6.0E-1	6.5E-3	4.9E-3	1.4E-2
FQE	1.6E-1	1.1E-3	6.7E-4	1.1E-3
R( $\lambda$ )	3.2E-3	4.1E-4	3.8E-4	1.3E-3
Q $^\pi$ ( $\lambda$ )	8.8E-3	5.6E-4	2.4E-4	1.4E-3
TREE	<b>2.7E-3</b>	3.8E-4	<b>1.8E-4</b>	5.3E-4
IH	5.5E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	9.8E-2	1.2E-1
WIS	<b>2.9E-3</b>	1.1E-2
NAIVE	1.3E0	-

Table 465: Pixel-Gridworld, relative MSE.  $T = 25, N = 64, \pi_b = 0.40$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.9E1	4.6E8	5.0E1	1.5E1
Q-REG	2.1E0	1.1E14	5.1E-1	5.0E-1
MRDR	2.3E0	4.2E19	4.3E0	1.5E0
FQE	1.5E-1	5.6E15	7.8E-2	3.8E-2
R( $\lambda$ )	<b>2.0E-3</b>	1.2E16	1.8E-3	<b>1.1E-3</b>
Q $^\pi$ ( $\lambda$ )	6.6E-3	6.2E15	3.4E-2	1.3E-2
TREE	4.5E-3	2.7E15	3.6E-2	8.7E-3
IH	7.0E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	3.0E19	1.9E8
WIS	3.5E-1	<b>2.1E-1</b>
NAIVE	1.1E0	-

Table 464: Pixel-Gridworld, relative MSE.  $T = 25, N = 512, \pi_b = 0.40$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  known. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.4E0	1.9E-1	2.0E-1	3.0E-1
Q-REG	8.1E-2	7.2E-4	6.0E-4	1.0E-3
MRDR	2.5E-1	4.6E-3	3.9E-3	3.5E-3
FQE	3.3E-3	7.8E-5	<b>6.9E-5</b>	2.5E-4
R( $\lambda$ )	2.8E-3	7.4E-5	9.3E-5	3.1E-4
Q $^\pi$ ( $\lambda$ )	1.8E-3	3.4E-4	2.2E-4	7.4E-4
TREE	<b>1.6E-3</b>	1.1E-4	8.2E-5	8.3E-5
IH	3.8E-2	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.9E-2	2.6E-2
WIS	<b>1.6E-3</b>	6.4E-3
NAIVE	1.2E0	-

Table 466: Pixel-Gridworld, relative MSE.  $T = 25, N = 128, \pi_b = 0.40$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.8E1	1.1E3	1.6E1	9.8E0
Q-REG	1.7E0	1.9E2	8.2E-1	9.4E-2
MRDR	1.9E0	4.2E2	4.6E0	8.6E-1
FQE	5.8E-2	2.2E1	2.0E-2	6.6E-3
R( $\lambda$ )	<b>1.1E-3</b>	2.5E0	1.9E-2	1.8E-3
Q $^\pi$ ( $\lambda$ )	3.2E-2	9.8E0	3.3E-3	2.4E-3
TREE	1.6E-3	6.7E0	3.7E-3	<b>1.7E-3</b>
IH	1.1E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.2E3	1.9E2
WIS	7.3E-1	<b>1.7E-1</b>
NAIVE	1.3E0	-

Table 467: Pixel-Gridworld, relative MSE.  $T = 25, N = 256, \pi_b = 0.40$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.5E1	1.8E2	5.5E0	5.4E1
Q-REG	8.3E-1	2.2E1	3.1E-1	9.7E-2
MRDR	8.5E-1	1.8E1	1.7E-1	1.7E-1
FQE	3.1E-2	8.9E-2	3.1E-3	2.4E-3
R( $\lambda$ )	<b>1.7E-3</b>	4.1E-2	2.9E-3	3.0E-3
Q $^\pi$ ( $\lambda$ )	2.1E-3	5.9E-2	3.0E-3	<b>1.6E-3</b>
TREE	5.9E-3	6.5E-2	6.0E-3	5.2E-3
IH	1.0E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	9.1E0	1.1E1
WIS	<b>4.1E-2</b>	5.1E-2
NAIVE	1.2E0	-

Table 469: Pixel-Gridworld, relative MSE.  $T = 25, N = 128, \pi_b = 0.40$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.3E3	1.4E6	1.4E3	3.7E3
Q-REG	4.1E1	2.6E3	2.1E0	6.4E0
MRDR	2.1E0	1.2E9	2.3E0	1.1E0
FQE	3.9E-1	2.1E5	4.1E-2	7.1E-2
R( $\lambda$ )	<b>5.2E-2</b>	6.0E7	8.5E-2	3.6E-2
Q $^\pi$ ( $\lambda$ )	1.1E-1	1.8E7	9.5E-2	8.3E-2
TREE	6.2E-2	1.9E7	6.9E-2	<b>3.2E-2</b>
IH	1.3E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.9E10	3.1E2
WIS	2.0E0	<b>1.2E0</b>
NAIVE	4.0E0	-

Table 468: Pixel-Gridworld, relative MSE.  $T = 25, N = 64, \pi_b = 0.40$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.2E2	3.3E10	3.3E3	6.6E2
Q-REG	4.9E0	1.2E6	1.4E1	2.7E0
MRDR	1.3E0	6.1E8	8.2E0	1.5E0
FQE	5.2E-1	1.1E6	8.4E-1	2.8E-1
R( $\lambda$ )	<b>7.0E-2</b>	1.3E9	5.6E-1	1.1E-1
Q $^\pi$ ( $\lambda$ )	1.6E-1	4.6E8	3.5E-1	9.2E-2
TREE	7.1E-2	5.3E8	1.3E0	<b>8.5E-2</b>
IH	1.4E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.2E11	1.9E5
WIS	<b>2.3E0</b>	6.9E0
NAIVE	4.3E0	-

Table 470: Pixel-Gridworld, relative MSE.  $T = 25, N = 256, \pi_b = 0.40$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.3E3	4.9E2	4.9E2	6.2E2
Q-REG	2.5E0	1.8E-1	<b>1.2E-2</b>	2.1E-1
MRDR	2.6E0	1.9E0	1.7E-1	5.1E-1
FQE	2.3E-1	7.0E-2	5.0E-2	2.6E-2
R( $\lambda$ )	<b>3.5E-2</b>	9.0E-2	5.9E-2	1.9E-2
Q $^\pi$ ( $\lambda$ )	6.5E-2	5.3E-2	6.6E-2	7.1E-2
TREE	6.2E-2	7.8E-2	4.2E-2	2.7E-2
IH	1.0E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.3E0	1.1E0
WIS	3.1E-1	<b>1.9E-1</b>
NAIVE	3.9E0	-

Table 471: Pixel-Gridworld, relative MSE.  $T = 25, N = 64, \pi_b = 0.60$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  known. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	6.0E1	2.6E2	5.1E1	5.8E1
Q-REG	1.9E0	3.2E0	6.4E-1	1.6E0
MRDR	7.7E-1	2.5E0	1.2E0	1.0E0
FQE	2.5E-1	4.1E-1	3.4E-2	6.9E-2
R( $\lambda$ )	<b>7.2E-3</b>	3.6E-2	2.7E-2	8.7E-3
Q $^\pi$ ( $\lambda$ )	3.0E-2	4.0E-2	1.6E-2	9.5E-3
TREE	9.0E-3	3.4E-1	2.7E-2	<b>6.2E-3</b>
IH	4.7E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.8E0	3.9E0
WIS	<b>2.0E-1</b>	5.6E-1
NAIVE	3.9E0	-

Table 473: Pixel-Gridworld, relative MSE.  $T = 25, N = 256, \pi_b = 0.60$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  known. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.4E1	1.5E1	1.9E1	5.2E1
Q-REG	3.1E-1	1.5E-1	3.8E-2	1.8E-1
MRDR	1.1E0	6.3E-1	3.7E-1	3.8E-1
FQE	7.4E-2	5.9E-3	6.5E-3	6.3E-3
R( $\lambda$ )	8.1E-3	5.5E-3	4.7E-3	2.3E-3
Q $^\pi$ ( $\lambda$ )	4.1E-2	1.1E-2	<b>2.1E-3</b>	1.9E-2
TREE	<b>3.2E-3</b>	3.3E-3	2.4E-3	2.9E-3
IH	5.2E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	3.1E-1	5.1E-1
WIS	<b>6.3E-2</b>	1.4E-1
NAIVE	4.3E0	-

Table 472: Pixel-Gridworld, relative MSE.  $T = 25, N = 128, \pi_b = 0.60$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  known. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.9E1	5.7E1	2.8E1	5.9E0
Q-REG	3.5E0	3.1E1	6.2E-1	8.6E-1
MRDR	7.3E-1	4.1E0	2.7E-1	1.1E0
FQE	9.3E-2	3.1E-2	7.4E-3	6.6E-3
R( $\lambda$ )	<b>3.5E-3</b>	5.3E-3	2.9E-3	<b>1.8E-3</b>
Q $^\pi$ ( $\lambda$ )	2.7E-2	1.1E-2	4.1E-3	1.2E-2
TREE	7.5E-3	5.1E-2	2.9E-3	6.5E-3
IH	4.9E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.3E0	5.1E0
WIS	2.3E-1	<b>1.7E-1</b>
NAIVE	4.0E0	-

Table 474: Pixel-Gridworld, relative MSE.  $T = 25, N = 512, \pi_b = 0.60$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  known. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.5E0	8.7E0	7.0E0	2.9E0
Q-REG	1.5E0	1.5E0	1.5E-1	6.4E-1
MRDR	2.7E0	3.8E0	1.0E0	1.5E0
FQE	3.4E-2	3.2E-3	1.3E-3	<b>1.2E-3</b>
R( $\lambda$ )	1.1E-2	1.6E-2	4.9E-3	6.3E-3
Q $^\pi$ ( $\lambda$ )	4.6E-2	2.0E-2	2.0E-2	7.1E-3
TREE	<b>8.3E-3</b>	9.8E-3	5.0E-3	2.2E-3
IH	5.2E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.2E0	1.7E0
WIS	<b>4.6E-2</b>	1.5E-1
NAIVE	4.3E0	-

Table 475: Pixel-Gridworld, relative MSE.  $T = 25, N = 64, \pi_b = 0.60$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	9.8E0	1.6E20	9.4E1	1.1E1
Q-REG	1.0E1	7.3E10	1.3E1	2.3E1
MRDR	2.2E0	2.9E20	2.8E0	3.5E0
FQE	3.9E-1	1.5E17	6.8E-2	1.7E-1
R( $\lambda$ )	2.7E-2	2.5E18	<b>1.2E-2</b>	2.2E-2
Q $^\pi$ ( $\lambda$ )	5.1E-2	6.0E18	3.5E-2	2.2E-2
TREE	<b>1.9E-2</b>	2.5E18	2.5E-2	2.3E-2
IH	4.8E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.5E21	2.4E5
WIS	<b>2.1E-1</b>	9.3E-1
NAIVE	4.5E0	-

Table 477: Pixel-Gridworld, relative MSE.  $T = 25, N = 256, \pi_b = 0.60$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.6E1	1.7E2	5.0E1	8.0E1
Q-REG	5.7E-1	6.4E-1	1.2E-1	4.9E-1
MRDR	4.3E-1	1.6E1	5.3E0	3.5E-1
FQE	4.1E-2	9.9E-2	3.3E-3	<b>1.8E-3</b>
R( $\lambda$ )	2.3E-2	9.6E-2	7.3E-3	3.6E-3
Q $^\pi$ ( $\lambda$ )	3.5E-2	1.0E-1	1.8E-2	1.1E-2
TREE	<b>7.9E-3</b>	4.2E-2	2.6E-3	3.6E-3
IH	5.7E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	7.5E-1	2.3E0
WIS	1.0E-1	<b>1.0E-1</b>
NAIVE	4.4E0	-

Table 476: Pixel-Gridworld, relative MSE.  $T = 25, N = 128, \pi_b = 0.60$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.4E1	5.5E14	8.8E1	1.2E1
Q-REG	6.2E-1	3.0E15	1.3E-1	6.5E-1
MRDR	9.1E-1	1.6E18	3.4E-1	5.6E-1
FQE	1.4E-1	3.4E15	2.5E-2	3.9E-2
R( $\lambda$ )	<b>9.0E-3</b>	3.0E15	1.1E-2	<b>3.3E-3</b>
Q $^\pi$ ( $\lambda$ )	2.2E-2	3.3E16	1.3E-2	2.2E-2
TREE	1.2E-2	9.1E16	1.9E-2	7.9E-3
IH	5.9E-1	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	6.1E18	4.4E5
WIS	<b>1.7E-1</b>	5.0E-1
NAIVE	4.5E0	-

Table 478: Pixel-Gridworld, relative MSE.  $T = 25, N = 64, \pi_b = 0.60$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.0E2	4.8E16	3.0E2	2.2E2
Q-REG	9.9E0	8.3E11	2.0E1	1.6E1
MRDR	1.5E0	6.5E18	4.0E1	2.7E1
FQE	9.0E-1	2.3E13	5.2E-1	5.9E-1
R( $\lambda$ )	2.9E-1	1.4E16	3.1E-1	2.2E-1
Q $^\pi$ ( $\lambda$ )	5.0E-1	1.3E16	2.8E-1	2.8E-1
TREE	<b>1.8E-1</b>	1.3E15	3.5E-1	<b>1.6E-1</b>
IH	2.9E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.6E18	2.2E5
WIS	2.2E0	<b>2.2E0</b>
NAIVE	1.2E1	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 479: Pixel-Gridworld, relative MSE.  $T = 25, N = 128, \pi_b = 0.60$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	7.4E1	2.9E12	1.1E4	9.7E1
Q-REG	1.6E0	7.0E7	1.6E0	2.4E0
MRDR	3.8E0	6.6E14	4.1E0	4.3E0
FQE	1.0E0	8.8E12	3.4E-1	4.5E-1
R( $\lambda$ )	4.0E-1	2.5E14	2.3E-1	<b>1.7E-1</b>
Q $^\pi$ ( $\lambda$ )	5.9E-1	8.7E13	3.2E-1	3.0E-1
TREE	<b>3.9E-1</b>	9.4E13	3.1E-1	2.5E-1
IH	3.1E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	5.6E16	6.2E1
WIS	2.4E0	<b>2.3E0</b>
NAIVE	1.2E1	-

Table 481: Pixel-Gridworld, relative MSE.  $T = 25, N = 64, \pi_b = 0.80$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  known. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.6E0	8.2E1	2.3E1	4.1E0
Q-REG	1.2E0	1.8E0	1.1E0	2.5E0
MRDR	1.2E0	8.8E-1	9.6E-1	2.1E0
FQE	2.8E0	1.5E0	4.1E-1	1.7E0
R( $\lambda$ )	<b>1.4E-1</b>	1.7E-1	9.5E-2	<b>7.1E-2</b>
Q $^\pi$ ( $\lambda$ )	6.2E0	4.3E0	7.5E-1	3.7E0
TREE	4.2E-1	3.7E-1	2.6E-1	3.7E-1
IH	1.6E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.0E0</b>	1.5E0
WIS	1.9E0	1.5E0
NAIVE	8.3E0	-

Table 480: Pixel-Gridworld, relative MSE.  $T = 25, N = 256, \pi_b = 0.60$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	9.2E3	3.3E5	6.5E4	9.4E3
Q-REG	1.4E0	3.6E2	1.4E0	1.7E0
MRDR	9.9E-1	2.3E11	1.4E0	1.5E0
FQE	3.9E-1	4.7E8	7.3E-2	7.2E-2
R( $\lambda$ )	<b>1.2E-1</b>	7.3E10	1.1E-1	<b>6.9E-2</b>
Q $^\pi$ ( $\lambda$ )	3.8E-1	3.0E10	3.7E-1	2.5E-1
TREE	1.7E-1	2.9E10	1.2E-1	9.9E-2
IH	3.2E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.2E13	3.1E0
WIS	<b>1.6E0</b>	1.8E0
NAIVE	1.2E1	-

Table 482: Pixel-Gridworld, relative MSE.  $T = 25, N = 128, \pi_b = 0.80$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  known. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	9.8E1	1.5E3	4.9E2	9.6E1
Q-REG	1.2E0	3.3E0	1.5E0	2.3E0
MRDR	8.4E-1	9.7E0	3.0E1	1.3E0
FQE	4.7E-1	2.0E-1	9.9E-2	1.7E-1
R( $\lambda$ )	7.3E-2	1.7E-1	7.9E-2	<b>3.9E-2</b>
Q $^\pi$ ( $\lambda$ )	1.8E-1	3.9E-1	6.2E-2	3.9E-2
TREE	<b>2.2E-2</b>	7.2E-1	6.4E-2	4.9E-2
IH	1.4E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	3.0E0	4.4E0
WIS	<b>1.2E0</b>	1.5E0
NAIVE	7.6E0	-

**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 483: Pixel-Gridworld, relative MSE.  $T = 25, N = 256, \pi_b = 0.80$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  known. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	7.4E0	1.2E2	1.4E2	5.2E0
Q-REG	1.4E0	3.1E0	3.0E0	2.0E0
MRDR	9.4E-1	3.1E0	3.2E0	1.1E0
FQE	3.6E-1	6.2E-2	4.6E-2	4.5E-2
R( $\lambda$ )	9.0E-2	1.9E-1	3.5E-2	<b>2.8E-2</b>
Q $^\pi$ ( $\lambda$ )	<b>8.3E-2</b>	1.8E-1	7.4E-2	5.4E-2
TREE	2.5E-1	1.1E-1	7.8E-2	1.3E-1
IH	1.5E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.4E0	1.9E0
WIS	<b>4.7E-1</b>	6.6E-1
NAIVE	8.1E0	-

Table 485: Pixel-Gridworld, relative MSE.  $T = 25, N = 64, \pi_b = 0.80$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.8E0	1.4E15	7.4E0	5.3E0
Q-REG	1.2E0	4.7E11	2.7E0	3.9E0
MRDR	1.3E0	2.5E16	1.6E0	3.6E0
FQE	1.3E0	1.1E12	2.2E-1	7.5E-1
R( $\lambda$ )	<b>6.5E-2</b>	4.7E14	8.6E-2	<b>4.9E-2</b>
Q $^\pi$ ( $\lambda$ )	6.2E-1	1.1E16	4.7E-1	4.3E-1
TREE	2.7E-1	1.4E15	1.4E-1	1.8E-1
IH	1.3E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.6E17	7.6E1
WIS	<b>1.1E0</b>	3.0E0
NAIVE	8.3E0	-

Table 484: Pixel-Gridworld, relative MSE.  $T = 25, N = 512, \pi_b = 0.80$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  known. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.5E1	1.1E5	4.2E2	7.3E1
Q-REG	7.1E0	4.2E2	4.9E-1	1.0E1
MRDR	2.2E0	5.1E1	1.6E0	9.1E0
FQE	9.6E-2	1.3E-1	<b>7.7E-3</b>	1.9E-2
R( $\lambda$ )	<b>1.3E-2</b>	1.7E0	2.0E-2	2.3E-2
Q $^\pi$ ( $\lambda$ )	1.6E-1	7.0E0	3.9E-2	6.3E-2
TREE	3.3E-2	2.6E0	4.0E-2	2.6E-2
IH	1.5E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.7E1	3.7E1
WIS	<b>1.4E-1</b>	4.1E-1
NAIVE	7.9E0	-

Table 486: Pixel-Gridworld, relative MSE.  $T = 25, N = 128, \pi_b = 0.80$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	1.0E1	2.2E17	3.9E1	8.0E0
Q-REG	1.6E0	4.1E14	3.6E0	2.9E0
MRDR	1.2E0	4.1E20	1.2E0	1.7E0
FQE	2.6E-1	2.4E18	4.5E-2	1.1E-1
R( $\lambda$ )	<b>6.0E-2</b>	2.0E19	4.1E-2	4.0E-2
Q $^\pi$ ( $\lambda$ )	5.0E-1	4.0E19	1.7E-1	1.3E-1
TREE	8.5E-2	1.6E19	<b>2.4E-2</b>	4.6E-2
IH	1.4E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.1E20	3.3E2
WIS	<b>4.4E-1</b>	2.1E0
NAIVE	8.2E0	-



**Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning**

---

Table 487: Pixel-Gridworld, relative MSE.  $T = 25, N = 256, \pi_b = 0.80$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	7.2E1	1.1E15	4.1E2	7.2E1
Q-REG	5.1E0	5.6E11	3.1E0	1.1E1
MRDR	4.7E0	4.5E18	1.1E0	5.1E0
FQE	7.8E-2	2.5E13	<b>1.5E-2</b>	2.7E-2
R( $\lambda$ )	<b>5.7E-2</b>	2.6E17	3.3E-2	2.1E-2
Q $^\pi$ ( $\lambda$ )	2.2E-1	2.3E17	1.7E-1	4.4E-2
TREE	2.5E-1	2.4E17	1.5E-2	6.1E-2
IH	1.4E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	6.1E18	2.8E1
WIS	<b>5.0E-1</b>	1.5E0
NAIVE	8.2E0	-

Table 489: Pixel-Gridworld, relative MSE.  $T = 25, N = 128, \pi_b = 0.80$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	6.0E0	8.8E15	5.2E1	5.3E0
Q-REG	1.9E0	3.6E8	2.9E0	4.4E0
MRDR	8.1E0	3.9E18	1.3E0	1.5E1
FQE	1.8E0	9.4E16	1.2E0	8.6E-1
R( $\lambda$ )	5.2E-1	3.4E16	6.3E-1	<b>4.8E-1</b>
Q $^\pi$ ( $\lambda$ )	<b>4.5E-1</b>	3.7E17	9.5E-1	6.6E-1
TREE	1.1E0	6.6E17	8.0E-1	5.4E-1
IH	5.0E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.6E19	7.4E1
WIS	5.3E0	<b>3.1E0</b>
NAIVE	2.3E1	-

Table 488: Pixel-Gridworld, relative MSE.  $T = 25, N = 64, \pi_b = 0.80$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	7.7E0	2.7E17	2.5E1	3.3E1
Q-REG	5.4E0	5.8E14	2.0E1	2.1E1
MRDR	<b>2.1E0</b>	8.4E21	1.3E1	6.4E0
FQE	1.0E1	1.9E19	4.0E0	5.8E0
R( $\lambda$ )	1.7E1	1.6E20	7.1E0	1.5E1
Q $^\pi$ ( $\lambda$ )	2.0E1	1.3E21	<b>3.4E0</b>	8.0E0
TREE	1.7E1	1.4E20	4.4E0	1.1E1
IH	5.9E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.1E22	9.6E2
WIS	7.2E0	<b>5.2E0</b>
NAIVE	2.4E1	-

Table 490: Pixel-Gridworld, relative MSE.  $T = 25, N = 256, \pi_b = 0.80$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	4.8E1	1.3E13	8.8E2	5.2E1
Q-REG	2.3E0	1.9E9	4.1E0	5.1E0
MRDR	1.4E0	5.5E14	3.8E0	3.7E0
FQE	<b>1.1E0</b>	2.9E13	3.1E-1	<b>1.5E-1</b>
R( $\lambda$ )	1.2E0	7.5E13	3.6E-1	5.6E-1
Q $^\pi$ ( $\lambda$ )	7.8E0	3.0E14	1.3E0	1.8E0
TREE	1.3E0	2.1E14	5.8E-1	8.0E-1
IH	5.7E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.9E16	1.1E1
WIS	8.9E0	<b>3.6E0</b>
NAIVE	2.4E1	-

Table 491: Pixel-Gridworld, relative MSE.  $T = 25, N = 64, \pi_b = 1.00$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  known. Stochastic environment.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	2.1E0	3.7E1	1.4E1	5.7E0	
Q-REG	1.1E0	1.8E0	3.2E0	4.0E0	
MRDR	<b>1.1E0</b>	<b>8.8E-1</b>	3.6E0	4.5E0	
FQE	2.3E1	2.1E1	6.1E0	1.4E1	
R( $\lambda$ )	9.1E2	2.0E3	7.6E4	9.1E2	
Q $^\pi$ ( $\lambda$ )	5.7E1	5.7E1	7.3E1	4.8E1	
TREE	2.0E1	2.1E2	4.2E2	6.9E1	
IH	2.0E0	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.0E0</b>	1.7E0
WIS	1.0E1	3.3E0
NAIVE	1.1E1	-

Table 493: Pixel-Gridworld, relative MSE.  $T = 25, N = 256, \pi_b = 1.00$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  known. Stochastic environment.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	3.1E0	6.2E1	3.2E1	3.5E0	
Q-REG	1.2E0	2.0E0	3.8E0	4.3E0	
MRDR	1.1E0	1.2E0	2.5E0	2.9E0	
FQE	<b>3.9E-1</b>	3.7E-1	3.3E-1	<b>2.9E-1</b>	
R( $\lambda$ )	7.7E0	1.1E1	5.9E0	5.8E0	
Q $^\pi$ ( $\lambda$ )	3.9E1	3.8E1	9.5E0	1.5E1	
TREE	5.8E-1	4.0E0	3.0E-1	3.1E-1	
IH	2.1E0	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.0E0</b>	1.7E0
WIS	4.1E0	4.2E0
NAIVE	1.1E1	-

Table 492: Pixel-Gridworld, relative MSE.  $T = 25, N = 128, \pi_b = 1.00$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  known. Stochastic environment.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	3.2E0	2.1E1	3.8E1	7.6E0	
Q-REG	1.3E0	1.6E0	5.9E0	6.1E0	
MRDR	<b>1.1E0</b>	<b>9.1E-1</b>	8.0E0	8.1E0	
FQE	1.1E1	1.1E1	2.8E0	6.5E0	
R( $\lambda$ )	5.1E0	4.8E0	2.7E0	2.8E0	
Q $^\pi$ ( $\lambda$ )	3.1E1	2.7E1	9.8E0	1.2E1	
TREE	3.2E0	3.2E0	1.4E0	2.0E0	
IH	2.1E0	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.0E0</b>	1.4E0
WIS	5.6E0	5.7E0
NAIVE	1.1E1	-

Table 494: Pixel-Gridworld, relative MSE.  $T = 25, N = 512, \pi_b = 1.00$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  known. Stochastic environment.

	DM		HYBRID		
	DIRECT	DR	WDR	MAGIC	
AM	4.9E0	1.1E3	3.7E2	6.3E0	
Q-REG	1.9E0	2.4E0	3.5E0	4.4E0	
MRDR	9.7E-1	2.5E1	4.7E0	2.7E0	
FQE	5.7E-1	7.3E0	<b>6.6E-2</b>	1.3E-1	
R( $\lambda$ )	<b>5.4E-1</b>	5.6E0	1.1E-1	2.2E-1	
Q $^\pi$ ( $\lambda$ )	3.6E1	3.2E1	6.4E0	1.2E1	
TREE	1.3E0	3.5E1	3.5E-1	6.5E-1	
IH	2.1E0	-	-	-	

	IPS	
	STANDARD	PER-DECISION
IS	<b>9.8E-1</b>	8.7E0
WIS	2.4E0	4.2E0
NAIVE	1.1E1	-

Table 495: Pixel-Gridworld, relative MSE.  $T = 25, N = 64, \pi_b = 1.00$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.1E0	7.8E0	9.7E0	4.8E0
Q-REG	1.2E0	1.6E0	6.5E0	4.8E0
MRDR	<b>1.1E0</b>	<b>9.6E-1</b>	8.1E0	7.5E0
FQE	2.8E1	2.3E1	1.2E1	1.3E1
$R(\lambda)$	1.5E1	1.6E1	8.7E0	9.0E0
$Q^\pi(\lambda)$	5.2E1	3.4E1	1.4E1	1.9E1
TREE	4.8E1	4.9E1	1.1E1	1.2E1
IH	2.1E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.0E0</b>	1.4E0
WIS	1.0E1	6.6E0
NAIVE	1.0E1	-

Table 497: Pixel-Gridworld, relative MSE.  $T = 25, N = 256, \pi_b = 1.00$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	2.6E0	6.8E15	3.5E1	2.6E0
Q-REG	1.2E0	2.0E8	3.8E0	4.5E0
MRDR	<b>1.2E0</b>	6.1E16	4.5E0	5.5E0
FQE	2.5E0	1.4E13	5.8E-1	1.1E0
$R(\lambda)$	1.4E0	3.0E16	<b>4.2E-1</b>	8.0E-1
$Q^\pi(\lambda)$	2.8E1	6.0E16	9.4E0	9.8E0
TREE	2.1E0	1.9E16	5.1E-1	7.7E-1
IH	2.0E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	8.8E15	<b>1.4E0</b>
WIS	2.9E0	3.9E0
NAIVE	1.0E1	-

Table 496: Pixel-Gridworld, relative MSE.  $T = 25, N = 128, \pi_b = 1.00$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	5.3E0	2.5E19	4.1E1	6.5E0
Q-REG	1.1E0	1.4E16	4.5E0	<b>3.9E0</b>
MRDR	<b>8.8E-1</b>	5.1E21	5.2E0	5.1E0
FQE	2.4E1	2.2E19	8.0E0	9.4E0
$R(\lambda)$	3.1E1	1.6E20	4.6E0	4.6E0
$Q^\pi(\lambda)$	2.7E1	1.7E20	8.7E0	1.2E1
TREE	1.4E2	2.1E19	7.1E0	7.6E0
IH	2.0E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	4.1E21	1.5E1
WIS	8.1E0	<b>4.2E0</b>
NAIVE	1.0E1	-

Table 498: Pixel-Gridworld, relative MSE.  $T = 25, N = 64, \pi_b = 1.00$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	3.4E0	5.4E1	1.7E1	1.5E1
Q-REG	1.7E0	3.6E0	1.9E1	1.8E1
MRDR	<b>1.2E0</b>	<b>8.5E-1</b>	3.0E1	2.7E1
FQE	1.2E2	1.1E2	3.4E1	6.6E1
$R(\lambda)$	1.1E3	2.5E3	1.4E4	5.1E3
$Q^\pi(\lambda)$	1.0E2	1.8E2	1.9E1	4.6E1
TREE	2.8E1	4.5E1	3.2E1	1.7E1
IH	7.8E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	<b>1.0E0</b>	3.3E0
WIS	2.8E1	1.7E1
NAIVE	3.6E1	-

Table 499: Pixel-Gridworld, relative MSE.  $T = 25, N = 128, \pi_b = 1.00$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	5.9E0	8.7E1	7.0E1	1.1E1
Q-REG	1.5E0	<b>2.1E0</b>	8.5E0	9.3E0
MRDR	<b>1.1E0</b>	6.7E0	1.2E1	1.4E1
FQE	6.1E1	6.0E1	1.6E1	2.4E1
R( $\lambda$ )	2.3E1	2.4E2	1.0E1	1.0E1
Q $^\pi$ ( $\lambda$ )	3.9E1	8.4E1	1.6E1	2.5E1
TREE	1.8E1	2.3E1	9.1E0	1.1E1
IH	8.4E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	2.6E3	<b>1.8E0</b>
WIS	3.6E1	1.2E1
NAIVE	3.8E1	-

Table 500: Pixel-Gridworld, relative MSE.  $T = 25, N = 256, \pi_b = 1.00$ -Greedy(V iter.),  $\pi_e = 0.10$ -Greedy(V iter.). Note: we use the same policy as in Gridworld.  $\pi_b$  unknown. Stochastic environment.

	DM	HYBRID		
	DIRECT	DR	WDR	MAGIC
AM	4.4E0	1.8E3	2.4E2	7.3E0
Q-REG	1.7E0	3.4E0	6.7E0	8.5E0
MRDR	<b>1.5E0</b>	8.8E2	8.6E0	8.9E0
FQE	6.5E0	5.0E1	<b>2.3E0</b>	2.7E0
R( $\lambda$ )	1.4E1	3.8E4	4.5E0	5.7E0
Q $^\pi$ ( $\lambda$ )	6.6E1	5.0E4	9.6E0	3.2E1
TREE	1.6E1	1.5E4	3.9E0	6.4E0
IH	7.8E0	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	3.7E5	<b>2.1E0</b>
WIS	2.2E1	8.2E0
NAIVE	3.5E1	-

## F.8 Detailed Results for Enduro

 Table 503: Enduro, relative MSE. Model Type: conv.  $T = 500, N = 512, \pi_b = 0.25$ -Greedy(DDQN),  $\pi_e = 0.10$ -Greedy(DDQN).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	-	-	-	-
Q-REG	9.0E-1	6.8E-1	7.5E-1	5.1E-1
MRDR	1.0E0	4.1E0	2.2E1	3.4E-1
FQE	6.5E-1	3.5E-1	8.6E-2	<b>4.5E-2</b>
R( $\lambda$ )	-	-	-	-
Q $^\pi$ ( $\lambda$ )	-	-	-	-
TREE	-	-	-	-
IH	<b>9.5E-2</b>	-	-	-

 Table 501: Enduro, relative MSE. Model Type: conv.  $T = 500, N = 512, \pi_b = 0.10$ -Greedy(DDQN),  $\pi_e = 0.00$ -Greedy(DDQN).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	-	-	-	-
Q-REG	8.9E-1	8.3E-1	9.8E-1	2.9E-1
MRDR	9.3E-1	1.5E0	1.5E0	3.1E-1
FQE	3.2E-1	8.5E-2	1.4E-1	<b>3.8E-2</b>
R( $\lambda$ )	-	-	-	-
Q $^\pi$ ( $\lambda$ )	-	-	-	-
TREE	-	-	-	-
IH	<b>1.0E-2</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	8.5E-1
WIS	8.9E-2	<b>8.2E-2</b>
NAIVE	1.1E-2	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	8.8E-1
WIS	1.5E-1	<b>8.9E-2</b>
NAIVE	9.9E-2	-

 Table 502: Enduro, relative MSE. Model Type: conv.  $T = 500, N = 512, \pi_b = 0.25$ -Greedy(DDQN),  $\pi_e = 0.00$ -Greedy(DDQN).

	DM		HYBRID	
	DIRECT	DR	WDR	MAGIC
AM	-	-	-	-
Q-REG	1.0E0	1.1E0	4.7E0	6.6E0
MRDR	1.0E0	1.0E0	1.3E-1	1.5E-1
FQE	7.1E-1	7.1E-1	<b>7.4E-2</b>	8.9E-2
R( $\lambda$ )	-	-	-	-
Q $^\pi$ ( $\lambda$ )	-	-	-	-
TREE	-	-	-	-
IH	<b>1.5E-1</b>	-	-	-

	IPS	
	STANDARD	PER-DECISION
IS	1.0E0	1.0E0
WIS	3.1E-1	<b>5.4E-2</b>
NAIVE	1.6E-1	-