Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning

Cameron Voloshin | Hoang M. Le | Nan Jiang | Yisong Yue
Caltech | Caltech | UIUC | 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 (Dudik et al., 2011b; Bottou et al., 2013; Swaminathan et al., 2017; Wang et al., 2017) and RL setting (Jiang & Li, 2016; Dudik 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 \((X, A, P, R, \gamma)\). \(X\) is the state space (or observation space in the POMDP case), \(A\) is the action space with finite cardinality, \(P : X \times A \times X \to [0, 1]\) is the transition function, \(R : X \times A \to \mathbb{R}\) 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-
In this paper, we propose to group OPE into three similar categories: Importance Sampling (IS), Per-Decision Importance Sampling (PDIS), and Weighted Per-Decision Importance Sampling (WPDIS). These categories are based on the historical data and the estimation of the action-value function.

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)

<table>
<thead>
<tr>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td></td>
</tr>
</tbody>
</table>

- IPS Methods
- Direct Methods
- Hybrid Methods
- Doubly Robust
- Weighted Doubly Robust
- MAGIC
- Double Robust
- Per-Decision IS
- Per-Decision WIS
- Infinite Horizon
- Tree Backup
- Model Fitting

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

These categories are based on the historical data and the estimation of the action-value function.

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(λ) (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 categories, but with expanded definition for each category. Figure 1 provides an overview of OPE methods that we consider. The relative position 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 \( \rho_{t,j}^e = \rho_{t,j}^e(\tau^t, \pi_e, \pi_b) = \prod_{i=1}^{\min(t-j, T-1)} \frac{\pi_i(x_t|a_t)}{\pi_i(x_t|a'_t)} \) be the cumulative importance weight between \( \pi_e \) and \( \pi_b \). The key idea is to reweight the rewards in the historical data by the importance sampling ratio between \( \pi_e \) and \( \pi_b \), 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 \( Q(\cdot) \) or \( V(\pi) \). 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 \( \hat{V}(\pi_e) = \frac{1}{N} \sum_{i=1}^{N} \sum_{a \in A} \pi_e(a|s_i) \hat{Q}(x_i, a; \theta) \). A simple example is Fitted Q Evaluation (FQE) (Le et al., 2019), which
Several model-free methods originated from off-policy learning settings, but are also natural for OPE. Several methods originated from off-policy estimation to that of directly minimizing the variance of the unbiased estimates produced by importance sampling techniques:

\[\hat{V}(x_i) = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T-1} \gamma^t \sum_{a \in A} \pi_t(a|x_{i+1})Q(x_{i+1}, a).\]

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.

### 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:

\[\hat{V}(x_i) = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T-1} \gamma^t \sum_{a \in A} \pi_t(a|x_{i+1})Q(x_{i+1}, a).\]

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

Several model-free methods originated from off-policy learning settings, but are also natural for OPE. \(Q^*(\lambda)\) (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., 2010) 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 Monos et al. (2016).

Direct Q Regression (Q-Reg) and More Robust Doubly-Robust (MRDR) (Farajtabar et al., 2015) are two recently proposed direct methods that make use of cumulative importance weights in deriving the regression estimate for \(Q^e\), 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 finite 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 \textit{state density distributions} induced by \(\pi_b\) and \(\pi_e\). This ratio replaces all but the final importance weights \(r_{T-1}\) in the IH estimate, which resembles IS.

### 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:

\[\frac{1}{10} \sum_{i=1}^{10} (\hat{V}(\pi_e)_i - V(\pi_e)_i)^2,\]

which allows a fair comparison across different conditions.

\[\text{MSE} = \frac{1}{10} \sum_{i=1}^{10} (\hat{V}(\pi_e)_i - V(\pi_e)_i)^2,\]
Table 3: High-level Guidelines

<table>
<thead>
<tr>
<th>Class</th>
<th>Recommended method</th>
<th>When to use</th>
<th>Prototypical example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td>FQE</td>
<td>Stochastic env, severe policy mismatch</td>
<td>Graph, MC, Pix-MC</td>
</tr>
<tr>
<td></td>
<td>Q(λ)</td>
<td>Compute non-issue, moderate policy mismatch</td>
<td>GW/Pix-GW</td>
</tr>
<tr>
<td></td>
<td>IH</td>
<td>Long horizon, mild policy mismatch, good kernel</td>
<td>Graph-MC</td>
</tr>
<tr>
<td>IPS</td>
<td>PDWIS</td>
<td>Short horizon, mild policy mismatch</td>
<td>Graph</td>
</tr>
<tr>
<td>Hybrid</td>
<td>MAGIC FQE</td>
<td>Severe model misspecification</td>
<td>Graph-POMDP, Enduro</td>
</tr>
<tr>
<td></td>
<td>MAGIC Q(λ)</td>
<td>Compute non-issue, severe model misspecification</td>
<td>Graph-POMDP</td>
</tr>
</tbody>
</table>

Figure 2: General Guideline Decision Tree

**Design.** We consider various domain characteristics (simple-complex, deterministic-stochastic, sparse-dense rewards, short-long horizon), π_b, π_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 3. 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:** π_b estimation quality depends on the state and action dimensionality, and historical data size. Poor π_b estimates cause HM and IPS to underperform simple DM.
- **Policy mismatch:** Large divergence between π_b and π_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 2 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 π_b and π_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.

---

2Poor estimation of π_b can be viewed as an instance of model misspecification. We distinguish the representation issue of π_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 6). 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 mismatch (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. 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).

3From correspondence with the authors.
Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning

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. Relatively large policy mismatch setting. 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. 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. 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 $Q^*(\lambda)$ are typically the most reliable: MAGIC with FQE or MAGIC with $Q^*(\lambda)$ 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. 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

4Applying IH in the indefinite horizon case requires setting up a separate absorbing state that loops over itself with zero terminal reward.
Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning

regime (e.g., pixel gridworld appendix figure 27–29). Similar phenomenon was observed in the statistics literature \cite{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 \( \sup_{a \in A, x \in X} \pi_b(a|x) \mathbb{E}_f [T^g | d_x] \) 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 6). 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 \cite{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 naive average of behavior policy (Figure 12). Quantifying biasedness, such as inherent Bellman error, for different function classes is currently an open problem \cite{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 \cite{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. They require order \( T \) times the number of \( Q_{k-1}(x,a) \) 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\footnote{Munos et al., 2016 limits the rolling-out horizon to 16 in Atari domains, but for policy learning scenario.} 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-
Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning

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.

dition, 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 (Dudık et al., 2011b). For contextual bandits, it has been shown that while DR is highly competitive, it is sometimes substantially outperformed 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


## Contents

1 Introduction

2 Preliminaries

3 Overview of OPE Methods
   3.1 Inverse Propensity Scoring (IPS)
   3.2 Direct Methods (DM)
   3.3 Hybrid Methods (HM)

4 Experiments

5 Results and Discussion
   5.1 High-Level Conclusions
   5.2 Method Selection Guideline
   5.3 Deeper Dive into Key Decision Factors
   5.4 Other Considerations

6 Limitations and Future Directions

7 Conclusion

A Glossary of Terms

B Methods
   B.1 Inverse Propensity Scoring (IPS) Methods
   B.2 Hybrid Methods
   B.3 Direct Methods (DM)
      B.3.1 Model-Based
      B.3.2 Model-Free

C Environments
   C.1 Environment Descriptions
      C.1.1 Graph
      C.1.2 Graph-POMDP
      C.1.3 Graph Mountain Car (Graph-MC)
      C.1.4 Mountain Car (MC)
      C.1.5 Pixel-based Mountain Car (Pix-MC)
      C.1.6 Enduro
      C.1.7 Gridworld (GW)
      C.1.8 Pixel-Gridworld (Pixel-GW)

D Experimental Setup
   D.1 Description of the policies
   D.2 Enumeration of Experiments
      D.2.1 Graph
      D.2.2 Graph-POMDP
      D.2.3 Gridworld
      D.2.4 Pixel-Gridworld (Pix-GW)
      D.2.5 Graph-MC
      D.2.6 Mountain Car (MC)
      D.2.7 Pixel-Mountain Car (Pix-MC)
      D.2.8 Enduro
      D.3 Representation and Function Class
      D.4 Choice of hyperparameters

E Additional Supporting Figures

F Tables of Results, per Environment
   F.1 Detailed Results for Graph
   F.2 Detailed Results for Graph-POMDP
   F.3 Detailed Results for Graph Mountain Car (Graph-MC)
   F.4 Detailed Results for Mountain Car (MC)
   F.5 Detailed Results for Pixel-Based Mountain Car (Pix-MC)
   F.6 Detailed Results for Gridworld
   F.7 Detailed Results for Pixel Gridworld
   F.8 Detailed Results for Enduro
A Glossary of Terms

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

Table 4: Glossary of terms

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

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

B.1 Inverse Propensity Scoring (IPS) Methods

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>$\sum_{i=1}^{N} \hat{p}<em>{it} \sum</em>{t=0}^{\hat{T}} \gamma^t r_t$</td>
<td>$\sum_{i=1}^{N} \sum_{t=0}^{\hat{T}} \gamma^t \frac{\hat{p}<em>{it}}{\hat{p}</em>{0t}} r_t$</td>
</tr>
<tr>
<td>WIS</td>
<td>$\sum_{i=1}^{N} \frac{\hat{p}<em>{it}}{\hat{p}</em>{0t}} \sum_{t=0}^{\hat{T}} \gamma^t r_t$</td>
<td>$\sum_{i=1}^{N} \sum_{t=0}^{\hat{T}} \gamma^t \frac{\hat{p}<em>{it}}{\hat{p}</em>{0t}} r_t$</td>
</tr>
</tbody>
</table>

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 methods 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)

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

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

$$V_{WDR} = \frac{1}{N} \sum_{i=1}^{N} \hat{V}(x_0^i) + \sum_{i=1}^{N} \sum_{t=0}^{\infty} \gamma^t \hat{p}_{it}(r_t - \hat{Q}(x_t^i, a_t^i) + \gamma \hat{V}(x_{t+1}^i))$$

MAGIC: (Thomas & Brunskill, 2016) Given $g_j = \{g^j|i \in J \subseteq \mathbb{N} \cup \{0\}\}$ where

$$g^j(D) = \sum_{i=1}^{N} \sum_{t=0}^{\hat{T}} \gamma^t \hat{p}_{it}(r_t - \hat{Q}(x_t^i, a_t^i))$$

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

$$\hat{x}^* = \arg\min_{x \in \Delta^{|J|}} \langle \hat{Q}_n + \hat{b}b \rangle x$$

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 $i \in J$, $g^j$ represents estimating the first $i$ steps of $V(\pi)$ 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, $V(\pi) = \mathbb{E}[\sum_{t=1}^{\hat{T}} \gamma^t r(x_t, a_t)|x_0 = x, a_0 = \pi(x_0)]$ 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

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

Note that

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

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

$$\hat{Q}(\cdot, \theta) = \min \theta \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{\hat{T}} \gamma^t \hat{p}_{it} (R_{i:t} - \hat{Q}(x_t^i, a_t^i; \theta))^2$$

Fitted Q Evaluation (FQE): (Le et al., 2019)

$$\hat{Q}(\cdot, \theta) = \lim_{k \to \infty} \hat{Q}_k$$

where

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

$$y_t^i = r_t^i + \gamma \mathbb{E}_{\pi_\epsilon}(\hat{Q}_{k-1}(x_{t+1}^i, \cdot; \theta) - y_{t+1}^i)$$
Retrace($\lambda$) (R($\lambda$)), Tree-Backup (Tree), \(Q^\pi(\lambda)\): 

\[ Q^\pi(\lambda) = \lim_{k \to \infty} Q_k \] 

where \(Q_k(x, a; \theta) = \hat{Q}_{k-1}(x, a; \theta) + \sum_{t=0}^{\infty} \gamma^t \sum_{s=1}^{N} C_s y_t | x_0 = x, a_0 = a \)

and

\[
y_t = r^t + \gamma \mathbb{E}_{\pi} \hat{Q}_{k-1}(x_{t+1}, \cdot; \theta) - \hat{Q}_{k-1}(x_t, a_t; \theta)
\]

where

\[
y_t = \begin{cases} 
\lambda \min(1, \frac{\pi(a_t|x_t)}{\pi_0(a_t|x_t)}) R(\lambda) \\
\lambda \pi_0(a_t|x_t) Tree \\
\lambda Q^\pi(\lambda)
\end{cases}
\]

More Robust Doubly-Robust (MRDR): \(\text{Farajtabar et al., 2018}\) 

Given

\[
\Omega_{\pi_0}(x) = \text{diag}[1/\pi_0(a|x)]_{a \in A} - ee^T 
\]

\[
e = [1, \ldots, 1]^T
\]

\[
R_{t:T}^i = \sum_{j=t}^{T} \gamma^{j-t} \rho_{t+1:j} r(x_t^i, a_t^i) 
\]

and

\[
q_0(x, a, r) = \text{diag}[\pi_0(a^i|x)]_{a^i \in A} [\hat{Q}(x, a^i|\theta)]_{a^i \in A} - r^i [\mathbb{I}\{a' = a\}]_{a' \in A}
\]

where \(\mathbb{I}\) is the indicator function, then

\[
\hat{Q}(. , \theta) = \min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{\hat{T}} \gamma^{2t} (\rho_{0,t}^i)^2 \times 
\]

\[
\rho_{0,t}^i q_0(x_t^i, a_t^i, R_{t:T}^i)^T \Omega_{\pi_0}(x_t^i) q_0(x_t^i, a_t^i, R_{t:T}^i)
\]

State Density Ratio Estimation (IH): \(\text{Liu et al., 2018}\)

\[
V_{1H} = \sum_{i=1}^{N} \sum_{t=0}^{\hat{T}} \gamma^t \omega(s_t^i) \rho_{0:t}^i r_t^i 
\]

\[
\omega(s_t^i) = \lim_{l \to \infty} \frac{\sum_{t=0}^{T} \gamma^t d_{\pi_0}(s_t^i)}{\sum_{t=0}^{T} \gamma^t d_{\pi_0}(s_t^i)}
\]

where \(\pi_0\) is assumed to be a fixed data-generating policy, and \(d_{\pi_0}\) 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 \(\text{Liu et al., 2018}\).

### 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 0.25. 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 \(\text{Thomas & Brunskill, 2016}\) 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 agent terminates and the agent transitions to the absorbing state.
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 \([-0.6, -0.4]\) to being one of \([-0.6, -0.5, -0.4]\), with no velocity. The environment is initialized with a horizon \( T \) and absorbing state \( x_{abs} = [0.5, 0] \), 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 5. 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 $-0.005$ and holes (denoted H) which give $-0.5$. The remaining states give a reward of $-0.01$.

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 $\epsilon - \text{Greedy}(Q^*)$ implies we follow a mixed policy $\pi = \text{arg max}_a Q^*(x, a)$ 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.

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.

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
experiments are the Cartesian product of the table.

Table 9: Pix-GW parameters

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$</td>
<td>.96</td>
</tr>
<tr>
<td>$N$</td>
<td>20</td>
</tr>
<tr>
<td>$T$</td>
<td>25</td>
</tr>
<tr>
<td>$\epsilon$ Greedy, $\pi_b$</td>
<td>{2, 4, 6, 8, 1}</td>
</tr>
<tr>
<td>$\epsilon$ Greedy, $\pi_e$</td>
<td>1</td>
</tr>
<tr>
<td>Stochastic Env</td>
<td>True, False</td>
</tr>
<tr>
<td>Stochastic Rew</td>
<td>False</td>
</tr>
<tr>
<td>Sparse Rew</td>
<td>False</td>
</tr>
<tr>
<td>Seed</td>
<td>10 of random(0 : 210)</td>
</tr>
<tr>
<td>ModelType Tabular</td>
<td></td>
</tr>
<tr>
<td>Regress $\pi_b$</td>
<td>True, False</td>
</tr>
</tbody>
</table>

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 10: Graph-MC parameters

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$</td>
<td>.99</td>
</tr>
<tr>
<td>$N$</td>
<td>27.11</td>
</tr>
<tr>
<td>$T$</td>
<td>250</td>
</tr>
<tr>
<td>($\pi_b(a = 0), \pi_e(a = 0)$</td>
<td>{{.45,.45}, (.6,.6), (.45,.6), (.6,.45), (.8,.2), (.2,.8)}</td>
</tr>
<tr>
<td>Stochastic Env</td>
<td>False</td>
</tr>
<tr>
<td>Stochastic Rew</td>
<td>False</td>
</tr>
<tr>
<td>Sparse Rew</td>
<td>False</td>
</tr>
<tr>
<td>Seed</td>
<td>10 of random(0 : 210)</td>
</tr>
<tr>
<td>ModelType Tabular</td>
<td></td>
</tr>
<tr>
<td>Regress $\pi_b$</td>
<td>False</td>
</tr>
</tbody>
</table>

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 11: MC parameters

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$</td>
<td>.99</td>
</tr>
<tr>
<td>$N$</td>
<td>27.10</td>
</tr>
<tr>
<td>$T$</td>
<td>250</td>
</tr>
<tr>
<td>$\epsilon$ Greedy, ($\pi_b$, $\pi_e$)</td>
<td>{(1,0), (1,0), (1,1), (1,1)}</td>
</tr>
<tr>
<td>Stochastic Env</td>
<td>False</td>
</tr>
<tr>
<td>Stochastic Rew</td>
<td>False</td>
</tr>
<tr>
<td>Sparse Rew</td>
<td>False</td>
</tr>
<tr>
<td>Seed</td>
<td>10 of random(0 : 210)</td>
</tr>
<tr>
<td>ModelType Tabular</td>
<td>{Tabular, NN}</td>
</tr>
<tr>
<td>Regress $\pi_b$</td>
<td>False</td>
</tr>
</tbody>
</table>

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 12: Pix-MC parameters

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$</td>
<td>.97</td>
</tr>
<tr>
<td>$N$</td>
<td>512</td>
</tr>
<tr>
<td>$T$</td>
<td>500</td>
</tr>
<tr>
<td>$\epsilon$ Greedy, ($\pi_b$, $\pi_e$)</td>
<td>{(25,0), (.1,.0), (.25,.1)}</td>
</tr>
<tr>
<td>Stochastic Env</td>
<td>False</td>
</tr>
<tr>
<td>Stochastic Rew</td>
<td>False</td>
</tr>
<tr>
<td>Sparse Rew</td>
<td>False</td>
</tr>
<tr>
<td>Seed</td>
<td>10 of random(0 : 210)</td>
</tr>
<tr>
<td>ModelType Tabular</td>
<td>{Tabular, NN}</td>
</tr>
<tr>
<td>Regress $\pi_b$</td>
<td>False</td>
</tr>
</tbody>
</table>

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 dynamic programming with Q tables. MRDR and Q-Reg used the Sherman Morrison (Sherman & Morrison 1950) method to solve the weighted-least-squares 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.
Table 14: Hyperparameters for each model by Environment

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameter</th>
<th>Graph</th>
<th>TMC</th>
<th>MC</th>
<th>Pix-MC</th>
<th>Enduro</th>
<th>Graph-POMDP</th>
<th>GW</th>
<th>Pix-GW</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>Max Traj Len</td>
<td>T</td>
<td>T</td>
<td>50</td>
<td>50</td>
<td>-</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td></td>
<td>NN Fit Epochs</td>
<td>-</td>
<td>-</td>
<td>100</td>
<td>100</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>NN Batchsize</td>
<td>-</td>
<td>-</td>
<td>32</td>
<td>32</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>NN Train size</td>
<td>-</td>
<td>-</td>
<td>.8</td>
<td>8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>.8</td>
</tr>
<tr>
<td></td>
<td>NN Val size</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>NN Early Stop delta</td>
<td>-</td>
<td>-</td>
<td>1e-4</td>
<td>1e-4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1e-4</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>Omega regul.</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>NN Fit Epochs</td>
<td>-</td>
<td>-</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>-</td>
<td>-</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>NN Batchsize</td>
<td>-</td>
<td>-</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>-</td>
<td>-</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>NN Train size</td>
<td>-</td>
<td>-</td>
<td>.8</td>
<td>.8</td>
<td>.8</td>
<td>-</td>
<td>-</td>
<td>.8</td>
</tr>
<tr>
<td></td>
<td>NN Val size</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>NN Early Stop delta</td>
<td>-</td>
<td>-</td>
<td>1e-4</td>
<td>1e-4</td>
<td>1e-4</td>
<td>-</td>
<td>-</td>
<td>1e-4</td>
</tr>
<tr>
<td>FQE</td>
<td>Convergence $\epsilon$</td>
<td>1e-5</td>
<td>1e-5</td>
<td>1e-4</td>
<td>1e-4</td>
<td>1e-5</td>
<td>1e-5</td>
<td>4e-4</td>
<td>1e-4</td>
</tr>
<tr>
<td></td>
<td>Max Iter</td>
<td>-</td>
<td>-</td>
<td>160</td>
<td>160</td>
<td>600</td>
<td>-</td>
<td>50</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>NN Batchsize</td>
<td>-</td>
<td>-</td>
<td>32</td>
<td>32</td>
<td>32</td>
<td>-</td>
<td>-</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>Optimizer Clipnorm</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>IH</td>
<td>Quad. prog. regular.</td>
<td>1e-3</td>
<td>1e-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1e-3</td>
<td>1e-3</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>NN Fit Epochs</td>
<td>-</td>
<td>-</td>
<td>10001</td>
<td>10001</td>
<td>10001</td>
<td>-</td>
<td>-</td>
<td>1001</td>
</tr>
<tr>
<td></td>
<td>NN Batchsize</td>
<td>-</td>
<td>-</td>
<td>1024</td>
<td>128</td>
<td>128</td>
<td>-</td>
<td>-</td>
<td>128</td>
</tr>
<tr>
<td>MRDR</td>
<td>Omega regul.</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>NN Fit Epochs</td>
<td>-</td>
<td>-</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>-</td>
<td>-</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>NN Batchsize</td>
<td>-</td>
<td>-</td>
<td>1024</td>
<td>1024</td>
<td>1024</td>
<td>-</td>
<td>-</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>NN Train size</td>
<td>-</td>
<td>-</td>
<td>.8</td>
<td>.8</td>
<td>.8</td>
<td>-</td>
<td>-</td>
<td>.8</td>
</tr>
<tr>
<td></td>
<td>NN Val size</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>NN Early Stop delta</td>
<td>-</td>
<td>-</td>
<td>1e-4</td>
<td>1e-4</td>
<td>1e-4</td>
<td>-</td>
<td>-</td>
<td>1e-4</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>Convergence $\epsilon$</td>
<td>1e-3</td>
<td>2e-3</td>
<td>1e-3</td>
<td>1e-3</td>
<td>1e-3</td>
<td>2e-3</td>
<td>2e-3</td>
<td>1e-3</td>
</tr>
<tr>
<td></td>
<td>Max Iter</td>
<td>500</td>
<td>500</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>500</td>
<td>50</td>
<td>-</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>NN Batchsize</td>
<td>-</td>
<td>-</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>NN Train Size</td>
<td>-</td>
<td>-</td>
<td>.03</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>.03</td>
</tr>
<tr>
<td></td>
<td>NN ClipNorm</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>$Q^*(\lambda)$</td>
<td>Convergence $\epsilon$</td>
<td>1e-3</td>
<td>2e-3</td>
<td>1e-3</td>
<td>1e-3</td>
<td>1e-3</td>
<td>2e-3</td>
<td>2e-3</td>
<td>1e-3</td>
</tr>
<tr>
<td></td>
<td>Max Iter</td>
<td>500</td>
<td>500</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>500</td>
<td>50</td>
<td>-</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>NN Batchsize</td>
<td>-</td>
<td>-</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>NN Train Size</td>
<td>-</td>
<td>-</td>
<td>.03</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>.03</td>
</tr>
<tr>
<td></td>
<td>NN ClipNorm</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>Tree</td>
<td>Convergence $\epsilon$</td>
<td>1e-3</td>
<td>2e-3</td>
<td>1e-3</td>
<td>1e-3</td>
<td>1e-3</td>
<td>2e-3</td>
<td>2e-3</td>
<td>1e-3</td>
</tr>
<tr>
<td></td>
<td>Max Iter</td>
<td>500</td>
<td>500</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>500</td>
<td>50</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>NN Batchsize</td>
<td>-</td>
<td>-</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>NN Train Size</td>
<td>-</td>
<td>-</td>
<td>.03</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>.03</td>
</tr>
<tr>
<td></td>
<td>NN ClipNorm</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
</tbody>
</table>
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.
Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning

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

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

Figure 22: Hybrid Method comparison. $\pi_b(a = 0) = .2, \pi_e(a = 0) = .8$. Min error per class. (Graph-MC)
Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning

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

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 26: Hybrid Method comparison. $\pi_b = .8$–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)
Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning

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

Figure 37: DR comparison with $\pi_b = 0.2$–Greedy(optimal), $\pi_e = 1.0$–Greedy(optimal). (Pixel Gridworld)

Figure 38: WDR comparison with $\pi_b = 0.2$–Greedy(optimal), $\pi_e = 1.0$–Greedy(optimal). (Pixel Gridworld)

Figure 39: MAGIC comparison with $\pi_b = 0.2$–Greedy(optimal), $\pi_e = 1.0$–Greedy(optimal). (Pixel Gridworld)

Figure 40: DR comparison with $\pi_b = 0.8$–Greedy(optimal), $\pi_e = 1.0$–Greedy(optimal). (Pixel Gridworld)
Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning

Figure 41: WDR comparison with $\pi_b = 0.8$-Greedy(optimal), $\pi_c = 1.0$-Greedy(optimal). (Pixel Gridworld)

Figure 42: MAGIC comparison with $\pi_b = 0.8$-Greedy(optimal), $\pi_c = 1.0$-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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.9E-1</td>
<td>5.3E-1</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>2.0E0</td>
<td>5.1E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.7E0</td>
<td>7.0E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.8E-1</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>4.8E-1</td>
<td>4.8E-1</td>
</tr>
<tr>
<td>Q^\pi(\lambda)</td>
<td>4.8E-1</td>
<td>4.8E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>4.8E-1</td>
<td>4.8E-1</td>
</tr>
<tr>
<td>IH</td>
<td>2.9E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.3E-1</td>
<td>6.7E-1</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>4.3E0</td>
<td>7.0E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.4E0</td>
<td>7.1E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>3.9E-1</td>
<td>3.9E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>3.9E-1</td>
<td>3.9E-1</td>
</tr>
<tr>
<td>Q^\pi(\lambda)</td>
<td>3.9E-1</td>
<td>3.9E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>3.9E-1</td>
<td>4.0E-1</td>
</tr>
<tr>
<td>IH</td>
<td>4.8E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.7E-1</td>
<td>2.8E-1</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>5.0E-1</td>
<td>2.4E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>7.6E-1</td>
<td>3.1E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>1.5E-1</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.5E-1</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>Q^\pi(\lambda)</td>
<td>1.5E-1</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>1.5E-1</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>IH</td>
<td>3.9E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.8E-2</td>
<td>1.1E-1</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>4.6E-1</td>
<td>1.2E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.4E-1</td>
<td>3.2E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>3.4E-2</td>
<td>3.4E-2</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>3.4E-2</td>
<td>3.4E-2</td>
</tr>
<tr>
<td>Q^\pi(\lambda)</td>
<td>3.4E-2</td>
<td>3.4E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>3.4E-2</td>
<td>3.4E-2</td>
</tr>
<tr>
<td>IH</td>
<td>4.6E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.8E-2</td>
<td>1.1E-1</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>4.6E-1</td>
<td>1.2E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.4E-1</td>
<td>3.2E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>3.4E-2</td>
<td>3.4E-2</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>3.4E-2</td>
<td>3.4E-2</td>
</tr>
<tr>
<td>Q^\pi(\lambda)</td>
<td>3.4E-2</td>
<td>3.4E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>3.4E-2</td>
<td>3.4E-2</td>
</tr>
<tr>
<td>IH</td>
<td>4.6E-2</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 19: Graph, relative MSE. $T = 4$, $N = 128$, $\pi_b(a = 0) = 0.2$, $\pi_e(a = 0) = 0.8$. Dense rewards.

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td></td>
<td>2.5E-3</td>
<td>6.1E-2</td>
<td>5.2E-2</td>
<td>5.8E-3</td>
</tr>
<tr>
<td>Q-REG</td>
<td></td>
<td>4.3E-1</td>
<td>9.7E-2</td>
<td>9.9E-3</td>
<td>1.4E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td></td>
<td>3.9E-1</td>
<td>2.5E-1</td>
<td>6.9E-2</td>
<td>1.3E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>$1.2E-5$</td>
<td>1.2E-5</td>
<td>1.2E-5</td>
<td>1.2E-5</td>
<td>1.2E-5</td>
</tr>
<tr>
<td>$R(\lambda)$ $Q^b(\lambda)$</td>
<td>$1.2E-5$</td>
<td>$1.2E-5$</td>
<td>$9.0E-6$</td>
<td>$1.2E-5$</td>
<td>$1.2E-5$</td>
</tr>
<tr>
<td>TREE</td>
<td>$1.4E-5$</td>
<td>$1.2E-5$</td>
<td>$1.1E-5$</td>
<td>$1.4E-5$</td>
<td>$1.4E-5$</td>
</tr>
<tr>
<td>IH</td>
<td>2.5E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th>Standard</th>
<th>Per-Decision</th>
<th>IS</th>
<th>4.7E0</th>
<th>5.9E-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>WIS</td>
<td>2.2E-1</td>
<td>4.5E-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.9E0</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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

IPS

<table>
<thead>
<tr>
<th>Standard</th>
<th>Per-Decision</th>
<th>IS</th>
<th>5.0E-1</th>
<th>5.5E-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>WIS</td>
<td>1.5E-1</td>
<td>1.5E-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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

IPS

<table>
<thead>
<tr>
<th>Standard</th>
<th>Per-Decision</th>
<th>IS</th>
<th>5.5E-1</th>
<th>9.7E-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>WIS</td>
<td>2.4E-1</td>
<td>5.4E-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.1E0</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td></td>
<td>4.7E-4</td>
<td>6.5E-3</td>
<td>6.2E-3</td>
<td>4.8E-3</td>
</tr>
<tr>
<td>Q-REG</td>
<td></td>
<td>4.8E-2</td>
<td>6.2E-4</td>
<td>3.7E-4</td>
<td>3.6E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td></td>
<td>2.9E-2</td>
<td>4.6E-3</td>
<td>2.9E-2</td>
<td>3.0E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>$4.4E-5$</td>
<td>$4.4E-5$</td>
<td>$4.4E-5$</td>
<td>$4.4E-5$</td>
<td>$4.4E-5$</td>
</tr>
<tr>
<td>$R(\lambda)$ $Q^b(\lambda)$</td>
<td>$4.4E-5$</td>
<td>$4.4E-5$</td>
<td>$4.4E-5$</td>
<td>$4.4E-5$</td>
<td>$4.4E-5$</td>
</tr>
<tr>
<td>TREE</td>
<td>$4.0E-5$</td>
<td>$4.4E-5$</td>
<td>$4.3E-5$</td>
<td>$4.1E-5$</td>
<td>$4.1E-5$</td>
</tr>
<tr>
<td>IH</td>
<td>1.8E-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th>Standard</th>
<th>Per-Decision</th>
<th>IS</th>
<th>3.6E-1</th>
<th>3.7E-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>WIS</td>
<td>5.2E-2</td>
<td>1.2E-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>6.3E-1</td>
<td>7.9E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>7.4E-1</td>
<td>1.0E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>6.4E-1</td>
<td>1.0E0</td>
</tr>
<tr>
<td>FQE</td>
<td>5.6E-1</td>
<td>5.8E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>5.5E-1</td>
<td>6.0E-1</td>
</tr>
<tr>
<td>Q^\pi(\lambda)</td>
<td>5.5E-1</td>
<td>8.4E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>5.5E-1</td>
<td>5.9E-1</td>
</tr>
<tr>
<td>IH</td>
<td>1.3E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>2.5E-1</td>
<td>4.7E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.3E-1</td>
<td>4.5E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.0E-1</td>
<td>2.4E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>2.2E-1</td>
<td>2.7E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>2.2E-1</td>
<td>2.7E-1</td>
</tr>
<tr>
<td>Q^\pi(\lambda)</td>
<td>2.7E-1</td>
<td>2.8E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>2.2E-1</td>
<td>2.8E-1</td>
</tr>
<tr>
<td>IH</td>
<td>1.6E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>2.1E-1</td>
<td>3.1E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>7.7E-1</td>
<td>2.5E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>6.1E-1</td>
<td>2.2E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>2.0E-1</td>
<td>2.3E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>2.2E-1</td>
<td>2.1E-1</td>
</tr>
<tr>
<td>Q^\pi(\lambda)</td>
<td>2.0E-1</td>
<td>1.7E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>2.2E-1</td>
<td>2.1E-1</td>
</tr>
<tr>
<td>IH</td>
<td>1.5E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>9.0E0</td>
<td>5.8E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>6.6E1</td>
<td>1.7E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.8E0</td>
<td>-</td>
</tr>
<tr>
<td>FQE</td>
<td>9.0E0</td>
<td>5.8E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>6.6E1</td>
<td>1.7E-1</td>
</tr>
<tr>
<td>Q^\pi(\lambda)</td>
<td>3.8E0</td>
<td>-</td>
</tr>
<tr>
<td>Tree</td>
<td>9.0E0</td>
<td>5.8E-1</td>
</tr>
<tr>
<td>IH</td>
<td>6.6E1</td>
<td>1.7E-1</td>
</tr>
<tr>
<td>DM</td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>--------</td>
<td>--------</td>
<td>-------</td>
</tr>
<tr>
<td>AM</td>
<td>2.5E-2</td>
<td>5.3E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>8.0E-1</td>
<td>1.4E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.8E-1</td>
<td>1.4E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>2.8E-2</td>
<td>3.5E-1</td>
</tr>
<tr>
<td>R(λ)</td>
<td>6.3E-2</td>
<td>2.7E-1</td>
</tr>
<tr>
<td>Q^θ(λ)</td>
<td>1.0E-1</td>
<td>3.5E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>5.6E-2</td>
<td>2.7E-1</td>
</tr>
<tr>
<td>IH</td>
<td>2.7E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>4.0E0</td>
<td>7.0E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>5.7E-1</td>
<td>1.4E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.8E0</td>
<td>-</td>
</tr>
</tbody>
</table>

**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.
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.

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid</th>
<th>DIRECT</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.2E-1</td>
<td>4.5E-1</td>
<td>1.2E0</td>
<td>6.0E-1</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.9E-1</td>
<td>6.9E-1</td>
<td>2.0E0</td>
<td>4.5E-1</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>4.4E-1</td>
<td>9.3E-1</td>
<td>2.1E0</td>
<td>3.9E-1</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>2.7E-1</td>
<td>2.8E-1</td>
<td><strong>2.5E-1</strong></td>
<td>2.7E-1</td>
<td></td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>2.8E-1</td>
<td>2.9E-1</td>
<td>2.5E-1</td>
<td>2.8E-1</td>
<td></td>
</tr>
<tr>
<td>Q$^\lambda$($\lambda$)</td>
<td>3.9E-1</td>
<td>3.3E-1</td>
<td>3.6E-1</td>
<td>3.9E-1</td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>2.8E-1</td>
<td>2.8E-1</td>
<td>2.5E-1</td>
<td>2.8E-1</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>6.2E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.4E0</td>
<td>4.7E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>2.7E0</td>
<td>1.4E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>7.4E1</td>
<td>1.1E1</td>
</tr>
<tr>
<td>WIS</td>
<td>1.6E0</td>
<td><strong>5.7E-1</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.5E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.3E0</td>
<td>9.3E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>2.0E0</td>
<td><strong>7.8E-1</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.2E0</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.1E-1</td>
<td>8.8E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.4E0</td>
<td>6.8E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>6.5E-1</td>
<td>4.3E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>8.8E-2</td>
<td>9.1E-1</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>9.8E-2</td>
<td>8.1E-1</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td><strong>7.2E-2</strong></td>
<td>9.6E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>9.9E-2</td>
<td>8.1E-1</td>
</tr>
<tr>
<td>IH</td>
<td>1.8E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>5.3E0</td>
</tr>
<tr>
<td>Per-Decision</td>
<td>8.2E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>7.5E-2</td>
<td>2.8E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.8E-1</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.6E-1</td>
<td>1.0E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>6.6E-2</td>
<td>1.9E-1</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.2E-1</td>
<td>2.0E-1</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>1.1E-1</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>1.1E-1</td>
<td>2.0E-1</td>
</tr>
<tr>
<td>IH</td>
<td><strong>5.7E-2</strong></td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>7.2E-1</td>
</tr>
<tr>
<td>Per-Decision</td>
<td>5.8E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.3E0</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.4E-2</td>
<td>1.3E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>5.5E-2</td>
<td>1.2E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.3E-2</td>
<td>7.5E-2</td>
</tr>
<tr>
<td>FQE</td>
<td><strong>1.3E-2</strong></td>
<td>8.7E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>2.2E-2</td>
<td>9.2E-2</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>1.8E-2</td>
<td>8.9E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>2.2E-2</td>
<td>9.2E-2</td>
</tr>
<tr>
<td>IH</td>
<td>1.5E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>3.2E-1</td>
</tr>
<tr>
<td>Per-Decision</td>
<td>1.4E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.1E-2</td>
<td>8.9E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>6.7E-2</td>
<td>6.6E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.4E-2</td>
<td>7.3E-2</td>
</tr>
<tr>
<td>FQE</td>
<td><strong>9.2E-3</strong></td>
<td>6.2E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.7E-2</td>
<td>6.4E-2</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>3.0E-2</td>
<td>6.4E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>1.6E-2</td>
<td>6.4E-2</td>
</tr>
<tr>
<td>IH</td>
<td>1.8E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>2.6E-1</td>
</tr>
<tr>
<td>Per-Decision</td>
<td>2.0E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.9E0</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.0E0</td>
<td>3.4E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>8.4E1</td>
<td>9.4E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.1E1</td>
<td>4.3E1</td>
</tr>
<tr>
<td>FQE</td>
<td>9.0E-1</td>
<td>2.5E0</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>8.9E-1</td>
<td>1.2E0</td>
</tr>
<tr>
<td>$Q^e(\lambda)$</td>
<td>1.1E0</td>
<td>2.1E0</td>
</tr>
<tr>
<td>TREE</td>
<td>9.0E-1</td>
<td>1.2E0</td>
</tr>
<tr>
<td>IH</td>
<td>1.7E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>7.3E-1</td>
<td>8.9E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.4E0</td>
<td>1.8E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.1E0</td>
<td>9.4E-1</td>
</tr>
<tr>
<td>FQE</td>
<td><strong>7.1E-1</strong></td>
<td>4.3E0</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>9.7E-1</td>
<td>3.1E0</td>
</tr>
<tr>
<td>$Q^e(\lambda)$</td>
<td>2.6E0</td>
<td>5.4E0</td>
</tr>
<tr>
<td>TREE</td>
<td>8.8E-1</td>
<td>3.1E0</td>
</tr>
<tr>
<td>IH</td>
<td>1.6E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>STANDARD</td>
</tr>
<tr>
<td>IS</td>
<td>1.6E1</td>
</tr>
<tr>
<td>WIS</td>
<td>3.9E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>6.3E0</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>STANDARD</td>
</tr>
<tr>
<td>IS</td>
<td><strong>5.8E0</strong></td>
</tr>
<tr>
<td>WIS</td>
<td>5.8E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.8E0</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.3E-1</td>
<td>3.0E0</td>
<td>1.9E0</td>
<td>1.2E0</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>2.5E0</td>
<td>1.1E0</td>
<td>7.3E-1</td>
<td>2.3E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.5E0</td>
<td>1.8E0</td>
<td>9.4E-1</td>
<td>3.1E0</td>
</tr>
<tr>
<td>FQE</td>
<td>3.7E-1</td>
<td>1.3E0</td>
<td>9.3E-1</td>
<td>3.8E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>3.8E-1</td>
<td>1.4E0</td>
<td>9.4E-1</td>
<td>3.7E-1</td>
</tr>
<tr>
<td>Q^b(\lambda)</td>
<td>4.4E-1</td>
<td>1.2E0</td>
<td>8.4E-1</td>
<td>2.8E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>3.7E-1</td>
<td>1.4E0</td>
<td>9.5E-1</td>
<td>3.7E-1</td>
</tr>
<tr>
<td>IH</td>
<td>3.5E-1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>7.3E-2</td>
<td>5.1E-1</td>
<td>4.0E-1</td>
<td>1.3E-1</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>3.3E-1</td>
<td>1.8E-1</td>
<td>1.7E-1</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.8E-1</td>
<td>7.8E-2</td>
<td>6.6E-2</td>
<td>2.2E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.1E-2</td>
<td>2.0E-1</td>
<td>1.9E-1</td>
<td>4.3E-2</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.1E-1</td>
<td>2.1E-1</td>
<td>2.0E-1</td>
<td>9.5E-2</td>
</tr>
<tr>
<td>Q^b(\lambda)</td>
<td>1.1E-1</td>
<td>2.0E-1</td>
<td>1.9E-1</td>
<td>8.1E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>1.0E-1</td>
<td>2.1E-1</td>
<td>2.0E-1</td>
<td>9.5E-2</td>
</tr>
<tr>
<td>IH</td>
<td>3.3E-2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.4E-2</td>
<td>3.0E-1</td>
<td>3.4E-1</td>
<td>8.5E-2</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>4.9E-1</td>
<td>2.4E-1</td>
<td>2.6E-1</td>
<td>2.3E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.5E-1</td>
<td>3.4E-1</td>
<td>2.9E-1</td>
<td>4.1E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>2.1E-2</td>
<td>2.1E-1</td>
<td>2.2E-1</td>
<td>2.3E-2</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>3.2E-2</td>
<td>2.0E-1</td>
<td>2.2E-1</td>
<td>3.0E-2</td>
</tr>
<tr>
<td>Q^b(\lambda)</td>
<td>4.5E-2</td>
<td>2.1E-1</td>
<td>2.3E-1</td>
<td>3.7E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>3.1E-2</td>
<td>2.0E-1</td>
<td>2.2E-1</td>
<td>3.0E-2</td>
</tr>
<tr>
<td>IH</td>
<td>2.3E-2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.4E-2</td>
<td>3.0E-1</td>
<td>3.4E-1</td>
<td>8.5E-2</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>4.9E-1</td>
<td>2.4E-1</td>
<td>2.6E-1</td>
<td>2.3E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.5E-1</td>
<td>3.4E-1</td>
<td>2.9E-1</td>
<td>4.1E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>2.1E-2</td>
<td>2.1E-1</td>
<td>2.2E-1</td>
<td>2.3E-2</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>3.2E-2</td>
<td>2.0E-1</td>
<td>2.2E-1</td>
<td>3.0E-2</td>
</tr>
<tr>
<td>Q^b(\lambda)</td>
<td>4.5E-2</td>
<td>2.1E-1</td>
<td>2.3E-1</td>
<td>3.7E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>3.1E-2</td>
<td>2.0E-1</td>
<td>2.2E-1</td>
<td>3.0E-2</td>
</tr>
<tr>
<td>IH</td>
<td>2.3E-2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 47: Graph, relative MSE. $T = 4, N = 8, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$. Sparse rewards.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.2E0</td>
<td><strong>1.2E0</strong></td>
<td>1.6E0</td>
<td>1.2E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-Reg</td>
<td>2.5E0</td>
<td>2.5E0</td>
<td>2.3E0</td>
<td>2.5E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>2.3E0</td>
<td>4.2E0</td>
<td>1.7E0</td>
<td>2.1E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>1.3E0</td>
<td>1.3E0</td>
<td>1.3E0</td>
<td>1.3E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R(λ)</td>
<td>1.3E0</td>
<td>1.3E0</td>
<td>1.3E0</td>
<td>1.3E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q^*(λ)</td>
<td>1.3E0</td>
<td>1.3E0</td>
<td>1.3E0</td>
<td>1.3E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>1.2E0</td>
<td>1.3E0</td>
<td>1.3E0</td>
<td>1.3E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td><strong>1.1E0</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>7.4E-1</td>
<td>1.3E0</td>
<td>8.8E-1</td>
<td>7.5E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-Reg</td>
<td>2.5E0</td>
<td>1.7E0</td>
<td>1.6E0</td>
<td>1.0E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>2.7E0</td>
<td>1.8E0</td>
<td><strong>5.7E-1</strong></td>
<td>1.3E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>6.8E-1</td>
<td>6.8E-1</td>
<td>6.8E-1</td>
<td>6.8E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R(λ)</td>
<td>6.8E-1</td>
<td>6.8E-1</td>
<td>7.0E-1</td>
<td>6.8E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q^*(λ)</td>
<td>6.8E-1</td>
<td>6.8E-1</td>
<td>6.8E-1</td>
<td>6.8E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td><strong>6.8E-1</strong></td>
<td>6.8E-1</td>
<td>7.2E-1</td>
<td>6.8E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>7.4E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.8E-1</td>
<td>1.1E0</td>
<td>7.2E-1</td>
<td>4.8E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.5E1</td>
<td>2.2E1</td>
<td>2.2E0</td>
<td>1.5E1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>1.4E1</td>
<td>4.2E1</td>
<td>3.0E0</td>
<td>1.5E1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>4.3E-1</td>
<td>4.3E-1</td>
<td>4.3E-1</td>
<td>4.3E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R(λ)</td>
<td>4.2E-1</td>
<td>4.3E-1</td>
<td>4.3E-1</td>
<td>4.3E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q^*(λ)</td>
<td>4.3E-1</td>
<td>4.3E-1</td>
<td>4.3E-1</td>
<td>4.3E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>4.2E-1</td>
<td>4.2E-1</td>
<td>4.3E-1</td>
<td><strong>4.2E-1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td><strong>4.0E-1</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 49: Graph, relative MSE. $T = 4, N = 64, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$. Sparse rewards.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td><strong>2.7E-1</strong></td>
<td>4.8E-1</td>
<td>6.0E-1</td>
<td>2.9E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-Reg</td>
<td>7.2E-1</td>
<td>6.4E-1</td>
<td>4.3E-1</td>
<td>4.8E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>9.3E-1</td>
<td>5.6E-1</td>
<td>5.9E-1</td>
<td>5.1E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>2.7E-1</td>
<td>2.7E-1</td>
<td>2.7E-1</td>
<td>2.7E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R(λ)</td>
<td>2.7E-1</td>
<td>2.7E-1</td>
<td>2.7E-1</td>
<td>2.7E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q^*(λ)</td>
<td>2.7E-1</td>
<td>2.7E-1</td>
<td>2.7E-1</td>
<td>2.7E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>2.7E-1</td>
<td>2.7E-1</td>
<td>2.7E-1</td>
<td><strong>2.7E-1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>4.1E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>3.8E0</td>
<td>3.8E0</td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>3.9E0</td>
<td>3.9E0</td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.9E0</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>3.8E0</td>
<td>3.9E0</td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>3.9E0</td>
<td>3.9E0</td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.9E0</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>6.7E-1</td>
<td>6.7E-1</td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>1.1E0</td>
<td>1.1E0</td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.9E0</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
Table 51: Graph, relative MSE. $T = 4, N = 128, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$. Sparse rewards.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>Hybrid DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>5.9E-3</td>
<td>1.4E-1</td>
<td>1.6E-1</td>
<td>3.3E-2</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>5.8E0</td>
<td>7.9E-1</td>
<td>1.1E-1</td>
<td>6.7E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.2E0</td>
<td>2.2E0</td>
<td>7.5E-1</td>
<td>4.1E0</td>
</tr>
<tr>
<td>FQE</td>
<td>9.0E-6</td>
<td>9.0E-6</td>
<td>9.0E-6</td>
<td>9.0E-6</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>$\mathbf{7.0E-6}$</td>
<td>7.0E-6</td>
<td>1.0E-5</td>
<td>7.0E-6</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>9.0E-6</td>
<td>9.0E-6</td>
<td>9.0E-6</td>
<td>9.0E-6</td>
</tr>
<tr>
<td>Tree</td>
<td>1.5E-5</td>
<td>2.3E-5</td>
<td>1.4E-5</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>2.4E-1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>4.8E0</td>
<td>4.8E0</td>
</tr>
<tr>
<td>WIS</td>
<td>1.0E-1</td>
<td>$\mathbf{1.0E-1}$</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>Hybrid DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.6E-3</td>
<td>4.3E-2</td>
<td>6.0E-2</td>
<td>4.0E-2</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>8.3E-1</td>
<td>2.5E0</td>
<td>1.8E0</td>
<td>6.8E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>7.9E-1</td>
<td>5.9E-1</td>
<td>6.3E-1</td>
<td>8.0E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>3.0E-5</td>
<td>3.0E-5</td>
<td>3.0E-5</td>
<td>3.0E-5</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>$\mathbf{3.0E-5}$</td>
<td>3.0E-5</td>
<td>2.1E-5</td>
<td>3.0E-5</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>3.0E-5</td>
<td>3.0E-5</td>
<td>3.0E-5</td>
<td>3.0E-5</td>
</tr>
<tr>
<td>Tree</td>
<td>4.5E-5</td>
<td>3.2E-5</td>
<td>$\mathbf{2.1E-5}$</td>
<td>4.3E-5</td>
</tr>
<tr>
<td>IH</td>
<td>1.0E-1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>7.8E-1</td>
<td>7.8E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>3.6E-1</td>
<td>$\mathbf{3.6E-1}$</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.1E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>Hybrid DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.1E-3</td>
<td>5.1E-2</td>
<td>4.0E-2</td>
<td>2.0E-2</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>4.8E-1</td>
<td>3.9E-2</td>
<td>1.2E-2</td>
<td>1.9E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.7E-1</td>
<td>2.7E-1</td>
<td>2.7E-1</td>
<td>3.6E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>9.3E-7</td>
<td>9.3E-7</td>
<td>9.3E-7</td>
<td>9.3E-7</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>8.3E-7</td>
<td>8.4E-7</td>
<td>3.0E-6</td>
<td>8.8E-7</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>9.7E-7</td>
<td>9.3E-7</td>
<td>9.3E-7</td>
<td>9.5E-7</td>
</tr>
<tr>
<td>Tree</td>
<td>$\mathbf{1.8E-7}$</td>
<td>9.3E-7</td>
<td>8.5E-6</td>
<td>$\mathbf{1.5E-7}$</td>
</tr>
<tr>
<td>IH</td>
<td>5.6E-2</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>4.9E-1</td>
<td>4.9E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>$\mathbf{2.8E-1}$</td>
<td>2.8E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.9E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>Hybrid DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.5E-3</td>
<td>1.1E-2</td>
<td>1.1E-2</td>
<td>6.8E-3</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>5.1E-1</td>
<td>1.8E-2</td>
<td>4.3E-3</td>
<td>3.6E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.3E-1</td>
<td>1.8E-1</td>
<td>4.7E-1</td>
<td>7.5E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>1.5E-5</td>
<td>1.5E-5</td>
<td>1.5E-5</td>
<td>1.5E-5</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.6E-5</td>
<td>1.6E-5</td>
<td>$\mathbf{1.5E-5}$</td>
<td>1.6E-5</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>$\mathbf{1.5E-5}$</td>
<td>1.5E-5</td>
<td>1.5E-5</td>
<td>1.5E-5</td>
</tr>
<tr>
<td>Tree</td>
<td>2.7E-5</td>
<td>1.7E-5</td>
<td>1.6E-5</td>
<td>2.6E-5</td>
</tr>
<tr>
<td>IH</td>
<td>2.1E-2</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>4.5E1</td>
<td>4.5E1</td>
</tr>
<tr>
<td>WIS</td>
<td>$\mathbf{1.2E-1}$</td>
<td>1.2E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.1E0</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>5.5E0</td>
<td>1.1E1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.2E1</td>
<td>1.1E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.1E1</td>
<td>6.5E0</td>
</tr>
<tr>
<td>FQE</td>
<td>5.7E0</td>
<td>5.0E0</td>
</tr>
<tr>
<td>R(λ)</td>
<td>5.4E0</td>
<td>5.6E0</td>
</tr>
<tr>
<td>$Q^*(λ)$</td>
<td><strong>3.9E0</strong></td>
<td>1.9E1</td>
</tr>
<tr>
<td>TREE</td>
<td>5.5E0</td>
<td>5.4E0</td>
</tr>
<tr>
<td>IH</td>
<td>1.5E1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>2.7E0</td>
<td>5.3E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.5E0</td>
<td>3.9E0</td>
</tr>
<tr>
<td>MRDR</td>
<td><strong>1.3E0</strong></td>
<td><strong>1.8E0</strong></td>
</tr>
<tr>
<td>FQE</td>
<td>2.6E0</td>
<td>2.8E0</td>
</tr>
<tr>
<td>R(λ)</td>
<td>2.6E0</td>
<td>2.7E0</td>
</tr>
<tr>
<td>$Q^*(λ)$</td>
<td>3.8E0</td>
<td>2.6E0</td>
</tr>
<tr>
<td>TREE</td>
<td>2.6E0</td>
<td>2.7E0</td>
</tr>
<tr>
<td>IH</td>
<td>2.3E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>3.3E0</td>
<td>9.0E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>5.4E1</td>
<td>7.6E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.8E1</td>
<td>1.3E1</td>
</tr>
<tr>
<td>FQE</td>
<td>2.8E0</td>
<td>6.4E0</td>
</tr>
<tr>
<td>R(λ)</td>
<td>3.5E0</td>
<td>5.1E0</td>
</tr>
<tr>
<td>$Q^*(λ)$</td>
<td>4.8E0</td>
<td>6.8E0</td>
</tr>
<tr>
<td>TREE</td>
<td>3.4E0</td>
<td>5.1E0</td>
</tr>
<tr>
<td>IH</td>
<td><strong>1.8E0</strong></td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>3.5E0</td>
<td><strong>9.8E0</strong></td>
</tr>
<tr>
<td>Q-REG</td>
<td>5.8E0</td>
<td>7.6E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.8E1</td>
<td>1.3E1</td>
</tr>
<tr>
<td>FQE</td>
<td>2.8E0</td>
<td>6.4E0</td>
</tr>
<tr>
<td>R(λ)</td>
<td>3.5E0</td>
<td>5.1E0</td>
</tr>
<tr>
<td>$Q^*(λ)$</td>
<td>4.8E0</td>
<td>6.8E0</td>
</tr>
<tr>
<td>TREE</td>
<td>3.4E0</td>
<td>5.1E0</td>
</tr>
<tr>
<td>IH</td>
<td><strong>1.8E0</strong></td>
<td>-</td>
</tr>
</tbody>
</table>

Table 59: Graph, relative MSE. $T = 4, N = 64, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$. Stochastic rewards. Sparse rewards.

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>STANDARD</td>
</tr>
<tr>
<td>IS</td>
<td>1.4E1</td>
</tr>
<tr>
<td>WIS</td>
<td>1.3E1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>8.6E0</td>
</tr>
</tbody>
</table>

Table 60: Graph, relative MSE. $T = 4, N = 64, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$. Stochastic rewards. Sparse rewards.

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>STANDARD</td>
</tr>
<tr>
<td>IS</td>
<td>2.0E0</td>
</tr>
<tr>
<td>WIS</td>
<td>4.7E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.7E0</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td>1.5E0</td>
<td>4.2E0</td>
</tr>
<tr>
<td>WDR</td>
<td>2.5E0</td>
<td>1.5E0</td>
</tr>
<tr>
<td>MAGIC</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.2E1</td>
<td>1.2E1</td>
<td>1.5E1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.9E1</td>
<td>2.9E1</td>
<td>3.6E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.5E0</td>
<td>1.9E0</td>
<td>2.4E0</td>
</tr>
<tr>
<td>FQE</td>
<td>2.1E0</td>
<td>1.8E0</td>
<td>2.4E0</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.4E0</td>
<td>1.9E0</td>
<td>2.3E0</td>
</tr>
<tr>
<td>$Q^*(\lambda)$</td>
<td>2.1E0</td>
<td>1.8E0</td>
<td>2.4E0</td>
</tr>
<tr>
<td>Tree</td>
<td>1.4E0</td>
<td>2.0E0</td>
<td>2.3E0</td>
</tr>
<tr>
<td>IH</td>
<td>2.2E0</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td>1.0E0</td>
<td>1.2E0</td>
</tr>
<tr>
<td>WDR</td>
<td>6.1E0</td>
<td>1.5E0</td>
</tr>
<tr>
<td>MAGIC</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>9.5E0</td>
<td>7.2E0</td>
<td>5.6E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>6.7E0</td>
<td>4.2E0</td>
<td>4.8E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.1E0</td>
<td>8.0E0</td>
<td>5.2E0</td>
</tr>
<tr>
<td>FQE</td>
<td>2.1E0</td>
<td>7.5E0</td>
<td>5.2E0</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.6E0</td>
<td>8.4E0</td>
<td>5.4E0</td>
</tr>
<tr>
<td>$Q^*(\lambda)$</td>
<td>2.0E0</td>
<td>7.4E0</td>
<td>5.1E0</td>
</tr>
<tr>
<td>Tree</td>
<td>1.3E0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IPS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard</td>
<td>3.7E1</td>
<td>3.0E1</td>
<td></td>
</tr>
<tr>
<td>Per-Decision</td>
<td>3.1E0</td>
<td>3.9E0</td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.1E0</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td>2.5E-1</td>
<td>2.4E0</td>
</tr>
<tr>
<td>WDR</td>
<td>2.1E0</td>
<td>3.1E-1</td>
</tr>
<tr>
<td>MAGIC</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.6E0</td>
<td>1.2E0</td>
<td>1.2E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.8E0</td>
<td>1.4E0</td>
<td>1.0E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.8E-1</td>
<td>1.1E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>FQE</td>
<td>4.2E-1</td>
<td>1.2E0</td>
<td>1.2E0</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>4.9E-1</td>
<td>1.1E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>$Q^*(\lambda)$</td>
<td>4.0E-1</td>
<td>1.2E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>Tree</td>
<td>2.9E-1</td>
<td>4.0E-1</td>
<td>4.0E-1</td>
</tr>
<tr>
<td>IPS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard</td>
<td>5.5E-1</td>
<td>1.5E0</td>
<td></td>
</tr>
<tr>
<td>Per-Decision</td>
<td>3.8E-1</td>
<td>9.3E-1</td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.9E0</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td>5.1E-2</td>
<td>8.5E-1</td>
</tr>
<tr>
<td>WDR</td>
<td>7.2E-1</td>
<td>1.7E-1</td>
</tr>
<tr>
<td>MAGIC</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>5.2E-1</td>
<td>7.1E-1</td>
<td>4.8E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.9E-1</td>
<td>5.3E-1</td>
<td>4.9E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>6.8E-2</td>
<td>4.1E-1</td>
<td>4.2E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>8.8E-2</td>
<td>4.3E-1</td>
<td>4.4E-1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>7.8E-2</td>
<td>4.1E-1</td>
<td>4.1E-1</td>
</tr>
<tr>
<td>$Q^*(\lambda)$</td>
<td>8.5E-2</td>
<td>4.3E-1</td>
<td>4.4E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>5.1E-2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IPS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard</td>
<td>8.7E-1</td>
<td>5.2E-1</td>
<td></td>
</tr>
<tr>
<td>Per-Decision</td>
<td>1.1E0</td>
<td>4.6E-1</td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.9E0</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>3.8E0</td>
<td>7.7E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>6.4E2</td>
<td>1.8E3</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.2E2</td>
<td>3.7E3</td>
</tr>
<tr>
<td>FQE</td>
<td>3.5E0</td>
<td>2.9E0</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>3.4E0</td>
<td>2.8E0</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>5.1E0</td>
<td>2.9E0</td>
</tr>
<tr>
<td>Tree</td>
<td>3.4E0</td>
<td>2.8E0</td>
</tr>
<tr>
<td>IH</td>
<td>2.3E0</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 64: Graph, relative MSE. $T = 4$, $N = 8$, $\pi_b(a = 0) = 0.2$, $\pi_e(a = 0) = 0.8$. Stochastic environment. Sparse rewards.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>2.3E0</td>
<td>8.7E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.7E0</td>
<td>4.1E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.3E0</td>
<td>5.6E0</td>
</tr>
<tr>
<td>FQE</td>
<td>1.8E0</td>
<td>3.9E0</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.5E0</td>
<td>1.4E0</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>3.6E0</td>
<td>4.2E0</td>
</tr>
<tr>
<td>Tree</td>
<td>1.5E0</td>
<td>1.4E0</td>
</tr>
<tr>
<td>IH</td>
<td>7.9E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>3.0E0</td>
<td>2.6E1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>9.1E0</td>
<td>2.2E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.5E1</td>
<td>2.4E0</td>
</tr>
<tr>
<td>FQE</td>
<td>2.6E0</td>
<td>2.1E1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>3.5E0</td>
<td>1.2E1</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>5.6E0</td>
<td>2.1E1</td>
</tr>
<tr>
<td>Tree</td>
<td>3.3E0</td>
<td>1.2E1</td>
</tr>
<tr>
<td>IH</td>
<td>5.8E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>3.2E-1</td>
<td>1.6E1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.2E1</td>
<td>4.3E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.7E1</td>
<td>9.4E0</td>
</tr>
<tr>
<td>FQE</td>
<td>2.4E-1</td>
<td>5.3E0</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>9.6E-1</td>
<td>5.4E0</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>7.6E-1</td>
<td>5.5E0</td>
</tr>
<tr>
<td>Tree</td>
<td>7.7E-1</td>
<td>5.4E0</td>
</tr>
<tr>
<td>IH</td>
<td>2.6E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 63: Graph, relative MSE. $T = 4$, $N = 16$, $\pi_b(a = 0) = 0.2$, $\pi_e(a = 0) = 0.8$. Stochastic environment. Sparse rewards.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>2.3E0</td>
<td>8.7E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.7E0</td>
<td>4.1E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.3E0</td>
<td>5.6E0</td>
</tr>
<tr>
<td>FQE</td>
<td>1.8E0</td>
<td>3.9E0</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.5E0</td>
<td>1.4E0</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>3.6E0</td>
<td>4.2E0</td>
</tr>
<tr>
<td>Tree</td>
<td>1.5E0</td>
<td>1.4E0</td>
</tr>
<tr>
<td>IH</td>
<td>7.9E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>2.3E0</td>
<td>8.7E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.7E0</td>
<td>4.1E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.3E0</td>
<td>5.6E0</td>
</tr>
<tr>
<td>FQE</td>
<td>1.8E0</td>
<td>3.9E0</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.5E0</td>
<td>1.4E0</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>3.6E0</td>
<td>4.2E0</td>
</tr>
<tr>
<td>Tree</td>
<td>1.5E0</td>
<td>1.4E0</td>
</tr>
<tr>
<td>IH</td>
<td>7.9E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 65: Graph, relative MSE. $T = 4$, $N = 16$, $\pi_b(a = 0) = 0.2$, $\pi_e(a = 0) = 0.8$. Stochastic environment. Sparse rewards.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>2.3E0</td>
<td>8.7E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.7E0</td>
<td>4.1E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.3E0</td>
<td>5.6E0</td>
</tr>
<tr>
<td>FQE</td>
<td>1.8E0</td>
<td>3.9E0</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.5E0</td>
<td>1.4E0</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>3.6E0</td>
<td>4.2E0</td>
</tr>
<tr>
<td>Tree</td>
<td>1.5E0</td>
<td>1.4E0</td>
</tr>
<tr>
<td>IH</td>
<td>7.9E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 66: Graph, relative MSE. $T = 4$, $N = 16$, $\pi_b(a = 0) = 0.2$, $\pi_e(a = 0) = 0.8$. Stochastic environment. Sparse rewards.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>2.3E0</td>
<td>8.7E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.7E0</td>
<td>4.1E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.3E0</td>
<td>5.6E0</td>
</tr>
<tr>
<td>FQE</td>
<td>1.8E0</td>
<td>3.9E0</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.5E0</td>
<td>1.4E0</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>3.6E0</td>
<td>4.2E0</td>
</tr>
<tr>
<td>Tree</td>
<td>1.5E0</td>
<td>1.4E0</td>
</tr>
<tr>
<td>IH</td>
<td>7.9E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 63: Graph, relative MSE. $T = 4$, $N = 16$, $\pi_b(a = 0) = 0.2$, $\pi_e(a = 0) = 0.8$. Stochastic environment. Sparse rewards.
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM Direct</th>
<th>DM Hybrid</th>
<th>DR WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.2E-1</td>
<td>2.1E0</td>
<td>1.5E0</td>
<td>6.0E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.7E1</td>
<td>2.8E0</td>
<td>2.9E0</td>
<td>1.5E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.4E1</td>
<td>1.1E1</td>
<td>9.8E0</td>
<td>2.1E1</td>
</tr>
<tr>
<td>FQE</td>
<td>3.6E-1</td>
<td>2.3E0</td>
<td>1.8E0</td>
<td>3.6E-1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>6.8E-1</td>
<td>2.1E0</td>
<td>1.8E0</td>
<td>6.8E-1</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>4.5E-1</td>
<td>2.5E0</td>
<td>1.9E0</td>
<td>4.8E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>6.5E-1</td>
<td>2.1E0</td>
<td>1.8E0</td>
<td>6.5E-1</td>
</tr>
<tr>
<td>IH</td>
<td>3.0E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM Direct</th>
<th>DM Hybrid</th>
<th>DR WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.6E-1</td>
<td>1.9E0</td>
<td>2.3E0</td>
<td>4.9E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.4E-1</td>
<td>7.5E-1</td>
<td>5.7E-1</td>
<td>2.7E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.8E-1</td>
<td>5.3E-1</td>
<td>2.1E0</td>
<td>1.9E0</td>
</tr>
<tr>
<td>FQE</td>
<td>1.4E-1</td>
<td>6.5E-1</td>
<td>5.6E-1</td>
<td>1.3E-1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>2.7E-1</td>
<td>7.1E-1</td>
<td>5.9E-1</td>
<td>2.8E-1</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>2.5E-1</td>
<td>6.6E-1</td>
<td>5.5E-1</td>
<td>2.2E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>2.7E-1</td>
<td>7.1E-1</td>
<td>5.9E-1</td>
<td>2.8E-1</td>
</tr>
<tr>
<td>IH</td>
<td>2.0E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM Direct</th>
<th>DM Hybrid</th>
<th>DR WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>5.1E-2</td>
<td>1.4E0</td>
<td>1.3E0</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.4E0</td>
<td>4.7E-1</td>
<td>3.7E-1</td>
<td>9.8E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.8E0</td>
<td>5.1E-1</td>
<td>9.4E-1</td>
<td>1.8E0</td>
</tr>
<tr>
<td>FQE</td>
<td>6.1E-2</td>
<td>3.2E-1</td>
<td>3.1E-1</td>
<td>6.4E-2</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>9.8E-2</td>
<td>3.3E-1</td>
<td>3.3E-1</td>
<td>1.0E-1</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>2.2E-1</td>
<td>3.3E-1</td>
<td>3.3E-1</td>
<td>1.9E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>9.0E-2</td>
<td>3.3E-1</td>
<td>3.3E-1</td>
<td>9.4E-2</td>
</tr>
<tr>
<td>IH</td>
<td>1.2E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM Direct</th>
<th>DM Hybrid</th>
<th>DR WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.1E-2</td>
<td>2.7E-1</td>
<td>3.1E-1</td>
<td>4.5E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.6E-1</td>
<td>2.4E-1</td>
<td>2.3E-1</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.1E0</td>
<td>3.1E-1</td>
<td>3.0E-1</td>
<td>6.5E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>2.5E-2</td>
<td>2.1E-1</td>
<td>2.0E-1</td>
<td>2.5E-2</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>4.0E-2</td>
<td>2.1E-1</td>
<td>2.0E-1</td>
<td>3.9E-2</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>4.9E-2</td>
<td>2.2E-1</td>
<td>2.1E-1</td>
<td>2.3E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>4.0E-2</td>
<td>2.1E-1</td>
<td>2.0E-1</td>
<td>3.9E-2</td>
</tr>
<tr>
<td>IH</td>
<td>7.6E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 71: Graph, relative MSE. $T = 4, N = 2048, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$. Stochastic environment. Sparse rewards.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM Direct</th>
<th>DM Hybrid</th>
<th>DR WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.0E-1</td>
<td>3.0E-1</td>
<td>3.0E-1</td>
<td>3.0E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.4E-1</td>
<td>3.4E-1</td>
<td>3.4E-1</td>
<td>3.4E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.9E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FQE</td>
<td>3.6E-1</td>
<td>3.6E-1</td>
<td>3.6E-1</td>
<td>3.6E-1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>4.0E-2</td>
<td>4.0E-2</td>
<td>4.0E-2</td>
<td>4.0E-2</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>4.9E-2</td>
<td>4.9E-2</td>
<td>4.9E-2</td>
<td>4.9E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>4.0E-2</td>
<td>4.0E-2</td>
<td>4.0E-2</td>
<td>4.0E-2</td>
</tr>
<tr>
<td>IH</td>
<td>7.6E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DM WDR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.6E1</td>
<td>8.6E1</td>
<td>5.5E1</td>
<td>1.6E1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.7E0</td>
<td>6.2E1</td>
<td>3.3E1</td>
<td>3.7E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.5E0</td>
<td>9.6E1</td>
<td>2.6E1</td>
<td>3.5E0</td>
</tr>
<tr>
<td>FQE</td>
<td>1.1E1</td>
<td>2.3E1</td>
<td>1.7E1</td>
<td>1.1E1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>9.5E0</td>
<td>9.4E0</td>
<td>1.1E1</td>
<td>9.5E0</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>1.1E1</td>
<td>1.8E1</td>
<td>1.2E1</td>
<td>1.1E1</td>
</tr>
<tr>
<td>Tree</td>
<td>9.7E0</td>
<td>9.6E0</td>
<td>1.2E1</td>
<td>9.7E0</td>
</tr>
<tr>
<td>IH</td>
<td>2.0E1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DM WDR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.6E1</td>
<td>7.1E1</td>
<td>2.4E1</td>
<td>1.4E1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>8.0E2</td>
<td>1.8E3</td>
<td>8.5E1</td>
<td>7.9E2</td>
</tr>
<tr>
<td>MRDR</td>
<td>7.2E2</td>
<td>4.1E3</td>
<td>1.2E2</td>
<td>6.9E2</td>
</tr>
<tr>
<td>FQE</td>
<td>1.3E1</td>
<td>1.7E2</td>
<td>1.8E1</td>
<td>1.3E1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.4E1</td>
<td>5.2E1</td>
<td>1.5E1</td>
<td>1.3E1</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>2.4E1</td>
<td>1.9E2</td>
<td>1.8E1</td>
<td>2.4E1</td>
</tr>
<tr>
<td>Tree</td>
<td>1.3E1</td>
<td>5.3E1</td>
<td>1.5E1</td>
<td>1.3E1</td>
</tr>
<tr>
<td>IH</td>
<td>2.0E1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DM WDR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>8.8E0</td>
<td>3.2E1</td>
<td>1.5E1</td>
<td>8.4E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.0E1</td>
<td>5.6E1</td>
<td>1.8E1</td>
<td>4.0E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.3E1</td>
<td>7.1E1</td>
<td>2.1E1</td>
<td>2.7E1</td>
</tr>
<tr>
<td>FQE</td>
<td>9.6E0</td>
<td>2.0E1</td>
<td>1.1E1</td>
<td>9.6E0</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.3E1</td>
<td>2.6E1</td>
<td>1.6E1</td>
<td>1.3E1</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>1.3E1</td>
<td>2.2E1</td>
<td>1.5E1</td>
<td>1.3E1</td>
</tr>
<tr>
<td>Tree</td>
<td>1.3E1</td>
<td>2.6E1</td>
<td>1.6E1</td>
<td>1.3E1</td>
</tr>
<tr>
<td>IH</td>
<td>1.5E1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DM WDR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>6.4E0</td>
<td>3.6E1</td>
<td>3.7E1</td>
<td>7.5E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>6.4E1</td>
<td>2.3E1</td>
<td>1.0E1</td>
<td>6.6E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.3E1</td>
<td>4.3E1</td>
<td>6.8E0</td>
<td>5.4E1</td>
</tr>
<tr>
<td>FQE</td>
<td>6.4E0</td>
<td>8.7E0</td>
<td>8.6E0</td>
<td>6.4E0</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>7.1E0</td>
<td>7.4E0</td>
<td>7.0E0</td>
<td>7.1E0</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>8.0E0</td>
<td>1.2E1</td>
<td>9.8E0</td>
<td>8.1E0</td>
</tr>
<tr>
<td>Tree</td>
<td>7.1E0</td>
<td>7.2E0</td>
<td>6.9E0</td>
<td>7.1E0</td>
</tr>
<tr>
<td>IH</td>
<td>5.1E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>6.6E1</td>
<td>1.1E1</td>
</tr>
<tr>
<td>WIS</td>
<td>2.6E1</td>
<td>1.4E1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>1.2E1</td>
<td>-</td>
</tr>
</tbody>
</table>

### IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>2.0E1</td>
<td>3.7E1</td>
</tr>
<tr>
<td>WIS</td>
<td>1.6E1</td>
<td>1.2E1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.3E0</td>
<td>-</td>
</tr>
</tbody>
</table>

### IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>9.3E1</td>
<td>8.0E1</td>
</tr>
<tr>
<td>WIS</td>
<td>1.4E1</td>
<td>9.1E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>6.5E0</td>
<td>-</td>
</tr>
</tbody>
</table>
### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>Hybrid DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.8E0</td>
<td>5.3E1</td>
<td>2.0E1</td>
<td>2.8E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.4E1</td>
<td>1.2E1</td>
<td>1.4E1</td>
<td>3.9E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.5E1</td>
<td>2.1E1</td>
<td>1.6E1</td>
<td>4.6E1</td>
</tr>
<tr>
<td>FQE</td>
<td>2.8E0</td>
<td>4.4E1</td>
<td>1.4E1</td>
<td>2.8E0</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>5.2E1</td>
<td>2.9E1</td>
<td>1.4E1</td>
<td>4.4E0</td>
</tr>
<tr>
<td>Q^T(\lambda)</td>
<td>6.4E0</td>
<td>4.2E1</td>
<td>1.5E1</td>
<td>5.3E0</td>
</tr>
<tr>
<td>Tree</td>
<td>4.9E0</td>
<td>3.0E1</td>
<td>1.4E1</td>
<td>4.3E0</td>
</tr>
<tr>
<td>IH</td>
<td><strong>2.6E0</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 76: IPS, Standard Per-Decision

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.3E2</td>
<td>5.7E1</td>
</tr>
<tr>
<td>WIS</td>
<td>3.1E1</td>
<td><strong>1.5E1</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.4E0</td>
<td>-</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>Hybrid DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>6.4E-1</td>
<td>7.3E0</td>
<td>6.0E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.0E0</td>
<td>1.9E0</td>
<td>1.8E0</td>
<td>2.6E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.8E0</td>
<td>1.8E0</td>
<td>2.5E0</td>
<td>2.5E0</td>
</tr>
<tr>
<td>FQE</td>
<td><strong>5.5E-1</strong></td>
<td>2.0E0</td>
<td>1.5E0</td>
<td><strong>5.4E-1</strong></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>7.0E-1</td>
<td>1.9E0</td>
<td>1.5E0</td>
<td>6.2E-1</td>
</tr>
<tr>
<td>Q^T(\lambda)</td>
<td>1.1E0</td>
<td>1.9E0</td>
<td>1.4E1</td>
<td>7.6E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>6.7E-1</td>
<td>2.0E0</td>
<td>1.5E0</td>
<td>6.1E-1</td>
</tr>
<tr>
<td>IH</td>
<td>8.0E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 78: IPS, Standard Per-Decision

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.3E1</td>
<td>4.1E0</td>
</tr>
<tr>
<td>WIS</td>
<td>7.4E0</td>
<td><strong>2.7E0</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 79: 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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>Hybrid DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td><strong>1.0E-1</strong></td>
<td>6.5E0</td>
<td>4.4E0</td>
<td>3.5E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.6E0</td>
<td>1.4E0</td>
<td>1.5E0</td>
<td>2.1E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.0E0</td>
<td>1.8E0</td>
<td>3.0E0</td>
<td>2.4E0</td>
</tr>
<tr>
<td>FQE</td>
<td>1.2E-1</td>
<td>2.2E0</td>
<td>1.6E0</td>
<td><strong>1.2E-1</strong></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>2.0E-1</td>
<td>2.1E0</td>
<td>1.5E0</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>Q^T(\lambda)</td>
<td>7.9E-1</td>
<td>2.2E0</td>
<td>1.6E0</td>
<td>4.2E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>1.8E-1</td>
<td>2.1E0</td>
<td>1.6E0</td>
<td>1.4E-1</td>
</tr>
<tr>
<td>IH</td>
<td>1.7E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 80: IPS, Standard Per-Decision

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.4E1</td>
<td>4.0E0</td>
</tr>
<tr>
<td>WIS</td>
<td>1.2E1</td>
<td><strong>2.9E0</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.6E0</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 79: Graph, relative MSE. $T = 16, N = 8, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$. Dense rewards.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>8.5E-1</td>
<td>8.7E-1</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>6.8E-1</td>
<td>9.0E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>7.2E-1</td>
<td>9.8E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>8.5E-1</td>
<td>8.5E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>8.5E-1</td>
<td>8.4E-1</td>
</tr>
<tr>
<td>Q^\phi(\lambda)</td>
<td>8.5E-1</td>
<td>8.5E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>8.5E-1</td>
<td>8.3E-1</td>
</tr>
<tr>
<td>IH</td>
<td>7.5E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>6.6E-1</td>
<td>7.4E-1</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>4.4E-1</td>
<td>9.4E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.3E-1</td>
<td>8.7E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>6.5E-1</td>
<td>6.5E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>6.6E-1</td>
<td>6.5E-1</td>
</tr>
<tr>
<td>Q^\phi(\lambda)</td>
<td>6.5E-1</td>
<td>6.5E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>6.7E-1</td>
<td>6.5E-1</td>
</tr>
<tr>
<td>IH</td>
<td>1.1E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>5.7E-1</td>
<td>5.2E-1</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>5.9E-1</td>
<td>5.0E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.9E-1</td>
<td>8.3E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>5.4E-1</td>
<td>5.4E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>6.1E-1</td>
<td>6.0E-1</td>
</tr>
<tr>
<td>Q^\phi(\lambda)</td>
<td>5.5E-1</td>
<td>5.4E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>6.4E-1</td>
<td>6.1E-1</td>
</tr>
<tr>
<td>IH</td>
<td>1.7E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.7E-1</td>
<td>2.5E-1</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>4.9E-1</td>
<td>9.2E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.7E-1</td>
<td>7.4E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>1.7E-1</td>
<td>1.7E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>3.4E-1</td>
<td>3.3E-1</td>
</tr>
<tr>
<td>Q^\phi(\lambda)</td>
<td>1.7E-1</td>
<td>1.7E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>4.0E-1</td>
<td>3.6E-1</td>
</tr>
<tr>
<td>IH</td>
<td>5.9E-3</td>
<td>-</td>
</tr>
</tbody>
</table>

 IPS

<table>
<thead>
<tr>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E0</td>
</tr>
<tr>
<td>WIS</td>
<td>2.6E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.2E0</td>
</tr>
</tbody>
</table>

 IPS

<table>
<thead>
<tr>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E0</td>
</tr>
<tr>
<td>WIS</td>
<td>1.7E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.1E0</td>
</tr>
</tbody>
</table>
### Table 83: Graph, relative MSE. $T = 16, N = 128, \pi_b(a = 0) = 0.2, \pi_c(a = 0) = 0.8$. Dense rewards.

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.1E-2</td>
<td>1.5E-1</td>
<td>2.5E-1</td>
<td>4.0E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.0E-1</td>
<td>7.1E0</td>
<td>1.7E0</td>
<td>3.8E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.5E-1</td>
<td>5.6E-1</td>
<td>6.0E0</td>
<td>6.0E0</td>
</tr>
<tr>
<td>FQE</td>
<td>2.0E-2</td>
<td>2.0E-2</td>
<td>2.0E-2</td>
<td>2.0E-2</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>2.2E-1</td>
<td>1.9E-1</td>
<td>1.0E-1</td>
<td>2.3E-1</td>
</tr>
<tr>
<td>$Q^*(\lambda)$</td>
<td>2.0E-2</td>
<td>2.0E-2</td>
<td>2.0E-2</td>
<td>2.0E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>3.0E-1</td>
<td>2.6E-1</td>
<td>1.3E-1</td>
<td>3.0E-1</td>
</tr>
<tr>
<td>IH</td>
<td>3.2E-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.3E-3</td>
<td>1.7E1</td>
<td>2.5E-1</td>
<td>5.5E-4</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.5E1</td>
<td>4.0E2</td>
<td>2.3E1</td>
<td>9.9E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.9E-1</td>
<td>1.2E3</td>
<td>1.9E1</td>
<td>2.2E1</td>
</tr>
<tr>
<td>FQE</td>
<td>9.9E-8</td>
<td>9.9E-8</td>
<td>9.9E-8</td>
<td>9.9E-8</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>9.7E-2</td>
<td>2.7E0</td>
<td>7.1E-2</td>
<td>1.2E-1</td>
</tr>
<tr>
<td>$Q^*(\lambda)$</td>
<td>1.0E-7</td>
<td>1.1E-7</td>
<td>9.9E-8</td>
<td>1.0E-7</td>
</tr>
<tr>
<td>TREE</td>
<td>1.8E-1</td>
<td>1.2E1</td>
<td>9.7E-2</td>
<td>1.9E-1</td>
</tr>
<tr>
<td>IH</td>
<td>6.9E-4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.2E-4</td>
<td>5.9E-1</td>
<td>1.8E-1</td>
<td>6.2E-4</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.1E0</td>
<td>3.4E1</td>
<td>1.8E1</td>
<td>3.3E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.6E0</td>
<td>1.4E2</td>
<td>4.1E1</td>
<td>1.0E1</td>
</tr>
<tr>
<td>FQE</td>
<td>3.6E-7</td>
<td>3.6E-7</td>
<td>3.6E-7</td>
<td>3.6E-7</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.0E-1</td>
<td>8.3E-1</td>
<td>8.7E-2</td>
<td>1.3E-1</td>
</tr>
<tr>
<td>$Q^*(\lambda)$</td>
<td>3.6E-7</td>
<td>3.6E-7</td>
<td>3.6E-7</td>
<td>3.6E-7</td>
</tr>
<tr>
<td>TREE</td>
<td>1.8E-1</td>
<td>2.2E0</td>
<td>1.3E-1</td>
<td>2.1E-1</td>
</tr>
<tr>
<td>IH</td>
<td>5.5E-4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>5.5E-5</td>
<td>1.2E-1</td>
<td>5.6E-2</td>
<td>3.8E-4</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.8E-1</td>
<td>1.4E0</td>
<td>3.0E-1</td>
<td>2.4E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.7E-1</td>
<td>6.0E-1</td>
<td>3.9E0</td>
<td>3.9E0</td>
</tr>
<tr>
<td>FQE</td>
<td>1.0E-6</td>
<td>1.0E-6</td>
<td>1.0E-6</td>
<td>1.0E-6</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>9.3E-2</td>
<td>7.5E-2</td>
<td>5.8E-2</td>
<td>9.3E-2</td>
</tr>
<tr>
<td>$Q^*(\lambda)$</td>
<td>1.0E-6</td>
<td>1.0E-6</td>
<td>1.0E-6</td>
<td>1.0E-6</td>
</tr>
<tr>
<td>TREE</td>
<td>1.7E-1</td>
<td>1.2E-1</td>
<td>7.6E-2</td>
<td>1.7E-1</td>
</tr>
<tr>
<td>IH</td>
<td>8.5E-4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 87: Graph, relative MSE. $T = 16, N = 1024, \pi_b(a = 0) = 0.2, \pi_c(a = 0) = 0.8$. Dense rewards.

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>2.2E1</td>
<td>2.8E1</td>
</tr>
<tr>
<td>WIS</td>
<td>7.0E-1</td>
<td>2.2E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>9.3E-1</td>
<td>2.7E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>6.9E-1</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM Direct</th>
<th>Hybrid DR</th>
<th>Hybrid WDR</th>
<th>Hybrid MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>7.8E-1</td>
<td>2.0E0</td>
<td>1.2E0</td>
<td>7.8E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>9.5E-1</td>
<td>2.5E1</td>
<td>2.2E1</td>
<td>1.1E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.1E0</td>
<td>6.5E1</td>
<td>2.5E1</td>
<td>2.6E0</td>
</tr>
<tr>
<td>FQE</td>
<td>7.6E-1</td>
<td>7.1E-1</td>
<td>9.5E-1</td>
<td>7.6E-1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>7.7E-1</td>
<td>7.6E-1</td>
<td>1.1E0</td>
<td>8.7E-1</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>7.6E-1</td>
<td>7.2E-1</td>
<td>7.8E-1</td>
<td>7.6E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>7.7E-1</td>
<td>7.0E-1</td>
<td>1.2E0</td>
<td>9.0E-1</td>
</tr>
<tr>
<td>IH</td>
<td>2.9E-1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM Direct</th>
<th>Hybrid DR</th>
<th>Hybrid WDR</th>
<th>Hybrid MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>5.7E-1</td>
<td>7.8E-1</td>
<td>1.3E0</td>
<td>5.5E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>6.9E-1</td>
<td>1.2E0</td>
<td>4.2E0</td>
<td>1.6E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.5E0</td>
<td>2.5E0</td>
<td>8.6E0</td>
<td>8.1E0</td>
</tr>
<tr>
<td>FQE</td>
<td>5.8E-1</td>
<td>7.1E-1</td>
<td>5.7E-1</td>
<td>5.8E-1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>6.6E-1</td>
<td>6.7E-1</td>
<td>1.0E0</td>
<td>1.0E0</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>6.7E-1</td>
<td>1.3E0</td>
<td>5.8E-1</td>
<td>6.7E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>6.8E-1</td>
<td>6.8E-1</td>
<td>1.1E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>IH</td>
<td>6.5E-2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>9.7E-1</td>
<td>4.8E0</td>
</tr>
<tr>
<td>WIS</td>
<td>1.9E0</td>
<td>1.4E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.7E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E0</td>
<td>6.9E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>2.3E0</td>
<td>1.2E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.1E0</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>9.9E-1</td>
<td>6.0E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>1.5E0</td>
<td>9.9E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.8E0</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E0</td>
<td>5.4E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>1.1E0</td>
<td>3.4E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.9E0</td>
<td>-</td>
</tr>
</tbody>
</table>
### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>6.4E-2</td>
<td>6.1E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.3E-1</td>
<td>3.1E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.7E-1</td>
<td>5.8E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>6.5E-2</td>
<td>8.3E-2</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.9E-1</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>Q^p(\lambda)</td>
<td>5.8E-2</td>
<td>5.1E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>2.6E-1</td>
<td>2.2E-1</td>
</tr>
<tr>
<td>IPS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td>9.8E-1</td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>1.3E0</td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
<td></td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>5.4E-3</td>
<td>3.2E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.1E-1</td>
<td>8.0E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.7E-1</td>
<td>5.0E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.0E-3</td>
<td>4.3E-2</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.1E-1</td>
<td>1.7E-1</td>
</tr>
<tr>
<td>Q^p(\lambda)</td>
<td>1.4E-2</td>
<td>6.2E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>1.9E-1</td>
<td>2.5E-1</td>
</tr>
<tr>
<td>IPS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td>1.0E0</td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>1.0E0</td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
<td></td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>6.0E-2</td>
<td>7.0E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.5E-1</td>
<td>3.2E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.6E-1</td>
<td>6.4E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>8.4E-3</td>
<td>9.4E-2</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.4E-1</td>
<td>1.9E-1</td>
</tr>
<tr>
<td>Q^p(\lambda)</td>
<td>5.7E-2</td>
<td>1.4E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>2.3E-1</td>
<td>2.9E-1</td>
</tr>
<tr>
<td>IPS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td>9.7E-1</td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>1.1E0</td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.9E0</td>
<td></td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.3E-3</td>
<td>1.5E1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.5E0</td>
<td>9.5E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.7E-1</td>
<td>5.0E1</td>
</tr>
<tr>
<td>FQE</td>
<td>8.6E-4</td>
<td>3.0E0</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.1E-1</td>
<td>5.8E-1</td>
</tr>
<tr>
<td>Q^p(\lambda)</td>
<td>7.7E-3</td>
<td>3.1E0</td>
</tr>
<tr>
<td>TREE</td>
<td>2.0E-1</td>
<td>1.7E0</td>
</tr>
<tr>
<td>IPS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td>1.4E2</td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>8.0E-1</td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
<td></td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>7.5E-1</td>
<td>1.0E0</td>
<td>1.2E0</td>
<td>7.1E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>7.6E-1</td>
<td>8.3E-1</td>
<td>2.5E0</td>
<td>7.3E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>7.9E-1</td>
<td>1.6E0</td>
<td>3.0E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>FQE</td>
<td>7.3E-1</td>
<td>6.8E-1</td>
<td>6.6E-1</td>
<td>7.3E-1</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>7.2E-1</td>
<td>6.9E-1</td>
<td>1.1E0</td>
<td>8.7E-1</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>6.2E-1</td>
<td>8.5E-1</td>
<td>8.7E-1</td>
<td>6.2E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>7.4E-1</td>
<td>6.9E-1</td>
<td>1.3E0</td>
<td>9.6E-1</td>
</tr>
<tr>
<td>IH</td>
<td><strong>2.3E-1</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td><strong>1.0E0</strong></td>
<td>1.1E0</td>
</tr>
<tr>
<td>WIS</td>
<td>2.7E0</td>
<td>1.9E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.2E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.3E-1</td>
<td>2.3E0</td>
<td>1.3E0</td>
<td>3.3E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>7.6E-1</td>
<td>5.0E-1</td>
<td>5.9E0</td>
<td>8.5E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.9E-1</td>
<td>7.8E-1</td>
<td>2.2E1</td>
<td>1.2E0</td>
</tr>
<tr>
<td>FQE</td>
<td>1.9E-1</td>
<td>2.7E-1</td>
<td>2.7E-1</td>
<td>1.9E-1</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>3.6E-1</td>
<td>3.6E-1</td>
<td>5.7E-1</td>
<td>6.0E-1</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>4.4E-1</td>
<td>4.6E-1</td>
<td>3.2E-1</td>
<td>4.3E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>4.1E-1</td>
<td>4.5E-1</td>
<td>6.7E-1</td>
<td>6.8E-1</td>
</tr>
<tr>
<td>IH</td>
<td><strong>4.8E-2</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td><strong>1.0E0</strong></td>
<td><strong>6.9E-1</strong></td>
</tr>
<tr>
<td>WIS</td>
<td>2.4E0</td>
<td>8.1E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.2E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>5.3E-1</td>
<td>3.9E-1</td>
<td>1.8E0</td>
<td>5.4E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>5.4E-1</td>
<td>4.1E-1</td>
<td>7.6E0</td>
<td>8.5E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.0E-1</td>
<td>8.0E-1</td>
<td>9.5E0</td>
<td>8.7E0</td>
</tr>
<tr>
<td>FQE</td>
<td>5.1E-1</td>
<td>4.7E-1</td>
<td>7.3E-1</td>
<td>5.1E-1</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>5.0E-1</td>
<td>5.1E-1</td>
<td>1.4E0</td>
<td>7.6E-1</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>3.9E-1</td>
<td>3.3E-1</td>
<td><strong>3.1E-1</strong></td>
<td>3.9E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>5.4E-1</td>
<td>5.6E-1</td>
<td>1.7E0</td>
<td>1.0E0</td>
</tr>
<tr>
<td>IH</td>
<td><strong>1.6E-1</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E0</td>
<td><strong>5.3E-1</strong></td>
</tr>
<tr>
<td>WIS</td>
<td>2.3E0</td>
<td>2.0E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.1E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>5.3E-2</td>
<td>2.3E1</td>
<td>1.6E0</td>
<td>6.2E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.0E0</td>
<td>2.3E0</td>
<td>3.7E0</td>
<td>3.9E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.1E0</td>
<td>2.5E1</td>
<td>1.1E1</td>
<td>1.1E1</td>
</tr>
<tr>
<td>FQE</td>
<td>5.1E-2</td>
<td>1.8E0</td>
<td>3.2E-1</td>
<td>5.1E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>2.0E-1</td>
<td>1.7E-1</td>
<td>2.7E-1</td>
<td>3.2E-1</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>3.2E-1</td>
<td>1.9E0</td>
<td>7.7E-1</td>
<td>3.4E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>2.5E-1</td>
<td>7.3E-1</td>
<td>2.5E-1</td>
<td>3.2E-1</td>
</tr>
<tr>
<td>IH</td>
<td><strong>1.3E-2</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.1E0</td>
<td><strong>3.1E0</strong></td>
</tr>
<tr>
<td>WIS</td>
<td>1.9E0</td>
<td><strong>3.1E-1</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.2E0</td>
<td>-</td>
</tr>
</tbody>
</table>
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          |                  |
| **DM**          | **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             | **1.8E-2**      | 1.4E-1          | 7.4E-2          | **1.7E-2**      |
| $R(\lambda)$   | 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             |                  |                  |
| **IS**          | **9.3E-1**       |                  |
| **WIS**         | **8.3E-1**       | 2.9E-1           |
| **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          |                  |
| **DM**          | **DR**          | **WDR**         | **MAGIC**       |
| AM              | 1.1E-2          | 3.9E0           | 1.1E0           | **1.0E-2**      |
| 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              | **6.8E-3**      | -               | -               | -               |

| IPS             |                  |                  |
| **IS**          | 9.3E-1           |                  |
| **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          |                  |
| **DM**          | **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             | **4.9E-3**      | 5.5E-2          | 1.3E-1          | **4.8E-3**      |
| $R(\lambda)$   | 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             |                  |                  |
| **IS**          | 9.9E-1           |                  |
| **WIS**         | 1.3E0           | 3.1E-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          |                  |
| **DM**          | **DR**          | **WDR**         | **MAGIC**       |
| AM              | **1.8E-3**      | 9.5E-1          | 6.8E-1          | **3.8E-3**      |
| 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             |                  |                  |
| **IS**          | 9.3E-1           |                  |
| **WIS**         | 6.4E-1           | 5.4E-2           |
| **NAIVE**       | 4.0E0           |                  |
Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>WDR</td>
</tr>
<tr>
<td>AM</td>
<td>9.2E-1</td>
<td>1.4E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.3E0</td>
<td>1.8E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.3E0</td>
<td>8.1E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>9.3E-1</td>
<td>8.0E-1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>8.6E-1</td>
<td>7.7E-1</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>8.6E-1</td>
<td>7.8E-1</td>
</tr>
<tr>
<td>IH</td>
<td>1.2E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>WDR</td>
</tr>
<tr>
<td>AM</td>
<td>7.4E-1</td>
<td>1.5E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>8.6E-1</td>
<td>2.4E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.6E-1</td>
<td>1.4E0</td>
</tr>
<tr>
<td>FQE</td>
<td>7.7E-1</td>
<td>9.1E-1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>8.8E-1</td>
<td>8.8E-1</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>1.4E0</td>
<td>1.4E0</td>
</tr>
<tr>
<td>Tree</td>
<td>9.2E-1</td>
<td>9.2E-1</td>
</tr>
<tr>
<td>IH</td>
<td>4.3E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>WDR</td>
</tr>
<tr>
<td>AM</td>
<td>2.2E1</td>
<td>4.2E1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.4E0</td>
<td>6.8E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.6E0</td>
<td>1.3E2</td>
</tr>
<tr>
<td>FQE</td>
<td>1.9E-1</td>
<td>2.2E-1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>4.0E-1</td>
<td>1.3E0</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>1.8E0</td>
<td>3.1E0</td>
</tr>
<tr>
<td>Tree</td>
<td>4.4E-1</td>
<td>3.1E0</td>
</tr>
<tr>
<td>IH</td>
<td>2.1E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>WDR</td>
</tr>
<tr>
<td>AM</td>
<td>6.6E-2</td>
<td>2.6E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>7.1E-1</td>
<td>4.0E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>8.1E-1</td>
<td>1.4E0</td>
</tr>
<tr>
<td>FQE</td>
<td>5.5E-2</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.4E-1</td>
<td>2.5E-1</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>1.2E0</td>
<td>9.1E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>2.0E-1</td>
<td>3.5E-1</td>
</tr>
<tr>
<td>IH</td>
<td>6.2E-2</td>
<td>-</td>
</tr>
</tbody>
</table>
### 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.

<table>
<thead>
<tr>
<th>DM</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>5.7E-2</td>
<td>3.1E1</td>
<td>3.4E0</td>
<td>5.4E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.9E0</td>
<td>1.1E1</td>
<td>4.9E0</td>
<td>1.8E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.0E1</td>
<td>2.3E1</td>
<td>1.6E1</td>
<td>1.8E1</td>
</tr>
<tr>
<td>FQE</td>
<td>5.1E-2</td>
<td>7.0E0</td>
<td>4.8E-1</td>
<td>5.1E-2</td>
</tr>
<tr>
<td>( R(\lambda) )</td>
<td>2.3E-1</td>
<td>1.3E0</td>
<td>5.9E-1</td>
<td>4.9E-1</td>
</tr>
<tr>
<td>( \text{Q}^\pi(\lambda) )</td>
<td>2.0E-1</td>
<td>1.7E0</td>
<td>5.0E-1</td>
<td>2.1E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>3.0E-1</td>
<td>2.9E0</td>
<td>6.6E-1</td>
<td>6.9E-1</td>
</tr>
<tr>
<td>IH</td>
<td>4.0E-2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th>IPS</th>
<th>STANDARD</th>
<th>PER-DECISION</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.7E0</td>
<td>3.0E0</td>
</tr>
<tr>
<td>WIS</td>
<td>2.2E0</td>
<td>7.9E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.1E0</td>
<td>-</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th>DM</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>6.3E-3</td>
<td>2.6E0</td>
<td>1.4E0</td>
<td>1.1E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.0E0</td>
<td>9.8E0</td>
<td>1.5E1</td>
<td>5.1E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.1E1</td>
<td>3.2E1</td>
<td>3.1E2</td>
<td>3.0E2</td>
</tr>
<tr>
<td>FQE</td>
<td>7.9E-3</td>
<td>1.5E0</td>
<td>2.4E-1</td>
<td>7.9E-3</td>
</tr>
<tr>
<td>( R(\lambda) )</td>
<td>1.5E-1</td>
<td>7.9E-1</td>
<td>2.8E-1</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>( \text{Q}^\pi(\lambda) )</td>
<td>1.0E-1</td>
<td>3.1E0</td>
<td>2.9E-1</td>
<td>9.3E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>2.4E-1</td>
<td>1.1E0</td>
<td>3.5E-1</td>
<td>2.4E-1</td>
</tr>
<tr>
<td>IH</td>
<td>8.1E-3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 109: 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.

<table>
<thead>
<tr>
<th>IPS</th>
<th>STANDARD</th>
<th>PER-DECISION</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.6E0</td>
<td>3.5E0</td>
</tr>
<tr>
<td>WIS</td>
<td>1.5E0</td>
<td>4.0E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th>DM</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>5.3E-3</td>
<td>1.3E2</td>
<td>1.3E0</td>
<td>9.0E-3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>8.1E0</td>
<td>2.6E0</td>
<td>1.4E1</td>
<td>4.0E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.7E0</td>
<td>1.2E2</td>
<td>3.2E1</td>
<td>8.8E0</td>
</tr>
<tr>
<td>FQE</td>
<td>6.4E-3</td>
<td>1.3E1</td>
<td>2.4E-1</td>
<td>6.5E-3</td>
</tr>
<tr>
<td>( R(\lambda) )</td>
<td>2.7E-1</td>
<td>3.0E0</td>
<td>2.5E-1</td>
<td>2.7E-1</td>
</tr>
<tr>
<td>( \text{Q}^\pi(\lambda) )</td>
<td>3.3E-2</td>
<td>2.0E1</td>
<td>2.4E-1</td>
<td>3.4E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>3.2E-1</td>
<td>4.9E0</td>
<td>2.9E-1</td>
<td>4.1E-1</td>
</tr>
<tr>
<td>IH</td>
<td>5.8E-3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 111: Graph, relative MSE. $T = 16, N = 8, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$. Sparse rewards.

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.0E0</td>
<td>9.1E-1</td>
<td>9.1E-1</td>
<td>1.0E0</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>9.7E-1</td>
<td>1.2E1</td>
<td>1.2E1</td>
<td>1.0E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.7E-1</td>
<td>1.3E0</td>
<td>1.2E1</td>
<td>1.0E0</td>
</tr>
<tr>
<td>FQE</td>
<td>1.0E0</td>
<td>1.0E0</td>
<td>1.0E0</td>
<td>1.0E0</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.0E0</td>
<td>9.7E-1</td>
<td>3.4E0</td>
<td>1.0E0</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>1.0E0</td>
<td>1.0E0</td>
<td>1.0E0</td>
<td>1.0E0</td>
</tr>
<tr>
<td>Tree</td>
<td>1.0E0</td>
<td>9.7E-1</td>
<td>4.3E0</td>
<td>3.4E0</td>
</tr>
<tr>
<td>IH</td>
<td>1.2E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.0E0</td>
<td>2.1E0</td>
<td>2.6E0</td>
<td>1.2E0</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>9.9E-1</td>
<td>1.0E0</td>
<td>1.2E1</td>
<td>1.1E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.8E-1</td>
<td>1.0E0</td>
<td>1.3E1</td>
<td>5.7E0</td>
</tr>
<tr>
<td>FQE</td>
<td>9.8E-1</td>
<td>9.8E-1</td>
<td>9.8E-1</td>
<td>9.8E-1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.0E0</td>
<td>9.9E-1</td>
<td>1.1E0</td>
<td>1.5E0</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>9.8E-1</td>
<td>9.8E-1</td>
<td>9.8E-1</td>
<td>9.8E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>1.0E0</td>
<td>9.9E-1</td>
<td>1.1E0</td>
<td>1.5E0</td>
</tr>
<tr>
<td>IH</td>
<td>9.6E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.0E0</td>
<td>2.1E0</td>
<td>2.6E0</td>
<td>1.2E0</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>9.9E-1</td>
<td>1.0E0</td>
<td>1.2E1</td>
<td>1.1E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.8E-1</td>
<td>1.0E0</td>
<td>1.3E1</td>
<td>5.7E0</td>
</tr>
<tr>
<td>FQE</td>
<td>9.8E-1</td>
<td>9.8E-1</td>
<td>9.8E-1</td>
<td>9.8E-1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.0E0</td>
<td>9.9E-1</td>
<td>1.1E0</td>
<td>1.5E0</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>9.8E-1</td>
<td>9.8E-1</td>
<td>9.8E-1</td>
<td>9.8E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>1.0E0</td>
<td>9.9E-1</td>
<td>1.1E0</td>
<td>1.5E0</td>
</tr>
<tr>
<td>IH</td>
<td>9.6E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>5.3E-1</td>
<td>2.2E0</td>
<td>9.4E0</td>
<td>1.2E0</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.0E0</td>
<td>9.6E-1</td>
<td>7.3E0</td>
<td>1.0E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.9E-1</td>
<td>9.7E-1</td>
<td>5.2E1</td>
<td>3.5E1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.9E-1</td>
<td>4.9E-1</td>
<td>4.9E-1</td>
<td>4.9E-1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.0E0</td>
<td>1.0E0</td>
<td>1.8E0</td>
<td>1.8E0</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>4.9E-1</td>
<td>4.9E-1</td>
<td>4.9E-1</td>
<td>4.9E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>1.0E0</td>
<td>1.0E0</td>
<td>1.8E0</td>
<td>1.8E0</td>
</tr>
<tr>
<td>IH</td>
<td>8.4E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 115: Graph, relative MSE. $T = 16, N = 128, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$. Sparse rewards.

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

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

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

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

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

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

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

Table 119: Graph, relative MSE. $T = 16, N = 2048, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$. Sparse rewards.

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

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>6.7E0</td>
<td>1.3E1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>7.7E2</td>
<td>6.4E3</td>
</tr>
<tr>
<td>MRDR</td>
<td>8.2E2</td>
<td>2.8E4</td>
</tr>
<tr>
<td>FQE</td>
<td>6.8E0</td>
<td>1.1E1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>6.9E0</td>
<td>3.2E2</td>
</tr>
<tr>
<td>Q^T(\lambda)</td>
<td>5.8E0</td>
<td>4.2E1</td>
</tr>
<tr>
<td>Tree</td>
<td>6.9E0</td>
<td>7.1E2</td>
</tr>
<tr>
<td>IH</td>
<td>9.9E1</td>
<td></td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>5.7E0</td>
<td>9.4E1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>7.4E1</td>
<td>5.8E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>6.2E1</td>
<td>3.2E2</td>
</tr>
<tr>
<td>FQE</td>
<td>5.1E0</td>
<td>1.5E1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>5.7E0</td>
<td>8.7E0</td>
</tr>
<tr>
<td>Q^T(\lambda)</td>
<td>1.9E1</td>
<td>3.1E2</td>
</tr>
<tr>
<td>Tree</td>
<td>5.5E0</td>
<td>9.9E0</td>
</tr>
<tr>
<td>IH</td>
<td>2.6E1</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>5.7E0</td>
<td>8.7E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>7.4E1</td>
<td>5.8E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>6.2E1</td>
<td>3.2E2</td>
</tr>
<tr>
<td>FQE</td>
<td>5.1E0</td>
<td>1.5E1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>5.7E0</td>
<td>8.7E0</td>
</tr>
<tr>
<td>Q^T(\lambda)</td>
<td>1.9E1</td>
<td>3.1E2</td>
</tr>
<tr>
<td>Tree</td>
<td>5.5E0</td>
<td>9.9E0</td>
</tr>
<tr>
<td>IH</td>
<td>2.6E1</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>4.6E0</td>
<td>4.4E3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.1E2</td>
<td>1.4E3</td>
</tr>
<tr>
<td>MRDR</td>
<td>7.1E1</td>
<td>1.2E3</td>
</tr>
<tr>
<td>FQE</td>
<td>3.4E0</td>
<td>8.8E1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>5.1E0</td>
<td>1.9E1</td>
</tr>
<tr>
<td>Q^T(\lambda)</td>
<td>4.5E1</td>
<td>8.2E1</td>
</tr>
<tr>
<td>Tree</td>
<td>4.2E0</td>
<td>2.3E1</td>
</tr>
<tr>
<td>IH</td>
<td>1.3E1</td>
<td></td>
</tr>
</tbody>
</table>

### Table 123: Graph, relative MSE. $T = 16, N = 64, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$. Stochastic rewards.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>4.6E0</td>
<td>4.4E3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.1E2</td>
<td>1.4E3</td>
</tr>
<tr>
<td>MRDR</td>
<td>7.1E1</td>
<td>1.2E3</td>
</tr>
<tr>
<td>FQE</td>
<td>3.4E0</td>
<td>8.8E1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>5.1E0</td>
<td>1.9E1</td>
</tr>
<tr>
<td>Q^T(\lambda)</td>
<td>4.5E1</td>
<td>8.2E1</td>
</tr>
<tr>
<td>Tree</td>
<td>4.2E0</td>
<td>2.3E1</td>
</tr>
<tr>
<td>IH</td>
<td>1.3E1</td>
<td></td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>5.7E0</td>
<td>9.8E1</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.4E1</td>
<td>4.0E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.9E1</td>
<td>4.2E0</td>
</tr>
<tr>
<td>FQE</td>
<td>4.6E0</td>
<td>6.3E0</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>7.4E0</td>
<td>8.2E0</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>4.7E1</td>
<td>4.5E0</td>
</tr>
<tr>
<td>Tree</td>
<td>5.7E0</td>
<td>6.5E0</td>
</tr>
<tr>
<td>IH</td>
<td>5.2E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.5E0</td>
<td>1.3E5</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.4E4</td>
<td>1.8E5</td>
</tr>
<tr>
<td>MRDR</td>
<td>8.9E3</td>
<td>3.8E5</td>
</tr>
<tr>
<td>FQE</td>
<td>1.1E0</td>
<td>2.4E4</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>5.1E0</td>
<td>8.1E0</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>6.5E0</td>
<td>2.2E4</td>
</tr>
<tr>
<td>Tree</td>
<td>4.4E0</td>
<td>8.7E0</td>
</tr>
<tr>
<td>IH</td>
<td>8.1E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.0E0</td>
<td>6.9E0</td>
</tr>
<tr>
<td>WIS</td>
<td>4.7E1</td>
<td>3.1E1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>5.7E0</td>
<td>1.7E4</td>
</tr>
<tr>
<td>WIS</td>
<td>8.1E1</td>
<td>1.7E1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.7E0</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td>1.6E0</td>
<td>5.3E1</td>
</tr>
<tr>
<td>WIS</td>
<td>6.7E1</td>
<td>9.9E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td>7.9E-1</td>
<td>8.5E0</td>
</tr>
<tr>
<td>WIS</td>
<td>1.9E1</td>
<td>9.9E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.7E0</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.0E0</td>
<td>9.9E-1</td>
<td>8.5E0</td>
<td>1.0E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>9.9E-1</td>
<td>1.3E0</td>
<td>1.8E3</td>
<td>4.0E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.9E-1</td>
<td>1.3E0</td>
<td>1.4E3</td>
<td>1.3E1</td>
</tr>
<tr>
<td>FQE</td>
<td>1.0E0</td>
<td>1.0E0</td>
<td>2.5E0</td>
<td>1.0E0</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.0E0</td>
<td>9.9E-1</td>
<td>6.2E0</td>
<td>1.0E0</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>1.1E0</td>
<td>2.2E0</td>
<td>9.9E-1</td>
<td>9.7E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>1.0E0</td>
<td>9.9E-1</td>
<td>6.2E0</td>
<td>1.0E0</td>
</tr>
<tr>
<td>IH</td>
<td><strong>8.0E-1</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**IPS**

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td><strong>9.9E-1</strong></td>
<td>9.9E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>6.2E0</td>
<td>6.2E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.5E0</td>
<td>1.2E2</td>
<td>4.0E1</td>
<td>1.5E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.8E0</td>
<td>9.1E0</td>
<td>8.0E1</td>
<td>1.8E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.8E0</td>
<td>3.7E1</td>
<td>2.1E2</td>
<td>2.0E2</td>
</tr>
<tr>
<td>FQE</td>
<td>1.3E0</td>
<td>1.6E0</td>
<td>6.4E0</td>
<td>1.3E0</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.0E0</td>
<td>2.8E0</td>
<td>1.1E1</td>
<td>3.7E0</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>2.8E0</td>
<td>2.6E3</td>
<td>2.6E0</td>
<td>2.7E0</td>
</tr>
<tr>
<td>Tree</td>
<td><strong>1.0E0</strong></td>
<td>2.8E0</td>
<td>1.1E1</td>
<td>3.7E0</td>
</tr>
<tr>
<td>IH</td>
<td>1.3E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**IPS**

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td><strong>2.8E0</strong></td>
<td>2.8E0</td>
</tr>
<tr>
<td>WIS</td>
<td>1.1E1</td>
<td>1.1E1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.2E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.8E0</td>
<td>4.9E0</td>
<td>5.5E1</td>
<td>1.7E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>9.8E-1</td>
<td><strong>9.0E-1</strong></td>
<td>6.9E1</td>
<td>9.8E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td><strong>9.6E-1</strong></td>
<td>1.1E0</td>
<td>1.5E2</td>
<td>1.5E2</td>
</tr>
<tr>
<td>FQE</td>
<td>1.5E0</td>
<td>1.4E0</td>
<td>5.5E0</td>
<td>1.5E0</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.0E0</td>
<td>9.8E-1</td>
<td>6.4E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>8.5E0</td>
<td>1.1E1</td>
<td>9.8E0</td>
<td>7.0E0</td>
</tr>
<tr>
<td>Tree</td>
<td>1.0E0</td>
<td>9.8E-1</td>
<td>6.4E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>IH</td>
<td>1.0E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**IPS**

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>9.8E-1</td>
<td>9.8E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>6.4E0</td>
<td>6.4E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.5E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.3E0</td>
<td>8.9E2</td>
<td>8.3E1</td>
<td>1.8E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.0E0</td>
<td>1.0E1</td>
<td>8.1E4</td>
<td>3.3E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.0E0</td>
<td>1.3E1</td>
<td>4.2E5</td>
<td>4.2E5</td>
</tr>
<tr>
<td>FQE</td>
<td>2.1E0</td>
<td>2.1E0</td>
<td>6.4E0</td>
<td>2.2E0</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.0E0</td>
<td><strong>1.0E0</strong></td>
<td>1.1E1</td>
<td>2.5E0</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>1.0E1</td>
<td>1.3E1</td>
<td>6.3E0</td>
<td>9.2E0</td>
</tr>
<tr>
<td>Tree</td>
<td><strong>1.0E0</strong></td>
<td>1.0E0</td>
<td>1.1E1</td>
<td>2.5E0</td>
</tr>
<tr>
<td>IH</td>
<td>1.0E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**IPS**

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td><strong>1.0E0</strong></td>
<td>1.0E0</td>
</tr>
<tr>
<td>WIS</td>
<td>1.1E1</td>
<td>1.1E1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.3E0</td>
<td>-</td>
</tr>
</tbody>
</table>
### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>4.4E-1</td>
<td>3.2E3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>5.0E0</td>
<td>4.9E2</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.2E1</td>
<td>4.6E1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.1E-1</td>
<td>6.4E-1</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.0E0</td>
<td>4.3E0</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>3.8E0</td>
<td>4.8E1</td>
</tr>
<tr>
<td>Tree</td>
<td>1.0E0</td>
<td>4.3E0</td>
</tr>
<tr>
<td>IH</td>
<td>6.8E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

**IPS**

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>4.3E0</td>
<td>4.3E0</td>
</tr>
<tr>
<td>WIS</td>
<td>9.0E0</td>
<td>9.0E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.9E0</td>
<td>-</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.1E-1</td>
<td>1.4E3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.2E0</td>
<td>6.1E2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.3E1</td>
<td>5.1E1</td>
</tr>
<tr>
<td>FQE</td>
<td>9.2E-2</td>
<td>2.1E0</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.0E0</td>
<td>1.3E0</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>1.3E0</td>
<td>3.1E1</td>
</tr>
<tr>
<td>Tree</td>
<td>1.0E0</td>
<td>1.3E0</td>
</tr>
<tr>
<td>IH</td>
<td>1.3E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

**IPS**

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.3E0</td>
<td>1.3E0</td>
</tr>
<tr>
<td>WIS</td>
<td>3.9E0</td>
<td>3.9E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.1E0</td>
<td>-</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>8.1E-2</td>
<td>2.1E1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.6E0</td>
<td>3.5E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.4E0</td>
<td>4.0E0</td>
</tr>
<tr>
<td>FQE</td>
<td>5.3E-2</td>
<td>9.1E0</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.0E0</td>
<td>1.7E0</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>1.1E0</td>
<td>1.0E1</td>
</tr>
<tr>
<td>Tree</td>
<td>1.0E0</td>
<td>1.7E0</td>
</tr>
<tr>
<td>IH</td>
<td>1.9E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

**IPS**

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.7E0</td>
<td>1.7E0</td>
</tr>
<tr>
<td>WIS</td>
<td>6.8E0</td>
<td>6.8E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.9E0</td>
<td>-</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>3.6E-2</td>
<td>5.0E1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.1E0</td>
<td>2.8E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>8.7E-1</td>
<td>1.7E0</td>
</tr>
<tr>
<td>FQE</td>
<td>3.4E-2</td>
<td>3.9E-1</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.0E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>6.6E-1</td>
<td>2.1E0</td>
</tr>
<tr>
<td>Tree</td>
<td>1.0E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>IH</td>
<td>9.3E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

**IPS**

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.1E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>WIS</td>
<td>4.9E0</td>
<td>4.9E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.6E0</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.6E1</td>
<td>8.4E3</td>
<td>3.2E2</td>
<td>1.7E1</td>
<td></td>
</tr>
<tr>
<td>Q-Reg</td>
<td>9.9E3</td>
<td>6.2E5</td>
<td>4.0E4</td>
<td>9.9E3</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>7.8E3</td>
<td>1.4E6</td>
<td>1.3E4</td>
<td>7.9E3</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>1.2E1</td>
<td>2.2E3</td>
<td>7.8E1</td>
<td>1.2E1</td>
<td></td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>2.2E1</td>
<td>9.1E3</td>
<td>7.9E1</td>
<td>2.5E1</td>
<td></td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>3.3E1</td>
<td>2.7E3</td>
<td>7.8E1</td>
<td>3.2E1</td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>2.0E1</td>
<td>1.3E4</td>
<td>8.8E1</td>
<td>2.0E1</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>3.2E2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>5.5E1</td>
<td>3.4E4</td>
<td>4.4E2</td>
<td>4.8E1</td>
<td></td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.4E4</td>
<td>9.1E4</td>
<td>8.4E4</td>
<td>1.4E4</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>1.0E5</td>
<td>3.1E5</td>
<td>4.0E5</td>
<td>1.5E5</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>5.2E1</td>
<td>2.0E4</td>
<td>1.5E2</td>
<td>5.2E1</td>
<td></td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>4.3E1</td>
<td>7.0E1</td>
<td>7.9E1</td>
<td>4.3E1</td>
<td></td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>1.5E2</td>
<td>1.5E4</td>
<td>1.6E2</td>
<td>1.5E2</td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>3.6E1</td>
<td>2.7E2</td>
<td>9.3E1</td>
<td>3.7E1</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>3.8E1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>8.9E2</td>
<td>1.8E4</td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>1.3E2</td>
<td>6.6E1</td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>1.8E1</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E0</td>
<td>2.1E4</td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>9.3E1</td>
<td>1.3E2</td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>1.8E1</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.3E1</td>
<td>8.9E2</td>
<td>8.5E2</td>
<td>1.4E1</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.6E1</td>
<td>3.8E4</td>
<td>9.3E3</td>
<td>2.4E1</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>5.7E1</td>
<td>4.3E2</td>
<td>7.7E2</td>
<td>5.6E2</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>1.1E1</td>
<td>6.0E1</td>
<td>1.1E2</td>
<td>1.1E1</td>
<td></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>9.9E0</td>
<td>3.2E1</td>
<td>9.8E1</td>
<td>1.2E1</td>
<td></td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>4.2E1</td>
<td>2.4E2</td>
<td>1.5E2</td>
<td>4.4E1</td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td><strong>7.8E0</strong></td>
<td>3.0E1</td>
<td>9.8E1</td>
<td><strong>1.0E1</strong></td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>1.6E1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.0E0</td>
<td>2.3E2</td>
<td>4.1E2</td>
<td>4.9E0</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.6E1</td>
<td>5.4E1</td>
<td>1.7E2</td>
<td>8.8E0</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>2.7E1</td>
<td>4.2E1</td>
<td>3.0E2</td>
<td>1.4E2</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>2.4E0</td>
<td>1.6E1</td>
<td>1.0E1</td>
<td>5.3E0</td>
<td></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>3.4E0</td>
<td>2.0E1</td>
<td>1.7E1</td>
<td><strong>3.2E0</strong></td>
<td></td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>1.3E1</td>
<td>2.9E1</td>
<td>6.4E0</td>
<td>1.2E1</td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>4.3E0</td>
<td>2.0E1</td>
<td>1.7E1</td>
<td>4.0E0</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td><strong>2.3E0</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.6E1</td>
<td>3.3E1</td>
<td>8.1E1</td>
<td>1.5E1</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>5.2E0</td>
<td>3.1E2</td>
<td>2.3E1</td>
<td>8.7E0</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>2.5E3</td>
<td>4.1E3</td>
<td>1.5E4</td>
<td>8.3E3</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td><strong>1.5E0</strong></td>
<td>5.0E2</td>
<td>1.6E1</td>
<td><strong>3.6E0</strong></td>
<td></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>3.4E0</td>
<td>3.1E2</td>
<td>2.3E1</td>
<td>8.6E0</td>
<td></td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>9.1E0</td>
<td>5.0E2</td>
<td>1.5E1</td>
<td>7.6E0</td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>2.1E0</td>
<td>3.7E2</td>
<td>2.1E1</td>
<td>6.5E0</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>1.5E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.5E0</td>
<td>1.0E3</td>
<td>1.6E2</td>
<td>4.7E0</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.2E2</td>
<td>9.7E3</td>
<td>2.3E3</td>
<td>3.2E2</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>6.6E3</td>
<td>2.3E3</td>
<td>1.4E4</td>
<td>8.8E3</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td><strong>1.5E0</strong></td>
<td>5.0E2</td>
<td>1.6E1</td>
<td><strong>3.6E0</strong></td>
<td></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>3.4E0</td>
<td>3.1E2</td>
<td>2.3E1</td>
<td>8.6E0</td>
<td></td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>9.1E0</td>
<td>5.0E2</td>
<td>1.5E1</td>
<td>7.6E0</td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>2.1E0</td>
<td>3.7E2</td>
<td>2.1E1</td>
<td>6.5E0</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>1.5E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

### Table 143: 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.

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td><strong>1.0E0</strong></td>
<td>3.1E1</td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>8.3E1</td>
<td>1.1E2</td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>7.8E0</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

### Table 144: 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.

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td><strong>1.2E0</strong></td>
<td>1.7E1</td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>4.2E1</td>
<td>1.6E1</td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.5E0</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
Table 143: Graph, relative MSE. $T = 4, N = 8, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$. Dense rewards.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>7.7E-2</td>
<td>4.9E-2</td>
<td>4.5E-2</td>
<td>4.1E-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-Reg</td>
<td>2.2E-1</td>
<td>9.6E-2</td>
<td>9.4E-2</td>
<td>1.8E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>3.6E-1</td>
<td>1.6E-1</td>
<td>1.6E-1</td>
<td>3.6E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>3.3E-2</td>
<td>3.3E-2</td>
<td>3.3E-2</td>
<td>3.3E-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>3.3E-2</td>
<td>3.3E-2</td>
<td>3.3E-2</td>
<td>3.3E-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q^\beta(\lambda)</td>
<td>3.3E-2</td>
<td>3.3E-2</td>
<td>3.3E-2</td>
<td>3.3E-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TREE</td>
<td>3.3E-2</td>
<td>3.4E-2</td>
<td>3.4E-2</td>
<td>3.3E-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>1.2E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.0E-2</td>
<td>4.0E-2</td>
<td>3.7E-2</td>
<td>2.7E-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-Reg</td>
<td>4.4E-2</td>
<td>5.5E-3</td>
<td>4.3E-3</td>
<td>7.1E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>7.9E-2</td>
<td>1.1E-2</td>
<td>7.6E-3</td>
<td>3.6E-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>4.2E-3</td>
<td>4.2E-3</td>
<td>4.2E-3</td>
<td>4.2E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>4.2E-3</td>
<td>4.2E-3</td>
<td>4.2E-3</td>
<td>4.2E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q^\beta(\lambda)</td>
<td>4.2E-3</td>
<td>4.2E-3</td>
<td>4.2E-3</td>
<td>4.2E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TREE</td>
<td>4.2E-3</td>
<td>4.2E-3</td>
<td>4.2E-3</td>
<td>4.2E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>2.5E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>7.0E-3</td>
<td>1.5E-2</td>
<td>1.4E-2</td>
<td>2.7E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.5E-2</td>
<td>1.2E-3</td>
<td>7.2E-4</td>
<td>4.2E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>6.7E-2</td>
<td>3.7E-3</td>
<td>2.0E-3</td>
<td>1.8E-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>4.2E-7</td>
<td>4.2E-7</td>
<td>4.2E-7</td>
<td>4.2E-7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>4.2E-7</td>
<td>4.2E-7</td>
<td>4.2E-7</td>
<td>4.2E-7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q^\beta(\lambda)</td>
<td>4.2E-7</td>
<td>4.2E-7</td>
<td>4.2E-7</td>
<td>4.2E-7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TREE</td>
<td>1.6E-6</td>
<td>5.3E-7</td>
<td>5.4E-7</td>
<td>1.6E-6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>1.1E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.1E-3</td>
<td>9.5E-3</td>
<td>9.6E-3</td>
<td>2.3E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-Reg</td>
<td>2.0E-2</td>
<td>4.2E-4</td>
<td>1.6E-4</td>
<td>1.3E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>3.2e-2</td>
<td>1.6E-3</td>
<td>8.6E-4</td>
<td>2.1E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>1.8E-5</td>
<td>1.8E-5</td>
<td>1.8E-5</td>
<td>1.8E-5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.8E-5</td>
<td>1.8E-5</td>
<td>1.8E-5</td>
<td>1.8E-5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q^\beta(\lambda)</td>
<td>1.8E-5</td>
<td>1.8E-5</td>
<td>1.8E-5</td>
<td>1.8E-5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TREE</td>
<td>1.3E-5</td>
<td>1.6E-5</td>
<td>1.6E-5</td>
<td>1.3E-5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>9.4E-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.8E-2</td>
<td>2.5E-2</td>
<td>2.5E-2</td>
<td>1.8E-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.4E-2</td>
<td>8.5E-3</td>
<td>8.5E-3</td>
<td>1.8E-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>3.9E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 148: Graph, relative MSE. $T = 4, N = 256, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$. Dense rewards.
### Table 147: Graph, relative MSE. $T = 4, N = 128, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$. Dense rewards.

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.3E-3</td>
<td>3.6E-3</td>
<td>3.8E-3</td>
<td>3.0E-3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>6.4E-3</td>
<td>3.1E-4</td>
<td>2.3E-4</td>
<td>1.4E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.3E-2</td>
<td>2.0E-3</td>
<td>1.2E-3</td>
<td>1.5E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>7.4E-8</td>
<td>7.4E-8</td>
<td>7.4E-8</td>
<td>7.4E-8</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td><strong>7.3E-8</strong></td>
<td>7.3E-8</td>
<td>7.2E-8</td>
<td>7.2E-8</td>
</tr>
<tr>
<td>Q$^\theta$(\lambda)</td>
<td>7.6E-8</td>
<td>7.5E-8</td>
<td>7.5E-8</td>
<td>7.5E-8</td>
</tr>
<tr>
<td>TREE</td>
<td>1.0E-7</td>
<td><strong>6.7E-8</strong></td>
<td>6.8E-8</td>
<td>9.1E-8</td>
</tr>
<tr>
<td>IH</td>
<td>3.7E-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**IPS**

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>2.2E-2</td>
<td>1.3E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>7.3E-3</td>
<td><strong>4.5E-3</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.7E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>7.4E-4</td>
<td>3.0E-3</td>
<td>3.0E-3</td>
<td>1.8E-3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.5E-3</td>
<td>7.6E-6</td>
<td>6.5E-6</td>
<td>2.5E-5</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.7E-2</td>
<td>1.7E-4</td>
<td>1.0E-4</td>
<td>2.7E-4</td>
</tr>
<tr>
<td>FQE</td>
<td>8.4E-7</td>
<td>8.4E-7</td>
<td>8.4E-7</td>
<td>8.4E-7</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>8.4E-7</td>
<td>8.4E-7</td>
<td>8.4E-7</td>
<td>8.4E-7</td>
</tr>
<tr>
<td>Q$^\theta$(\lambda)</td>
<td>8.4E-7</td>
<td>8.4E-7</td>
<td>8.4E-7</td>
<td>8.5E-7</td>
</tr>
<tr>
<td>TREE</td>
<td><strong>1.2E-7</strong></td>
<td>7.9E-7</td>
<td>8.0E-7</td>
<td><strong>1.3E-7</strong></td>
</tr>
<tr>
<td>IH</td>
<td>4.6E-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**IPS**

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.2E-2</td>
<td>9.2E-3</td>
</tr>
<tr>
<td>WIS</td>
<td><strong>2.9E-3</strong></td>
<td>3.3E-3</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.6E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.7E-4</td>
<td>2.5E-3</td>
<td>2.4E-3</td>
<td>1.6E-3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.3E-3</td>
<td>2.9E-5</td>
<td>2.0E-5</td>
<td>1.1E-4</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.7E-2</td>
<td>1.2E-4</td>
<td>8.9E-5</td>
<td>1.4E-4</td>
</tr>
<tr>
<td>FQE</td>
<td>1.4E-5</td>
<td>1.4E-5</td>
<td>1.4E-5</td>
<td>1.4E-5</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.4E-5</td>
<td>1.4E-5</td>
<td>1.4E-5</td>
<td>1.4E-5</td>
</tr>
<tr>
<td>Q$^\theta$(\lambda)</td>
<td>1.4E-5</td>
<td>1.4E-5</td>
<td>1.4E-5</td>
<td>1.4E-5</td>
</tr>
<tr>
<td>TREE</td>
<td><strong>1.0E-5</strong></td>
<td>1.5E-5</td>
<td>1.5E-5</td>
<td><strong>1.1E-5</strong></td>
</tr>
<tr>
<td>IH</td>
<td>1.7E-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**IPS**

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E-2</td>
<td>5.8E-3</td>
</tr>
<tr>
<td>WIS</td>
<td>2.6E-3</td>
<td><strong>1.7E-3</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.2E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>5.0E-4</td>
<td>5.5E-4</td>
<td>5.5E-4</td>
<td>2.2E-4</td>
</tr>
<tr>
<td>Q-REG</td>
<td>9.3E-4</td>
<td>1.6E-5</td>
<td>1.6E-5</td>
<td>1.7E-5</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.6E-2</td>
<td>1.8E-4</td>
<td>2.0E-4</td>
<td>2.4E-4</td>
</tr>
<tr>
<td>FQE</td>
<td>1.6E-5</td>
<td>1.6E-5</td>
<td>1.6E-5</td>
<td>1.6E-5</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.6E-5</td>
<td>1.6E-5</td>
<td>1.6E-5</td>
<td>1.6E-5</td>
</tr>
<tr>
<td>Q$^\theta$(\lambda)</td>
<td>1.6E-5</td>
<td>1.6E-5</td>
<td>1.6E-5</td>
<td>1.6E-5</td>
</tr>
<tr>
<td>TREE</td>
<td><strong>1.2E-5</strong></td>
<td>1.6E-5</td>
<td>1.6E-5</td>
<td><strong>1.2E-5</strong></td>
</tr>
<tr>
<td>IH</td>
<td>5.1E-4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**IPS**

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>2.4E-3</td>
<td>1.8E-3</td>
</tr>
<tr>
<td>WIS</td>
<td>8.6E-4</td>
<td><strong>6.1E-4</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.3E-1</td>
<td>-</td>
</tr>
</tbody>
</table>


### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.9E-1</td>
<td>1.7E-1</td>
<td>1.7E-1</td>
<td>1.8E-1</td>
<td></td>
</tr>
<tr>
<td>Q-Reg</td>
<td>3.1E-1</td>
<td>2.7E-1</td>
<td>2.6E-1</td>
<td>2.8E-1</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>3.2E-1</td>
<td>2.1E-1</td>
<td>2.3E-1</td>
<td>3.2E-1</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>1.9E-1</td>
<td>2.5E-1</td>
<td>2.4E-1</td>
<td>1.9E-1</td>
<td></td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>2.3E-1</td>
<td>2.5E-1</td>
<td>2.4E-1</td>
<td>2.3E-1</td>
<td></td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>2.2E-1</td>
<td>2.2E-1</td>
<td>2.2E-1</td>
<td>2.2E-1</td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>2.4E-1</td>
<td>2.6E-1</td>
<td>2.6E-1</td>
<td>2.4E-1</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>3.0E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>5.5E-2</td>
<td>4.9E-2</td>
<td>4.7E-2</td>
<td>4.3E-2</td>
<td></td>
</tr>
<tr>
<td>Q-Reg</td>
<td>3.1E-2</td>
<td>3.0E-2</td>
<td>3.0E-2</td>
<td>3.1E-2</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>5.6E-2</td>
<td>3.1E-2</td>
<td>3.1E-2</td>
<td>4.9E-2</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>2.3E-2</td>
<td>2.6E-2</td>
<td>2.6E-2</td>
<td>2.3E-2</td>
<td></td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>2.7E-2</td>
<td>2.9E-2</td>
<td>2.8E-2</td>
<td>2.7E-2</td>
<td></td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>2.5E-2</td>
<td>3.0E-2</td>
<td>3.0E-2</td>
<td>2.5E-2</td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>2.8E-2</td>
<td>2.7E-2</td>
<td>2.7E-2</td>
<td>2.8E-2</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>2.0E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.0E0</td>
<td>5.8E-1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-Reg</td>
<td>5.1E-1</td>
<td>3.7E-1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>5.6E-1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.6E-2</td>
<td>3.4E-2</td>
<td>3.2E-2</td>
<td>1.5E-2</td>
<td></td>
</tr>
<tr>
<td>Q-Reg</td>
<td>2.9E-2</td>
<td>7.3E-3</td>
<td>7.3E-3</td>
<td>7.3E-3</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>2.6E-2</td>
<td>6.4E-3</td>
<td>6.6E-3</td>
<td>2.3E-2</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>8.0E-3</td>
<td>8.6E-3</td>
<td>8.6E-3</td>
<td>7.9E-3</td>
<td></td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>7.8E-3</td>
<td>7.9E-3</td>
<td>7.9E-3</td>
<td>7.9E-3</td>
<td></td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>8.1E-3</td>
<td>8.4E-3</td>
<td>8.3E-3</td>
<td>8.2E-3</td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>8.1E-3</td>
<td>7.5E-3</td>
<td>7.7E-3</td>
<td>8.0E-3</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>1.6E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>2.8E-2</td>
<td>4.8E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.8E-2</td>
<td>1.0E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.5E-2</td>
<td>8.6E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>7.8E-3</td>
<td>1.0E-2</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.0E-2</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>Q'(\lambda)</td>
<td>1.0E-2</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>9.7E-3</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>IH</td>
<td>9.8E-3</td>
<td>-</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>5.2E-3</td>
<td>1.9E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.2E-2</td>
<td>8.0E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.7E-2</td>
<td>7.7E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>5.9E-3</td>
<td>7.3E-3</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>7.0E-3</td>
<td>7.3E-3</td>
</tr>
<tr>
<td>Q'(\lambda)</td>
<td>7.2E-3</td>
<td>7.3E-3</td>
</tr>
<tr>
<td>Tree</td>
<td>7.0E-3</td>
<td>7.3E-3</td>
</tr>
<tr>
<td>IH</td>
<td>7.7E-3</td>
<td>-</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>2.9E-3</td>
<td>3.4E-3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.4E-3</td>
<td>9.3E-4</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.5E-2</td>
<td>1.4E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>6.7E-4</td>
<td>1.0E-3</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>8.5E-4</td>
<td>1.0E-3</td>
</tr>
<tr>
<td>Q'(\lambda)</td>
<td>6.8E-4</td>
<td>1.0E-3</td>
</tr>
<tr>
<td>Tree</td>
<td>9.0E-4</td>
<td>1.0E-3</td>
</tr>
<tr>
<td>IH</td>
<td>3.1E-3</td>
<td>-</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.1E-3</td>
<td>1.8E-3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>5.3E-4</td>
<td>1.0E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.9E-2</td>
<td>1.4E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>9.5E-4</td>
<td>1.0E-3</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.0E-3</td>
<td>1.1E-3</td>
</tr>
<tr>
<td>Q'(\lambda)</td>
<td>1.1E-3</td>
<td>1.0E-3</td>
</tr>
<tr>
<td>Tree</td>
<td>1.0E-3</td>
<td>1.1E-3</td>
</tr>
<tr>
<td>IH</td>
<td>8.7E-4</td>
<td>-</td>
</tr>
</tbody>
</table>

### IPS

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard</td>
<td>Per-Decision</td>
</tr>
<tr>
<td>IS</td>
<td>5.4E-2</td>
<td>1.6E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>1.9E-2</td>
<td><strong>1.0E-2</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.7E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

### IPS

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard</td>
<td>Per-Decision</td>
</tr>
<tr>
<td>IS</td>
<td>8.3E-3</td>
<td>5.4E-3</td>
</tr>
<tr>
<td>WIS</td>
<td>2.9E-3</td>
<td><strong>2.6E-3</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.7E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

### IPS

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard</td>
<td>Per-Decision</td>
</tr>
<tr>
<td>IS</td>
<td>2.3E-2</td>
<td>1.5E-2</td>
</tr>
<tr>
<td>WIS</td>
<td><strong>9.3E-3</strong></td>
<td>9.6E-3</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.2E-1</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>7.7E-1</td>
<td>8.5E-1</td>
<td>9.0E-1</td>
<td>9.6E-1</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>9.0E-1</td>
<td>1.2E0</td>
<td>1.1E0</td>
<td>1.0E0</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>6.6E-1</td>
<td>6.9E-1</td>
<td>7.3E-1</td>
<td>6.4E-1</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>6.3E-1</td>
<td>9.1E-1</td>
<td>9.0E-1</td>
<td>6.3E-1</td>
<td></td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>9.5E-1</td>
<td>1.0E0</td>
<td>1.0E0</td>
<td>9.5E-1</td>
<td></td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>8.6E-1</td>
<td>9.3E-1</td>
<td>9.3E-1</td>
<td>8.6E-1</td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>9.4E-1</td>
<td>9.9E-1</td>
<td>9.9E-1</td>
<td>9.4E-1</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>4.1E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.2E0</td>
<td>9.5E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>9.4E-1</td>
<td>8.5E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>1.2E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.1E-1</td>
<td>1.3E-1</td>
<td>1.6E-1</td>
<td>1.5E-1</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.6E-1</td>
<td>2.3E-1</td>
<td>2.2E-1</td>
<td>2.8E-1</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>1.2E-1</td>
<td>1.7E-1</td>
<td>1.6E-1</td>
<td>1.3E-1</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>2.0E-1</td>
<td>1.4E-1</td>
<td>1.5E-1</td>
<td>2.0E-1</td>
<td></td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.7E-1</td>
<td>1.8E-1</td>
<td>1.8E-1</td>
<td>1.7E-1</td>
<td></td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>2.1E-1</td>
<td>1.8E-1</td>
<td>1.8E-1</td>
<td>2.1E-1</td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>1.5E-1</td>
<td>1.7E-1</td>
<td>1.7E-1</td>
<td>1.5E-1</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>1.9E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>5.8E-1</td>
<td>3.0E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>2.9E-1</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>6.2E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.1E-1</td>
<td>2.0E-1</td>
<td>2.0E-1</td>
<td>1.9E-1</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.6E-1</td>
<td>1.0E-1</td>
<td>1.1E-1</td>
<td>1.5E-1</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>1.6E-1</td>
<td>1.1E-1</td>
<td>1.1E-1</td>
<td>1.6E-1</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>1.3E-1</td>
<td>1.5E-1</td>
<td>1.4E-1</td>
<td>1.3E-1</td>
<td></td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.2E-1</td>
<td>1.3E-1</td>
<td>1.2E-1</td>
<td>1.2E-1</td>
<td></td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>1.1E-1</td>
<td>1.3E-1</td>
<td>1.2E-1</td>
<td>1.1E-1</td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>1.4E-1</td>
<td>1.3E-1</td>
<td>1.3E-1</td>
<td>1.3E-1</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>1.5E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>2.1E-1</td>
<td>1.6E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>1.8E-1</td>
<td>1.4E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.8E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.3E-1</td>
<td>2.4E-1</td>
<td>2.4E-1</td>
<td>1.5E-1</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.4E-1</td>
<td>9.1E-2</td>
<td>9.4E-2</td>
<td>1.0E-1</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>1.2E-1</td>
<td>7.7E-2</td>
<td>8.0E-2</td>
<td>7.7E-2</td>
<td></td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>8.8E-2</td>
<td>1.0E-1</td>
<td>1.0E-1</td>
<td>7.9E-2</td>
<td></td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>9.6E-2</td>
<td>1.0E-1</td>
<td>1.0E-1</td>
<td>8.7E-2</td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>7.9E-2</td>
<td>9.6E-2</td>
<td>9.6E-2</td>
<td>7.5E-2</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>1.3E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>3.9E-1</td>
<td>2.1E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>1.9E-1</td>
<td>1.4E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.8E-1</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>9.7E-2</td>
<td>1.1E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>9.1E-2</td>
<td>9.0E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.3E-1</td>
<td>8.8E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>7.8E-2</td>
<td>8.4E-2</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>9.3E-2</td>
<td>8.9E-2</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>1.0E-1</td>
<td>8.8E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>1.0E-1</td>
<td>9.2E-2</td>
</tr>
<tr>
<td>IH</td>
<td>1.2E-1</td>
<td>9.7E-2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.1E-1</td>
<td>9.7E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>9.8E-2</td>
<td>8.8E-2</td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.3E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>3.0E-2</td>
<td>2.9E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.1E-2</td>
<td>4.1E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>8.6E-2</td>
<td>4.3E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>2.4E-2</td>
<td>3.8E-2</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>3.7E-2</td>
<td>3.9E-2</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>3.9E-2</td>
<td>3.9E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>3.6E-2</td>
<td>3.9E-2</td>
</tr>
<tr>
<td>IH</td>
<td>3.4E-2</td>
<td>3.4E-2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>5.3E-2</td>
<td>4.6E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>4.3E-2</td>
<td>4.0E-2</td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.7E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.2E-2</td>
<td>2.0E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.4E-2</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>6.6E-3</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>5.3E-3</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>9.0E-3</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>9.3E-3</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>8.7E-3</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>IH</td>
<td>8.6E-3</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.4E-2</td>
<td>1.6E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>1.2E-2</td>
<td>1.4E-2</td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.9E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>4.2E-3</td>
<td>6.8E-3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.6E-3</td>
<td>1.5E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.1E-2</td>
<td>1.4E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>1.7E-3</td>
<td>1.8E-3</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.8E-3</td>
<td>1.8E-3</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>2.5E-3</td>
<td>1.8E-3</td>
</tr>
<tr>
<td>Tree</td>
<td>1.8E-3</td>
<td>1.8E-3</td>
</tr>
<tr>
<td>IH</td>
<td>2.2E-3</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>4.8E-3</td>
<td>3.3E-3</td>
</tr>
<tr>
<td>WIS</td>
<td>3.2E-3</td>
<td>2.7E-3</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.6E-1</td>
<td>-</td>
</tr>
</tbody>
</table>
### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>Hybrid DR</th>
<th>Hybrid WDR</th>
<th>Hybrid MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.3E-1</td>
<td>5.9E-1</td>
<td>7.1E-1</td>
<td>4.3E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.9E-1</td>
<td>3.7E-1</td>
<td>3.6E-1</td>
<td>3.9E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.8E-1</td>
<td><strong>2.8E-1</strong></td>
<td>3.4E-1</td>
<td>3.8E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.1E-1</td>
<td>3.0E-1</td>
<td>3.3E-1</td>
<td>4.1E-1</td>
</tr>
<tr>
<td>R(λ)</td>
<td>3.2E-1</td>
<td>3.4E-1</td>
<td>3.6E-1</td>
<td>3.2E-1</td>
</tr>
<tr>
<td>Q^π(λ)</td>
<td>3.7E-1</td>
<td>3.1E-1</td>
<td>3.4E-1</td>
<td>3.7E-1</td>
</tr>
<tr>
<td>TREE</td>
<td><strong>3.2E-1</strong></td>
<td>3.5E-1</td>
<td>3.6E-1</td>
<td>3.2E-1</td>
</tr>
<tr>
<td>IH</td>
<td>7.9E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.4E0</td>
<td>9.7E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>6.9E-1</td>
<td><strong>6.8E-1</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>6.6E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>Hybrid DR</th>
<th>Hybrid WDR</th>
<th>Hybrid MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.2E0</td>
<td>1.7E0</td>
<td>1.9E0</td>
<td>1.4E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>7.7E-1</td>
<td>9.8E-1</td>
<td>9.7E-1</td>
<td>7.7E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td><strong>6.6E-1</strong></td>
<td>6.9E-1</td>
<td><strong>7.4E-1</strong></td>
<td><strong>6.5E-1</strong></td>
</tr>
<tr>
<td>FQE</td>
<td>9.8E-1</td>
<td>9.3E-1</td>
<td>1.0E0</td>
<td>9.8E-1</td>
</tr>
<tr>
<td>R(λ)</td>
<td>9.9E-1</td>
<td>1.0E0</td>
<td>9.9E-1</td>
<td></td>
</tr>
<tr>
<td>Q^π(λ)</td>
<td>9.6E-1</td>
<td>9.6E-1</td>
<td>1.0E0</td>
<td>9.6E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>1.0E0</td>
<td>1.0E0</td>
<td>1.1E0</td>
<td>1.0E0</td>
</tr>
<tr>
<td>IH</td>
<td>7.1E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>3.8E-1</td>
<td>3.6E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>6.8E-1</td>
<td><strong>3.1E-1</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>9.0E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>Hybrid DR</th>
<th>Hybrid WDR</th>
<th>Hybrid MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.5E-1</td>
<td>4.6E-1</td>
<td>4.6E-1</td>
<td>2.6E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.4E-1</td>
<td>2.9E-1</td>
<td>3.0E-1</td>
<td>3.4E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.0E-1</td>
<td><strong>2.4E-1</strong></td>
<td>2.5E-1</td>
<td>3.0E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>2.4E-1</td>
<td>3.0E-1</td>
<td>3.1E-1</td>
<td>2.4E-1</td>
</tr>
<tr>
<td>R(λ)</td>
<td>3.0E-1</td>
<td>3.1E-1</td>
<td>3.1E-1</td>
<td>3.0E-1</td>
</tr>
<tr>
<td>Q^π(λ)</td>
<td>3.4E-1</td>
<td>3.3E-1</td>
<td>3.3E-1</td>
<td>3.4E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>2.9E-1</td>
<td>3.0E-1</td>
<td>2.9E-1</td>
<td>2.9E-1</td>
</tr>
<tr>
<td>IH</td>
<td>2.7E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>2.0E1</td>
<td>1.1E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>6.8E-2</td>
<td><strong>8.4E-2</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>2.0E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>Hybrid DR</th>
<th>Hybrid WDR</th>
<th>Hybrid MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.0E-1</td>
<td>1.1E-1</td>
<td>1.2E-1</td>
<td>9.8E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>6.8E-2</td>
<td><strong>1.0E-1</strong></td>
<td>1.1E-1</td>
<td>8.4E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.1E-1</td>
<td>8.4E-2</td>
<td><strong>8.8E-2</strong></td>
<td><strong>8.8E-2</strong></td>
</tr>
<tr>
<td>FQE</td>
<td>8.8E-2</td>
<td>9.1E-2</td>
<td>9.3E-2</td>
<td>8.8E-2</td>
</tr>
<tr>
<td>R(λ)</td>
<td>8.9E-2</td>
<td><strong>1.1E-1</strong></td>
<td>1.0E-1</td>
<td>8.7E-2</td>
</tr>
<tr>
<td>Q^π(λ)</td>
<td>8.8E-2</td>
<td><strong>1.1E-1</strong></td>
<td><strong>1.1E-1</strong></td>
<td><strong>8.5E-2</strong></td>
</tr>
<tr>
<td>TREE</td>
<td>8.3E-2</td>
<td><strong>1.0E-1</strong></td>
<td>9.7E-2</td>
<td><strong>8.4E-2</strong></td>
</tr>
<tr>
<td>IH</td>
<td>7.5E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td><strong>5.9E-2</strong></td>
<td>6.0E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>8.1E-2</td>
<td><strong>7.5E-2</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.3E-1</td>
<td>-</td>
</tr>
</tbody>
</table>
### 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM DR</th>
<th>Hybrid DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.9E-2</td>
<td>8.8E-2</td>
<td>9.3E-2</td>
<td>6.4E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>9.3E-2</td>
<td>7.4E-2</td>
<td>7.4E-2</td>
<td>7.0E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.3E-1</td>
<td>8.1E-2</td>
<td>7.9E-2</td>
<td>1.3E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.0E-2</td>
<td>6.3E-2</td>
<td>6.8E-2</td>
<td>4.0E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>6.8E-2</td>
<td>7.1E-2</td>
<td>7.2E-2</td>
<td>4.7E-2</td>
</tr>
<tr>
<td>Q($^\pi$(A))</td>
<td>5.9E-2</td>
<td>7.0E-2</td>
<td>7.0E-2</td>
<td>5.3E-2</td>
</tr>
<tr>
<td>IH</td>
<td>3.3E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM DR</th>
<th>Hybrid DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>8.9E-2</td>
<td>1.7E-1</td>
<td>1.6E-1</td>
<td>9.5E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.3E-1</td>
<td>7.6E-2</td>
<td>7.7E-2</td>
<td>1.1E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.3E-1</td>
<td>7.5E-2</td>
<td>7.0E-2</td>
<td>9.6E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>4.9E-2</td>
<td>8.0E-2</td>
<td>7.9E-2</td>
<td>5.0E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>7.0E-2</td>
<td>7.8E-2</td>
<td>7.8E-2</td>
<td>6.1E-2</td>
</tr>
<tr>
<td>Q($^\pi$(A))</td>
<td>7.1E-2</td>
<td>7.9E-2</td>
<td>7.8E-2</td>
<td>6.2E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>7.0E-2</td>
<td>7.8E-2</td>
<td>7.8E-2</td>
<td>6.3E-2</td>
</tr>
<tr>
<td>IH</td>
<td>7.2E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM DR</th>
<th>Hybrid DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.7E-2</td>
<td>2.3E-2</td>
<td>2.3E-2</td>
<td>2.8E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.1E-2</td>
<td>3.5E-2</td>
<td>3.5E-2</td>
<td>3.4E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.6E-2</td>
<td>3.9E-2</td>
<td>3.9E-2</td>
<td>5.3E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>2.4E-2</td>
<td>3.5E-2</td>
<td>3.5E-2</td>
<td>2.5E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>3.3E-2</td>
<td>3.5E-2</td>
<td>3.5E-2</td>
<td>2.8E-2</td>
</tr>
<tr>
<td>Q($^\pi$(A))</td>
<td>3.3E-2</td>
<td>3.4E-2</td>
<td>3.5E-2</td>
<td>2.8E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>3.2E-2</td>
<td>3.5E-2</td>
<td>3.5E-2</td>
<td>2.9E-2</td>
</tr>
<tr>
<td>IH</td>
<td>2.8E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM DR</th>
<th>Hybrid DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.0E-2</td>
<td>2.2E-2</td>
<td>2.3E-2</td>
<td>1.7E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.5E-2</td>
<td>1.5E-2</td>
<td>1.5E-2</td>
<td>1.5E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.1E-2</td>
<td>1.6E-2</td>
<td>1.7E-2</td>
<td>2.6E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>1.3E-2</td>
<td>1.5E-2</td>
<td>1.5E-2</td>
<td>1.3E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.5E-2</td>
<td>1.5E-2</td>
<td>1.5E-2</td>
<td>1.4E-2</td>
</tr>
<tr>
<td>Q($^\pi$(A))</td>
<td>1.5E-2</td>
<td>1.5E-2</td>
<td>1.5E-2</td>
<td>1.4E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>1.5E-2</td>
<td>1.5E-2</td>
<td>1.5E-2</td>
<td>1.5E-2</td>
</tr>
<tr>
<td>IH</td>
<td>1.4E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 175: Graph, relative MSE. $T = 4, N = 1024$. Stochastic environment. Stochastic rewards. Dense rewards.

<table>
<thead>
<tr>
<th>Method</th>
<th>IPS Standard</th>
<th>IPS Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.8E-1</td>
<td>9.7E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>1.6E-1</td>
<td>8.9E-2</td>
</tr>
<tr>
<td>NAIVE</td>
<td>7.0E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 176: Graph, relative MSE. $T = 4, N = 1024$. Stochastic environment. Stochastic rewards. Dense rewards.

<table>
<thead>
<tr>
<th>Method</th>
<th>IPS Standard</th>
<th>IPS Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>4.9E-2</td>
<td>4.2E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>4.1E-2</td>
<td>3.8E-2</td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.1E-1</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>DM</td>
<td>Hybrid</td>
</tr>
<tr>
<td>----------------</td>
<td>-----</td>
<td>--------</td>
</tr>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>8.6E-2</td>
<td>3.4E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.6E-1</td>
<td>9.6E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.7E-1</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>1.9E-2</td>
<td>1.9E-2</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.9E-2</td>
<td>1.9E-2</td>
</tr>
<tr>
<td>Q^p(\lambda)</td>
<td>1.9E-2</td>
<td>1.9E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>1.9E-2</td>
<td>1.9E-2</td>
</tr>
<tr>
<td>IH</td>
<td>1.7E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>7.6E-2</td>
<td>1.1E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>8.3E-2</td>
<td>1.3E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.4E-1</td>
<td>3.1E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>5.6E-4</td>
<td>5.6E-4</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>5.6E-4</td>
<td>5.6E-4</td>
</tr>
<tr>
<td>Q^p(\lambda)</td>
<td>5.6E-4</td>
<td>5.6E-4</td>
</tr>
<tr>
<td>TREE</td>
<td>5.9E-4</td>
<td>5.3E-4</td>
</tr>
<tr>
<td>IH</td>
<td>1.4E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.1E-2</td>
<td>4.8E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.2E-1</td>
<td>2.1E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.8E-2</td>
<td>2.2E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>4.1E-4</td>
<td>4.1E-4</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>4.2E-4</td>
<td>4.1E-4</td>
</tr>
<tr>
<td>Q^p(\lambda)</td>
<td>4.1E-4</td>
<td>4.1E-4</td>
</tr>
<tr>
<td>TREE</td>
<td>4.3E-4</td>
<td>4.2E-4</td>
</tr>
<tr>
<td>IH</td>
<td>1.0E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>3.3E-2</td>
<td>3.9E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>8.1E-2</td>
<td>2.0E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>8.5E-2</td>
<td>1.5E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>1.1E-9</td>
<td>1.1E-9</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>8.2E-9</td>
<td>5.3E-9</td>
</tr>
<tr>
<td>Q^p(\lambda)</td>
<td>7.8E-10</td>
<td>1.1E-9</td>
</tr>
<tr>
<td>TREE</td>
<td>6.0E-6</td>
<td>5.9E-7</td>
</tr>
<tr>
<td>IH</td>
<td>1.2E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>6.0E-1</td>
<td>6.0E-1</td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>2.4E-1</td>
<td>2.4E-1</td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.6E-1</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>6.0E-1</td>
<td>6.0E-1</td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>5.1E-2</td>
<td>5.1E-2</td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>6.2E-1</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
### Table 179: Graph, relative MSE. $T = 4, N = 128, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$. Sparse rewards.

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.3E-2</td>
<td>1.2E-2</td>
<td>1.2E-2</td>
<td>2.2E-2</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>3.0E-2</td>
<td>3.4E-4</td>
<td>2.1E-4</td>
<td>5.9E-4</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.9E-2</td>
<td>1.1E-3</td>
<td>2.4E-3</td>
<td>9.5E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>6.9E-7</td>
<td>6.9E-7</td>
<td>6.9E-7</td>
<td>6.9E-7</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td><strong>6.7E-7</strong></td>
<td>6.8E-7</td>
<td>6.7E-7</td>
<td><strong>6.7E-7</strong></td>
</tr>
<tr>
<td>Q^\pi(\lambda)</td>
<td>7.0E-7</td>
<td>6.9E-7</td>
<td>6.9E-7</td>
<td>7.0E-7</td>
</tr>
<tr>
<td>Tree</td>
<td>2.6E-6</td>
<td>9.0E-7</td>
<td>9.8E-7</td>
<td>2.4E-6</td>
</tr>
<tr>
<td>IH</td>
<td>1.2E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.7E-3</td>
<td>6.8E-3</td>
<td>6.9E-3</td>
<td>7.2E-3</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>2.2E-2</td>
<td>1.0E-4</td>
<td>7.4E-5</td>
<td>9.7E-5</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.3E-2</td>
<td>4.8E-4</td>
<td>5.5E-4</td>
<td>3.3E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>4.9E-7</td>
<td>4.9E-7</td>
<td>4.9E-7</td>
<td>4.9E-7</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>5.1E-7</td>
<td>5.1E-7</td>
<td>5.1E-7</td>
<td>5.2E-7</td>
</tr>
<tr>
<td>Q^\pi(\lambda)</td>
<td><strong>4.9E-7</strong></td>
<td>4.9E-7</td>
<td>4.9E-7</td>
<td><strong>4.9E-7</strong></td>
</tr>
<tr>
<td>Tree</td>
<td>9.3E-6</td>
<td>7.4E-7</td>
<td>7.6E-7</td>
<td>8.8E-6</td>
</tr>
<tr>
<td>IH</td>
<td>1.4E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.0E-3</td>
<td>5.5E-3</td>
<td>5.8E-3</td>
<td>4.0E-3</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>2.5E-3</td>
<td>1.9E-5</td>
<td>2.0E-5</td>
<td>5.9E-4</td>
</tr>
<tr>
<td>MRDR</td>
<td>6.1E-3</td>
<td>5.4E-4</td>
<td>7.9E-4</td>
<td>2.0E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>3.2E-5</td>
<td>3.2E-5</td>
<td>3.2E-5</td>
<td>3.2E-5</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>3.2E-5</td>
<td>3.2E-5</td>
<td>3.2E-5</td>
<td>3.2E-5</td>
</tr>
<tr>
<td>Q^\pi(\lambda)</td>
<td>3.2E-5</td>
<td>3.2E-5</td>
<td>3.2E-5</td>
<td>3.2E-5</td>
</tr>
<tr>
<td>Tree</td>
<td><strong>1.1E-5</strong></td>
<td>3.2E-5</td>
<td>3.2E-5</td>
<td><strong>1.2E-5</strong></td>
</tr>
<tr>
<td>IH</td>
<td>3.4E-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.1E-3</td>
<td>1.7E-3</td>
<td>1.6E-3</td>
<td>1.4E-3</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>3.0E-3</td>
<td>6.3E-5</td>
<td>6.4E-5</td>
<td>5.8E-4</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.6E-3</td>
<td>6.0E-4</td>
<td>8.6E-4</td>
<td>2.0E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>6.3E-5</td>
<td>6.3E-5</td>
<td>6.3E-5</td>
<td>6.3E-5</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td><strong>6.3E-5</strong></td>
<td>6.3E-5</td>
<td>6.3E-5</td>
<td><strong>6.3E-5</strong></td>
</tr>
<tr>
<td>Q^\pi(\lambda)</td>
<td>6.3E-5</td>
<td>6.3E-5</td>
<td>6.3E-5</td>
<td>6.3E-5</td>
</tr>
<tr>
<td>Tree</td>
<td>1.1E-4</td>
<td><strong>6.2E-5</strong></td>
<td>6.2E-5</td>
<td>1.0E-4</td>
</tr>
<tr>
<td>IH</td>
<td>2.2E-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 183: Graph, relative MSE. $T = 4, N = 1024, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$. Sparse rewards.

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.6E-3</td>
<td>3.6E-3</td>
<td>3.6E-3</td>
<td>3.6E-3</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.5E-3</td>
<td><strong>1.5E-3</strong></td>
<td>1.5E-3</td>
<td>1.5E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.5E-3</td>
<td>1.5E-3</td>
<td>1.5E-3</td>
<td>1.5E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>4.5E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### IPS

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>2.9E-2</td>
<td>2.9E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>1.9E-2</td>
<td><strong>1.9E-2</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.4E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>2.8E-3</td>
<td>2.8E-3</td>
</tr>
<tr>
<td>WIS</td>
<td><strong>1.2E-3</strong></td>
<td><strong>1.2E-3</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.6E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>3.6E-3</td>
<td>3.6E-3</td>
</tr>
<tr>
<td>WIS</td>
<td><strong>1.5E-3</strong></td>
<td><strong>1.5E-3</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.5E-1</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>3.3E0</td>
<td>3.9E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.0E0</td>
<td>3.8E0</td>
</tr>
<tr>
<td>MRDR</td>
<td><strong>1.5E0</strong></td>
<td>2.5E0</td>
</tr>
<tr>
<td>FQE</td>
<td>2.8E0</td>
<td>3.8E0</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>4.3E0</td>
<td>4.6E0</td>
</tr>
<tr>
<td>$Q^8(\lambda)$</td>
<td>4.6E0</td>
<td>4.8E0</td>
</tr>
<tr>
<td>Tree</td>
<td>3.7E0</td>
<td>4.1E0</td>
</tr>
<tr>
<td>IH</td>
<td>3.4E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>6.1E-1</td>
<td>6.6E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>5.2E-1</td>
<td>4.2E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.5E-1</td>
<td>4.1E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>3.8E-1</td>
<td>3.2E-1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>3.3E-1</td>
<td>3.2E-1</td>
</tr>
<tr>
<td>$Q^8(\lambda)$</td>
<td>3.2E-1</td>
<td>3.1E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>3.7E-1</td>
<td>3.3E-1</td>
</tr>
<tr>
<td>IH</td>
<td>7.4E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>4.9E-1</td>
<td>2.2E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.8E-1</td>
<td>4.0E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.0E-1</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>FQE</td>
<td><strong>1.7E-1</strong></td>
<td>1.9E-1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>2.3E-1</td>
<td>2.3E-1</td>
</tr>
<tr>
<td>$Q^8(\lambda)$</td>
<td>3.1E-1</td>
<td>2.6E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>2.3E-1</td>
<td>2.2E-1</td>
</tr>
<tr>
<td>IH</td>
<td>4.1E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>2.7E-1</td>
<td>6.6E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.0E-1</td>
<td>2.5E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.8E-1</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>3.5E-1</td>
<td>2.6E-1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>2.6E-1</td>
<td>2.5E-1</td>
</tr>
<tr>
<td>$Q^8(\lambda)$</td>
<td>2.8E-1</td>
<td>2.5E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>2.7E-1</td>
<td>2.5E-1</td>
</tr>
<tr>
<td>IH</td>
<td>2.5E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

Ips

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard</td>
</tr>
<tr>
<td>IS</td>
<td>3.8E0</td>
</tr>
<tr>
<td>WIS</td>
<td>3.5E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>2.7E0</td>
</tr>
</tbody>
</table>

Ips

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard</td>
</tr>
<tr>
<td>IS</td>
<td>4.8E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>4.7E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>8.0E-1</td>
</tr>
</tbody>
</table>

Ips

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard</td>
</tr>
<tr>
<td>IS</td>
<td>5.3E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>3.4E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.1E-1</td>
</tr>
</tbody>
</table>

Ips

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard</td>
</tr>
<tr>
<td>IS</td>
<td>2.4E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>3.4E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.1E-1</td>
</tr>
</tbody>
</table>
### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.4E-1</td>
<td>5.8E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.3E-1</td>
<td>9.3E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.6E-2</td>
<td>1.0E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>5.4E-2</td>
<td>8.3E-2</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>7.1E-2</td>
<td>8.3E-2</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>7.2E-2</td>
<td>8.2E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>6.7E-2</td>
<td>8.2E-2</td>
</tr>
<tr>
<td>IH</td>
<td>2.1E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>5.0E-2</td>
<td>9.7E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.8E-2</td>
<td>1.9E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.4E-2</td>
<td>1.7E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>1.6E-2</td>
<td>1.9E-2</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.8E-2</td>
<td>1.8E-2</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>1.8E-2</td>
<td>1.9E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>1.8E-2</td>
<td>1.8E-2</td>
</tr>
<tr>
<td>IH</td>
<td>1.7E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.5E-2</td>
<td>4.1E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>8.7E-3</td>
<td>1.0E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.0E-2</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>9.3E-3</td>
<td>9.9E-3</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>9.8E-3</td>
<td>1.0E-2</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>1.0E-2</td>
<td>1.0E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>9.7E-3</td>
<td>9.9E-3</td>
</tr>
<tr>
<td>IH</td>
<td>9.7E-3</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 180: Graph, relative MSE. $T = 4, N = 1024, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$. Stochastic rewards. Sparse rewards.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.7E-2</td>
<td>4.1E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.8E-2</td>
<td>1.9E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.4E-2</td>
<td>1.7E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>1.6E-2</td>
<td>1.9E-2</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.8E-2</td>
<td>1.8E-2</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>1.8E-2</td>
<td>1.9E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>1.8E-2</td>
<td>1.8E-2</td>
</tr>
<tr>
<td>IH</td>
<td>1.7E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>5.0E-2</td>
<td>9.7E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.8E-2</td>
<td>1.9E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.4E-2</td>
<td>1.7E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>1.6E-2</td>
<td>1.9E-2</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.8E-2</td>
<td>1.8E-2</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>1.8E-2</td>
<td>1.9E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>1.8E-2</td>
<td>1.8E-2</td>
</tr>
<tr>
<td>IH</td>
<td>1.7E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.7E-2</td>
<td>4.1E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.8E-2</td>
<td>1.9E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.4E-2</td>
<td>1.7E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>1.6E-2</td>
<td>1.9E-2</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.8E-2</td>
<td>1.8E-2</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>1.8E-2</td>
<td>1.9E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>1.8E-2</td>
<td>1.8E-2</td>
</tr>
<tr>
<td>IH</td>
<td>1.7E-2</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.5E0</td>
<td>3.5E0</td>
<td>3.2E0</td>
<td>2.3E0</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.3E0</td>
<td>2.1E0</td>
<td>2.2E0</td>
<td>2.3E0</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>1.7E0</td>
<td><strong>1.6E0</strong></td>
<td>1.8E0</td>
<td>1.7E0</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>1.6E0</td>
<td>2.1E0</td>
<td>2.2E0</td>
<td>1.7E0</td>
<td></td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>2.3E0</td>
<td>2.6E0</td>
<td>2.6E0</td>
<td>2.4E0</td>
<td></td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>2.5E0</td>
<td>2.7E0</td>
<td>2.7E0</td>
<td>2.5E0</td>
<td></td>
</tr>
<tr>
<td>TREE</td>
<td>2.0E0</td>
<td>2.5E0</td>
<td>2.5E0</td>
<td>2.2E0</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>6.0E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.7E0</td>
<td>2.4E0</td>
<td>2.3E0</td>
<td>2.2E0</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.9E0</td>
<td>1.8E0</td>
<td>1.9E0</td>
<td>2.0E0</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>1.4E0</td>
<td><strong>9.2E-1</strong></td>
<td>9.8E-1</td>
<td>1.3E0</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>1.8E0</td>
<td>1.9E0</td>
<td>1.8E0</td>
<td>2.0E0</td>
<td></td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>2.1E0</td>
<td>2.2E0</td>
<td>2.1E0</td>
<td>2.2E0</td>
<td></td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>2.1E0</td>
<td>2.3E0</td>
<td>2.2E0</td>
<td>2.2E0</td>
<td></td>
</tr>
<tr>
<td>TREE</td>
<td>2.0E0</td>
<td>2.0E0</td>
<td>2.0E0</td>
<td>2.2E0</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td><strong>1.0E0</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

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.

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

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.7E-1</td>
<td>3.2E-1</td>
<td>3.3E-1</td>
<td>2.7E-1</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.4E-1</td>
<td>2.1E-1</td>
<td>2.1E-1</td>
<td>2.1E-1</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>9.0E-2</td>
<td>1.8E-1</td>
<td>2.1E-1</td>
<td>2.0E-1</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td><strong>4.2E-2</strong></td>
<td>1.7E-1</td>
<td>1.8E-1</td>
<td><strong>1.4E-1</strong></td>
<td></td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.2E-1</td>
<td>1.8E-1</td>
<td>1.9E-1</td>
<td>1.7E-1</td>
<td></td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>1.4E-1</td>
<td>1.8E-1</td>
<td>1.9E-1</td>
<td>1.8E-1</td>
<td></td>
</tr>
<tr>
<td>TREE</td>
<td>1.4E-1</td>
<td>1.8E-1</td>
<td>1.9E-1</td>
<td>1.9E-1</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>6.5E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS</td>
<td><strong>2.6E0</strong></td>
</tr>
<tr>
<td>WS</td>
<td>2.6E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>2.1E0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS</td>
<td>7.8E-1</td>
</tr>
<tr>
<td>WS</td>
<td><strong>5.8E-1</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>2.5E-1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS</td>
<td>1.7E-1</td>
</tr>
<tr>
<td>WS</td>
<td><strong>1.6E-1</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>6.1E-1</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>2.0E-1</td>
<td>2.3E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.1E-1</td>
<td>9.2E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.5E-1</td>
<td>1.1E-1</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.0E-1</td>
<td>9.6E-2</td>
</tr>
<tr>
<td>Q$^\phi$($\lambda$)</td>
<td>1.1E-1</td>
<td>9.9E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>1.2E-1</td>
<td>9.5E-2</td>
</tr>
<tr>
<td>IH</td>
<td>7.5E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>8.3E-2</td>
<td>5.4E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.7E-2</td>
<td>3.1E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.4E-2</td>
<td>3.2E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>3.5E-2</td>
<td>3.0E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>3.4E-2</td>
<td>3.1E-2</td>
</tr>
<tr>
<td>Q$^\phi$($\lambda$)</td>
<td>3.2E-2</td>
<td>3.0E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>3.7E-2</td>
<td>3.0E-2</td>
</tr>
<tr>
<td>IH</td>
<td>6.7E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>4.4E-2</td>
<td>3.2E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.8E-2</td>
<td>1.7E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.5E-2</td>
<td>2.7E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>1.4E-2</td>
<td>1.6E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.6E-2</td>
<td>1.7E-2</td>
</tr>
<tr>
<td>Q$^\phi$($\lambda$)</td>
<td>1.6E-2</td>
<td>1.7E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>1.6E-2</td>
<td>1.7E-2</td>
</tr>
<tr>
<td>IH</td>
<td>2.3E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>8.0E-3</td>
<td>2.8E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>8.4E-3</td>
<td>8.3E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.3E-2</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>9.3E-3</td>
<td>8.5E-3</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>7.9E-3</td>
<td>8.5E-3</td>
</tr>
<tr>
<td>Q$^\phi$($\lambda$)</td>
<td>8.5E-3</td>
<td>8.5E-3</td>
</tr>
<tr>
<td>Tree</td>
<td>7.7E-3</td>
<td>8.5E-3</td>
</tr>
<tr>
<td>IH</td>
<td>7.8E-3</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 199: Graph, relative MSE. $T = 4, N = 512, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$. Stochastic environment. Sparse rewards.

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard</td>
</tr>
<tr>
<td>IS</td>
<td>1.1E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>1.0E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.2E-1</td>
</tr>
</tbody>
</table>

Table 200: Graph, relative MSE. $T = 4, N = 256, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$. Stochastic environment. Sparse rewards.

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard</td>
</tr>
<tr>
<td>IS</td>
<td>2.7E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>2.4E-2</td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.2E-1</td>
</tr>
</tbody>
</table>

Table 201: Graph, relative MSE. $T = 4, N = 1024, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$. Stochastic environment. Sparse rewards.

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard</td>
</tr>
<tr>
<td>IS</td>
<td>9.0E-3</td>
</tr>
<tr>
<td>WIS</td>
<td>9.1E-3</td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.1E-1</td>
</tr>
</tbody>
</table>
### 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.4E1</td>
<td>2.5E1</td>
<td>2.4E1</td>
<td>1.7E1</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.7E1</td>
<td>2.1E1</td>
<td>2.0E1</td>
<td>1.7E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.7E0</td>
<td>1.8E1</td>
<td>1.8E1</td>
<td>9.7E0</td>
</tr>
<tr>
<td>FQE</td>
<td>1.7E1</td>
<td>1.8E1</td>
<td>1.8E1</td>
<td>1.7E1</td>
</tr>
<tr>
<td>R(\alpha)</td>
<td>1.8E1</td>
<td>1.9E1</td>
<td>1.8E1</td>
<td>1.8E1</td>
</tr>
<tr>
<td>Q^\alpha(\lambda)</td>
<td>1.7E1</td>
<td>1.8E1</td>
<td>1.8E1</td>
<td>1.7E1</td>
</tr>
<tr>
<td>Tree</td>
<td>1.9E1</td>
<td>1.9E1</td>
<td>1.8E1</td>
<td>1.9E1</td>
</tr>
<tr>
<td>IH</td>
<td>1.8E1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>5.6E0</td>
<td>9.8E0</td>
<td>1.1E1</td>
<td>5.8E0</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>4.7E0</td>
<td>3.3E0</td>
<td>3.3E0</td>
<td>4.7E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.1E0</td>
<td>3.0E0</td>
<td>3.0E0</td>
<td>3.1E0</td>
</tr>
<tr>
<td>FQE</td>
<td>2.1E0</td>
<td>3.7E0</td>
<td>3.6E0</td>
<td>2.1E0</td>
</tr>
<tr>
<td>R(\alpha)</td>
<td>2.7E0</td>
<td>3.4E0</td>
<td>3.3E0</td>
<td>2.7E0</td>
</tr>
<tr>
<td>Q^\alpha(\lambda)</td>
<td>3.6E0</td>
<td>4.0E0</td>
<td>3.8E0</td>
<td>3.6E0</td>
</tr>
<tr>
<td>Tree</td>
<td>2.4E0</td>
<td>3.2E0</td>
<td>3.0E0</td>
<td>2.4E0</td>
</tr>
<tr>
<td>IH</td>
<td>3.1E0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.4E0</td>
<td>7.3E0</td>
<td>6.8E0</td>
<td>5.6E0</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.9E0</td>
<td>1.5E0</td>
<td>1.5E0</td>
<td>1.9E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.2E0</td>
<td>1.5E0</td>
<td>1.5E0</td>
<td>1.2E0</td>
</tr>
<tr>
<td>FQE</td>
<td>1.1E0</td>
<td>1.3E0</td>
<td>1.3E0</td>
<td>1.2E0</td>
</tr>
<tr>
<td>R(\alpha)</td>
<td>1.3E0</td>
<td>1.4E0</td>
<td>1.4E0</td>
<td>1.3E0</td>
</tr>
<tr>
<td>Q^\alpha(\lambda)</td>
<td>1.4E0</td>
<td>1.4E0</td>
<td>1.4E0</td>
<td>1.4E0</td>
</tr>
<tr>
<td>Tree</td>
<td>1.4E0</td>
<td>1.4E0</td>
<td>1.4E0</td>
<td>1.4E0</td>
</tr>
<tr>
<td>IH</td>
<td>7.4E-1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>9.7E-1</td>
<td>1.8E0</td>
<td>1.7E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.6E0</td>
<td>1.9E0</td>
<td>1.9E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>8.0E-1</td>
<td>1.8E0</td>
<td>1.9E0</td>
<td>7.7E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>5.3E-1</td>
<td>1.9E0</td>
<td>1.9E0</td>
<td>5.6E-1</td>
</tr>
<tr>
<td>R(\alpha)</td>
<td>1.5E0</td>
<td>1.9E0</td>
<td>2.0E0</td>
<td>1.0E0</td>
</tr>
<tr>
<td>Q^\alpha(\lambda)</td>
<td>1.4E0</td>
<td>1.9E0</td>
<td>1.9E0</td>
<td>1.0E0</td>
</tr>
<tr>
<td>Tree</td>
<td>1.4E0</td>
<td>2.0E0</td>
<td>2.0E0</td>
<td>1.2E0</td>
</tr>
<tr>
<td>IH</td>
<td>4.9E-1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>9.7E-1</td>
<td>1.8E0</td>
<td>1.7E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.6E0</td>
<td>1.9E0</td>
<td>1.9E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>8.0E-1</td>
<td>1.8E0</td>
<td>1.9E0</td>
<td>7.7E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>5.3E-1</td>
<td>1.9E0</td>
<td>1.9E0</td>
<td>5.6E-1</td>
</tr>
<tr>
<td>R(\alpha)</td>
<td>1.5E0</td>
<td>1.9E0</td>
<td>2.0E0</td>
<td>1.0E0</td>
</tr>
<tr>
<td>Q^\alpha(\lambda)</td>
<td>1.4E0</td>
<td>1.9E0</td>
<td>1.9E0</td>
<td>1.0E0</td>
</tr>
<tr>
<td>Tree</td>
<td>1.4E0</td>
<td>2.0E0</td>
<td>2.0E0</td>
<td>1.2E0</td>
</tr>
<tr>
<td>IH</td>
<td>4.9E-1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### IPS

#### 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>IPS Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>6.4E0</td>
<td>1.9E1</td>
</tr>
<tr>
<td>WIS</td>
<td>6.9E0</td>
<td>1.9E1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>1.3E1</td>
<td>-</td>
</tr>
</tbody>
</table>

#### 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>IPS Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.7E0</td>
<td>1.7E0</td>
</tr>
<tr>
<td>WIS</td>
<td>1.5E0</td>
<td>1.3E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>1.2E0</td>
<td>-</td>
</tr>
</tbody>
</table>

#### 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>IPS Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.6E0</td>
<td>1.7E0</td>
</tr>
<tr>
<td>WIS</td>
<td>1.8E0</td>
<td>1.8E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>2.0E-1</td>
<td>-</td>
</tr>
<tr>
<td>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.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>DM</strong></td>
<td><strong>HYBRID</strong></td>
<td></td>
</tr>
<tr>
<td><strong>DIRECT</strong></td>
<td><strong>DR</strong></td>
<td><strong>WDR</strong></td>
</tr>
<tr>
<td>AM</td>
<td>7.1E-1</td>
<td>1.2E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.5E-1</td>
<td>4.5E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.5E-1</td>
<td>3.9E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>2.1E-1</td>
<td>4.8E-1</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>3.8E-1</td>
<td>4.7E-1</td>
</tr>
<tr>
<td>Q$^\pi$($\lambda$)</td>
<td>2.8E-1</td>
<td>4.6E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>4.5E-1</td>
<td>4.7E-1</td>
</tr>
<tr>
<td>IH</td>
<td><strong>1.3E-1</strong></td>
<td>-</td>
</tr>
</tbody>
</table>

| 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 | 1.7E-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 | **1.4E-1** | 2.1E-1 | 2.1E-1 | **1.4E-1** |
| 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** | - | - | - |

| 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 | 1.9E-1 | 3.9E-1 | 4.0E-1 | **4.5E-1** |
| Q-REG | 2.8E-1 | 2.2E-1 | 2.2E-1 | 2.2E-1 |
| MRDR | 2.0E-1 | 2.3E-1 | 2.3E-1 | 1.9E-1 |
| FQE | **2.3E-1** | 2.1E-1 | 2.1E-1 | **1.4E-1** |
| 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** | - | - | - |

| 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 | **3.6E-2** |
| 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 | **2.6E-2** | - | - | - |

<table>
<thead>
<tr>
<th>IPS</th>
<th><strong>STANDARD</strong></th>
<th><strong>PER-DECISION</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>4.8E-1</td>
<td><strong>3.3E-1</strong></td>
</tr>
<tr>
<td>WIS</td>
<td>5.1E-1</td>
<td>3.6E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.5E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th><strong>STANDARD</strong></th>
<th><strong>PER-DECISION</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>2.8E-1</td>
<td>2.3E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>2.7E-1</td>
<td><strong>2.3E-1</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.8E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th><strong>STANDARD</strong></th>
<th><strong>PER-DECISION</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>9.3E-2</td>
<td>6.6E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>8.9E-2</td>
<td><strong>6.2E-2</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.8E-1</td>
<td>-</td>
</tr>
</tbody>
</table>
### Table 207: Graph, relative MSE. $T = 16, N = 8, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$. Dense rewards.

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>Hybrid</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.1E-1</td>
<td>2.2E-1</td>
<td>1.8E-1</td>
<td><strong>1.0E-1</strong></td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.6E-1</td>
<td>2.0E-1</td>
<td>2.3E-1</td>
<td>2.8E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.4E-1</td>
<td>1.3E0</td>
<td>8.6E-1</td>
<td>8.1E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>1.1E-1</td>
<td>1.1E-1</td>
<td>1.1E-1</td>
<td>1.1E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.1E-1</td>
<td>1.1E-1</td>
<td>1.1E-1</td>
<td>1.1E-1</td>
</tr>
<tr>
<td>Q^g(\lambda)</td>
<td>1.1E-1</td>
<td>1.1E-1</td>
<td>1.1E-1</td>
<td>1.1E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>1.8E-1</td>
<td>1.4E-1</td>
<td>1.2E-1</td>
<td>1.7E-1</td>
</tr>
<tr>
<td>IH</td>
<td><strong>3.1E-2</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 209: Graph, relative MSE. $T = 16, N = 32, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$. Dense rewards.

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>Hybrid</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.0E-3</td>
<td>4.1E-2</td>
<td>2.7E-2</td>
<td>1.1E-3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>7.7E-2</td>
<td>8.8E-3</td>
<td>2.6E-3</td>
<td>4.8E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>7.4E-2</td>
<td>2.4E-1</td>
<td>1.4E-1</td>
<td>1.1E-1</td>
</tr>
<tr>
<td>FQE</td>
<td><strong>5.0E-6</strong></td>
<td>5.0E-6</td>
<td>5.0E-6</td>
<td>5.0E-6</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>7.0E-6</td>
<td>7.0E-6</td>
<td>7.0E-6</td>
<td>7.0E-6</td>
</tr>
<tr>
<td>Q^g(\lambda)</td>
<td>5.0E-6</td>
<td>5.0E-6</td>
<td>5.0E-6</td>
<td>5.0E-6</td>
</tr>
<tr>
<td>TREE</td>
<td>5.6E-2</td>
<td>9.1E-3</td>
<td>6.5E-3</td>
<td>1.4E-2</td>
</tr>
<tr>
<td>IH</td>
<td>4.1E-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 208: Graph, relative MSE. $T = 16, N = 16, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$. Dense rewards.

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.1E0</td>
<td>3.0E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>1.3E-1</td>
<td><strong>9.6E-2</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>6.2E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 210: Graph, relative MSE. $T = 16, N = 64, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$. Dense rewards.

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>4.2E-1</td>
<td>1.1E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>3.1E-2</td>
<td><strong>6.7E-3</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.1E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 211: Graph, relative MSE. $T = 16, N = 64, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$. Dense rewards.

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>4.7E-1</td>
<td>1.9E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>7.5E-2</td>
<td><strong>2.4E-2</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.6E-1</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AM</td>
<td>5.6E-4</td>
<td>7.1E-3</td>
<td>6.6E-3</td>
<td>2.6E-4</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.6E-2</td>
<td>3.6E-3</td>
<td>2.3E-3</td>
<td>2.7E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.0E-2</td>
<td>1.8E-2</td>
<td>1.5E-2</td>
<td>2.8E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>3.0E-6</td>
<td>3.0E-6</td>
<td>3.0E-6</td>
<td>3.0E-6</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>3.0E-6</td>
<td>3.0E-6</td>
<td>3.0E-6</td>
<td>3.0E-6</td>
</tr>
<tr>
<td>Q^b(\lambda)</td>
<td>3.0E-6</td>
<td>3.0E-6</td>
<td>3.0E-6</td>
<td>3.0E-6</td>
</tr>
<tr>
<td>TREE</td>
<td>4.8E-2</td>
<td>2.6E-3</td>
<td>1.2E-3</td>
<td>1.2E-3</td>
</tr>
<tr>
<td>IH</td>
<td>6.7E-4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 212: Graph, relative MSE. \( T = 16, N = 256, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8 \). Dense rewards.

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AM</td>
<td>5.8E-4</td>
<td>1.5E-3</td>
<td>1.6E-3</td>
<td>6.5E-4</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.8E-2</td>
<td>2.3E-3</td>
<td>9.4E-4</td>
<td>9.9E-4</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.4E-2</td>
<td>3.7E-2</td>
<td>2.5E-2</td>
<td>1.7E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>1.0E-6</td>
<td>1.0E-6</td>
<td>1.0E-6</td>
<td>1.0E-6</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.0E-6</td>
<td>1.0E-6</td>
<td>1.0E-6</td>
<td>1.0E-6</td>
</tr>
<tr>
<td>Q^b(\lambda)</td>
<td>1.0E-6</td>
<td>1.0E-6</td>
<td>1.0E-6</td>
<td>1.0E-6</td>
</tr>
<tr>
<td>TREE</td>
<td>3.0E-3</td>
<td>2.6E-4</td>
<td>2.4E-4</td>
<td>2.4E-4</td>
</tr>
<tr>
<td>IH</td>
<td>3.4E-4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 213: Graph, relative MSE. \( T = 16, N = 512, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8 \). Dense rewards.

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AM</td>
<td>5.0E-5</td>
<td>3.3E-3</td>
<td>3.3E-3</td>
<td>1.5E-3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.5E-3</td>
<td>7.7E-5</td>
<td>3.3E-5</td>
<td>4.5E-5</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.0E-3</td>
<td>2.6E-3</td>
<td>2.7E-3</td>
<td>5.7E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>7.7E-7</td>
<td>7.7E-7</td>
<td>7.7E-7</td>
<td>7.7E-7</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>8.1E-7</td>
<td>8.0E-7</td>
<td>7.9E-7</td>
<td>8.1E-7</td>
</tr>
<tr>
<td>Q^b(\lambda)</td>
<td>7.6E-7</td>
<td>7.7E-7</td>
<td>7.7E-7</td>
<td>7.6E-7</td>
</tr>
<tr>
<td>TREE</td>
<td>2.2E-3</td>
<td>1.2E-4</td>
<td>1.1E-4</td>
<td>1.1E-4</td>
</tr>
<tr>
<td>IH</td>
<td>1.7E-4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 214: Graph, relative MSE. \( T = 16, N = 1024, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8 \). Dense rewards.

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AM</td>
<td>4.8E-5</td>
<td>4.7E-4</td>
<td>4.8E-4</td>
<td>7.2E-5</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.1E-3</td>
<td>1.4E-5</td>
<td>1.0E-5</td>
<td>2.9E-5</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.9E-3</td>
<td>5.0E-4</td>
<td>4.6E-4</td>
<td>9.6E-4</td>
</tr>
<tr>
<td>FQE</td>
<td>2.7E-7</td>
<td>2.7E-7</td>
<td>2.7E-7</td>
<td>2.7E-7</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>2.9E-7</td>
<td>2.8E-7</td>
<td>2.8E-7</td>
<td>2.9E-7</td>
</tr>
<tr>
<td>Q^b(\lambda)</td>
<td>2.7E-7</td>
<td>2.7E-7</td>
<td>2.7E-7</td>
<td>2.7E-7</td>
</tr>
<tr>
<td>TREE</td>
<td>1.8E-3</td>
<td>7.7E-5</td>
<td>5.8E-5</td>
<td>5.8E-5</td>
</tr>
<tr>
<td>IH</td>
<td>9.6E-5</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td>5.7E-2</td>
<td>2.0E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>6.0E-3</td>
<td><strong>1.8E-3</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.6E-1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td>4.0E-2</td>
<td><strong>4.7E-3</strong></td>
</tr>
<tr>
<td>WIS</td>
<td>4.0E-3</td>
<td><strong>6.0E-4</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.5E-1</td>
<td></td>
</tr>
</tbody>
</table>

Table 215: Graph, relative MSE. \( T = 16, N = 512, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8 \). Dense rewards.

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AM</td>
<td>5.8E-5</td>
<td>3.3E-3</td>
<td>3.3E-3</td>
<td>1.5E-3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.5E-3</td>
<td>7.7E-5</td>
<td>3.3E-5</td>
<td>4.5E-5</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.0E-3</td>
<td>2.6E-3</td>
<td>2.7E-3</td>
<td>5.7E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>7.7E-7</td>
<td>7.7E-7</td>
<td>7.7E-7</td>
<td>7.7E-7</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>8.1E-7</td>
<td>8.0E-7</td>
<td>7.9E-7</td>
<td>8.1E-7</td>
</tr>
<tr>
<td>Q^b(\lambda)</td>
<td>7.6E-7</td>
<td>7.7E-7</td>
<td>7.7E-7</td>
<td>7.6E-7</td>
</tr>
<tr>
<td>TREE</td>
<td>2.2E-3</td>
<td>1.2E-4</td>
<td>1.1E-4</td>
<td>1.1E-4</td>
</tr>
<tr>
<td>IH</td>
<td>1.7E-4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td>1.7E-2</td>
<td>2.1E-3</td>
</tr>
<tr>
<td>WIS</td>
<td>1.8E-3</td>
<td><strong>2.9E-4</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.4E-1</td>
<td></td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.5E-1</td>
<td>5.0E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.3E-1</td>
<td>3.1E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.8E-1</td>
<td>1.1E0</td>
</tr>
<tr>
<td>FQE</td>
<td>1.3E-1</td>
<td>1.4E-1</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.4E-1</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>Q$^\pi$(4)</td>
<td>1.5E-1</td>
<td>1.6E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>2.1E-1</td>
<td>1.2E-1</td>
</tr>
<tr>
<td>IH</td>
<td>4.2E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>IS</th>
<th>WIS</th>
<th>NAIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td>4.9E-1</td>
<td>1.1E-1</td>
<td>4.4E-1</td>
</tr>
<tr>
<td>HYBRID</td>
<td>6.0E-2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.0E-2</td>
<td>8.5E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>7.3E-2</td>
<td>4.8E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.1E-1</td>
<td>2.0E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>9.8E-3</td>
<td>2.6E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>2.9E-2</td>
<td>3.5E-2</td>
</tr>
<tr>
<td>Q$^\pi$(4)</td>
<td>2.9E-2</td>
<td>4.4E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>1.1E-1</td>
<td>4.7E-2</td>
</tr>
<tr>
<td>IH</td>
<td>8.4E-3</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>IS</th>
<th>WIS</th>
<th>NAIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td>3.0E-1</td>
<td>5.4E-2</td>
<td>4.4E-1</td>
</tr>
<tr>
<td>HYBRID</td>
<td>3.6E-2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

IPS

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.8E-2</td>
<td>3.2E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.4E-1</td>
<td>4.3E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.4E-1</td>
<td>2.3E0</td>
</tr>
<tr>
<td>FQE</td>
<td>1.4E-2</td>
<td>3.6E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.6E-2</td>
<td>1.4E-2</td>
</tr>
<tr>
<td>Q$^\pi$(4)</td>
<td>2.8E-2</td>
<td>2.3E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>6.7E-2</td>
<td>3.2E-2</td>
</tr>
<tr>
<td>IH</td>
<td>1.5E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>IS</th>
<th>WIS</th>
<th>NAIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td>1.8E1</td>
<td>4.5E-1</td>
<td></td>
</tr>
<tr>
<td>HYBRID</td>
<td>3.2E-2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>IS</th>
<th>WIS</th>
<th>NAIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td>1.7E-1</td>
<td>6.4E-2</td>
<td></td>
</tr>
<tr>
<td>HYBRID</td>
<td>2.4E-2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>6.6E-3</td>
<td>2.6E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.5E-2</td>
<td>5.1E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.0E-2</td>
<td>2.8E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>2.1E-3</td>
<td>4.9E-3</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>3.5E-3</td>
<td>4.3E-3</td>
</tr>
<tr>
<td>Q^\pi(\lambda)</td>
<td>4.7E-3</td>
<td>5.0E-3</td>
</tr>
<tr>
<td>Tree</td>
<td>4.7E-2</td>
<td>4.9E-3</td>
</tr>
<tr>
<td>IH</td>
<td>3.3E-3</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.9E-3</td>
<td>1.2E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.0E-2</td>
<td>3.1E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.7E-2</td>
<td>1.4E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>1.1E-3</td>
<td>4.3E-3</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>2.2E-3</td>
<td>3.7E-3</td>
</tr>
<tr>
<td>Q^\pi(\lambda)</td>
<td>1.2E-3</td>
<td>3.8E-3</td>
</tr>
<tr>
<td>Tree</td>
<td>5.5E-3</td>
<td>4.2E-3</td>
</tr>
<tr>
<td>IH</td>
<td>2.1E-3</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>9.6E-4</td>
<td>5.6E-3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>7.0E-3</td>
<td>8.2E-4</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.1E-3</td>
<td>4.5E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>5.5E-4</td>
<td>9.0E-4</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>6.5E-4</td>
<td>8.1E-4</td>
</tr>
<tr>
<td>Q^\pi(\lambda)</td>
<td>7.3E-4</td>
<td>8.8E-4</td>
</tr>
<tr>
<td>Tree</td>
<td>3.0E-3</td>
<td>7.8E-4</td>
</tr>
<tr>
<td>IH</td>
<td>7.5E-4</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>2.9E-4</td>
<td>3.4E-3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.8E-3</td>
<td>9.9E-4</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.8E-3</td>
<td>1.1E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>2.9E-4</td>
<td>9.4E-4</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>6.8E-4</td>
<td>9.7E-4</td>
</tr>
<tr>
<td>Q^\pi(\lambda)</td>
<td>5.4E-4</td>
<td>9.7E-4</td>
</tr>
<tr>
<td>Tree</td>
<td>1.9E-3</td>
<td>8.3E-4</td>
</tr>
<tr>
<td>IH</td>
<td>1.7E-4</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 223: Graph, relative MSE. $T = 16, N = 512, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$. Stochastic rewards. Dense rewards.

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard</td>
</tr>
<tr>
<td>IS</td>
<td>7.6E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>9.2E-3</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.5E-1</td>
</tr>
</tbody>
</table>

Table 224: Graph, relative MSE. $T = 16, N = 1024, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$. Stochastic rewards. Dense rewards.

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard</td>
</tr>
<tr>
<td>IS</td>
<td>3.0E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>5.7E-3</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.5E-1</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.5E-1</td>
<td>7.2E-1</td>
<td>4.7E-1</td>
<td>1.5E-1</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>5.4E-1</td>
<td>5.1E-1</td>
<td>4.9E-1</td>
<td>5.4E-1</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>4.1E-1</td>
<td>1.2E0</td>
<td>6.6E-1</td>
<td>4.2E-1</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>1.7E-1</td>
<td>3.8E-1</td>
<td>2.2E-1</td>
<td>1.7E-1</td>
<td></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.7E-1</td>
<td>1.7E-1</td>
<td>1.7E-1</td>
<td>1.7E-1</td>
<td></td>
</tr>
<tr>
<td>Q'(\lambda)</td>
<td>1.9E-1</td>
<td>2.6E-1</td>
<td>2.0E-1</td>
<td>1.9E-1</td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>2.4E-1</td>
<td>2.1E-1</td>
<td>2.1E-1</td>
<td>2.3E-1</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>1.5E-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.2E-2</td>
<td>6.0E-1</td>
<td>4.5E-1</td>
<td>3.4E-2</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.8E-1</td>
<td>1.5E-1</td>
<td>2.2E-1</td>
<td>1.8E-1</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>1.3E-1</td>
<td>1.2E0</td>
<td>1.1E0</td>
<td>2.3E-1</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>2.3E-2</td>
<td>8.7E-2</td>
<td>7.4E-2</td>
<td>2.3E-2</td>
<td></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>3.2E-2</td>
<td>3.9E-2</td>
<td>4.3E-2</td>
<td>3.2E-2</td>
<td></td>
</tr>
<tr>
<td>Q'(\lambda)</td>
<td>6.0E-2</td>
<td>5.8E-2</td>
<td>6.2E-2</td>
<td>5.9E-2</td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>1.0E-1</td>
<td>9.5E-2</td>
<td>8.1E-2</td>
<td>8.7E-2</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>3.2E-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.1E-2</td>
<td>9.7E-2</td>
<td>1.4E-1</td>
<td>3.5E-2</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.4E-1</td>
<td>1.0E-1</td>
<td>1.0E-1</td>
<td>1.4E-1</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>1.4E-1</td>
<td>4.5E-1</td>
<td>3.6E-1</td>
<td>1.7E-1</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>2.0E-2</td>
<td>7.2E-2</td>
<td>5.7E-2</td>
<td>2.0E-2</td>
<td></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>4.0E-2</td>
<td>3.8E-2</td>
<td>3.9E-2</td>
<td>4.1E-2</td>
<td></td>
</tr>
<tr>
<td>Q'(\lambda)</td>
<td>3.7E-2</td>
<td>6.4E-2</td>
<td>4.3E-2</td>
<td>3.7E-2</td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>5.9E-2</td>
<td>5.1E-2</td>
<td>5.2E-2</td>
<td>6.2E-2</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>1.2E-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.5E-2</td>
<td>9.3E-2</td>
<td>1.2E-1</td>
<td>2.1E-2</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>6.1E-2</td>
<td>4.5E-2</td>
<td>4.5E-2</td>
<td>6.4E-2</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>7.1E-2</td>
<td>7.7E-2</td>
<td>6.9E-2</td>
<td>9.1E-2</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>1.3E-2</td>
<td>3.3E-2</td>
<td>4.1E-2</td>
<td>1.3E-2</td>
<td></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>2.9E-2</td>
<td>3.4E-2</td>
<td>3.9E-2</td>
<td>3.1E-2</td>
<td></td>
</tr>
<tr>
<td>Q'(\lambda)</td>
<td>3.0E-2</td>
<td>2.9E-2</td>
<td>3.9E-2</td>
<td>3.0E-2</td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>6.5E-2</td>
<td>3.6E-2</td>
<td>5.1E-2</td>
<td>6.3E-2</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>1.0E-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 227: Graph, relative MSE. $T = 16, N = 64, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$. Stochastic environment. Dense rewards.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.5E-1</td>
<td>4.8E-1</td>
<td>1.6E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.4E-1</td>
<td>5.7E-2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>4.5E-1</td>
<td>4.5E-1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>1.5E-1</td>
<td>5.7E-2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.4E-1</td>
<td>1.4E-1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q'(\lambda)</td>
<td>1.4E-1</td>
<td>1.4E-1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>1.4E-1</td>
<td>1.4E-1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>1.4E-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.3E-2</td>
<td>4.5E-2</td>
<td>3.7E-2</td>
<td>1.0E-2</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.9E-2</td>
<td>1.1E-2</td>
<td>1.2E-2</td>
<td>1.9E-2</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>4.1E-2</td>
<td>2.9E-2</td>
<td>2.5E-2</td>
<td>5.2E-2</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td><strong>6.5E-3</strong></td>
<td>9.9E-3</td>
<td>9.6E-3</td>
<td>6.7E-3</td>
<td></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>6.7E-3</td>
<td>9.3E-3</td>
<td>9.6E-3</td>
<td>7.3E-3</td>
<td></td>
</tr>
<tr>
<td>Q'(\lambda)</td>
<td>8.7E-3</td>
<td>8.5E-3</td>
<td>9.2E-3</td>
<td><strong>5.9E-3</strong></td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>5.4E-2</td>
<td>1.2E-2</td>
<td>1.1E-2</td>
<td>3.8E-2</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>8.9E-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.6E-1</td>
<td>2.8E-2</td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>6.2E-2</td>
<td><strong>1.2E-2</strong></td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.5E-1</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.0E-3</td>
<td>4.6E-2</td>
<td>5.3E-2</td>
<td>3.6E-2</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.3E-2</td>
<td>6.0E-3</td>
<td>6.2E-3</td>
<td>8.9E-3</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>2.8E-2</td>
<td>1.7E-2</td>
<td>1.8E-2</td>
<td>3.0E-2</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td><strong>2.5E-3</strong></td>
<td>7.1E-3</td>
<td>7.4E-3</td>
<td><strong>2.5E-3</strong></td>
<td></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>7.0E-3</td>
<td>7.1E-3</td>
<td>7.0E-3</td>
<td>7.0E-3</td>
<td></td>
</tr>
<tr>
<td>Q'(\lambda)</td>
<td>3.8E-3</td>
<td>7.7E-3</td>
<td>7.9E-3</td>
<td>3.8E-3</td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>3.4E-2</td>
<td>1.1E-2</td>
<td>1.1E-2</td>
<td>1.9E-2</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>3.7E-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>5.2E-2</td>
<td>1.5E-2</td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>1.7E-2</td>
<td><strong>1.2E-2</strong></td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.8E-1</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>8.8E-4</td>
<td>1.2E-2</td>
<td>1.2E-2</td>
<td>1.7E-3</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.5E-3</td>
<td>5.9E-3</td>
<td>5.9E-3</td>
<td>5.7E-3</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>4.1E-3</td>
<td>6.1E-3</td>
<td>6.2E-3</td>
<td>8.0E-3</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>1.1E-3</td>
<td>6.8E-3</td>
<td>6.4E-3</td>
<td><strong>1.2E-3</strong></td>
<td></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>3.3E-3</td>
<td>5.9E-3</td>
<td>5.7E-3</td>
<td>3.3E-3</td>
<td></td>
</tr>
<tr>
<td>Q'(\lambda)</td>
<td>3.7E-3</td>
<td>7.1E-3</td>
<td>6.8E-3</td>
<td>3.8E-3</td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>1.2E-2</td>
<td>5.7E-3</td>
<td>5.7E-3</td>
<td>1.2E-2</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>9.3E-4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>2.6E-2</td>
<td><strong>5.3E-3</strong></td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>1.3E-2</td>
<td>6.3E-3</td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.3E-1</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.1E-3</td>
<td>8.3E-3</td>
<td>8.7E-3</td>
<td>2.1E-3</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.8E-3</td>
<td>1.6E-3</td>
<td>1.6E-3</td>
<td>2.1E-3</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>3.1E-3</td>
<td>1.6E-3</td>
<td>1.5E-3</td>
<td>3.0E-3</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>7.7E-4</td>
<td>1.7E-3</td>
<td>1.7E-3</td>
<td><strong>7.5E-4</strong></td>
<td></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.6E-3</td>
<td>1.8E-3</td>
<td>1.8E-3</td>
<td>1.6E-3</td>
<td></td>
</tr>
<tr>
<td>Q'(\lambda)</td>
<td>9.0E-4</td>
<td>1.4E-3</td>
<td>1.5E-3</td>
<td>1.0E-3</td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>2.8E-3</td>
<td>1.9E-3</td>
<td>2.0E-3</td>
<td>9.0E-4</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>9.4E-4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.9E-2</td>
<td>5.0E-3</td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>5.4E-3</td>
<td><strong>1.7E-3</strong></td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.3E-1</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>3.4E-1</td>
<td>1.3E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>6.3E-1</td>
<td>4.7E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.6E-1</td>
<td>2.6E0</td>
</tr>
<tr>
<td>FQE</td>
<td>3.6E-1</td>
<td>4.3E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>4.5E-1</td>
<td>4.7E-1</td>
</tr>
<tr>
<td>Q^\theta(\lambda)</td>
<td>5.8E-1</td>
<td>6.4E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>4.7E-1</td>
<td>3.9E-1</td>
</tr>
<tr>
<td>IH</td>
<td><strong>2.3E-1</strong></td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E0</td>
<td>7.5E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>6.5E-1</td>
<td><strong>4.4E-1</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.2E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>2.3E-1</td>
<td>8.6E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.7E-1</td>
<td>4.6E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.3E-1</td>
<td>1.1E0</td>
</tr>
<tr>
<td>FQE</td>
<td>1.5E-1</td>
<td>5.2E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>2.5E-1</td>
<td>2.8E-1</td>
</tr>
<tr>
<td>Q^\theta(\lambda)</td>
<td>3.0E-1</td>
<td>5.6E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>3.1E-1</td>
<td>2.3E-1</td>
</tr>
<tr>
<td>IH</td>
<td><strong>9.8E-2</strong></td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.5E0</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>5.1E-1</td>
<td><strong>1.5E-1</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>6.2E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>6.8E-2</td>
<td>7.0E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.3E-1</td>
<td>8.7E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.5E-1</td>
<td>3.5E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>3.0E-2</td>
<td>2.5E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>7.0E-2</td>
<td>1.1E-1</td>
</tr>
<tr>
<td>Q^\theta(\lambda)</td>
<td>7.4E-2</td>
<td>3.0E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>1.3E-1</td>
<td>8.5E-2</td>
</tr>
<tr>
<td>IH</td>
<td>5.5E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.7E0</td>
<td>3.5E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>5.4E-1</td>
<td><strong>1.4E-1</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.2E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>2.6E-2</td>
<td>2.9E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>6.5E-2</td>
<td>3.7E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>7.7E-2</td>
<td>1.6E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>1.6E-2</td>
<td>2.7E-2</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>2.0E-2</td>
<td>2.6E-2</td>
</tr>
<tr>
<td>Q^\theta(\lambda)</td>
<td>2.3E-2</td>
<td>3.8E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>8.8E-2</td>
<td>2.4E-2</td>
</tr>
<tr>
<td>IH</td>
<td><strong>1.4E-2</strong></td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>6.6E-1</td>
<td>6.9E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>1.3E-1</td>
<td><strong>2.2E-2</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.8E-1</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>6.0E-3</td>
<td>1.2E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.2E-1</td>
<td>1.2E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.7E-2</td>
<td>2.0E-1</td>
</tr>
<tr>
<td>FQE</td>
<td><strong>2.6E-3</strong></td>
<td>5.2E-2</td>
</tr>
<tr>
<td>R(λ)</td>
<td>1.2E-2</td>
<td>2.6E-2</td>
</tr>
<tr>
<td>Q(λ)</td>
<td>1.1E-2</td>
<td>4.2E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>4.5E-2</td>
<td>3.6E-2</td>
</tr>
<tr>
<td>IH</td>
<td>4.1E-3</td>
<td>-</td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>3.0E-1</td>
<td><strong>1.1E-1</strong></td>
</tr>
<tr>
<td>WIS</td>
<td>4.4E-2</td>
<td><strong>4.1E-2</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.5E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>4.3E-3</td>
<td>3.0E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.4E-2</td>
<td>1.4E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>8.3E-3</td>
<td>6.6E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>5.0E-3</td>
<td>1.2E-2</td>
</tr>
<tr>
<td>R(λ)</td>
<td>8.8E-3</td>
<td>1.3E-2</td>
</tr>
<tr>
<td>Q(λ)</td>
<td>1.3E-2</td>
<td>1.3E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>9.4E-3</td>
<td>1.3E-2</td>
</tr>
<tr>
<td>IH</td>
<td><strong>4.2E-3</strong></td>
<td>-</td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E-1</td>
<td>1.5E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>2.1E-2</td>
<td><strong>1.2E-2</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.2E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>4.6E-3</td>
<td>4.6E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>8.3E-3</td>
<td>1.4E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.3E-2</td>
<td>2.5E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>3.8E-3</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>R(λ)</td>
<td>9.0E-3</td>
<td>1.3E-2</td>
</tr>
<tr>
<td>Q(λ)</td>
<td>6.3E-3</td>
<td>1.4E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>3.1E-2</td>
<td>1.5E-2</td>
</tr>
<tr>
<td>IH</td>
<td><strong>3.5E-3</strong></td>
<td>-</td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.4E-1</td>
<td><strong>1.1E-2</strong></td>
</tr>
<tr>
<td>WIS</td>
<td>6.6E-2</td>
<td>1.2E-2</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.4E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>2.3E-3</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>6.0E-3</td>
<td>5.0E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.2E-3</td>
<td>3.4E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>1.1E-3</td>
<td>5.3E-3</td>
</tr>
<tr>
<td>R(λ)</td>
<td>2.7E-3</td>
<td>5.1E-3</td>
</tr>
<tr>
<td>Q(λ)</td>
<td>2.0E-3</td>
<td>5.0E-3</td>
</tr>
<tr>
<td>Tree</td>
<td>1.3E-2</td>
<td>5.0E-3</td>
</tr>
<tr>
<td>IH</td>
<td><strong>9.8E-4</strong></td>
<td>-</td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>9.2E-3</td>
<td>5.4E-3</td>
</tr>
<tr>
<td>WIS</td>
<td>7.3E-3</td>
<td><strong>5.4E-3</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.6E-1</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>6.5E-1</td>
<td>1.3E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.7E0</td>
<td>3.2E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>8.7E-1</td>
<td>6.3E0</td>
</tr>
<tr>
<td>FQE</td>
<td>4.3E-1</td>
<td>4.3E-1</td>
</tr>
<tr>
<td>$\mathbb{R}(\lambda)$</td>
<td>4.6E-1</td>
<td>4.5E-1</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td><strong>4.3E-1</strong></td>
<td><strong>4.3E-1</strong></td>
</tr>
<tr>
<td>TREE</td>
<td>9.8E-1</td>
<td>7.1E-1</td>
</tr>
<tr>
<td>IH</td>
<td>6.9E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 240: Graph, relative MSE. $T = 16, N = 16$, $\pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$. Sparse rewards.

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard</td>
</tr>
<tr>
<td>IS</td>
<td>1.0E0</td>
</tr>
<tr>
<td>WIS</td>
<td><strong>7.0E-1</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>2.6E-1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.2E-1</td>
<td>2.0E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>5.6E-1</td>
<td>1.8E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.4E-1</td>
<td>2.2E0</td>
</tr>
<tr>
<td>FQE</td>
<td><strong>4.6E-2</strong></td>
<td>4.6E-2</td>
</tr>
<tr>
<td>$\mathbb{R}(\lambda)$</td>
<td>4.6E-2</td>
<td>4.6E-2</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>4.6E-2</td>
<td>4.6E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>9.9E-1</td>
<td>5.7E-1</td>
</tr>
<tr>
<td>IH</td>
<td>6.2E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 241: Graph, relative MSE. $T = 16, N = 32$, $\pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$. Sparse rewards.

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard</td>
</tr>
<tr>
<td>IS</td>
<td>1.0E0</td>
</tr>
<tr>
<td>WIS</td>
<td><strong>7.0E-1</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>2.6E-1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>3.4E-2</td>
<td>5.0E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.6E0</td>
<td>5.3E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.1E0</td>
<td>8.7E0</td>
</tr>
<tr>
<td>FQE</td>
<td>3.6E-3</td>
<td>3.6E-3</td>
</tr>
<tr>
<td>$\mathbb{R}(\lambda)$</td>
<td>4.6E-3</td>
<td>4.5E-3</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td><strong>3.6E-3</strong></td>
<td>3.6E-3</td>
</tr>
<tr>
<td>TREE</td>
<td>9.1E-1</td>
<td>8.5E-1</td>
</tr>
<tr>
<td>IH</td>
<td>6.1E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 242: Graph, relative MSE. $T = 16, N = 64$, $\pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.8$. Sparse rewards.

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard</td>
</tr>
<tr>
<td>IS</td>
<td>1.9E0</td>
</tr>
<tr>
<td>WIS</td>
<td><strong>1.6E-1</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.3E-1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.3E-2</td>
<td>3.2E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.8E-1</td>
<td>1.1E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.8E-1</td>
<td>7.2E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>2.0E-6</td>
<td>2.0E-6</td>
</tr>
<tr>
<td>$\mathbb{R}(\lambda)$</td>
<td>4.4E-5</td>
<td>3.2E-5</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td><strong>2.0E-6</strong></td>
<td>2.0E-6</td>
</tr>
<tr>
<td>TREE</td>
<td>9.7E-1</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>IH</td>
<td>6.2E-1</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 243: Graph, relative MSE. $T = 16$, $N = 128$, $\pi_b(a = 0) = 0.6$, $\pi_e(a = 0) = 0.8$. Sparse rewards.

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>8.8E-3</td>
<td>6.9E-1</td>
<td>6.1E-1</td>
<td>3.1E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.7E-1</td>
<td>4.7E-2</td>
<td>2.4E-2</td>
<td>1.1E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.5E-1</td>
<td>4.0E-1</td>
<td>2.9E-1</td>
<td>3.1E-1</td>
</tr>
<tr>
<td>FQE</td>
<td><strong>2.2E-7</strong></td>
<td>2.2E-7</td>
<td>2.2E-7</td>
<td>2.2E-7</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.2E-5</td>
<td>9.7E-6</td>
<td>1.0E-5</td>
<td>1.2E-5</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>2.3E-7</td>
<td>2.2E-7</td>
<td>2.2E-7</td>
<td>2.3E-7</td>
</tr>
<tr>
<td>TREE</td>
<td>9.4E-1</td>
<td>1.6E-1</td>
<td>7.0E-2</td>
<td>1.9E-1</td>
</tr>
<tr>
<td>INF</td>
<td>6.1E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 245: Graph, relative MSE. $T = 16$, $N = 256$, $\pi_b(a = 0) = 0.6$, $\pi_e(a = 0) = 0.8$. Sparse rewards.

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.2E-3</td>
<td>6.4E-2</td>
<td>5.8E-2</td>
<td>3.4E-3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.8E-2</td>
<td>2.4E-3</td>
<td>1.7E-3</td>
<td>4.4E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.3E-2</td>
<td>5.0E-2</td>
<td>6.0E-2</td>
<td>6.9E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>2.2E-5</td>
<td>2.2E-5</td>
<td>2.2E-5</td>
<td>2.2E-5</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>2.3E-5</td>
<td>2.0E-5</td>
<td>2.0E-5</td>
<td>2.3E-5</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>2.2E-5</td>
<td>2.2E-5</td>
<td>2.2E-5</td>
<td>2.2E-5</td>
</tr>
<tr>
<td>TREE</td>
<td>9.6E-1</td>
<td>4.5E-2</td>
<td>2.2E-2</td>
<td>2.2E-2</td>
</tr>
<tr>
<td>INF</td>
<td>3.8E-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 244: Graph, relative MSE. $T = 16$, $N = 512$, $\pi_b(a = 0) = 0.6$, $\pi_e(a = 0) = 0.8$. Sparse rewards.

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.2E-3</td>
<td>6.4E-2</td>
<td>5.8E-2</td>
<td>3.4E-3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.8E-2</td>
<td>2.4E-3</td>
<td>1.7E-3</td>
<td>4.4E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.3E-2</td>
<td>5.0E-2</td>
<td>6.0E-2</td>
<td>6.9E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>2.2E-5</td>
<td>2.2E-5</td>
<td>2.2E-5</td>
<td>2.2E-5</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>2.3E-5</td>
<td>2.0E-5</td>
<td>2.0E-5</td>
<td>2.3E-5</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>2.2E-5</td>
<td>2.2E-5</td>
<td>2.2E-5</td>
<td>2.2E-5</td>
</tr>
<tr>
<td>TREE</td>
<td>9.6E-1</td>
<td>4.5E-2</td>
<td>2.2E-2</td>
<td>2.2E-2</td>
</tr>
<tr>
<td>INF</td>
<td>3.8E-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 246: Graph, relative MSE. $T = 16$, $N = 1024$, $\pi_b(a = 0) = 0.6$, $\pi_e(a = 0) = 0.8$. Sparse rewards.

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.3E-3</td>
<td>3.0E-2</td>
<td>2.6E-2</td>
<td>8.3E-4</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.4E-2</td>
<td>5.7E-4</td>
<td>3.2E-4</td>
<td>1.1E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.0E-2</td>
<td>1.6E-2</td>
<td>1.1E-2</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>FQE</td>
<td><strong>2.5E-5</strong></td>
<td>2.5E-5</td>
<td>2.5E-5</td>
<td>2.5E-5</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>2.5E-5</td>
<td>2.7E-5</td>
<td>2.8E-5</td>
<td>2.5E-5</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>2.5E-5</td>
<td><strong>2.5E-5</strong></td>
<td>2.5E-5</td>
<td>2.5E-5</td>
</tr>
<tr>
<td>TREE</td>
<td>1.0E0</td>
<td>2.5E-2</td>
<td>1.4E-2</td>
<td>1.4E-2</td>
</tr>
<tr>
<td>INF</td>
<td>2.0E-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 247: Graph, relative MSE. $T = 16$, $N = 2048$, $\pi_b(a = 0) = 0.6$, $\pi_e(a = 0) = 0.8$. Sparse rewards.

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.3E-3</td>
<td>3.0E-2</td>
<td>2.6E-2</td>
<td>8.3E-4</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.4E-2</td>
<td>5.7E-4</td>
<td>3.2E-4</td>
<td>1.1E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.0E-2</td>
<td>1.6E-2</td>
<td>1.1E-2</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>FQE</td>
<td><strong>2.5E-5</strong></td>
<td>2.5E-5</td>
<td>2.5E-5</td>
<td>2.5E-5</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>2.5E-5</td>
<td>2.7E-5</td>
<td>2.8E-5</td>
<td>2.5E-5</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>2.5E-5</td>
<td><strong>2.5E-5</strong></td>
<td>2.5E-5</td>
<td>2.5E-5</td>
</tr>
<tr>
<td>TREE</td>
<td>1.0E0</td>
<td>2.5E-2</td>
<td>1.4E-2</td>
<td>1.4E-2</td>
</tr>
<tr>
<td>INF</td>
<td>2.0E-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>8.4E0</td>
<td>9.8E1</td>
<td>4.9E1</td>
<td>8.9E0</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.5E1</td>
<td>2.1E1</td>
<td>1.5E1</td>
<td>1.5E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>6.6E0</td>
<td>1.2E1</td>
<td>8.7E0</td>
<td>6.6E0</td>
</tr>
<tr>
<td>FQE</td>
<td>5.8E0</td>
<td>1.1E1</td>
<td>1.1E1</td>
<td>5.8E0</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>8.3E0</td>
<td>8.3E0</td>
<td>8.3E0</td>
<td>8.3E0</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>1.0E1</td>
<td>1.1E1</td>
<td>9.3E0</td>
<td>1.0E1</td>
</tr>
<tr>
<td>Tree</td>
<td>7.8E0</td>
<td>9.3E0</td>
<td>1.0E1</td>
<td>7.8E0</td>
</tr>
<tr>
<td>IH</td>
<td>7.8E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.1E1</td>
<td>1.5E1</td>
</tr>
<tr>
<td>WIS</td>
<td>2.0E1</td>
<td>-</td>
</tr>
<tr>
<td>NAIVE</td>
<td>6.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>6.4E0</td>
<td>5.9E1</td>
<td>4.9E1</td>
<td>1.3E1</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>2.8E0</td>
<td>3.8E0</td>
<td>3.2E0</td>
<td>2.8E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.9E0</td>
<td>1.9E0</td>
<td>1.1E0</td>
<td>1.9E0</td>
</tr>
<tr>
<td>FQE</td>
<td>5.9E0</td>
<td>1.1E1</td>
<td>9.5E0</td>
<td>5.9E0</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>6.3E0</td>
<td>5.5E0</td>
<td>6.1E0</td>
<td>6.3E0</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>7.1E0</td>
<td>8.0E0</td>
<td>7.9E0</td>
<td>7.1E0</td>
</tr>
<tr>
<td>Tree</td>
<td>3.8E0</td>
<td>3.7E0</td>
<td>4.7E0</td>
<td>4.1E0</td>
</tr>
<tr>
<td>IH</td>
<td>2.9E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>4.1E0</td>
<td>3.2E0</td>
</tr>
<tr>
<td>WIS</td>
<td>9.0E0</td>
<td>3.2E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.5E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>5.9E0</td>
<td>2.1E1</td>
<td>2.1E1</td>
<td>6.5E0</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.4E0</td>
<td>7.3E0</td>
<td>4.9E0</td>
<td>1.4E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.0E0</td>
<td>8.8E0</td>
<td>4.6E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>FQE</td>
<td>5.2E0</td>
<td>7.4E0</td>
<td>3.7E0</td>
<td>5.2E0</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>3.3E0</td>
<td>4.1E0</td>
<td>3.3E0</td>
<td>3.3E0</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>3.8E0</td>
<td>9.6E0</td>
<td>3.6E0</td>
<td>3.8E0</td>
</tr>
<tr>
<td>Tree</td>
<td>4.0E0</td>
<td>2.1E0</td>
<td>2.4E0</td>
<td>4.0E0</td>
</tr>
<tr>
<td>IH</td>
<td>4.6E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>6.8E0</td>
<td>1.5E0</td>
</tr>
<tr>
<td>WIS</td>
<td>9.5E0</td>
<td>1.9E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>2.3E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.8E0</td>
<td>1.3E1</td>
<td>1.2E1</td>
<td>3.7E0</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>2.3E0</td>
<td>2.6E0</td>
<td>2.5E0</td>
<td>2.2E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.7E0</td>
<td>2.9E0</td>
<td>2.8E0</td>
<td>1.6E0</td>
</tr>
<tr>
<td>FQE</td>
<td>1.1E0</td>
<td>2.8E0</td>
<td>2.5E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.2E0</td>
<td>2.0E0</td>
<td>2.0E0</td>
<td>1.2E0</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>1.7E0</td>
<td>2.6E0</td>
<td>2.6E0</td>
<td>1.7E0</td>
</tr>
<tr>
<td>Tree</td>
<td>1.4E0</td>
<td>2.1E0</td>
<td>2.0E0</td>
<td>1.4E0</td>
</tr>
<tr>
<td>IH</td>
<td>1.9E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>7.0E0</td>
<td>2.3E0</td>
</tr>
<tr>
<td>WIS</td>
<td>7.7E0</td>
<td>2.0E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>2.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>
### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>9.5E-1</td>
<td>1.9E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>9.8E-1</td>
<td>1.4E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.1E0</td>
<td>2.8E0</td>
</tr>
<tr>
<td>FQE</td>
<td>9.1E-1</td>
<td>1.5E0</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>9.9E-1</td>
<td>1.2E0</td>
</tr>
<tr>
<td>$Q^*(\lambda)$</td>
<td>7.5E-1</td>
<td>1.3E0</td>
</tr>
<tr>
<td>TREE</td>
<td>1.2E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>IH</td>
<td>1.5E0</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IS</th>
<th>WIS</th>
<th>NAIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>3.7E0</td>
<td>4.0E0</td>
<td>1.2E0</td>
</tr>
<tr>
<td>Per-Decision</td>
<td>1.1E0</td>
<td>8.9E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.6E-1</td>
<td>4.4E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.1E0</td>
<td>1.4E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.4E-1</td>
<td>7.8E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>3.8E-1</td>
<td>1.3E0</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>6.6E-1</td>
<td>1.1E0</td>
</tr>
<tr>
<td>$Q^*(\lambda)$</td>
<td>8.0E-1</td>
<td>1.4E0</td>
</tr>
<tr>
<td>TREE</td>
<td>1.0E0</td>
<td>9.8E-1</td>
</tr>
<tr>
<td>IH</td>
<td>3.6E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IS</th>
<th>WIS</th>
<th>NAIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>3.1E0</td>
<td>2.5E0</td>
<td>7.1E-1</td>
</tr>
<tr>
<td>Per-Decision</td>
<td>1.1E0</td>
<td>1.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.7E-1</td>
<td>1.2E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.2E-1</td>
<td>2.5E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.9E-1</td>
<td>2.8E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>1.0E-1</td>
<td>2.2E-1</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>2.1E-1</td>
<td>2.3E-1</td>
</tr>
<tr>
<td>$Q^*(\lambda)$</td>
<td>1.9E-1</td>
<td>2.3E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>9.8E-1</td>
<td>3.0E-1</td>
</tr>
<tr>
<td>IH</td>
<td>1.1E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IS</th>
<th>WIS</th>
<th>NAIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>1.4E0</td>
<td>1.2E0</td>
<td>6.3E-1</td>
</tr>
<tr>
<td>Per-Decision</td>
<td>3.3E-1</td>
<td>2.9E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.2E-1</td>
<td>9.3E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.0E-1</td>
<td>4.2E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.9E-1</td>
<td>3.2E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>1.0E-1</td>
<td>4.0E-1</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.7E-1</td>
<td>4.0E-1</td>
</tr>
<tr>
<td>$Q^*(\lambda)$</td>
<td>1.5E-1</td>
<td>4.0E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>7.1E-1</td>
<td>3.7E-1</td>
</tr>
<tr>
<td>IH</td>
<td>1.0E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IS</th>
<th>WIS</th>
<th>NAIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>8.7E-1</td>
<td>9.6E-1</td>
<td>3.5E-1</td>
</tr>
<tr>
<td>Per-Decision</td>
<td>3.8E-1</td>
<td>4.2E-1</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.0E0</td>
<td>1.4E1</td>
<td>1.5E1</td>
<td>1.5E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.9E1</td>
<td>5.1E1</td>
<td>4.1E1</td>
<td>3.0E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>3.0E0</td>
<td>2.7E1</td>
<td>2.0E1</td>
<td>8.9E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>1.1E0</td>
<td>6.4E0</td>
<td>4.8E0</td>
<td>1.3E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.2E0</td>
<td>1.2E0</td>
<td>1.3E0</td>
<td>1.2E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q$^2$($\lambda$)</td>
<td>2.0E0</td>
<td>4.3E0</td>
<td>3.0E0</td>
<td>1.4E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>1.0E0</td>
<td>4.0E0</td>
<td>3.8E0</td>
<td>1.3E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>6.5E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td>8.7E0</td>
<td>8.7E0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>5.2E0</td>
<td>5.2E0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>1.7E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>5.8E-1</td>
<td>1.5E1</td>
<td>1.0E1</td>
<td>3.8E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.1E0</td>
<td>3.5E0</td>
<td>2.7E0</td>
<td>2.8E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>1.6E0</td>
<td>4.5E0</td>
<td>3.1E0</td>
<td>3.4E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>4.4E-1</td>
<td>1.9E0</td>
<td>2.0E0</td>
<td>4.5E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.4E0</td>
<td>1.7E0</td>
<td>1.7E0</td>
<td>1.4E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q$^2$($\lambda$)</td>
<td>9.7E-1</td>
<td>1.9E0</td>
<td>1.6E0</td>
<td>1.0E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>9.4E-1</td>
<td>1.9E0</td>
<td>2.7E0</td>
<td>9.4E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>7.4E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td>2.1E0</td>
<td>2.1E0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>2.5E0</td>
<td>2.5E0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>1.7E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td>1.0E0</td>
<td>1.0E0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>8.7E-1</td>
<td>8.7E-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.9E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.3E-1</td>
<td>4.4E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>9.4E-1</td>
<td>7.4E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>7.9E-1</td>
<td>3.9E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>6.3E-2</td>
<td>1.2E0</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>3.6E-1</td>
<td>7.2E-1</td>
</tr>
<tr>
<td>Q$^\theta$(\lambda)</td>
<td>2.5E-1</td>
<td>1.1E0</td>
</tr>
<tr>
<td>TREE</td>
<td>9.6E-1</td>
<td>8.0E-1</td>
</tr>
<tr>
<td>IH</td>
<td>5.3E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>6.2E-2</td>
<td>4.7E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.4E-1</td>
<td>1.4E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.8E-1</td>
<td>4.8E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>3.5E-2</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>8.3E-2</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>Q$^\theta$(\lambda)</td>
<td>1.0E-1</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>9.9E-1</td>
<td>2.5E-1</td>
</tr>
<tr>
<td>IH</td>
<td>3.4E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>0.0E-1</td>
<td>0.0E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>0.0E-1</td>
<td>0.0E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>0.0E-1</td>
<td>0.0E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>0.0E-1</td>
<td>0.0E-1</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>0.0E-1</td>
<td>0.0E-1</td>
</tr>
<tr>
<td>Q$^\theta$(\lambda)</td>
<td>0.0E-1</td>
<td>0.0E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>0.0E-1</td>
<td>0.0E-1</td>
</tr>
<tr>
<td>IH</td>
<td>0.0E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>0.0E-1</td>
<td>0.0E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>0.0E-1</td>
<td>0.0E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>0.0E-1</td>
<td>0.0E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>0.0E-1</td>
<td>0.0E-1</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>0.0E-1</td>
<td>0.0E-1</td>
</tr>
<tr>
<td>Q$^\theta$(\lambda)</td>
<td>0.0E-1</td>
<td>0.0E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>0.0E-1</td>
<td>0.0E-1</td>
</tr>
<tr>
<td>IH</td>
<td>0.0E-1</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>4.8E1</td>
<td>2.2E2</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.3E2</td>
<td>3.2E2</td>
</tr>
<tr>
<td>MRDR</td>
<td>6.7E1</td>
<td>4.4E2</td>
</tr>
<tr>
<td>FQE</td>
<td><strong>2.6E1</strong></td>
<td>1.1E2</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>4.2E1</td>
<td>5.2E1</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>3.8E1</td>
<td>8.1E1</td>
</tr>
<tr>
<td>Tree</td>
<td>3.3E1</td>
<td>7.7E1</td>
</tr>
<tr>
<td>IH</td>
<td>5.8E1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard</td>
</tr>
<tr>
<td>IS</td>
<td>1.3E2</td>
</tr>
<tr>
<td>WIS</td>
<td>1.5E2</td>
</tr>
<tr>
<td>NAIVE</td>
<td>2.8E1</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>5.2E0</td>
<td>1.5E2</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>6.5E1</td>
<td>5.4E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.3E1</td>
<td>3.2E2</td>
</tr>
<tr>
<td>FQE</td>
<td>3.9E0</td>
<td>3.8E1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>9.8E0</td>
<td>1.5E1</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>1.4E1</td>
<td>3.1E1</td>
</tr>
<tr>
<td>Tree</td>
<td>2.3E1</td>
<td>2.6E1</td>
</tr>
<tr>
<td>IH</td>
<td><strong>3.8E0</strong></td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard</td>
</tr>
<tr>
<td>IS</td>
<td>5.5E1</td>
</tr>
<tr>
<td>WIS</td>
<td>3.6E1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.0E0</td>
</tr>
</tbody>
</table>

Table 267: 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.

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard</td>
</tr>
<tr>
<td>IS</td>
<td>1.3E1</td>
</tr>
<tr>
<td>WIS</td>
<td>2.0E1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.9E0</td>
</tr>
</tbody>
</table>
### 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.

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>6.1E0</td>
<td>2.7E1</td>
<td>3.0E1</td>
<td>3.8E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td><strong>2.5E0</strong></td>
<td>6.2E0</td>
<td>5.2E0</td>
<td>2.4E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.9E0</td>
<td>6.3E0</td>
<td>5.8E0</td>
<td><strong>1.6E0</strong></td>
</tr>
<tr>
<td>FQE</td>
<td>2.7E0</td>
<td>3.5E0</td>
<td>3.3E0</td>
<td>2.6E0</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>3.1E0</td>
<td>3.7E0</td>
<td>3.5E0</td>
<td>2.9E0</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>5.1E0</td>
<td>5.0E0</td>
<td>4.4E0</td>
<td>5.1E0</td>
</tr>
<tr>
<td>TREE</td>
<td>3.1E0</td>
<td>2.4E0</td>
<td>2.3E0</td>
<td>2.9E0</td>
</tr>
<tr>
<td>IH</td>
<td>3.2E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.5E0</td>
<td>7.6E0</td>
<td>8.0E0</td>
<td>1.6E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.1E0</td>
<td>3.0E0</td>
<td>3.0E0</td>
<td>1.7E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.9E0</td>
<td>1.7E0</td>
<td>1.8E0</td>
<td>1.7E0</td>
</tr>
<tr>
<td>FQE</td>
<td><strong>4.5E-1</strong></td>
<td>2.2E0</td>
<td>2.5E0</td>
<td><strong>4.8E-1</strong></td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.8E0</td>
<td>2.7E0</td>
<td>2.8E0</td>
<td>1.8E0</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>1.4E0</td>
<td>2.3E0</td>
<td>2.5E0</td>
<td>1.3E0</td>
</tr>
<tr>
<td>TREE</td>
<td>1.2E0</td>
<td>3.0E0</td>
<td>3.1E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>IH</td>
<td><strong>4.5E-1</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>8.5E-1</td>
<td>1.0E1</td>
<td>1.0E1</td>
<td>5.1E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.4E0</td>
<td>1.2E0</td>
<td>1.3E0</td>
<td>1.3E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>7.9E-1</td>
<td>1.5E0</td>
<td>1.4E0</td>
<td>7.6E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.4E-1</td>
<td>1.1E0</td>
<td>1.1E0</td>
<td><strong>4.3E-1</strong></td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>4.3E-1</td>
<td>1.1E0</td>
<td>1.1E0</td>
<td>4.4E-1</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>5.0E-1</td>
<td>1.1E0</td>
<td>1.1E0</td>
<td>4.6E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>1.2E0</td>
<td>1.0E0</td>
<td>1.1E0</td>
<td>7.9E-1</td>
</tr>
<tr>
<td>IH</td>
<td><strong>4.2E-1</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 269: 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.

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>5.5E-1</td>
<td>9.0E-1</td>
<td>8.9E-1</td>
<td>6.2E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>5.9E-1</td>
<td>5.3E-1</td>
<td>5.6E-1</td>
<td>5.7E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>6.7E-1</td>
<td>5.7E-1</td>
<td>6.2E-1</td>
<td>6.5E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>3.6E-1</td>
<td>5.9E-1</td>
<td>5.9E-1</td>
<td><strong>3.4E-1</strong></td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>5.9E-1</td>
<td>6.0E-1</td>
<td>5.8E-1</td>
<td>5.9E-1</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>4.2E-1</td>
<td>5.8E-1</td>
<td>5.6E-1</td>
<td>4.3E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>1.5E0</td>
<td>6.0E-1</td>
<td>5.8E-1</td>
<td>1.4E0</td>
</tr>
<tr>
<td>IH</td>
<td><strong>3.2E-1</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>5.8E0</td>
<td>5.9E-1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Q-REG</td>
<td>6.2E0</td>
<td>5.8E-1</td>
<td><strong>5.6E-1</strong></td>
<td>-</td>
</tr>
<tr>
<td>MRDR</td>
<td>8.9E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FQE</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TREE</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IH</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
F.2 Detailed Results for Graph-POMDP

Table 271: Graph-POMDP, relative MSE. $T = 2, N = 256, H = 2, \pi_b(a = 0) = 0.2, \pi_c(a = 0) = 0.8$. Dense rewards.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
<td>DR</td>
</tr>
<tr>
<td>DM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AM</td>
<td>9.4E-1</td>
<td>1.2E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.5E-1</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.4E0</td>
<td>1.2E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>8.7E-1</td>
<td>2.3E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>5.1E-1</td>
<td>1.0E-2</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>5.3E-2</td>
<td>9.2E-3</td>
</tr>
<tr>
<td>TREE</td>
<td>3.8E-1</td>
<td>8.2E-3</td>
</tr>
<tr>
<td>IH</td>
<td>8.4E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 272: Graph-POMDP, relative MSE. $T = 2, N = 512, H = 2, \pi_b(a = 0) = 0.2, \pi_c(a = 0) = 0.8$. Stochastic rewards. Dense rewards.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.1E0</td>
<td>1.9E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.8E-1</td>
<td>5.7E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.7E0</td>
<td>5.4E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>1.0E0</td>
<td>8.1E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>6.7E-1</td>
<td>6.4E-2</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>1.0E-1</td>
<td>5.7E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>5.3E-1</td>
<td>6.2E-2</td>
</tr>
<tr>
<td>IH</td>
<td>1.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 273: Graph-POMDP, relative MSE. $T = 2, N = 1024, H = 2, \pi_b(a = 0) = 0.2, \pi_c(a = 0) = 0.8$. Dense rewards.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>9.7E-1</td>
<td>6.7E-3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.7E-1</td>
<td>3.1E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.5E0</td>
<td>3.0E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>1.0E0</td>
<td>7.8E-3</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>6.3E-1</td>
<td>5.0E-3</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>4.0E-1</td>
<td>4.5E-3</td>
</tr>
</tbody>
</table>

Table 274: Graph-POMDP, relative MSE. $T = 2, N = 256, H = 2, \pi_b(a = 0) = 0.2, \pi_c(a = 0) = 0.8$. Stochastic rewards. Dense rewards.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.1E0</td>
<td>1.9E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.8E-1</td>
<td>5.7E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.7E0</td>
<td>5.4E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>1.0E0</td>
<td>8.1E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>6.7E-1</td>
<td>6.4E-2</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>1.0E-1</td>
<td>5.7E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>5.3E-1</td>
<td>6.2E-2</td>
</tr>
<tr>
<td>IH</td>
<td>1.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 275: Graph-POMDP, relative MSE. $T = 2, N = 512, H = 2, \pi_b(a = 0) = 0.2, \pi_c(a = 0) = 0.8$. Stochastic rewards. Dense rewards.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>9.8E-2</td>
<td>6.0E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>9.7E-2</td>
<td>5.3E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.0E0</td>
<td>-</td>
</tr>
<tr>
<td>FQE</td>
<td>6.3E-2</td>
<td>2.0E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.5E-2</td>
<td>4.3E-3</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>4.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.1E0</td>
<td>4.7E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.0E-1</td>
<td>1.4E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.5E0</td>
<td>1.6E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>1.2E0</td>
<td>1.4E-2</td>
</tr>
<tr>
<td>R(λ)</td>
<td>7.3E-1</td>
<td>9.8E-3</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td><strong>1.3E-1</strong></td>
<td>1.2E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>5.9E-1</td>
<td>9.5E-3</td>
</tr>
<tr>
<td>IH</td>
<td>1.2E0</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IPS Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>5.5E-2</td>
<td>1.7E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>2.7E-2</td>
<td><strong>8.6E-3</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.1E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.6E0</td>
<td>6.5E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>5.7E-1</td>
<td>3.1E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.3E0</td>
<td>3.0E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>1.7E0</td>
<td>4.2E-1</td>
</tr>
<tr>
<td>R(λ)</td>
<td>1.2E0</td>
<td>3.7E-1</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td><strong>4.0E-1</strong></td>
<td>2.9E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>1.1E0</td>
<td>3.5E-1</td>
</tr>
<tr>
<td>IH</td>
<td>1.8E0</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IPS Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>8.2E-1</td>
<td>4.1E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>7.3E-1</td>
<td><strong>3.6E-1</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.6E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>9.6E-1</td>
<td>4.2E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.7E-1</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.5E0</td>
<td><strong>1.0E-2</strong></td>
</tr>
<tr>
<td>FQE</td>
<td>1.0E0</td>
<td>2.1E-2</td>
</tr>
<tr>
<td>R(λ)</td>
<td>6.1E-1</td>
<td>1.6E-2</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td><strong>6.4E-2</strong></td>
<td>1.0E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>4.7E-1</td>
<td>1.4E-2</td>
</tr>
<tr>
<td>IH</td>
<td>1.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IPS Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>5.0E-2</td>
<td>2.3E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>3.3E-2</td>
<td><strong>1.4E-2</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.1E0</td>
<td>1.2E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.7E-1</td>
<td>5.5E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.8E0</td>
<td>5.6E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>1.2E0</td>
<td>5.7E-2</td>
</tr>
<tr>
<td>R(λ)</td>
<td>7.8E-1</td>
<td>5.4E-2</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td><strong>1.2E-1</strong></td>
<td>5.4E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>6.3E-1</td>
<td>5.3E-2</td>
</tr>
<tr>
<td>IH</td>
<td>1.2E0</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IPS Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.5E-1</td>
<td>5.9E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>1.4E-1</td>
<td><strong>5.6E-2</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.3E0</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.0E0</td>
<td>1.4E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.8E-1</td>
<td>6.7E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.7E0</td>
<td>6.6E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>1.1E0</td>
<td>1.1E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>6.9E-1</td>
<td>8.6E-2</td>
</tr>
<tr>
<td>Q^p(\lambda)</td>
<td><strong>9.0E-2</strong></td>
<td><strong>6.1E-2</strong></td>
</tr>
<tr>
<td>Tree</td>
<td>5.4E-1</td>
<td>8.1E-2</td>
</tr>
<tr>
<td>IH</td>
<td>1.1E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>7.4E-1</td>
<td>2.8E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.8E-1</td>
<td>1.7E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.4E0</td>
<td>1.7E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>6.5E-1</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>3.5E-1</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>Q^p(\lambda)</td>
<td><strong>1.2E-1</strong></td>
<td>2.3E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>2.4E-1</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>IH</td>
<td>6.8E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.2E0</td>
<td>4.5E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.2E-1</td>
<td>7.6E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.7E0</td>
<td>7.9E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>1.1E0</td>
<td>6.3E-2</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>6.8E-1</td>
<td>6.7E-2</td>
</tr>
<tr>
<td>Q^p(\lambda)</td>
<td><strong>1.1E-1</strong></td>
<td>7.8E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>5.5E-1</td>
<td>6.8E-2</td>
</tr>
<tr>
<td>IH</td>
<td>1.1E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.0E0</td>
<td>1.3E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.9E-1</td>
<td>7.4E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.5E0</td>
<td>7.2E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>1.0E0</td>
<td>6.5E-2</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>6.3E-1</td>
<td>6.7E-2</td>
</tr>
<tr>
<td>Q^p(\lambda)</td>
<td><strong>1.4E-1</strong></td>
<td>8.0E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>5.0E-1</td>
<td>6.9E-2</td>
</tr>
<tr>
<td>IH</td>
<td>1.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 283: 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.

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E-1</td>
<td>6.6E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>9.4E-2</td>
<td><strong>6.0E-2</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.3E0</td>
<td>-</td>
</tr>
</tbody>
</table>

IPS

Table 284: 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.

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E-1</td>
<td>6.6E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>9.4E-2</td>
<td><strong>6.0E-2</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.3E0</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DM DR</th>
<th>DM WDR</th>
<th>DM MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.9E0</td>
<td>3.0E-1</td>
<td>7.8E-2</td>
<td>7.8E-2</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.2E-1</td>
<td>1.5E-2</td>
<td>5.8E-3</td>
<td>3.1E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.0E0</td>
<td>5.3E-3</td>
<td>3.2E-2</td>
<td>3.2E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>3.8E0</td>
<td>2.7E-1</td>
<td>2.6E-2</td>
<td>2.6E-2</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.8E0</td>
<td>1.4E-1</td>
<td>1.8E-2</td>
<td>1.8E-2</td>
</tr>
<tr>
<td>Q^e(\lambda)</td>
<td>2.6E-1</td>
<td>2.5E-2</td>
<td>8.4E-3</td>
<td>1.1E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>1.8E0</td>
<td>1.4E-1</td>
<td>1.9E-2</td>
<td>1.9E-2</td>
</tr>
<tr>
<td>IH</td>
<td>3.8E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DM DR</th>
<th>DM WDR</th>
<th>DM MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.0E0</td>
<td>5.5E-2</td>
<td>1.7E-2</td>
<td>1.7E-2</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.0E-1</td>
<td>2.4E-3</td>
<td>9.6E-4</td>
<td>1.2E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.1E0</td>
<td>1.3E-3</td>
<td>1.2E-2</td>
<td>1.2E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>3.9E0</td>
<td>7.6E-2</td>
<td>5.7E-3</td>
<td>5.7E-3</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.8E0</td>
<td>3.7E-2</td>
<td>3.8E-3</td>
<td>3.8E-3</td>
</tr>
<tr>
<td>Q^e(\lambda)</td>
<td>2.7E-1</td>
<td>7.3E-3</td>
<td>1.6E-3</td>
<td>1.6E-3</td>
</tr>
<tr>
<td>TREE</td>
<td>1.8E0</td>
<td>3.8E-2</td>
<td>4.1E-3</td>
<td>4.1E-3</td>
</tr>
<tr>
<td>IH</td>
<td>3.9E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DM DR</th>
<th>DM WDR</th>
<th>DM MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.2E-1</td>
<td>1.2E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>2.6E-2</td>
<td>2.6E-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.9E0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DM DR</th>
<th>DM WDR</th>
<th>DM MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.0E0</td>
<td>5.6E-2</td>
<td>4.2E-2</td>
<td>4.2E-2</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.1E-1</td>
<td>6.1E-3</td>
<td>5.0E-3</td>
<td>3.7E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.1E0</td>
<td>5.1E-3</td>
<td>8.5E-3</td>
<td>8.5E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>4.0E0</td>
<td>3.7E-2</td>
<td>7.4E-3</td>
<td>7.4E-3</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.9E0</td>
<td>2.1E-2</td>
<td>6.5E-3</td>
<td>6.5E-3</td>
</tr>
<tr>
<td>Q^e(\lambda)</td>
<td>3.8E-1</td>
<td>8.4E-3</td>
<td>5.5E-3</td>
<td>2.6E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>1.9E0</td>
<td>2.2E-2</td>
<td>6.6E-3</td>
<td>6.6E-3</td>
</tr>
<tr>
<td>IH</td>
<td>4.1E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 287: 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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DM DR</th>
<th>DM WDR</th>
<th>DM MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>3.3E-2</td>
<td>3.3E-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>5.8E-3</td>
<td>5.8E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.9E0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 288: 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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DM DR</th>
<th>DM WDR</th>
<th>DM MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>8.2E-2</td>
<td>1.3E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>8.0E-2</td>
<td>9.5E-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.8E0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DM Hybrid</th>
<th>DR WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.7E0</td>
<td>3.3E-1</td>
<td>2.2E-1</td>
<td>2.2E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.3E-1</td>
<td>9.5E-2</td>
<td>9.9E-2</td>
<td>3.2E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.5E0</td>
<td>1.0E-1</td>
<td>1.4E-1</td>
<td>1.4E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.7E0</td>
<td>1.0E-1</td>
<td>9.0E-2</td>
<td>9.0E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>2.2E0</td>
<td>8.7E-2</td>
<td>9.3E-2</td>
<td>9.3E-2</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>2.7E-1</td>
<td>9.2E-2</td>
<td>9.7E-2</td>
<td>2.4E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>2.2E0</td>
<td>8.7E-2</td>
<td>9.2E-2</td>
<td>9.2E-2</td>
</tr>
<tr>
<td>IH</td>
<td>4.5E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DM Hybrid</th>
<th>DR WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.8E0</td>
<td>1.2E0</td>
<td>6.4E-1</td>
<td>2.0E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.3E-1</td>
<td>2.7E-1</td>
<td>2.6E-1</td>
<td>3.4E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.1E0</td>
<td>2.5E-1</td>
<td>2.8E-1</td>
<td>2.8E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>3.8E0</td>
<td>8.6E-1</td>
<td>3.3E-1</td>
<td>3.3E-1</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.7E0</td>
<td>6.0E-1</td>
<td>3.1E-1</td>
<td>3.4E-1</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>2.5E-1</td>
<td>3.9E-1</td>
<td>2.9E-1</td>
<td>2.5E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>1.7E0</td>
<td>6.1E-1</td>
<td>3.2E-1</td>
<td>3.4E-1</td>
</tr>
<tr>
<td>IH</td>
<td>3.7E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DM Hybrid</th>
<th>DR WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.0E0</td>
<td>2.4E-1</td>
<td>1.8E-1</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.8E-1</td>
<td>5.5E-2</td>
<td>5.2E-2</td>
<td>1.3E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.2E0</td>
<td>5.1E-2</td>
<td>5.4E-2</td>
<td>5.4E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>4.0E0</td>
<td>1.2E-1</td>
<td>5.9E-2</td>
<td>5.9E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.9E0</td>
<td>8.6E-2</td>
<td>5.6E-2</td>
<td>5.6E-2</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>2.7E-1</td>
<td>6.2E-2</td>
<td>5.3E-2</td>
<td>1.4E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>1.9E0</td>
<td>8.9E-2</td>
<td>5.7E-2</td>
<td>5.7E-2</td>
</tr>
<tr>
<td>IH</td>
<td>4.0E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DM Hybrid</th>
<th>DR WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.1E0</td>
<td>4.3E-1</td>
<td>3.5E-1</td>
<td>1.3E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.8E-1</td>
<td>8.7E-2</td>
<td>9.7E-2</td>
<td>2.1E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.3E0</td>
<td>1.1E-1</td>
<td>2.0E-1</td>
<td>2.0E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.1E0</td>
<td>1.6E-1</td>
<td>8.0E-2</td>
<td>8.0E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.9E0</td>
<td>1.0E-1</td>
<td>8.2E-2</td>
<td>3.8E-1</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>3.1E-1</td>
<td>7.9E-2</td>
<td>9.1E-2</td>
<td>2.9E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>1.9E0</td>
<td>1.0E-1</td>
<td>8.2E-2</td>
<td>5.3E-1</td>
</tr>
<tr>
<td>IH</td>
<td>4.1E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM Direct</th>
<th>Hybrid DR</th>
<th>Hybrid WDR</th>
<th>Hybrid MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.2E-1</td>
<td>2.6E-1</td>
<td>2.4E-1</td>
<td>2.4E-1</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.5E-1</td>
<td>1.3E-1</td>
<td>1.3E-1</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.0E0</td>
<td>1.3E-1</td>
<td>1.5E-1</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.0E0</td>
<td>1.6E-1</td>
<td>1.3E-1</td>
<td>1.3E-1</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.9E0</td>
<td>1.4E-1</td>
<td>1.3E-1</td>
<td>1.3E-1</td>
</tr>
<tr>
<td>Q$^\theta$(\lambda)</td>
<td>3.1E-1</td>
<td>1.2E-1</td>
<td>1.3E-1</td>
<td>2.8E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>1.9E0</td>
<td>1.4E-1</td>
<td>1.3E-1</td>
<td>1.3E-1</td>
</tr>
<tr>
<td>IH</td>
<td>4.1E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM Direct</th>
<th>Hybrid DR</th>
<th>Hybrid WDR</th>
<th>Hybrid MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.5E0</td>
<td>1.3E0</td>
<td>1.1E0</td>
<td>3.6E0</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.1E-1</td>
<td>1.7E-1</td>
<td>1.9E-1</td>
<td>1.9E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.0E-1</td>
<td>1.9E-1</td>
<td>2.8E-1</td>
<td>4.1E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.3E0</td>
<td>2.6E-1</td>
<td>1.7E-1</td>
<td>1.7E-1</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.8E0</td>
<td>1.9E-1</td>
<td>1.7E-1</td>
<td>8.7E-1</td>
</tr>
<tr>
<td>Q$^\theta$(\lambda)</td>
<td>2.6E-1</td>
<td>1.6E-1</td>
<td>1.8E-1</td>
<td>3.4E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>1.9E0</td>
<td>1.9E-1</td>
<td>1.7E-1</td>
<td>8.7E-1</td>
</tr>
<tr>
<td>IH</td>
<td>4.3E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM Direct</th>
<th>Hybrid DR</th>
<th>Hybrid WDR</th>
<th>Hybrid MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.5E0</td>
<td>1.3E0</td>
<td>1.1E0</td>
<td>3.6E0</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.1E-1</td>
<td>1.7E-1</td>
<td>1.9E-1</td>
<td>1.9E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.0E-1</td>
<td>1.9E-1</td>
<td>2.8E-1</td>
<td>4.1E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.3E0</td>
<td>2.6E-1</td>
<td>1.7E-1</td>
<td>1.7E-1</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.8E0</td>
<td>1.9E-1</td>
<td>1.7E-1</td>
<td>8.7E-1</td>
</tr>
<tr>
<td>Q$^\theta$(\lambda)</td>
<td>2.6E-1</td>
<td>1.6E-1</td>
<td>1.8E-1</td>
<td>3.4E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>1.9E0</td>
<td>1.9E-1</td>
<td>1.7E-1</td>
<td>8.7E-1</td>
</tr>
<tr>
<td>IH</td>
<td>4.3E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM Direct</th>
<th>Hybrid DR</th>
<th>Hybrid WDR</th>
<th>Hybrid MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.5E0</td>
<td>1.3E0</td>
<td>1.1E0</td>
<td>3.6E0</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.1E-1</td>
<td>1.7E-1</td>
<td>1.9E-1</td>
<td>1.9E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.0E-1</td>
<td>1.9E-1</td>
<td>2.8E-1</td>
<td>4.1E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.3E0</td>
<td>2.6E-1</td>
<td>1.7E-1</td>
<td>1.7E-1</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.8E0</td>
<td>1.9E-1</td>
<td>1.7E-1</td>
<td>8.7E-1</td>
</tr>
<tr>
<td>Q$^\theta$(\lambda)</td>
<td>2.6E-1</td>
<td>1.6E-1</td>
<td>1.8E-1</td>
<td>3.4E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>1.9E0</td>
<td>1.9E-1</td>
<td>1.7E-1</td>
<td>8.7E-1</td>
</tr>
<tr>
<td>IH</td>
<td>4.3E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.0E-1</td>
<td>2.4E-1</td>
<td>1.0E-1</td>
<td>9.6E-2</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.7E0</td>
<td>1.1E1</td>
<td>1.6E-1</td>
<td>1.8E0</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>1.6E0</td>
<td>1.3E1</td>
<td>5.6E1</td>
<td>5.6E1</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>1.4E-2</td>
<td>1.9E0</td>
<td>6.3E-3</td>
<td>4.9E-3</td>
<td></td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>2.7E-2</td>
<td>4.9E0</td>
<td>1.3E-1</td>
<td>1.3E-1</td>
<td></td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td><strong>7.1E-3</strong></td>
<td>3.2E0</td>
<td>7.5E-3</td>
<td><strong>2.5E-3</strong></td>
<td></td>
</tr>
<tr>
<td>TREE</td>
<td>8.4E-3</td>
<td>1.0E1</td>
<td>1.9E-1</td>
<td>1.9E-1</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>1.0E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>9.1E-1</td>
<td><strong>4.3E0</strong></td>
</tr>
<tr>
<td>WIS</td>
<td>8.8E-1</td>
<td><strong>2.8E-1</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.1E-1</td>
<td>1.6E-1</td>
<td>8.1E-2</td>
<td>9.2E-2</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.6E-1</td>
<td>3.7E-1</td>
<td>6.6E-2</td>
<td>1.5E-1</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>2.0E-2</td>
<td>2.6E-2</td>
<td>1.2E0</td>
<td>1.2E0</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>1.4E-2</td>
<td>2.9E-2</td>
<td>9.0E-3</td>
<td>4.1E-3</td>
<td></td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>2.0E-2</td>
<td>1.3E-1</td>
<td>9.4E-2</td>
<td>9.8E-2</td>
<td></td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>5.8E-3</td>
<td>2.6E-2</td>
<td>7.4E-3</td>
<td><strong>6.2E-4</strong></td>
<td></td>
</tr>
<tr>
<td>TREE</td>
<td><strong>2.0E-3</strong></td>
<td>1.6E-1</td>
<td>1.0E-1</td>
<td>1.0E-1</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>1.1E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td><strong>1.1E0</strong></td>
<td><strong>2.1E-1</strong></td>
</tr>
<tr>
<td>WIS</td>
<td><strong>7.0E-1</strong></td>
<td><strong>1.6E-1</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.1E-1</td>
<td>1.6E-1</td>
<td>1.1E-1</td>
<td>1.1E-1</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.5E-1</td>
<td>1.4E-1</td>
<td>1.3E-1</td>
<td>3.9E-1</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>4.2E-1</td>
<td>3.5E-1</td>
<td>2.9E0</td>
<td>2.9E0</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>2.1E-2</td>
<td>2.8E-2</td>
<td>6.1E-2</td>
<td><strong>1.1E-2</strong></td>
<td></td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>5.6E-2</td>
<td>1.5E-1</td>
<td>2.0E-1</td>
<td>2.0E-1</td>
<td></td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>2.0E-2</td>
<td>1.8E-2</td>
<td>5.1E-2</td>
<td>1.2E-2</td>
<td></td>
</tr>
<tr>
<td>TREE</td>
<td><strong>2.4E-3</strong></td>
<td>2.0E-1</td>
<td>2.2E-1</td>
<td>2.2E-1</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td><strong>1.7E-2</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td><strong>9.2E0</strong></td>
<td>3.8E-1</td>
</tr>
<tr>
<td>WIS</td>
<td><strong>1.2E0</strong></td>
<td><strong>3.4E-1</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.3E-1</td>
<td>2.2E-1</td>
<td>1.6E-1</td>
<td>1.0E-1</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.5E-1</td>
<td>1.4E-1</td>
<td>1.3E-1</td>
<td>3.9E-1</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>4.2E-1</td>
<td>3.5E-1</td>
<td>2.9E0</td>
<td>2.9E0</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>2.1E-2</td>
<td>2.8E-2</td>
<td>6.1E-2</td>
<td><strong>1.1E-2</strong></td>
<td></td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>5.6E-2</td>
<td>1.5E-1</td>
<td>2.0E-1</td>
<td>2.0E-1</td>
<td></td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>2.0E-2</td>
<td>1.8E-2</td>
<td>5.1E-2</td>
<td>1.2E-2</td>
<td></td>
</tr>
<tr>
<td>TREE</td>
<td><strong>2.4E-3</strong></td>
<td>2.0E-1</td>
<td>2.2E-1</td>
<td>2.2E-1</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td><strong>1.7E-2</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td><strong>9.2E0</strong></td>
<td>3.8E-1</td>
</tr>
<tr>
<td>WIS</td>
<td><strong>1.2E0</strong></td>
<td><strong>3.4E-1</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>
Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>4.1E-1</td>
<td>1.9E-1</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>4.2E-1</td>
<td>1.2E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.6E-1</td>
<td>3.3E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>1.8E-2</td>
<td>3.5E-2</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>2.8E-2</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>8.5E-3</td>
<td>2.7E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>1.3E-2</td>
<td>2.2E-1</td>
</tr>
<tr>
<td>IH</td>
<td>1.5E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>9.9E-1</td>
<td>3.4E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>1.1E0</td>
<td>2.9E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>4.5E-1</td>
<td>1.9E0</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>3.3E-1</td>
<td>2.1E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.1E-1</td>
<td>4.6E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.1E-2</td>
<td>2.9E-1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>4.9E-2</td>
<td>4.7E-1</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>6.0E-2</td>
<td>2.2E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>2.7E-2</td>
<td>7.1E-1</td>
</tr>
<tr>
<td>IH</td>
<td>3.8E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E0</td>
<td>5.6E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>2.1E0</td>
<td>4.5E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.1E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>4.1E-1</td>
<td>1.9E-1</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>4.2E-1</td>
<td>1.2E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.6E-1</td>
<td>3.3E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>1.8E-2</td>
<td>3.5E-2</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>2.8E-2</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>8.5E-3</td>
<td>2.7E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>1.3E-2</td>
<td>2.2E-1</td>
</tr>
<tr>
<td>IH</td>
<td>1.5E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>9.9E-1</td>
<td>3.4E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>1.1E0</td>
<td>2.9E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>4.3E-1</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>3.1E-1</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.7E-1</td>
<td>4.8E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>2.7E-2</td>
<td>1.4E-1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>6.0E-2</td>
<td>1.4E-1</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>4.1E-2</td>
<td>2.0E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>3.7E-2</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>IH</td>
<td>2.1E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E0</td>
<td>2.7E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>1.5E0</td>
<td>3.1E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.2E-1</td>
<td>4.7E-1</td>
<td>3.4E-1</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.6E-1</td>
<td>1.1E-1</td>
<td>1.5E-1</td>
<td>3.9E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>7.4E-1</td>
<td>3.3E-1</td>
<td>5.8E0</td>
<td>5.7E0</td>
</tr>
<tr>
<td>FQE</td>
<td>1.9E-2</td>
<td>5.2E-2</td>
<td>1.1E-1</td>
<td>8.6E-3</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>2.7E-2</td>
<td>1.2E-1</td>
<td>1.9E-1</td>
<td>2.0E-1</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>2.2E-2</td>
<td>6.0E-2</td>
<td>9.6E-2</td>
<td>1.0E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>$9.0E-3$</td>
<td>2.3E-1</td>
<td>1.9E-1</td>
<td>1.9E-1</td>
</tr>
<tr>
<td>IH</td>
<td>1.3E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.9E-1</td>
<td>1.6E0</td>
<td>4.9E-1</td>
<td>3.4E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>6.5E-1</td>
<td>2.7E-1</td>
<td>3.4E-1</td>
<td>4.8E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.1E-1</td>
<td>7.5E-1</td>
<td>2.3E0</td>
<td>2.3E0</td>
</tr>
<tr>
<td>FQE</td>
<td>$1.3E-2$</td>
<td>2.7E-1</td>
<td>2.1E-1</td>
<td>$9.4E-3$</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>3.6E-2</td>
<td>2.9E-1</td>
<td>2.8E-1</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>3.9E-2</td>
<td>2.6E-1</td>
<td>2.2E-1</td>
<td>3.8E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>3.2E-2</td>
<td>3.3E-1</td>
<td>3.2E-1</td>
<td>2.5E-1</td>
</tr>
<tr>
<td>IH</td>
<td>1.3E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.1E-1</td>
<td>1.5E0</td>
<td>5.9E-1</td>
<td>3.1E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.6E-1</td>
<td>2.4E-1</td>
<td>6.2E-1</td>
<td>7.0E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.6E-1</td>
<td>7.8E-1</td>
<td>8.9E0</td>
<td>9.0E0</td>
</tr>
<tr>
<td>FQE</td>
<td>3.7E-2</td>
<td>3.0E-1</td>
<td>2.5E-1</td>
<td>5.8E-2</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>4.4E-2</td>
<td>4.4E-1</td>
<td>4.3E-1</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>3.0E-2</td>
<td>2.5E-1</td>
<td>2.4E-1</td>
<td>$4.8E-2$</td>
</tr>
<tr>
<td>TREE</td>
<td>3.9E-2</td>
<td>5.1E-1</td>
<td>4.5E-1</td>
<td>2.3E-1</td>
</tr>
<tr>
<td>IH</td>
<td>3.2E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.2E-1</td>
<td>3.1E-1</td>
<td>1.2E0</td>
<td>3.1E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.1E-1</td>
<td>3.4E-1</td>
<td>4.6E-1</td>
<td>3.9E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>7.7E-1</td>
<td>3.9E-1</td>
<td>2.6E0</td>
<td>2.7E0</td>
</tr>
<tr>
<td>FQE</td>
<td>2.0E-2</td>
<td>4.8E-2</td>
<td>3.1E-1</td>
<td>3.4E-2</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>4.7E-2</td>
<td>1.2E-1</td>
<td>4.4E-1</td>
<td>8.2E-2</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>3.6E-2</td>
<td>3.5E-1</td>
<td>2.8E-1</td>
<td>5.0E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>2.2E-2</td>
<td>1.4E-1</td>
<td>4.4E-1</td>
<td>2.1E-1</td>
</tr>
<tr>
<td>IH</td>
<td>$1.6E-2$</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>9.8E-1</td>
<td>3.4E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>1.1E0</td>
<td>4.2E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E0</td>
<td>3.2E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>1.7E0</td>
<td>5.3E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>DIRECT</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.5E0</td>
<td>5.6E0</td>
<td>5.8E0</td>
<td>1.6E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>8.9E-1</td>
<td>6.1E-1</td>
<td>1.7E0</td>
<td>1.2E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>7.5E-1</td>
<td>1.5E0</td>
<td>1.2E2</td>
<td>1.2E2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>4.3E0</td>
<td>3.9E0</td>
<td>1.7E0</td>
<td>4.1E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.0E0</td>
<td>9.2E-1</td>
<td>1.6E0</td>
<td>1.4E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q^t(\lambda)</td>
<td><strong>1.6E-1</strong></td>
<td><strong>9.3E-2</strong></td>
<td>1.2E0</td>
<td>1.6E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TREE</td>
<td>1.0E0</td>
<td>9.2E-1</td>
<td>1.6E0</td>
<td>1.4E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>4.3E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>DIRECT</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>9.3E-1</td>
<td>8.7E1</td>
<td>3.8E1</td>
<td><strong>1.7E0</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>8.0E1</td>
<td>4.4E1</td>
<td>2.2E1</td>
<td>3.6E1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>4.0E2</td>
<td>5.3E0</td>
<td>1.8E2</td>
<td>2.1E2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>2.4E0</td>
<td>4.8E1</td>
<td>1.1E1</td>
<td>2.4E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>2.0E0</td>
<td>3.3E1</td>
<td>1.2E1</td>
<td>2.1E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q^t(\lambda)</td>
<td>2.9E0</td>
<td>3.7E1</td>
<td>1.1E1</td>
<td>2.5E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TREE</td>
<td>2.0E0</td>
<td>4.1E1</td>
<td>1.1E1</td>
<td>2.2E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>2.2E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>DIRECT</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.5E0</td>
<td>3.2E0</td>
<td>8.7E0</td>
<td>1.7E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>9.8E-1</td>
<td>7.0E-1</td>
<td>2.0E0</td>
<td>9.2E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>1.0E0</td>
<td>1.2E0</td>
<td>1.7E1</td>
<td>1.6E1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>4.1E0</td>
<td>3.8E0</td>
<td>2.6E0</td>
<td>3.2E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.0E0</td>
<td>9.6E-1</td>
<td>2.6E0</td>
<td>1.8E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q^t(\lambda)</td>
<td><strong>3.8E-2</strong></td>
<td><strong>3.8E-2</strong></td>
<td>1.2E0</td>
<td><strong>3.2E-2</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TREE</td>
<td>1.0E0</td>
<td>9.6E-1</td>
<td>2.6E0</td>
<td>1.8E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>4.0E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>DIRECT</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>8.2E-1</td>
<td>3.5E1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.5E1</td>
<td>1.1E1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>3.3E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>2.4E0</td>
<td>4.8E1</td>
<td>1.1E1</td>
<td>2.4E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>2.0E0</td>
<td>3.3E1</td>
<td>1.2E1</td>
<td>2.1E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q^t(\lambda)</td>
<td>2.9E0</td>
<td>3.7E1</td>
<td>1.1E1</td>
<td>2.5E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TREE</td>
<td>2.0E0</td>
<td>4.1E1</td>
<td>1.1E1</td>
<td>2.2E0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>2.2E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.2E0</td>
<td>1.2E2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>6.2E1</td>
<td>1.6E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.6E1</td>
<td>1.5E1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.8E0</td>
<td>1.1E2</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>2.8E0</td>
<td>1.0E2</td>
</tr>
<tr>
<td>Q^e(\lambda)</td>
<td>3.2E0</td>
<td>9.6E1</td>
</tr>
<tr>
<td>Tree</td>
<td>2.2E0</td>
<td>1.0E2</td>
</tr>
<tr>
<td>IH</td>
<td>5.1E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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 environment. Sparse rewards.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.9E0</td>
<td>5.9E2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>5.9E1</td>
<td>3.0E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.5E1</td>
<td>2.1E1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.9E0</td>
<td>9.5E1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>2.3E0</td>
<td>7.9E1</td>
</tr>
<tr>
<td>Q^e(\lambda)</td>
<td><strong>1.4E0</strong></td>
<td>7.0E1</td>
</tr>
<tr>
<td>Tree</td>
<td>2.7E0</td>
<td>8.3E1</td>
</tr>
<tr>
<td>IH</td>
<td>4.9E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.5E0</td>
<td>7.2E2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.0E0</td>
<td>7.9E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.3E0</td>
<td>2.0E0</td>
</tr>
<tr>
<td>FQE</td>
<td>3.8E0</td>
<td>3.5E0</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.0E0</td>
<td>1.0E0</td>
</tr>
<tr>
<td>Q^e(\lambda)</td>
<td>2.0E0</td>
<td>2.7E0</td>
</tr>
<tr>
<td>Tree</td>
<td><strong>1.0E0</strong></td>
<td>1.0E0</td>
</tr>
<tr>
<td>IH</td>
<td>3.6E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.5E0</td>
<td>1.9E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.1E0</td>
<td>1.3E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.4E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>FQE</td>
<td>4.4E0</td>
<td>4.2E0</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.0E0</td>
<td>1.0E0</td>
</tr>
<tr>
<td>Q^e(\lambda)</td>
<td><strong>9.0E-1</strong></td>
<td><strong>5.1E-1</strong></td>
</tr>
<tr>
<td>Tree</td>
<td>1.0E0</td>
<td>1.0E0</td>
</tr>
<tr>
<td>IH</td>
<td>4.3E0</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard</td>
</tr>
<tr>
<td>IS</td>
<td>2.0E1</td>
</tr>
<tr>
<td>WIS</td>
<td>4.5E1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard</td>
</tr>
<tr>
<td>IS</td>
<td>1.0E0</td>
</tr>
<tr>
<td>WIS</td>
<td>6.1E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.4E0</td>
<td>2.4E1</td>
<td>1.4E1</td>
<td>1.5E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.2E0</td>
<td>1.4E0</td>
<td>4.9E0</td>
<td>4.5E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.3E1</td>
<td>2.0E0</td>
<td>6.9E1</td>
<td>6.8E1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.1E0</td>
<td>4.1E0</td>
<td>4.7E0</td>
<td>3.9E0</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.0E0</td>
<td>1.2E0</td>
<td>4.8E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>3.0E-1</td>
<td>3.1E-1</td>
<td>3.4E0</td>
<td>2.9E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>1.0E0</td>
<td>1.2E0</td>
<td>4.8E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>IH</td>
<td>3.9E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**IPS**

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.2E0</td>
<td>1.2E0</td>
</tr>
<tr>
<td>WIS</td>
<td>4.8E0</td>
<td>4.8E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.9E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.1E0</td>
<td>2.2E3</td>
<td>2.5E2</td>
<td>5.5E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.7E1</td>
<td>6.3E1</td>
<td>5.0E1</td>
<td>5.9E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.4E1</td>
<td>3.7E1</td>
<td>1.5E2</td>
<td>1.4E2</td>
</tr>
<tr>
<td>FQE</td>
<td>1.2E1</td>
<td>9.4E1</td>
<td>7.6E1</td>
<td>1.0E1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>2.6E1</td>
<td>1.2E2</td>
<td>8.2E1</td>
<td>2.5E1</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>1.2E1</td>
<td>6.8E1</td>
<td>7.0E1</td>
<td>1.1E1</td>
</tr>
<tr>
<td>TREE</td>
<td>2.3E1</td>
<td>1.1E2</td>
<td>8.4E1</td>
<td>1.8E1</td>
</tr>
<tr>
<td>IH</td>
<td>1.2E1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**IPS**

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>9.9E-1</td>
<td>6.0E1</td>
</tr>
<tr>
<td>WIS</td>
<td>6.5E1</td>
<td>7.3E1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>7.5E0</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 317: 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.3E0</td>
<td>8.8E2</td>
<td>2.6E2</td>
<td>3.6E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.8E2</td>
<td>9.7E1</td>
<td>6.5E1</td>
<td>1.4E2</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.9E2</td>
<td>7.4E1</td>
<td>2.2E2</td>
<td>4.0E2</td>
</tr>
<tr>
<td>FQE</td>
<td>9.0E0</td>
<td>1.7E2</td>
<td>6.7E1</td>
<td>1.1E1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.2E1</td>
<td>1.6E2</td>
<td>7.1E1</td>
<td>1.4E1</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>1.4E1</td>
<td>2.2E2</td>
<td>7.1E1</td>
<td>1.5E1</td>
</tr>
<tr>
<td>TREE</td>
<td>1.3E1</td>
<td>1.7E2</td>
<td>6.8E1</td>
<td>1.4E1</td>
</tr>
<tr>
<td>IH</td>
<td>8.8E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**IPS**

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.7E0</td>
<td>2.0E2</td>
</tr>
<tr>
<td>WIS</td>
<td>5.0E1</td>
<td>6.6E1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.9E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.9E0</td>
<td>1.7E2</td>
<td>1.9E2</td>
<td>4.1E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.9E2</td>
<td>3.1E2</td>
<td>2.0E2</td>
<td>1.8E2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.6E3</td>
<td>7.5E0</td>
<td>9.4E1</td>
<td>9.9E1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.8E0</td>
<td>5.6E1</td>
<td>5.6E1</td>
<td>5.0E0</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>2.8E0</td>
<td>4.6E1</td>
<td>5.4E1</td>
<td>2.8E0</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>2.7E0</td>
<td>4.4E1</td>
<td>4.7E1</td>
<td>3.0E0</td>
</tr>
<tr>
<td>TREE</td>
<td>2.6E0</td>
<td>4.6E1</td>
<td>5.4E1</td>
<td>2.8E0</td>
</tr>
<tr>
<td>IH</td>
<td>4.9E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**IPS**

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.2E0</td>
<td>4.7E1</td>
</tr>
<tr>
<td>WIS</td>
<td>5.0E1</td>
<td>5.4E1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.1E-1</td>
<td>3.6E-3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.1E-1</td>
<td>2.9E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.7E-1</td>
<td>2.5E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>1.2E-1</td>
<td>5.4E-3</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.7E-1</td>
<td>3.0E-3</td>
</tr>
<tr>
<td>Q^\mu(\lambda)</td>
<td>5.7E-2</td>
<td>2.7E-3</td>
</tr>
<tr>
<td>Tree</td>
<td>2.1E-1</td>
<td>2.9E-3</td>
</tr>
<tr>
<td>IH</td>
<td>1.2E-1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>STANDARD</td>
</tr>
<tr>
<td>IS</td>
<td>1.7E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>9.0E-3</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.6E-1</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.1E-1</td>
<td>2.9E-3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.1E-1</td>
<td>2.9E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.7E-1</td>
<td>2.5E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>1.2E-1</td>
<td>5.4E-3</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.7E-1</td>
<td>3.0E-3</td>
</tr>
<tr>
<td>Q^\mu(\lambda)</td>
<td>5.7E-2</td>
<td>2.7E-3</td>
</tr>
<tr>
<td>Tree</td>
<td>2.1E-1</td>
<td>2.9E-3</td>
</tr>
<tr>
<td>IH</td>
<td>1.2E-1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>STANDARD</td>
</tr>
<tr>
<td>IS</td>
<td>3.4E-3</td>
</tr>
<tr>
<td>WIS</td>
<td>1.6E-3</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.9E-1</td>
</tr>
</tbody>
</table>

Table 321: 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>7.4E-2</td>
<td>2.5E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>8.4E-2</td>
<td>8.2E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.0E-1</td>
<td>8.0E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>7.8E-2</td>
<td>9.9E-3</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.3E-1</td>
<td>8.3E-3</td>
</tr>
<tr>
<td>Q^\mu(\lambda)</td>
<td>3.9E-2</td>
<td>8.3E-3</td>
</tr>
<tr>
<td>Tree</td>
<td>1.6E-1</td>
<td>8.2E-3</td>
</tr>
<tr>
<td>IH</td>
<td>7.9E-2</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>STANDARD</td>
</tr>
<tr>
<td>IS</td>
<td>1.8E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>6.1E-2</td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.6E-1</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>7.4E-2</td>
<td>2.5E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>8.4E-2</td>
<td>8.2E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.0E-1</td>
<td>8.0E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>7.8E-2</td>
<td>9.9E-3</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.3E-1</td>
<td>8.3E-3</td>
</tr>
<tr>
<td>Q^\mu(\lambda)</td>
<td>3.9E-2</td>
<td>8.3E-3</td>
</tr>
<tr>
<td>Tree</td>
<td>1.6E-1</td>
<td>8.2E-3</td>
</tr>
<tr>
<td>IH</td>
<td>7.9E-2</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>STANDARD</td>
</tr>
<tr>
<td>IS</td>
<td>1.8E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.6E-1</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>HYBRID Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.5E-1</td>
<td>1.8E-2</td>
<td>1.7E-2</td>
<td>2.0E-2</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.2E-1</td>
<td>5.7E-3</td>
<td>5.5E-3</td>
<td><strong>4.3E-3</strong></td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>4.9E-1</td>
<td>5.6E-3</td>
<td>5.4E-3</td>
<td>1.0E-2</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>1.3E-1</td>
<td>6.4E-3</td>
<td>6.2E-3</td>
<td>4.4E-3</td>
<td></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.8E-1</td>
<td>5.6E-3</td>
<td>5.4E-3</td>
<td>1.1E-2</td>
<td></td>
</tr>
<tr>
<td>Q^\omega(\lambda)</td>
<td><strong>6.5E-2</strong></td>
<td>5.5E-3</td>
<td>5.4E-3</td>
<td>5.4E-3</td>
<td></td>
</tr>
<tr>
<td>TREE</td>
<td>2.2E-1</td>
<td>5.6E-3</td>
<td>5.3E-3</td>
<td>1.2E-2</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>1.3E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>7.6E-3</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.7E-1</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>HYBRID Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.0E-1</td>
<td>4.7E-3</td>
<td>4.4E-3</td>
<td>4.4E-3</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.1E-1</td>
<td>2.1E-3</td>
<td>1.9E-3</td>
<td>1.9E-3</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>5.0E-1</td>
<td>2.1E-3</td>
<td>2.2E-3</td>
<td>2.2E-3</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>1.2E-1</td>
<td>2.2E-3</td>
<td>1.9E-3</td>
<td>1.9E-3</td>
<td></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.8E-1</td>
<td>2.1E-3</td>
<td>1.9E-3</td>
<td>4.7E-3</td>
<td></td>
</tr>
<tr>
<td>Q^\omega(\lambda)</td>
<td><strong>6.5E-2</strong></td>
<td>2.2E-3</td>
<td>1.9E-3</td>
<td>1.9E-3</td>
<td></td>
</tr>
<tr>
<td>TREE</td>
<td>2.2E-1</td>
<td>2.1E-3</td>
<td>1.9E-3</td>
<td>2.3E-3</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>1.2E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>5.5E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>5.4E-2</td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.8E-1</td>
</tr>
</tbody>
</table>

Table 325: 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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>HYBRID Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.9E-1</td>
<td>5.8E-2</td>
<td>5.8E-2</td>
<td>2.4E-1</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.7E-1</td>
<td>4.9E-2</td>
<td>4.9E-2</td>
<td>1.5E-1</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>5.6E-1</td>
<td>4.8E-2</td>
<td>4.9E-2</td>
<td>1.7E-1</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>2.2E-1</td>
<td>4.8E-2</td>
<td>4.8E-2</td>
<td>1.6E-1</td>
<td></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>2.6E-1</td>
<td>4.8E-2</td>
<td>4.8E-2</td>
<td>1.5E-1</td>
<td></td>
</tr>
<tr>
<td>Q^\omega(\lambda)</td>
<td><strong>1.3E-1</strong></td>
<td><strong>4.8E-2</strong></td>
<td>4.8E-2</td>
<td>1.6E-1</td>
<td></td>
</tr>
<tr>
<td>TREE</td>
<td>3.1E-1</td>
<td>4.8E-2</td>
<td>4.8E-2</td>
<td>1.5E-1</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>2.1E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>2.7E-3</td>
</tr>
<tr>
<td>WIS</td>
<td>2.1E-3</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.8E-1</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>HYBRID Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.4E-1</td>
<td>2.7E-2</td>
<td>2.5E-2</td>
<td>5.6E-2</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.2E-1</td>
<td>2.5E-2</td>
<td>2.3E-2</td>
<td>6.1E-2</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>4.8E-1</td>
<td>2.4E-2</td>
<td><strong>2.1E-2</strong></td>
<td>2.9E-2</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>1.3E-1</td>
<td>2.8E-2</td>
<td>2.6E-2</td>
<td>4.8E-2</td>
<td></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.8E-1</td>
<td>2.5E-2</td>
<td>2.3E-2</td>
<td>4.2E-2</td>
<td></td>
</tr>
<tr>
<td>Q^\omega(\lambda)</td>
<td><strong>6.7E-2</strong></td>
<td><strong>2.4E-2</strong></td>
<td><strong>2.3E-2</strong></td>
<td><strong>5.9E-2</strong></td>
<td></td>
</tr>
<tr>
<td>TREE</td>
<td>2.3E-1</td>
<td>2.5E-2</td>
<td>2.2E-2</td>
<td>4.4E-2</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>1.3E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td><strong>4.5E-2</strong></td>
</tr>
<tr>
<td>WIS</td>
<td><strong>3.8E-2</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.7E-1</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.6E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.4E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.1E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>1.6E-1</td>
</tr>
<tr>
<td>R(λ)</td>
<td>2.1E-1</td>
</tr>
<tr>
<td>$Q^b(λ)$</td>
<td><strong>8.7E-2</strong></td>
</tr>
<tr>
<td>TREE</td>
<td>2.6E-1</td>
</tr>
<tr>
<td>IH</td>
<td>1.5E-1</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>2.7E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.9E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.6E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>2.5E-1</td>
</tr>
<tr>
<td>R(λ)</td>
<td>2.6E-1</td>
</tr>
<tr>
<td>$Q^b(λ)$</td>
<td><strong>1.3E-1</strong></td>
</tr>
<tr>
<td>TREE</td>
<td>3.2E-1</td>
</tr>
<tr>
<td>IH</td>
<td>2.5E-1</td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>3.5E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>2.9E-2</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.7E-1</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.4E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.5E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>R(λ)</td>
<td>2.3E-1</td>
</tr>
<tr>
<td>$Q^b(λ)$</td>
<td><strong>1.0E-1</strong></td>
</tr>
<tr>
<td>TREE</td>
<td>2.8E-1</td>
</tr>
<tr>
<td>IH</td>
<td>1.8E-1</td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>3.2E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>3.4E-2</td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.6E-1</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.3E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.3E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>1.7E-1</td>
</tr>
<tr>
<td>R(λ)</td>
<td>2.2E-1</td>
</tr>
<tr>
<td>$Q^b(λ)$</td>
<td><strong>9.6E-2</strong></td>
</tr>
<tr>
<td>TREE</td>
<td>2.6E-1</td>
</tr>
<tr>
<td>IH</td>
<td>1.6E-1</td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>2.2E-2</td>
</tr>
<tr>
<td>WIS</td>
<td><strong>2.0E-2</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.2E-1</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>5.4E-1</td>
<td>2.4E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.9E-2</td>
<td>2.9E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.0E-1</td>
<td>3.3E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>4.5E-1</td>
<td>8.0E-3</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.6E-1</td>
<td>2.7E-3</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>3.4E-2</td>
<td>2.5E-3</td>
</tr>
<tr>
<td>TREE</td>
<td>2.7E-1</td>
<td>3.1E-3</td>
</tr>
<tr>
<td>IH</td>
<td>4.5E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>4.4E-1</td>
<td>2.2E-3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.4E-2</td>
<td>4.4E-4</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.1E-1</td>
<td>5.0E-4</td>
</tr>
<tr>
<td>FQE</td>
<td>4.4E-1</td>
<td>7.9E-4</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.4E-1</td>
<td>3.8E-4</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>2.4E-2</td>
<td>4.6E-4</td>
</tr>
<tr>
<td>TREE</td>
<td>2.5E-1</td>
<td>3.6E-4</td>
</tr>
<tr>
<td>IH</td>
<td>4.4E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>4.2E-1</td>
<td>2.1E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.4E-2</td>
<td>4.4E-4</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.1E-1</td>
<td>5.0E-4</td>
</tr>
<tr>
<td>FQE</td>
<td>4.4E-1</td>
<td>7.9E-4</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.4E-1</td>
<td>3.8E-4</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>2.4E-2</td>
<td>4.6E-4</td>
</tr>
<tr>
<td>TREE</td>
<td>2.5E-1</td>
<td>3.6E-4</td>
</tr>
<tr>
<td>IH</td>
<td>4.4E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>6.0E-1</td>
<td>6.3E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.4E-2</td>
<td>1.6E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.6E-1</td>
<td>1.6E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>5.3E-1</td>
<td>1.9E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.9E-1</td>
<td>1.5E-2</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>5.3E-2</td>
<td>1.5E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>2.9E-1</td>
<td>1.5E-2</td>
</tr>
<tr>
<td>IH</td>
<td>5.3E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.1E-3</td>
<td>1.1E-3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.6E-4</td>
<td>4.6E-4</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.4E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.8E-2</td>
<td>1.8E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.6E-2</td>
<td>1.6E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.1E-1</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DM Hybrid</th>
<th>WDR MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.7E-1</td>
<td>2.7E-2</td>
<td>2.6E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.7E-2</td>
<td>1.7E-2</td>
<td>1.6E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.5E-1</td>
<td>1.6E-2</td>
<td>1.6E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>5.0E-1</td>
<td>1.9E-2</td>
<td>1.8E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>2.0E-1</td>
<td>1.7E-2</td>
<td>1.6E-2</td>
</tr>
<tr>
<td>Q^($\lambda$)</td>
<td>5.0E-2</td>
<td>1.6E-2</td>
<td>1.6E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>3.2E-1</td>
<td>1.7E-2</td>
<td>1.6E-2</td>
</tr>
<tr>
<td>IH</td>
<td>5.0E-1</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Stochastic environment. Sparse rewards. | IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.9E-2</td>
<td>2.0E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>1.8E-2</td>
<td>-</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.8E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DM Hybrid</th>
<th>WDR MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>5.0E-1</td>
<td>2.2E-1</td>
<td>1.9E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.8E-2</td>
<td>1.1E-2</td>
<td>1.0E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.5E-1</td>
<td>1.0E-2</td>
<td>9.1E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>4.9E-1</td>
<td>1.7E-2</td>
<td>1.4E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.7E-1</td>
<td>1.2E-2</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>Q^($\lambda$)</td>
<td>4.0E-2</td>
<td>1.1E-2</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>2.8E-1</td>
<td>1.3E-2</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>IH</td>
<td>4.9E-1</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Stochastic environment. Sparse rewards. | IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.9E-2</td>
<td>2.0E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>1.8E-2</td>
<td>-</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.8E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DM Hybrid</th>
<th>WDR MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>5.0E-1</td>
<td>1.4E-1</td>
<td>1.4E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>6.8E-2</td>
<td>3.7E-2</td>
<td>3.7E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.6E-1</td>
<td>3.7E-2</td>
<td>3.8E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>5.3E-1</td>
<td>3.5E-2</td>
<td>3.5E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>2.2E-1</td>
<td>3.6E-2</td>
<td>3.7E-2</td>
</tr>
<tr>
<td>Q^($\lambda$)</td>
<td>7.1E-2</td>
<td>3.6E-2</td>
<td>3.7E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>3.5E-1</td>
<td>3.6E-2</td>
<td>3.7E-2</td>
</tr>
<tr>
<td>IH</td>
<td>5.3E-1</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Stochastic environment. Sparse rewards. | IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>3.7E-2</td>
<td>3.7E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>3.6E-2</td>
<td>-</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.8E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DM Hybrid</th>
<th>WDR MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>5.8E-1</td>
<td>9.3E-2</td>
<td>8.9E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>6.2E-2</td>
<td>2.0E-2</td>
<td>2.0E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.9E-1</td>
<td>2.0E-2</td>
<td>2.0E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>5.5E-1</td>
<td>2.1E-2</td>
<td>2.0E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>2.2E-1</td>
<td>2.0E-2</td>
<td>2.0E-2</td>
</tr>
<tr>
<td>Q^($\lambda$)</td>
<td>5.9E-2</td>
<td>2.0E-2</td>
<td>2.0E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>3.5E-1</td>
<td>2.0E-2</td>
<td>2.0E-2</td>
</tr>
<tr>
<td>IH</td>
<td>5.5E-1</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Stochastic environment. Sparse rewards. | IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>3.7E-2</td>
<td>3.7E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>3.6E-2</td>
<td>-</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.8E-1</td>
<td>-</td>
</tr>
</tbody>
</table>
Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning

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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>Hybrid DR</th>
<th>Hybrid WDR</th>
<th>Hybrid MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.1E-1</td>
<td>2.1E-2</td>
<td>4.2E-2</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.7E-2</td>
<td>7.4E-3</td>
<td>1.8E-2</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>2.7E-1</td>
<td>7.2E-3</td>
<td>8.6E-3</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>4.0E-1</td>
<td>9.5E-3</td>
<td>8.6E-3</td>
<td></td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.2E-1</td>
<td>7.7E-3</td>
<td>7.4E-3</td>
<td>2.2E-2</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>2.1E-2</td>
<td>7.5E-3</td>
<td>7.4E-3</td>
<td>7.5E-3</td>
</tr>
<tr>
<td>TREE</td>
<td>2.2E-1</td>
<td>8.0E-3</td>
<td>7.5E-3</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>4.1E-1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>9.9E-3</td>
<td>9.9E-3</td>
</tr>
<tr>
<td>WIS</td>
<td>8.6E-3</td>
<td>8.6E-3</td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.9E-1</td>
<td></td>
</tr>
</tbody>
</table>

Table 340: Graph-POMDP, relative MSE. $T = 2$, $N = 256$, $H = 2$, $\pi_b(a = 0) = 0.6$, $\pi_e(a = 0) = 0.8$.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>Hybrid DR</th>
<th>Hybrid WDR</th>
<th>Hybrid MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>5.1E-1</td>
<td>3.1E-2</td>
<td>3.0E-2</td>
<td>3.5E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.7E-2</td>
<td>1.5E-2</td>
<td>1.5E-2</td>
<td>9.5E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.3E-1</td>
<td>1.5E-2</td>
<td>1.6E-2</td>
<td>1.2E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>5.7E-1</td>
<td>1.7E-2</td>
<td>1.5E-2</td>
<td>9.4E-2</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.7E-1</td>
<td>1.5E-2</td>
<td>1.5E-2</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>3.5E-2</td>
<td>1.5E-2</td>
<td>1.5E-2</td>
<td>8.7E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>3.1E-1</td>
<td>1.5E-2</td>
<td>1.5E-2</td>
<td>1.6E-1</td>
</tr>
<tr>
<td>IH</td>
<td>5.7E-1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>2.2E-2</td>
<td>1.7E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>1.8E-2</td>
<td>1.5E-2</td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.2E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 341: Graph-POMDP, relative MSE. $T = 2$, $N = 512$, $H = 2$, $\pi_b(a = 0) = 0.6$, $\pi_e(a = 0) = 0.8$.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>Hybrid DR</th>
<th>Hybrid WDR</th>
<th>Hybrid MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>7.0E-1</td>
<td>1.0E-1</td>
<td>1.0E-1</td>
<td>2.1E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>7.0E-2</td>
<td>5.7E-2</td>
<td>5.7E-2</td>
<td>1.2E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.5E-1</td>
<td>5.8E-2</td>
<td>5.9E-2</td>
<td>1.2E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>5.5E-1</td>
<td>5.6E-2</td>
<td>5.6E-2</td>
<td>5.6E-2</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.9E-1</td>
<td>5.6E-2</td>
<td>5.7E-2</td>
<td>2.0E-1</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>7.0E-2</td>
<td>5.7E-2</td>
<td>5.7E-2</td>
<td>1.2E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>3.2E-1</td>
<td>5.6E-2</td>
<td>5.7E-2</td>
<td>1.1E-1</td>
</tr>
<tr>
<td>IH</td>
<td>5.5E-1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>5.3E-2</td>
<td>5.7E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>5.3E-2</td>
<td>5.6E-2</td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.2E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 342: Graph-POMDP, relative MSE. $T = 2$, $N = 1024$, $H = 2$, $\pi_b(a = 0) = 0.6$, $\pi_e(a = 0) = 0.8$.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>Hybrid DR</th>
<th>Hybrid WDR</th>
<th>Hybrid MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>7.0E-1</td>
<td>1.0E-1</td>
<td>1.0E-1</td>
<td>2.1E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>7.0E-2</td>
<td>5.7E-2</td>
<td>5.7E-2</td>
<td>1.2E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.5E-1</td>
<td>5.8E-2</td>
<td>5.9E-2</td>
<td>1.2E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>5.5E-1</td>
<td>5.6E-2</td>
<td>5.6E-2</td>
<td>5.6E-2</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.9E-1</td>
<td>5.6E-2</td>
<td>5.7E-2</td>
<td>2.0E-1</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>7.0E-2</td>
<td>5.7E-2</td>
<td>5.7E-2</td>
<td>1.2E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>3.2E-1</td>
<td>5.6E-2</td>
<td>5.7E-2</td>
<td>1.1E-1</td>
</tr>
<tr>
<td>IH</td>
<td>5.5E-1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>5.3E-2</td>
<td>5.7E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>5.3E-2</td>
<td>5.6E-2</td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.2E-1</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.9E-1</td>
<td>5.4E-3</td>
<td>2.1E-3</td>
<td>2.1E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.5E-2</td>
<td>4.4E-4</td>
<td>1.2E-4</td>
<td>1.4E-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>1.5E-2</td>
<td>5.2E-3</td>
<td>3.6E-3</td>
<td>9.4E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>3.0E-3</td>
<td>2.2E-4</td>
<td>1.7E-4</td>
<td>1.2E-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>7.3E-4</td>
<td>2.4E-4</td>
<td>1.7E-4</td>
<td>4.9E-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>6.5E-3</td>
<td>2.9E-4</td>
<td>1.6E-4</td>
<td>2.6E-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TREE</td>
<td>4.8E-3</td>
<td>3.2E-3</td>
<td>4.8E-4</td>
<td>5.2E-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>8.7E-4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.9E-1</td>
<td>9.8E-4</td>
<td>8.9E-4</td>
<td>8.9E-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.2E-2</td>
<td>1.2E-4</td>
<td>1.1E-4</td>
<td>1.4E-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>1.2E-2</td>
<td>7.0E-4</td>
<td>7.1E-4</td>
<td>3.0E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>2.8E-3</td>
<td>1.4E-4</td>
<td>1.3E-4</td>
<td>6.3E-5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>6.5E-4</td>
<td>1.7E-4</td>
<td>1.1E-4</td>
<td>1.6E-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>6.3E-3</td>
<td>1.3E-4</td>
<td>1.1E-4</td>
<td>1.7E-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TREE</td>
<td>6.1E-3</td>
<td>4.3E-4</td>
<td>2.3E-4</td>
<td>5.9E-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>1.5E-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 345: Graph-POMDP, relative MSE. $T = 16$, $N = 256$, $H = 6$, $\pi_b(a = 0) = 0.6$, $\pi_e(a = 0) = 0.8$. Dense rewards.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.9E-1</td>
<td>9.5E-3</td>
<td>9.5E-3</td>
<td>9.5E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.1E-2</td>
<td>4.2E-3</td>
<td>3.1E-3</td>
<td>5.6E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>1.6E-2</td>
<td>6.6E-3</td>
<td>6.3E-3</td>
<td>1.2E-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>3.0E-3</td>
<td>2.7E-3</td>
<td>2.4E-3</td>
<td>1.0E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>2.8E-3</td>
<td>2.7E-3</td>
<td>2.5E-3</td>
<td>2.3E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>7.0E-3</td>
<td>2.8E-3</td>
<td>2.6E-3</td>
<td>2.1E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TREE</td>
<td>6.5E-3</td>
<td>1.8E-3</td>
<td>2.1E-3</td>
<td>2.4E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>1.8E-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.9E-1</td>
<td>1.8E-2</td>
<td>9.5E-3</td>
<td>9.5E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.1E-2</td>
<td>4.2E-3</td>
<td>3.1E-3</td>
<td>5.6E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>1.6E-2</td>
<td>6.6E-3</td>
<td>6.3E-3</td>
<td>1.2E-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>3.0E-3</td>
<td>2.7E-3</td>
<td>2.4E-3</td>
<td>1.0E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>2.8E-3</td>
<td>2.7E-3</td>
<td>2.5E-3</td>
<td>2.3E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>7.0E-3</td>
<td>2.8E-3</td>
<td>2.6E-3</td>
<td>2.1E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TREE</td>
<td>6.5E-3</td>
<td>1.8E-3</td>
<td>2.1E-3</td>
<td>2.4E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>1.8E-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 347: 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.
Table 347: Graph-POMDP, relative MSE. $T = 16$, $N = 512$, $H = 6$, $\pi_b(a = 0) = 0.6$, $\pi_c(a = 0) = 0.8$. Stochastic rewards. Dense rewards.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>3.9E-1</td>
<td>8.8E-3</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>4.7E-3</td>
<td>2.0E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.9E-3</td>
<td>1.7E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>2.1E-3</td>
<td>2.5E-3</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.1E-3</td>
<td>2.4E-3</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>5.4E-3</td>
<td>2.3E-3</td>
</tr>
<tr>
<td>Tree</td>
<td>2.5E-3</td>
<td>3.7E-3</td>
</tr>
<tr>
<td>IH</td>
<td><strong>5.9E-4</strong></td>
<td>-</td>
</tr>
</tbody>
</table>

Table 348: Graph-POMDP, relative MSE. $T = 16$, $N = 1024$, $H = 6$, $\pi_b(a = 0) = 0.6$, $\pi_c(a = 0) = 0.8$. Stochastic rewards. Dense rewards.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>3.7E-1</td>
<td>4.2E-3</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.9E-2</td>
<td>6.6E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.0E-2</td>
<td>1.3E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>3.6E-3</td>
<td>7.4E-3</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>4.5E-3</td>
<td>7.2E-3</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>1.2E-2</td>
<td>6.9E-3</td>
</tr>
<tr>
<td>Tree</td>
<td>6.1E-3</td>
<td>5.0E-3</td>
</tr>
<tr>
<td>IH</td>
<td>3.8E-3</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 349: Graph-POMDP, relative MSE. $T = 16$, $N = 256$, $H = 6$, $\pi_b(a = 0) = 0.6$, $\pi_c(a = 0) = 0.8$. Stochastic environment. Dense rewards.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>3.7E-1</td>
<td>1.2E-2</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.0E-2</td>
<td>2.7E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.0E-2</td>
<td>3.6E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>4.0E-3</td>
<td>2.9E-3</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>2.7E-3</td>
<td>2.8E-3</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td>6.9E-3</td>
<td>2.9E-3</td>
</tr>
<tr>
<td>Tree</td>
<td>7.7E-3</td>
<td>2.6E-3</td>
</tr>
<tr>
<td>IH</td>
<td><strong>1.6E-3</strong></td>
<td>-</td>
</tr>
</tbody>
</table>

Table 350: Graph-POMDP, relative MSE. $T = 16$, $N = 512$, $H = 6$, $\pi_b(a = 0) = 0.6$, $\pi_c(a = 0) = 0.8$. Stochastic environment. Dense rewards.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>3.4E-2</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>8.3E-3</td>
<td>2.7E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.4E-1</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
</tr>
<tr>
<td>AM</td>
<td>4.0E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>7.4E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.1E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>4.9E-3</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.5E-3</td>
</tr>
<tr>
<td>Q^\pi(\lambda)</td>
<td>6.9E-3</td>
</tr>
<tr>
<td>Tree</td>
<td>5.5E-3</td>
</tr>
<tr>
<td>IH</td>
<td>2.4E-3</td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>2.8E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>7.4E-3</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.6E-1</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
</tr>
<tr>
<td>AM</td>
<td>4.2E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.2E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.0E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>7.2E-3</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.2E-2</td>
</tr>
<tr>
<td>Q^\pi(\lambda)</td>
<td>1.7E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>1.2E-2</td>
</tr>
<tr>
<td>IH</td>
<td>6.9E-3</td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>8.6E-2</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.7E-1</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
</tr>
<tr>
<td>AM</td>
<td>3.7E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.5E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.2E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>4.5E-3</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>6.2E-3</td>
</tr>
<tr>
<td>Q^\pi(\lambda)</td>
<td>9.4E-3</td>
</tr>
<tr>
<td>Tree</td>
<td>8.5E-3</td>
</tr>
<tr>
<td>IH</td>
<td>2.8E-3</td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>9.1E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>3.2E-2</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.5E-1</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
</tr>
<tr>
<td>AM</td>
<td>4.3E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.8E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.7E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>7.5E-3</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>7.6E-3</td>
</tr>
<tr>
<td>Q^\pi(\lambda)</td>
<td>1.2E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>1.5E-2</td>
</tr>
<tr>
<td>IH</td>
<td>4.9E-3</td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.4E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>1.4E-2</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.6E-1</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>8.6E-1</td>
<td>6.2E-2</td>
<td>3.7E-2</td>
<td>2.5E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.2E-1</td>
<td>3.4E-2</td>
<td>3.2E-2</td>
<td>6.4E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.1E-1</td>
<td>1.9E-1</td>
<td>1.2E-1</td>
<td>1.2E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>5.2E-1</td>
<td>1.3E-1</td>
<td>5.1E-2</td>
<td>1.7E-1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>2.9E-2</td>
<td>3.5E-2</td>
<td>3.3E-2</td>
<td>3.1E-2</td>
</tr>
<tr>
<td>K$^q(\lambda)$</td>
<td>1.8E-2</td>
<td>3.5E-2</td>
<td>3.3E-2</td>
<td>1.9E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>1.0E0</td>
<td>1.6E-1</td>
<td>5.0E-2</td>
<td>5.0E-2</td>
</tr>
<tr>
<td>IH</td>
<td>5.1E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.1E0</td>
<td>3.2E0</td>
<td>3.0E0</td>
<td>7.4E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>6.0E-1</td>
<td>4.6E-1</td>
<td>4.4E-1</td>
<td>6.4E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.0E-1</td>
<td>4.8E-1</td>
<td>4.2E-1</td>
<td>4.8E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.1E-1</td>
<td>7.2E-1</td>
<td>5.5E-1</td>
<td>4.5E-1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>2.3E-1</td>
<td>6.4E-1</td>
<td>5.4E-1</td>
<td>3.1E-1</td>
</tr>
<tr>
<td>K$^q(\lambda)$</td>
<td>1.6E-1</td>
<td>6.6E-1</td>
<td>5.0E-1</td>
<td>1.6E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>1.5E0</td>
<td>7.3E-1</td>
<td>5.5E-1</td>
<td>1.0E0</td>
</tr>
<tr>
<td>IH</td>
<td>4.0E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>8.5E-1</td>
<td>4.5E-2</td>
<td>2.2E-2</td>
<td>2.2E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.7E-2</td>
<td>3.6E-3</td>
<td>3.8E-3</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.4E-2</td>
<td>1.7E-2</td>
<td>1.3E-2</td>
<td>2.6E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>4.4E-1</td>
<td>2.4E-2</td>
<td>8.4E-3</td>
<td>8.4E-3</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>5.1E-3</td>
<td>4.7E-3</td>
<td>5.1E-3</td>
<td>-</td>
</tr>
<tr>
<td>K$^q(\lambda)$</td>
<td>4.3E-3</td>
<td>5.6E-3</td>
<td>4.3E-3</td>
<td>2.9E-3</td>
</tr>
<tr>
<td>TREE</td>
<td>1.0E0</td>
<td>3.6E-2</td>
<td>8.3E-3</td>
<td>8.3E-3</td>
</tr>
<tr>
<td>IH</td>
<td>4.5E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.1E0</td>
<td>3.2E0</td>
<td>3.0E0</td>
<td>7.4E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>6.0E-1</td>
<td>4.6E-1</td>
<td>4.4E-1</td>
<td>6.4E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.0E-1</td>
<td>4.8E-1</td>
<td>4.2E-1</td>
<td>4.8E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.1E-1</td>
<td>7.2E-1</td>
<td>5.5E-1</td>
<td>4.5E-1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>2.3E-1</td>
<td>6.4E-1</td>
<td>5.4E-1</td>
<td>3.1E-1</td>
</tr>
<tr>
<td>K$^q(\lambda)$</td>
<td>1.6E-1</td>
<td>6.6E-1</td>
<td>5.0E-1</td>
<td>1.6E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>1.5E0</td>
<td>7.3E-1</td>
<td>5.5E-1</td>
<td>1.0E0</td>
</tr>
<tr>
<td>IH</td>
<td>4.0E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Standard Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.6E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>5.0E-2</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.9E-1</td>
</tr>
</tbody>
</table>

Table 359: 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.

<table>
<thead>
<tr>
<th></th>
<th>Standard Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>3.6E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>8.3E-3</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.3E-1</td>
</tr>
</tbody>
</table>
Table 359: Graph-POMDP, relative MSE. \( T = 16, N = 512, H = 6, \pi_b(a = 0) = 0.6, \pi_c(a = 0) = 0.8 \). Stochastic rewards. Sparse rewards.

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.3E0</td>
<td>6.4E-1</td>
<td>5.9E-1</td>
<td>1.1E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.6E-1</td>
<td>2.6E-1</td>
<td>2.6E-1</td>
<td>2.4E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.2E-1</td>
<td>2.6E-1</td>
<td>2.8E-1</td>
<td>2.0E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>6.7E-1</td>
<td>2.9E-1</td>
<td>2.8E-1</td>
<td>6.1E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.5E-1</td>
<td>2.6E-1</td>
<td>2.6E-1</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>Q(\lambda)</td>
<td>1.4E-1</td>
<td>2.6E-1</td>
<td>2.6E-1</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>1.1E0</td>
<td>3.1E-1</td>
<td>2.8E-1</td>
<td>1.0E0</td>
</tr>
<tr>
<td>IH</td>
<td>6.8E-1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 360: Graph-POMDP, relative MSE. \( T = 16, N = 256, H = 6, \pi_b(a = 0) = 0.6, \pi_c(a = 0) = 0.8 \). Stochastic environment. Sparse rewards.

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>9.5E-1</td>
<td>8.5E-1</td>
<td>8.5E-1</td>
<td>8.1E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.5E-1</td>
<td>2.0E-1</td>
<td>2.5E-1</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.5E-1</td>
<td>2.0E-1</td>
<td>2.8E-1</td>
<td>2.9E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>5.4E-1</td>
<td>1.7E-1</td>
<td>2.1E-1</td>
<td>4.7E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.4E-1</td>
<td>2.3E-1</td>
<td>2.5E-1</td>
<td>1.4E-1</td>
</tr>
<tr>
<td>Q(\lambda)</td>
<td>8.4E-2</td>
<td>2.0E-1</td>
<td>2.5E-1</td>
<td>8.6E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>1.0E0</td>
<td>1.7E-1</td>
<td>2.0E-1</td>
<td>4.9E-1</td>
</tr>
<tr>
<td>IH</td>
<td>5.5E-1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 361: Graph-POMDP, relative MSE. \( T = 16, N = 512, H = 6, \pi_b(a = 0) = 0.6, \pi_c(a = 0) = 0.8 \). Stochastic environment. Sparse rewards.

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>8.5E-1</td>
<td>4.8E-1</td>
<td>4.6E-1</td>
<td>5.6E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.4E-1</td>
<td>3.6E-1</td>
<td>3.6E-1</td>
<td>3.2E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.9E-1</td>
<td>3.6E-1</td>
<td>3.5E-1</td>
<td>2.7E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.8E-1</td>
<td>3.9E-1</td>
<td>3.8E-1</td>
<td>4.1E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>2.1E-1</td>
<td>3.6E-1</td>
<td>3.6E-1</td>
<td>2.1E-1</td>
</tr>
<tr>
<td>Q(\lambda)</td>
<td>1.4E-1</td>
<td>3.6E-1</td>
<td>3.6E-1</td>
<td>1.3E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>1.1E0</td>
<td>3.8E-1</td>
<td>3.8E-1</td>
<td>6.3E-1</td>
</tr>
<tr>
<td>IH</td>
<td>4.8E-1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 362: Graph-POMDP, relative MSE. \( T = 16, N = 512, H = 6, \pi_b(a = 0) = 0.6, \pi_c(a = 0) = 0.8 \). Stochastic environment. Sparse rewards.

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>8.5E-1</td>
<td>4.8E-1</td>
<td>4.6E-1</td>
<td>5.6E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.4E-1</td>
<td>3.6E-1</td>
<td>3.6E-1</td>
<td>3.2E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.9E-1</td>
<td>3.6E-1</td>
<td>3.5E-1</td>
<td>2.7E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.8E-1</td>
<td>3.9E-1</td>
<td>3.8E-1</td>
<td>4.1E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>2.1E-1</td>
<td>3.6E-1</td>
<td>3.6E-1</td>
<td>2.1E-1</td>
</tr>
<tr>
<td>Q(\lambda)</td>
<td>1.4E-1</td>
<td>3.6E-1</td>
<td>3.6E-1</td>
<td>1.3E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>1.1E0</td>
<td>3.8E-1</td>
<td>3.8E-1</td>
<td>6.3E-1</td>
</tr>
<tr>
<td>IH</td>
<td>4.8E-1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Direct</th>
<th>Hybrid</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>8.6E-1</td>
<td>3.1E-1</td>
<td>3.1E-1</td>
<td>5.2E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.2E-1</td>
<td>7.7E-2</td>
<td>7.5E-2</td>
<td>6.4E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.3E-2</td>
<td>6.4E-2</td>
<td>7.0E-2</td>
<td>8.1E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>4.6E-1</td>
<td>1.3E-1</td>
<td>9.2E-2</td>
<td>2.2E-1</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>3.0E-2</td>
<td>8.2E-2</td>
<td>7.6E-2</td>
<td>3.0E-2</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td><strong>2.6E-2</strong></td>
<td>8.5E-2</td>
<td>7.5E-2</td>
<td><strong>2.1E-2</strong></td>
</tr>
<tr>
<td>Tree</td>
<td>1.0E0</td>
<td>1.4E-1</td>
<td>9.3E-2</td>
<td>9.3E-2</td>
</tr>
<tr>
<td>IH</td>
<td>4.7E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 364: 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.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Direct</th>
<th>Hybrid</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.2E0</td>
<td>4.7E0</td>
<td>4.5E0</td>
<td>2.0E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>9.4E-1</td>
<td>1.3E0</td>
<td>1.3E0</td>
<td>1.0E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.0E0</td>
<td>1.6E0</td>
<td>1.6E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>FQE</td>
<td>1.3E0</td>
<td>1.2E0</td>
<td>1.2E0</td>
<td>1.3E0</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.0E0</td>
<td>1.2E0</td>
<td>1.3E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td><strong>9.0E-1</strong></td>
<td>1.2E0</td>
<td>1.2E0</td>
<td><strong>9.0E-1</strong></td>
</tr>
<tr>
<td>Tree</td>
<td>2.7E0</td>
<td>1.1E0</td>
<td>1.2E0</td>
<td>2.7E0</td>
</tr>
<tr>
<td>IH</td>
<td>1.3E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 365: 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.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Direct</th>
<th>Hybrid</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>8.8E-1</td>
<td>1.2E0</td>
<td>1.1E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>5.6E-1</td>
<td>5.1E-1</td>
<td>5.0E-1</td>
<td>4.7E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.9E-1</td>
<td>5.9E-1</td>
<td>5.8E-1</td>
<td>4.8E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>5.7E-1</td>
<td>5.6E-1</td>
<td>5.3E-1</td>
<td>5.0E-1</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>2.7E-1</td>
<td>5.1E-1</td>
<td>5.0E-1</td>
<td>2.2E-1</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td><strong>1.3E-1</strong></td>
<td>4.6E-1</td>
<td>4.6E-1</td>
<td><strong>1.5E-1</strong></td>
</tr>
<tr>
<td>Tree</td>
<td>9.9E-1</td>
<td>5.9E-1</td>
<td>5.4E-1</td>
<td>9.6E-1</td>
</tr>
<tr>
<td>IH</td>
<td>5.7E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Direct</th>
<th>Hybrid</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>7.3E-1</td>
<td>5.8E-1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Q-REG</td>
<td>7.8E-1</td>
<td>5.4E-1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MRDR</td>
<td>6.7E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

IPS

Table 367: 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.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.4E-1</td>
<td>1.4E-1</td>
</tr>
<tr>
<td>WIS</td>
<td><strong>9.3E-2</strong></td>
<td>9.3E-2</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.9E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 368: 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.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>2.2E0</td>
<td>1.2E0</td>
</tr>
<tr>
<td>WIS</td>
<td>2.3E0</td>
<td><strong>1.2E0</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>1.1E0</td>
<td>-</td>
</tr>
</tbody>
</table>

IPS

Table 369: 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.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>7.3E-1</td>
<td>7.8E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>7.8E-1</td>
<td><strong>5.4E-1</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>6.7E-1</td>
<td>-</td>
</tr>
</tbody>
</table>
F.3 Detailed Results for Graph Mountain Car (Graph-MC)

Table 367: Graph-MC, relative MSE. $T = 250, N = 128, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct DR WDR MAGIC</td>
<td>Direct DR WDR MAGIC</td>
</tr>
<tr>
<td>AM</td>
<td>5.4E-1</td>
<td>5.1E-1</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.5E2</td>
<td>1.3E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.7E2</td>
<td>1.3E1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.0E-1</td>
<td>4.0E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>4.4E-1</td>
<td>9.4E0</td>
</tr>
<tr>
<td>Q^s(\lambda)</td>
<td>1.9E130</td>
<td>1.9E129</td>
</tr>
<tr>
<td>Tree</td>
<td>4.4E-1</td>
<td>9.4E0</td>
</tr>
<tr>
<td>IH</td>
<td>2.0E1</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 368: Graph-MC, relative MSE. $T = 250, N = 256, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct DR WDR MAGIC</td>
<td>Direct DR WDR MAGIC</td>
</tr>
<tr>
<td>AM</td>
<td>4.8E-1</td>
<td>5.1E-1</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>3.3E1</td>
<td>3.7E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.8E-1</td>
<td>4.3E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>3.7E-1</td>
<td>3.7E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>3.7E-1</td>
<td>3.7E-1</td>
</tr>
<tr>
<td>Q^s(\lambda)</td>
<td>9.1E118</td>
<td>7.5E117</td>
</tr>
<tr>
<td>Tree</td>
<td>3.9E-1</td>
<td>3.7E-1</td>
</tr>
<tr>
<td>IH</td>
<td>2.1E1</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 369: Graph-MC, relative MSE. $T = 250, N = 512, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct DR WDR MAGIC</td>
<td>Direct DR WDR MAGIC</td>
</tr>
<tr>
<td>AM</td>
<td>3.9E-1</td>
<td>3.9E-1</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>2.7E-1</td>
<td>1.1E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.4E-1</td>
<td>1.6E0</td>
</tr>
<tr>
<td>FQE</td>
<td>2.9E-1</td>
<td>2.9E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>3.0E-1</td>
<td>3.0E-1</td>
</tr>
<tr>
<td>Q^s(\lambda)</td>
<td>4.9E146</td>
<td>9.2E145</td>
</tr>
<tr>
<td>Tree</td>
<td>3.1E-1</td>
<td>2.9E-1</td>
</tr>
<tr>
<td>IH</td>
<td>2.1E1</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 370: Graph-MC, relative MSE. $T = 250, N = 1024, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct DR WDR MAGIC</td>
<td>Direct DR WDR MAGIC</td>
</tr>
<tr>
<td>AM</td>
<td>3.6E-1</td>
<td>3.9E-1</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>2.8E-1</td>
<td>2.4E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.7E-1</td>
<td>2.8E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>2.6E-1</td>
<td>2.6E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>2.6E-1</td>
<td>2.6E-1</td>
</tr>
<tr>
<td>Q^s(\lambda)</td>
<td>9.6E121</td>
<td>1.5E122</td>
</tr>
<tr>
<td>Tree</td>
<td>2.8E-1</td>
<td>2.6E-1</td>
</tr>
<tr>
<td>IH</td>
<td>2.0E1</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 367: Graph-MC, relative MSE. $T = 250, N = 128, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$.

<table>
<thead>
<tr>
<th></th>
<th>IPS Standard Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E0</td>
</tr>
<tr>
<td>WIS</td>
<td>2.1E1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>2.1E1</td>
</tr>
</tbody>
</table>

Table 368: Graph-MC, relative MSE. $T = 250, N = 256, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$.

<table>
<thead>
<tr>
<th></th>
<th>IPS Standard Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E0</td>
</tr>
<tr>
<td>WIS</td>
<td>2.0E1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>2.0E1</td>
</tr>
</tbody>
</table>

Table 369: Graph-MC, relative MSE. $T = 250, N = 512, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$.

<table>
<thead>
<tr>
<th></th>
<th>IPS Standard Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E0</td>
</tr>
<tr>
<td>WIS</td>
<td>2.1E1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>2.1E1</td>
</tr>
</tbody>
</table>

Table 370: Graph-MC, relative MSE. $T = 250, N = 1024, \pi_b(a = 0) = 0.2, \pi_e(a = 0) = 0.8$.

<table>
<thead>
<tr>
<th></th>
<th>IPS Standard Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E0</td>
</tr>
<tr>
<td>WIS</td>
<td>2.0E1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>2.0E1</td>
</tr>
</tbody>
</table>
Table 371: Graph-MC, relative MSE. $T = 250$, $N = 128, \pi_b(a = 0) = 0.5, \pi_e(a = 0) = 0.5$.

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AM</td>
<td>3.5E-1</td>
<td>1.7E-3</td>
<td>1.7E-3</td>
<td>3.7E-3</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.7E-2</td>
<td>1.5E-4</td>
<td>1.5E-4</td>
<td>1.3E-4</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.0E-2</td>
<td>5.0E-3</td>
<td>5.0E-3</td>
<td>6.1E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>3.5E-2</td>
<td>6.0E-5</td>
<td>6.0E-5</td>
<td>6.5E-5</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.7E-1</td>
<td>7.1E-4</td>
<td>7.1E-4</td>
<td>1.6E-3</td>
</tr>
<tr>
<td>Q^b(\lambda)</td>
<td>1.6E-1</td>
<td>3.0E-4</td>
<td>5.4E-4</td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>7.9E-1</td>
<td>3.0E-4</td>
<td>3.0E-4</td>
<td>5.4E-4</td>
</tr>
<tr>
<td>IH</td>
<td>2.2E-4</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>HYBRID</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td>2.2E-4</td>
<td>2.2E-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>2.2E-4</td>
<td>2.2E-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>2.2E-4</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

Table 372: Graph-MC, relative MSE. $T = 250$, $N = 256, \pi_b(a = 0) = 0.5, \pi_e(a = 0) = 0.5$.

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AM</td>
<td>3.4E-1</td>
<td>2.8E-4</td>
<td>2.8E-4</td>
<td>6.7E-4</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.5E-2</td>
<td>2.6E-5</td>
<td>2.6E-5</td>
<td>2.3E-5</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.6E-2</td>
<td>1.5E-3</td>
<td>1.5E-3</td>
<td>1.4E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>3.4E-2</td>
<td>9.0E-6</td>
<td>9.0E-6</td>
<td>1.1E-5</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.7E-1</td>
<td>2.2E-4</td>
<td>2.2E-4</td>
<td>3.0E-4</td>
</tr>
<tr>
<td>Q^b(\lambda)</td>
<td>1.6E-1</td>
<td>2.2E-4</td>
<td>2.2E-4</td>
<td>3.0E-4</td>
</tr>
<tr>
<td>Tree</td>
<td>7.9E-1</td>
<td>1.1E-4</td>
<td>1.1E-4</td>
<td>1.5E-4</td>
</tr>
<tr>
<td>IH</td>
<td>1.0E-4</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>HYBRID</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td>2.2E-4</td>
<td>2.2E-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>2.2E-4</td>
<td>2.2E-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>2.2E-4</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

Table 373: Graph-MC, relative MSE. $T = 250$, $N = 512, \pi_b(a = 0) = 0.5, \pi_e(a = 0) = 0.5$.

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AM</td>
<td>3.5E-1</td>
<td>1.7E-3</td>
<td>1.7E-3</td>
<td>3.7E-3</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.7E-2</td>
<td>1.5E-4</td>
<td>1.5E-4</td>
<td>1.3E-4</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.0E-2</td>
<td>5.0E-3</td>
<td>5.0E-3</td>
<td>6.1E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>3.5E-2</td>
<td>6.0E-5</td>
<td>6.0E-5</td>
<td>6.5E-5</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.7E-1</td>
<td>7.0E-4</td>
<td>7.0E-4</td>
<td>1.5E-3</td>
</tr>
<tr>
<td>Q^b(\lambda)</td>
<td>1.6E-1</td>
<td>7.1E-4</td>
<td>7.1E-4</td>
<td>1.6E-3</td>
</tr>
<tr>
<td>Tree</td>
<td>7.9E-1</td>
<td>3.0E-4</td>
<td>3.0E-4</td>
<td>5.4E-4</td>
</tr>
<tr>
<td>IH</td>
<td>2.2E-4</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>HYBRID</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td>2.2E-4</td>
<td>2.2E-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>2.2E-4</td>
<td>2.2E-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>2.2E-4</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

Table 374: Graph-MC, relative MSE. $T = 250$, $N = 1024, \pi_b(a = 0) = 0.5, \pi_e(a = 0) = 0.5$.

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AM</td>
<td>3.4E-1</td>
<td>2.8E-4</td>
<td>2.8E-4</td>
<td>6.7E-4</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.5E-2</td>
<td>2.6E-5</td>
<td>2.6E-5</td>
<td>2.3E-5</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.6E-2</td>
<td>1.5E-3</td>
<td>1.5E-3</td>
<td>1.4E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>3.4E-2</td>
<td>9.0E-6</td>
<td>9.0E-6</td>
<td>1.1E-5</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.7E-1</td>
<td>2.2E-4</td>
<td>2.2E-4</td>
<td>3.0E-4</td>
</tr>
<tr>
<td>Q^b(\lambda)</td>
<td>1.6E-1</td>
<td>2.2E-4</td>
<td>2.2E-4</td>
<td>3.0E-4</td>
</tr>
<tr>
<td>Tree</td>
<td>7.9E-1</td>
<td>1.1E-4</td>
<td>1.1E-4</td>
<td>1.5E-4</td>
</tr>
<tr>
<td>IH</td>
<td>1.0E-4</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>HYBRID</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td>2.2E-4</td>
<td>2.2E-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>2.2E-4</td>
<td>2.2E-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>2.2E-4</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
Table 375: Graph-MC, relative MSE. $T = 250, N = 128, \pi_b(a = 0) = 0.5, \pi_e(a = 0) = 0.6.$

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>7.6E-3</td>
<td>9.9E-3</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>7.3E-3</td>
<td>4.1E-4</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.2E-3</td>
<td>3.4E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>4.0E-3</td>
<td>4.7E-5</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.1E-1</td>
<td>3.0E-3</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>1.4E-3</td>
<td>3.1E-4</td>
</tr>
<tr>
<td>Tree</td>
<td>5.8E-1</td>
<td>9.5E-3</td>
</tr>
<tr>
<td>IH</td>
<td>4.0E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

---

Table 377: Graph-MC, relative MSE. $T = 250, N = 512, \pi_b(a = 0) = 0.5, \pi_e(a = 0) = 0.6.$

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>5.2E-2</td>
<td>4.3E-3</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>5.9E-3</td>
<td>3.4E-5</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.5E-3</td>
<td>4.4E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>4.3E-3</td>
<td>8.0E-6</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>7.9E-2</td>
<td>4.9E-4</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>2.1E-4</td>
<td>6.0E-6</td>
</tr>
<tr>
<td>Tree</td>
<td>5.7E-1</td>
<td>2.4E-3</td>
</tr>
<tr>
<td>IH</td>
<td>3.4E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

---

Table 376: Graph-MC, relative MSE. $T = 250, N = 256, \pi_b(a = 0) = 0.5, \pi_e(a = 0) = 0.6.$

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>5.4E-2</td>
<td>4.8E-3</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>3.3E-3</td>
<td>1.8E-4</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.6E-3</td>
<td>4.3E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>2.8E-3</td>
<td>2.3E-4</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>8.7E-2</td>
<td>6.2E-4</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>7.0E-6</td>
<td>2.0E-4</td>
</tr>
<tr>
<td>Tree</td>
<td>5.7E-1</td>
<td>2.6E-3</td>
</tr>
<tr>
<td>IH</td>
<td>2.9E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

---

Table 378: Graph-MC, relative MSE. $T = 250, N = 1024, \pi_b(a = 0) = 0.5, \pi_e(a = 0) = 0.6.$

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>4.6E-2</td>
<td>7.3E-4</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>3.3E-3</td>
<td>5.3E-6</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.8E-3</td>
<td>1.3E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>4.3E-3</td>
<td>2.6E-6</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>7.7E-2</td>
<td>6.6E-5</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>1.6E-4</td>
<td>9.4E-7</td>
</tr>
<tr>
<td>Tree</td>
<td>5.7E-1</td>
<td>4.0E-4</td>
</tr>
<tr>
<td>IH</td>
<td>2.9E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

---

Table 379: Graph-MC, relative MSE. $T = 250, N = 1024, \pi_b(a = 0) = 0.5, \pi_e(a = 0) = 0.6.$

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>4.6E-2</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>1.4E-2</td>
<td>1.8E-2</td>
</tr>
<tr>
<td>NAIVE</td>
<td>1.6E0</td>
<td>-</td>
</tr>
</tbody>
</table>

---

Table 380: Graph-MC, relative MSE. $T = 250, N = 1024, \pi_b(a = 0) = 0.5, \pi_e(a = 0) = 0.6.$

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.7E-2</td>
<td>3.0E-3</td>
</tr>
<tr>
<td>WIS</td>
<td>2.9E-3</td>
<td>1.8E-3</td>
</tr>
<tr>
<td>NAIVE</td>
<td>1.5E0</td>
<td>-</td>
</tr>
</tbody>
</table>

---

Table 381: Graph-MC, relative MSE. $T = 250, N = 1024, \pi_b(a = 0) = 0.5, \pi_e(a = 0) = 0.6.$

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.2E-2</td>
<td>6.8E-4</td>
</tr>
<tr>
<td>WIS</td>
<td>1.2E-3</td>
<td>9.2E-4</td>
</tr>
<tr>
<td>NAIVE</td>
<td>1.5E0</td>
<td>-</td>
</tr>
</tbody>
</table>
Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning

Table 379: Graph-MC, relative MSE. $T = 250, N = 128, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.5.$

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>4.2E-1</td>
<td>4.3E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.0E0</td>
<td>1.2E3</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.2E0</td>
<td>4.6E0</td>
</tr>
<tr>
<td>FQE</td>
<td>2.9E-2</td>
<td>1.4E0</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>3.2E-1</td>
<td>6.0E-1</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>1.0E-1</td>
<td>5.3E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>7.2E-1</td>
<td>1.5E0</td>
</tr>
<tr>
<td>IH</td>
<td>3.6E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 380: Graph-MC, relative MSE. $T = 250, N = 256, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.5.$

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>3.8E-1</td>
<td>1.1E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>8.4E-2</td>
<td>2.4E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>8.3E-2</td>
<td>6.2E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>6.2E-3</td>
<td>4.8E-3</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>3.0E-1</td>
<td>1.1E-1</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>1.4E-2</td>
<td>2.5E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>7.1E-1</td>
<td>8.2E-2</td>
</tr>
<tr>
<td>IH</td>
<td>3.3E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 381: Graph-MC, relative MSE. $T = 250, N = 512, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.5.$

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>3.6E-1</td>
<td>1.0E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>6.1E-2</td>
<td>2.2E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.6E-2</td>
<td>4.8E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>5.6E-3</td>
<td>4.8E-3</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>2.7E-1</td>
<td>6.4E-2</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>2.1E-2</td>
<td>2.2E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>7.2E-1</td>
<td>5.7E-2</td>
</tr>
<tr>
<td>IH</td>
<td>1.4E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 382: Graph-MC, relative MSE. $T = 250, N = 1024, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.5.$

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>3.4E-1</td>
<td>7.4E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.2E-2</td>
<td>2.0E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.2E-2</td>
<td>2.6E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>5.4E-3</td>
<td>2.3E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>2.7E-1</td>
<td>2.4E-2</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>1.7E-2</td>
<td>1.5E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>7.2E-1</td>
<td>4.0E-2</td>
</tr>
<tr>
<td>IH</td>
<td>1.6E-3</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>2.5E2</td>
<td>1.5E0</td>
</tr>
<tr>
<td>WIS</td>
<td>1.5E-1</td>
<td>$\bf{3.1E-2}$</td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.1E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>6.9E-1</td>
<td>5.7E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>9.1E-2</td>
<td>$\bf{8.8E-3}$</td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.0E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>6.1E-1</td>
<td>9.1E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>7.4E-2</td>
<td>$\bf{9.5E-3}$</td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.0E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>3.3E0</td>
<td>4.0E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>5.7E-2</td>
<td>$\bf{3.1E-3}$</td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.1E-1</td>
<td>-</td>
</tr>
</tbody>
</table>
### Table 383: Graph-MC, relative MSE. $T = 250, N = 128, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.6.$

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>3.8E-2</td>
<td>7.7E-3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.2E-3</td>
<td>3.5E-4</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.3E-3</td>
<td>1.8E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>6.3E-5</td>
<td>3.7E-5</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>3.9E-4</td>
<td>1.4E-4</td>
</tr>
<tr>
<td>Q^\theta(\lambda)</td>
<td>3.6E-4</td>
<td>1.3E-4</td>
</tr>
<tr>
<td>TREE</td>
<td>4.7E-1</td>
<td>1.4E-3</td>
</tr>
<tr>
<td>IH</td>
<td>1.8E-3</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 384: Graph-MC, relative MSE. $T = 250, N = 256, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.6.$

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>3.3E-2</td>
<td>1.1E-3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>9.7E-4</td>
<td>8.9E-5</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.0E-3</td>
<td>3.1E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>4.8E-5</td>
<td>3.7E-5</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.4E-4</td>
<td>8.4E-5</td>
</tr>
<tr>
<td>Q^\theta(\lambda)</td>
<td>1.3E-4</td>
<td>8.3E-5</td>
</tr>
<tr>
<td>TREE</td>
<td>4.7E-1</td>
<td>2.8E-4</td>
</tr>
<tr>
<td>IH</td>
<td>3.6E-4</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 385: Graph-MC, relative MSE. $T = 250, N = 512, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.6.$

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>3.0E-2</td>
<td>1.2E-3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>6.6E-4</td>
<td>1.7E-5</td>
</tr>
<tr>
<td>MRDR</td>
<td>7.6E-4</td>
<td>6.6E-4</td>
</tr>
<tr>
<td>FQE</td>
<td>1.3E-5</td>
<td>1.8E-5</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.0E-5</td>
<td>1.5E-5</td>
</tr>
<tr>
<td>Q^\theta(\lambda)</td>
<td>1.1E-5</td>
<td>1.5E-5</td>
</tr>
<tr>
<td>TREE</td>
<td>4.6E-1</td>
<td>3.4E-4</td>
</tr>
<tr>
<td>IH</td>
<td>3.4E-4</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 386: Graph-MC, relative MSE. $T = 250, N = 1024, \pi_b(a = 0) = 0.6, \pi_e(a = 0) = 0.6.$

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>3.0E-2</td>
<td>6.0E-4</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.3E-4</td>
<td>2.0E-6</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.6E-4</td>
<td>2.1E-4</td>
</tr>
<tr>
<td>FQE</td>
<td>2.0E-6</td>
<td>2.0E-6</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>6.0E-6</td>
<td>3.0E-6</td>
</tr>
<tr>
<td>Q^\theta(\lambda)</td>
<td>5.0E-6</td>
<td>3.0E-6</td>
</tr>
<tr>
<td>TREE</td>
<td>4.5E-1</td>
<td>4.9E-4</td>
</tr>
<tr>
<td>IH</td>
<td>4.0E-4</td>
<td>-</td>
</tr>
</tbody>
</table>

### IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.8E-3</td>
<td>1.8E-3</td>
</tr>
<tr>
<td>WIS</td>
<td>1.8E-3</td>
<td>1.8E-3</td>
</tr>
<tr>
<td>NAIVE</td>
<td>1.8E-3</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>3.4E-4</td>
<td>3.4E-4</td>
</tr>
<tr>
<td>WIS</td>
<td>3.4E-4</td>
<td>3.4E-4</td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.4E-4</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>3.6E-4</td>
<td>3.6E-4</td>
</tr>
<tr>
<td>WIS</td>
<td>3.6E-4</td>
<td>3.6E-4</td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.6E-4</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>4.0E-4</td>
<td>4.0E-4</td>
</tr>
<tr>
<td>WIS</td>
<td>4.0E-4</td>
<td>4.0E-4</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E-4</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 387: Graph-MC, relative MSE. $T = 250, N = 128, \pi_b(a = 0) = 0.8, \pi_e(a = 0) = 0.2$.

<table>
<thead>
<tr>
<th></th>
<th>DM Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
</tr>
<tr>
<td>AM</td>
<td>9.0E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>8.2E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>7.4E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>8.7E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>8.7E-1</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>1.1E125</td>
</tr>
<tr>
<td>Tree</td>
<td>8.8E-1</td>
</tr>
<tr>
<td>IH</td>
<td>4.5E-1</td>
</tr>
</tbody>
</table>

Table 388: Graph-MC, relative MSE. $T = 250, N = 256, \pi_b(a = 0) = 0.8, \pi_e(a = 0) = 0.2$.

<table>
<thead>
<tr>
<th></th>
<th>DM Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
</tr>
<tr>
<td>AM</td>
<td>8.8E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>8.2E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>7.4E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>8.7E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>8.7E-1</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>3.4E114</td>
</tr>
<tr>
<td>Tree</td>
<td>8.5E-1</td>
</tr>
<tr>
<td>IH</td>
<td>4.3E-1</td>
</tr>
</tbody>
</table>

Table 389: Graph-MC, relative MSE. $T = 250, N = 512, \pi_b(a = 0) = 0.8, \pi_e(a = 0) = 0.2$.

<table>
<thead>
<tr>
<th></th>
<th>DM Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
</tr>
<tr>
<td>AM</td>
<td>8.7E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>8.4E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>7.5E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>8.5E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>8.5E-1</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>3.4E114</td>
</tr>
<tr>
<td>Tree</td>
<td>8.5E-1</td>
</tr>
<tr>
<td>IH</td>
<td>4.3E-1</td>
</tr>
</tbody>
</table>

Table 390: Graph-MC, relative MSE. $T = 250, N = 1024, \pi_b(a = 0) = 0.8, \pi_e(a = 0) = 0.2$.

<table>
<thead>
<tr>
<th></th>
<th>DM Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
</tr>
<tr>
<td>AM</td>
<td>8.5E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>7.5E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>6.9E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>8.2E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>8.2E-1</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>2.4E112</td>
</tr>
<tr>
<td>Tree</td>
<td>8.3E-1</td>
</tr>
<tr>
<td>IH</td>
<td>3.5E-1</td>
</tr>
</tbody>
</table>

Table 388: Graph-MC, relative MSE. $T = 250, N = 1024, \pi_b(a = 0) = 0.8, \pi_e(a = 0) = 0.2$.

<table>
<thead>
<tr>
<th></th>
<th>DM Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
</tr>
<tr>
<td>AM</td>
<td>8.5E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>7.5E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>6.9E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>8.2E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>8.2E-1</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>2.4E112</td>
</tr>
<tr>
<td>Tree</td>
<td>8.3E-1</td>
</tr>
<tr>
<td>IH</td>
<td>3.5E-1</td>
</tr>
</tbody>
</table>
### 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).

<table>
<thead>
<tr>
<th>DM</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>7.0E-2</td>
<td>1.3E-2</td>
<td>7.6E-3</td>
<td>1.0E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.2E-1</td>
<td>2.6E-3</td>
<td>7.9E-4</td>
<td>2.4E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.2E-1</td>
<td>8.2E-3</td>
<td>5.9E-4</td>
<td>5.9E-4</td>
</tr>
<tr>
<td>FQE</td>
<td>5.7E-1</td>
<td>7.5E-3</td>
<td>1.2E-3</td>
<td>1.1E-3</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.7E-1</td>
<td>1.8E-3</td>
<td>3.8E-4</td>
<td>3.9E-3</td>
</tr>
<tr>
<td>Q($\pi$)($\lambda$)</td>
<td>1.5E-1</td>
<td>1.9E-3</td>
<td>3.9E-4</td>
<td>5.1E-3</td>
</tr>
<tr>
<td>TREE</td>
<td>1.7E-1</td>
<td>1.8E-3</td>
<td>3.7E-4</td>
<td>4.3E-3</td>
</tr>
<tr>
<td>IH</td>
<td><strong>3.2E-2</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 392: MC, relative MSE. Model Type: linear. $T = 250, N = 256, \pi_b = 0.10$-Greedy(DDQN), $\pi_e = 0.00$-Greedy(DDQN).

<table>
<thead>
<tr>
<th>DM</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>5.4E-2</td>
<td>2.1E-3</td>
<td>2.0E-3</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.1E-1</td>
<td>1.1E-3</td>
<td>4.2E-4</td>
<td>4.5E-4</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.9E-1</td>
<td>2.5E-3</td>
<td>6.1E-4</td>
<td>1.2E-3</td>
</tr>
<tr>
<td>FQE</td>
<td><strong>8.6E-3</strong></td>
<td>2.2E-4</td>
<td><strong>1.8E-4</strong></td>
<td>1.9E-4</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.6E-1</td>
<td>1.5E-3</td>
<td>1.2E-3</td>
<td>2.5E-3</td>
</tr>
<tr>
<td>Q($\pi$)($\lambda$)</td>
<td>3.4E-2</td>
<td>1.2E-3</td>
<td>6.9E-4</td>
<td>2.6E-3</td>
</tr>
<tr>
<td>TREE</td>
<td>1.4E-2</td>
<td>1.9E-3</td>
<td>1.1E-3</td>
<td>1.5E-3</td>
</tr>
<tr>
<td>IH</td>
<td><strong>3.1E-2</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### IPS

<table>
<thead>
<tr>
<th>DM</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>7.5E-2</td>
<td>7.0E-3</td>
<td>4.7E-4</td>
<td>1.0E-3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.0E-1</td>
<td>2.5E-3</td>
<td>8.6E-4</td>
<td>6.5E-4</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.2E-1</td>
<td>2.9E-3</td>
<td>2.8E-4</td>
<td><strong>1.4E-4</strong></td>
</tr>
<tr>
<td>FQE</td>
<td>5.8E-1</td>
<td>5.3E-3</td>
<td>1.7E-3</td>
<td>1.7E-3</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.7E-1</td>
<td>2.2E-3</td>
<td>5.6E-4</td>
<td>2.2E-3</td>
</tr>
<tr>
<td>Q($\pi$)($\lambda$)</td>
<td>1.5E-1</td>
<td>2.3E-3</td>
<td>6.1E-4</td>
<td>1.5E-3</td>
</tr>
<tr>
<td>TREE</td>
<td>1.7E-1</td>
<td>2.2E-3</td>
<td>5.8E-4</td>
<td>1.6E-3</td>
</tr>
<tr>
<td>IH</td>
<td><strong>3.2E-2</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 393: MC, relative MSE. Model Type: NN. $T = 250, N = 128, \pi_b = 0.10$-Greedy(DDQN), $\pi_e = 0.00$-Greedy(DDQN).

<table>
<thead>
<tr>
<th>DM</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.0E-1</td>
<td>9.7E-3</td>
<td>9.5E-3</td>
<td>1.5E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.0E-1</td>
<td>2.7E-3</td>
<td>8.9E-4</td>
<td>3.2E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>8.8E-1</td>
<td>6.7E-3</td>
<td>8.4E-4</td>
<td><strong>5.2E-4</strong></td>
</tr>
<tr>
<td>FQE</td>
<td>1.2E-2</td>
<td>9.5E-4</td>
<td>7.6E-4</td>
<td>7.6E-4</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td><strong>8.9E-3</strong></td>
<td>6.4E-3</td>
<td>1.7E-3</td>
<td>5.2E-3</td>
</tr>
<tr>
<td>Q($\pi$)($\lambda$)</td>
<td>1.4E-1</td>
<td>5.7E-3</td>
<td>1.6E-3</td>
<td>2.8E-3</td>
</tr>
<tr>
<td>TREE</td>
<td>8.7E-2</td>
<td>5.7E-3</td>
<td>1.5E-3</td>
<td>6.6E-3</td>
</tr>
<tr>
<td>IH</td>
<td>3.3E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 394: MC, relative MSE. Model Type: NN. $T = 250, N = 256, \pi_b = 0.10$-Greedy(DDQN), $\pi_e = 0.00$-Greedy(DDQN).

<table>
<thead>
<tr>
<th>DM</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>5.4E-2</td>
<td>2.1E-3</td>
<td>2.0E-3</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.1E-1</td>
<td>1.1E-3</td>
<td>4.2E-4</td>
<td>4.5E-4</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.9E-1</td>
<td>2.5E-3</td>
<td>6.1E-4</td>
<td>1.2E-3</td>
</tr>
<tr>
<td>FQE</td>
<td><strong>8.6E-3</strong></td>
<td>2.2E-4</td>
<td><strong>1.8E-4</strong></td>
<td>1.9E-4</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.6E-1</td>
<td>1.5E-3</td>
<td>1.2E-3</td>
<td>2.5E-3</td>
</tr>
<tr>
<td>Q($\pi$)($\lambda$)</td>
<td>3.4E-2</td>
<td>1.2E-3</td>
<td>6.9E-4</td>
<td>2.6E-3</td>
</tr>
<tr>
<td>TREE</td>
<td>1.4E-2</td>
<td>1.9E-3</td>
<td>1.1E-3</td>
<td>1.5E-3</td>
</tr>
<tr>
<td>IH</td>
<td><strong>3.1E-2</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### IPS

<table>
<thead>
<tr>
<th>DM</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.6E-2</td>
<td>3.3E-3</td>
<td>3.3E-3</td>
<td>3.4E-4</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.7E-4</td>
<td>3.4E-4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.2E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
### Table 395: MC, relative MSE. Model Type: linear. $T = 250, N = 128, \pi_b = 0.10\text{-Greedy}(\text{DDQN}), \pi_e = 1.00\text{-Greedy}(\text{DDQN})$.

<table>
<thead>
<tr>
<th>DM Hybrid</th>
<th>Direct</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>5.2E-1</td>
<td>4.1E-1</td>
<td>2.4E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>9.2E-1</td>
<td>8.6E-1</td>
<td>2.0E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.9E-1</td>
<td>8.8E-1</td>
<td>4.3E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>6.0E-1</td>
<td>5.2E-1</td>
<td>2.3E-1</td>
</tr>
<tr>
<td>R(λ)</td>
<td>7.3E-1</td>
<td>6.4E-1</td>
<td>4.6E-1</td>
</tr>
<tr>
<td>Q^*(λ)</td>
<td>6.4E-1</td>
<td>5.9E-1</td>
<td>4.6E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>7.3E-1</td>
<td>6.4E-1</td>
<td>4.7E-1</td>
</tr>
<tr>
<td>IH</td>
<td>4.1E-1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 397: MC, relative MSE. Model Type: linear. $T = 250, N = 128, \pi_b = 0.10\text{-Greedy}(\text{DDQN}), \pi_e = 1.00\text{-Greedy}(\text{DDQN})$.

<table>
<thead>
<tr>
<th>DM Hybrid</th>
<th>Direct</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>9.5E-1</td>
<td>8.5E-1</td>
<td>7.1E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>9.2E-1</td>
<td>7.5E-1</td>
<td>4.3E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.0E0</td>
<td>8.5E-1</td>
<td>4.2E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>6.1E-2</td>
<td>4.9E-2</td>
<td>6.8E-2</td>
</tr>
<tr>
<td>R(λ)</td>
<td>3.5E-1</td>
<td>3.0E-1</td>
<td>2.0E-1</td>
</tr>
<tr>
<td>Q^*(λ)</td>
<td>1.8E0</td>
<td>1.4E0</td>
<td>3.1E0</td>
</tr>
<tr>
<td>Tree</td>
<td>2.4E-1</td>
<td>2.4E-1</td>
<td>2.1E-1</td>
</tr>
<tr>
<td>IH</td>
<td>4.1E-1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 396: MC, relative MSE. Model Type: NN. $T = 250, N = 256, \pi_b = 0.10\text{-Greedy}(\text{DDQN}), \pi_e = 1.00\text{-Greedy}(\text{DDQN})$.

<table>
<thead>
<tr>
<th>DM Hybrid</th>
<th>Direct</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>9.5E-1</td>
<td>8.5E-1</td>
<td>7.1E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>9.2E-1</td>
<td>7.5E-1</td>
<td>4.3E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.0E0</td>
<td>8.5E-1</td>
<td>4.2E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.3E-2</td>
<td>3.3E-2</td>
<td>1.6E-2</td>
</tr>
<tr>
<td>R(λ)</td>
<td>5.7E-1</td>
<td>4.6E-1</td>
<td>2.0E-1</td>
</tr>
<tr>
<td>Q^*(λ)</td>
<td>2.1E1</td>
<td>2.0E1</td>
<td>1.8E1</td>
</tr>
<tr>
<td>Tree</td>
<td>4.5E-1</td>
<td>3.4E-1</td>
<td>1.3E-1</td>
</tr>
<tr>
<td>IH</td>
<td>4.0E-1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 398: MC, relative MSE. Model Type: NN. $T = 250, N = 256, \pi_b = 0.10\text{-Greedy}(\text{DDQN}), \pi_e = 1.00\text{-Greedy}(\text{DDQN})$.

<table>
<thead>
<tr>
<th>DM Hybrid</th>
<th>Direct</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>9.5E-1</td>
<td>8.5E-1</td>
<td>7.1E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>9.2E-1</td>
<td>7.5E-1</td>
<td>4.3E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.0E0</td>
<td>8.5E-1</td>
<td>4.2E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.3E-2</td>
<td>3.3E-2</td>
<td>1.6E-2</td>
</tr>
<tr>
<td>R(λ)</td>
<td>5.7E-1</td>
<td>4.6E-1</td>
<td>2.0E-1</td>
</tr>
<tr>
<td>Q^*(λ)</td>
<td>2.1E1</td>
<td>2.0E1</td>
<td>1.8E1</td>
</tr>
<tr>
<td>Tree</td>
<td>4.5E-1</td>
<td>3.4E-1</td>
<td>1.3E-1</td>
</tr>
<tr>
<td>IH</td>
<td>4.0E-1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 399: MC, relative MSE. Model Type: NN. $T = 250, N = 256, \pi_b = 0.10\text{-Greedy}(\text{DDQN}), \pi_e = 1.00\text{-Greedy}(\text{DDQN})$.

<table>
<thead>
<tr>
<th>DM Hybrid</th>
<th>Direct</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>9.5E-1</td>
<td>8.5E-1</td>
<td>7.1E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>9.2E-1</td>
<td>7.5E-1</td>
<td>4.3E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.0E0</td>
<td>8.5E-1</td>
<td>4.2E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.3E-2</td>
<td>3.3E-2</td>
<td>1.6E-2</td>
</tr>
<tr>
<td>R(λ)</td>
<td>5.7E-1</td>
<td>4.6E-1</td>
<td>2.0E-1</td>
</tr>
<tr>
<td>Q^*(λ)</td>
<td>2.1E1</td>
<td>2.0E1</td>
<td>1.8E1</td>
</tr>
<tr>
<td>Tree</td>
<td>4.5E-1</td>
<td>3.4E-1</td>
<td>1.3E-1</td>
</tr>
<tr>
<td>IH</td>
<td>4.0E-1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 399: MC, relative MSE. Model Type: linear. $T = 250, N = 128, \pi_b = 1.00$-Greedy(DDQN), $\pi_e = 0.00$-Greedy(DDQN).

<table>
<thead>
<tr>
<th>DM Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.4E-1</td>
<td>3.2E-1</td>
<td>3.0E0</td>
<td>2.7E0</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>6.6E-1</td>
<td>5.1E2</td>
<td>7.9E26</td>
<td>6.7E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.4E-1</td>
<td>9.7E0</td>
<td>5.0E0</td>
<td>4.9E0</td>
</tr>
<tr>
<td>FQE</td>
<td>2.2E-1</td>
<td>1.8E0</td>
<td>2.9E0</td>
<td>2.7E0</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.6E-2</td>
<td>3.4E0</td>
<td>1.9E0</td>
<td>1.6E0</td>
</tr>
<tr>
<td>Q^8(\lambda)</td>
<td>9.1E-2</td>
<td>3.6E0</td>
<td>2.0E0</td>
<td>1.8E0</td>
</tr>
<tr>
<td>Tree</td>
<td>1.5E-2</td>
<td>3.4E0</td>
<td>1.9E0</td>
<td>1.7E0</td>
</tr>
<tr>
<td>IH</td>
<td>5.2E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E0</td>
</tr>
<tr>
<td>WIS</td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.2E0</td>
</tr>
</tbody>
</table>

Table 400: MC, relative MSE. Model Type: linear. $T = 250, N = 256, \pi_b = 1.00$-Greedy(DDQN), $\pi_e = 0.00$-Greedy(DDQN).

<table>
<thead>
<tr>
<th>DM Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.2E-1</td>
<td>1.7E-1</td>
<td>3.6E0</td>
<td>3.4E0</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>7.0E-1</td>
<td>2.5E23</td>
<td>1.5E27</td>
<td>1.9E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.6E-1</td>
<td>6.7E-1</td>
<td>5.1E0</td>
<td>5.0E0</td>
</tr>
<tr>
<td>FQE</td>
<td>2.1E-1</td>
<td>2.3E-1</td>
<td>3.0E0</td>
<td>2.8E0</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.6E-2</td>
<td>8.0E-3</td>
<td>1.8E0</td>
<td>1.7E0</td>
</tr>
<tr>
<td>Q^8(\lambda)</td>
<td>8.2E-2</td>
<td>2.2E-1</td>
<td>1.6E0</td>
<td>1.6E0</td>
</tr>
<tr>
<td>Tree</td>
<td>1.5E-2</td>
<td>8.4E-3</td>
<td>1.7E0</td>
<td>1.7E0</td>
</tr>
<tr>
<td>IH</td>
<td>5.1E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E0</td>
</tr>
<tr>
<td>WIS</td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.2E0</td>
</tr>
</tbody>
</table>

Table 401: MC, relative MSE. Model Type: NN. $T = 250, N = 128, \pi_b = 1.00$-Greedy(DDQN), $\pi_e = 0.00$-Greedy(DDQN).

<table>
<thead>
<tr>
<th>DM Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>8.6E-1</td>
<td>7.3E-1</td>
<td>1.7E1</td>
<td>4.4E0</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>8.4E-1</td>
<td>5.5E-1</td>
<td>1.1E1</td>
<td>1.1E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.4E-1</td>
<td>5.0E-1</td>
<td>7.5E0</td>
<td>7.3E0</td>
</tr>
<tr>
<td>FQE</td>
<td>1.2E-1</td>
<td>3.9E-1</td>
<td>2.3E0</td>
<td>2.2E0</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.5E0</td>
<td>1.7E0</td>
<td>4.9E0</td>
<td>4.7E0</td>
</tr>
<tr>
<td>Q^8(\lambda)</td>
<td>5.4E0</td>
<td>5.6E0</td>
<td>4.5E0</td>
<td>4.4E0</td>
</tr>
<tr>
<td>Tree</td>
<td>2.0E0</td>
<td>2.3E0</td>
<td>4.8E0</td>
<td>4.6E0</td>
</tr>
<tr>
<td>IH</td>
<td>5.2E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E0</td>
</tr>
<tr>
<td>WIS</td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.2E0</td>
</tr>
</tbody>
</table>

Table 402: MC, relative MSE. Model Type: NN. $T = 250, N = 256, \pi_b = 1.00$-Greedy(DDQN), $\pi_e = 0.00$-Greedy(DDQN).

<table>
<thead>
<tr>
<th>DM Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>8.6E-1</td>
<td>7.3E-1</td>
<td>1.7E1</td>
<td>4.4E0</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>8.4E-1</td>
<td>5.5E-1</td>
<td>1.1E1</td>
<td>1.1E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.0E-1</td>
<td>4.7E-1</td>
<td>7.0E0</td>
<td>6.9E0</td>
</tr>
<tr>
<td>FQE</td>
<td>2.7E-2</td>
<td>1.3E-1</td>
<td>3.2E-1</td>
<td>2.9E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.0E0</td>
<td>1.2E0</td>
<td>4.3E0</td>
<td>4.3E0</td>
</tr>
<tr>
<td>Q^8(\lambda)</td>
<td>5.6E0</td>
<td>5.6E0</td>
<td>4.4E0</td>
<td>4.5E0</td>
</tr>
<tr>
<td>Tree</td>
<td>2.1E0</td>
<td>2.5E0</td>
<td>4.7E0</td>
<td>4.7E0</td>
</tr>
<tr>
<td>IH</td>
<td>5.1E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E0</td>
</tr>
<tr>
<td>WIS</td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.1E0</td>
</tr>
</tbody>
</table>
Table 403: MC, relative MSE. Model Type: linear. $T = 250, N = 128, \pi_0 = 1.00$-Greedy(DDQN), $\pi_e = 0.10$-Greedy(DDQN).

<table>
<thead>
<tr>
<th>DM Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>9.3E-2</td>
<td>3.2E-1</td>
<td>2.9E0</td>
<td>1.0E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>6.3E-1</td>
<td>4.6E-1</td>
<td>4.7E0</td>
<td>1.6E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>8.5E-1</td>
<td>6.0E-1</td>
<td>3.5E-1</td>
<td>3.5E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>2.7E-1</td>
<td>1.4E-1</td>
<td>7.1E-1</td>
<td>6.1E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>6.5E-2</td>
<td>1.8E-2</td>
<td>3.5E-1</td>
<td>2.9E-1</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>9.4E-3</td>
<td>1.1E-1</td>
<td>4.9E-1</td>
<td>3.9E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>6.3E-2</td>
<td>1.8E-2</td>
<td>3.7E-1</td>
<td>3.1E-1</td>
</tr>
<tr>
<td>IH</td>
<td>3.0E0</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

Table 404: MC, relative MSE. Model Type: linear. $T = 250, N = 256, \pi_0 = 1.00$-Greedy(DDQN), $\pi_e = 0.10$-Greedy(DDQN).

<table>
<thead>
<tr>
<th>IPS Standard Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
</tr>
<tr>
<td>WIS</td>
</tr>
<tr>
<td>NAIVE</td>
</tr>
</tbody>
</table>

Table 405: MC, relative MSE. Model Type: NN. $T = 250, N = 128, \pi_0 = 1.00$-Greedy(DDQN), $\pi_e = 0.10$-Greedy(DDQN).

<table>
<thead>
<tr>
<th>DM Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>8.7E-1</td>
<td>6.1E-1</td>
<td>7.7E-1</td>
<td>7.5E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>8.1E-1</td>
<td>5.7E-1</td>
<td>3.7E-1</td>
<td>4.1E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.6E-1</td>
<td>6.9E-1</td>
<td>6.5E-1</td>
<td>6.6E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>1.7E-2</td>
<td>3.9E-3</td>
<td>3.9E-1</td>
<td>2.0E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>8.1E-1</td>
<td>9.0E-1</td>
<td>1.5E0</td>
<td>1.4E0</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>3.7E0</td>
<td>3.8E0</td>
<td>3.2E0</td>
<td>3.2E0</td>
</tr>
<tr>
<td>TREE</td>
<td>8.9E-1</td>
<td>9.8E-1</td>
<td>1.6E0</td>
<td>1.4E0</td>
</tr>
<tr>
<td>IH</td>
<td>3.1E0</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

Table 406: MC, relative MSE. Model Type: NN. $T = 250, N = 256, \pi_0 = 1.00$-Greedy(DDQN), $\pi_e = 0.10$-Greedy(DDQN).

<table>
<thead>
<tr>
<th>IPS Standard Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
</tr>
<tr>
<td>WIS</td>
</tr>
<tr>
<td>NAIVE</td>
</tr>
</tbody>
</table>

Table 407: MC, relative MSE. Model Type: NN. $T = 250, N = 128, \pi_0 = 1.00$-Greedy(DDQN), $\pi_e = 0.10$-Greedy(DDQN).

<table>
<thead>
<tr>
<th>IPS Standard Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
</tr>
<tr>
<td>WIS</td>
</tr>
<tr>
<td>NAIVE</td>
</tr>
</tbody>
</table>
Table 407: MC, relative MSE. Model Type: linear. $T = 250, N = 128, \pi_b = 0.10$-Greedy(DDQN), $\pi_e = 0.00$-Greedy(DDQN).

<table>
<thead>
<tr>
<th>DM Hybrid</th>
<th>DM Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Direct</strong></td>
<td><strong>DR</strong></td>
</tr>
<tr>
<td>AM</td>
<td>7.0E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.2E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.2E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>5.7E-1</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.7E-1</td>
</tr>
<tr>
<td>Q$^\pi$(\lambda)</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>1.7E-1</td>
</tr>
<tr>
<td>IH</td>
<td><strong>3.2E-2</strong></td>
</tr>
</tbody>
</table>

Table 408: MC, relative MSE. Model Type: linear. $T = 250, N = 256, \pi_b = 0.10$-Greedy(DDQN), $\pi_e = 0.00$-Greedy(DDQN).

<table>
<thead>
<tr>
<th>DM Hybrid</th>
<th>DM Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Direct</strong></td>
<td><strong>DR</strong></td>
</tr>
<tr>
<td>AM</td>
<td>2.0E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.0E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>8.8E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>1.2E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td><strong>8.9E-3</strong></td>
</tr>
<tr>
<td>Q$^\pi$(\lambda)</td>
<td>1.4E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>8.7E-2</td>
</tr>
<tr>
<td>IH</td>
<td><strong>3.3E-2</strong></td>
</tr>
</tbody>
</table>

Table 409: MC, relative MSE. Model Type: NN. $T = 250, N = 128, \pi_b = 0.10$-Greedy(DDQN), $\pi_e = 0.00$-Greedy(DDQN).

<table>
<thead>
<tr>
<th>DM Hybrid</th>
<th>DM Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Direct</strong></td>
<td><strong>DR</strong></td>
</tr>
<tr>
<td>AM</td>
<td>2.0E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.0E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>8.8E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>1.2E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td><strong>8.9E-3</strong></td>
</tr>
<tr>
<td>Q$^\pi$(\lambda)</td>
<td>1.4E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>8.7E-2</td>
</tr>
<tr>
<td>IH</td>
<td><strong>3.3E-2</strong></td>
</tr>
</tbody>
</table>

Table 410: MC, relative MSE. Model Type: NN. $T = 250, N = 256, \pi_b = 0.10$-Greedy(DDQN), $\pi_e = 0.00$-Greedy(DDQN).

<table>
<thead>
<tr>
<th>DM Hybrid</th>
<th>DM Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Direct</strong></td>
<td><strong>DR</strong></td>
</tr>
<tr>
<td>AM</td>
<td>5.4E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.1E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.9E-1</td>
</tr>
<tr>
<td>FQE</td>
<td><strong>8.6E-3</strong></td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.6E-1</td>
</tr>
<tr>
<td>Q$^\pi$(\lambda)</td>
<td>3.4E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>1.4E-2</td>
</tr>
<tr>
<td>IH</td>
<td><strong>3.1E-2</strong></td>
</tr>
</tbody>
</table>

Table 411: MC, relative MSE. Model Type: NN. $T = 250, N = 256, \pi_b = 0.10$-Greedy(DDQN), $\pi_e = 0.00$-Greedy(DDQN).

<table>
<thead>
<tr>
<th>DM Hybrid</th>
<th>DM Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Direct</strong></td>
<td><strong>DR</strong></td>
</tr>
<tr>
<td>AM</td>
<td>5.4E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.1E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.9E-1</td>
</tr>
<tr>
<td>FQE</td>
<td><strong>8.6E-3</strong></td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.6E-1</td>
</tr>
<tr>
<td>Q$^\pi$(\lambda)</td>
<td>3.4E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>1.4E-2</td>
</tr>
<tr>
<td>IH</td>
<td><strong>3.1E-2</strong></td>
</tr>
</tbody>
</table>
Table 411: MC, relative MSE. Model Type: linear. $T = 250, N = 128, \pi_b = 0.10$-Greedy(DDQN), $\pi_e = 1.00$-Greedy(DDQN).

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>5.2E-1</td>
<td>4.1E-1</td>
<td>2.4E0</td>
<td>6.5E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>9.2E-1</td>
<td>8.6E-1</td>
<td>2.0E0</td>
<td>5.5E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.9E-1</td>
<td>8.8E-1</td>
<td>4.3E-1</td>
<td>4.4E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>6.0E-1</td>
<td>5.2E-1</td>
<td>2.3E-1</td>
<td>2.7E-1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>7.3E-1</td>
<td>6.4E-1</td>
<td>4.6E-1</td>
<td>4.8E-1</td>
</tr>
<tr>
<td>$Q^*$($\lambda$)</td>
<td>6.4E-1</td>
<td>5.9E-1</td>
<td>4.6E-1</td>
<td>4.5E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>7.3E-1</td>
<td>6.4E-1</td>
<td>4.7E-1</td>
<td>4.8E-1</td>
</tr>
<tr>
<td>IH</td>
<td>4.1E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 412: MC, relative MSE. Model Type: linear. $T = 250, N = 256, \pi_b = 0.10$-Greedy(DDQN), $\pi_e = 1.00$-Greedy(DDQN).

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>9.5E-1</td>
<td>8.5E-1</td>
<td>7.1E-1</td>
<td>7.7E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>9.2E-1</td>
<td>7.5E-1</td>
<td>4.3E-1</td>
<td>4.8E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.0E0</td>
<td>8.5E-1</td>
<td>4.2E-1</td>
<td>4.5E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>$\bf{6.1E-2}$</td>
<td>4.9E-2</td>
<td>6.8E-2</td>
<td>$\bf{4.8E-2}$</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>3.5E-1</td>
<td>3.0E-1</td>
<td>2.0E-1</td>
<td>1.6E-1</td>
</tr>
<tr>
<td>$Q^*$($\lambda$)</td>
<td>1.8E0</td>
<td>1.4E0</td>
<td>3.1E0</td>
<td>3.8E0</td>
</tr>
<tr>
<td>Tree</td>
<td>2.4E-1</td>
<td>2.4E-1</td>
<td>2.1E-1</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>IH</td>
<td>4.1E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 413: MC, relative MSE. Model Type: NN. $T = 250, N = 128, \pi_b = 0.10$-Greedy(DDQN), $\pi_e = 1.00$-Greedy(DDQN).

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>9.5E-1</td>
<td>8.5E-1</td>
<td>7.1E-1</td>
<td>7.7E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>9.2E-1</td>
<td>7.5E-1</td>
<td>4.3E-1</td>
<td>4.8E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.0E0</td>
<td>8.5E-1</td>
<td>4.2E-1</td>
<td>4.5E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>$\bf{6.1E-2}$</td>
<td>4.9E-2</td>
<td>6.8E-2</td>
<td>$\bf{4.8E-2}$</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>3.5E-1</td>
<td>3.0E-1</td>
<td>2.0E-1</td>
<td>1.6E-1</td>
</tr>
<tr>
<td>$Q^*$($\lambda$)</td>
<td>1.8E0</td>
<td>1.4E0</td>
<td>3.1E0</td>
<td>3.8E0</td>
</tr>
<tr>
<td>Tree</td>
<td>2.4E-1</td>
<td>2.4E-1</td>
<td>2.1E-1</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>IH</td>
<td>4.1E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 414: MC, relative MSE. Model Type: NN. $T = 250, N = 256, \pi_b = 0.10$-Greedy(DDQN), $\pi_e = 1.00$-Greedy(DDQN).

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>9.5E-1</td>
<td>8.5E-1</td>
<td>7.1E-1</td>
<td>7.7E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>9.2E-1</td>
<td>7.5E-1</td>
<td>4.3E-1</td>
<td>4.8E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.0E0</td>
<td>8.5E-1</td>
<td>4.2E-1</td>
<td>4.5E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>$\bf{4.3E-2}$</td>
<td>3.3E-2</td>
<td>1.6E-1</td>
<td>$\bf{3.2E-2}$</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>5.7E-1</td>
<td>4.6E-1</td>
<td>2.0E-1</td>
<td>2.2E-1</td>
</tr>
<tr>
<td>$Q^*$($\lambda$)</td>
<td>2.1E1</td>
<td>2.0E1</td>
<td>1.8E1</td>
<td>2.7E1</td>
</tr>
<tr>
<td>Tree</td>
<td>4.5E-1</td>
<td>3.4E-1</td>
<td>1.3E-1</td>
<td>1.7E-1</td>
</tr>
<tr>
<td>IH</td>
<td>4.0E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 415: MC, relative MSE. Model Type: NN. $T = 250, N = 256, \pi_b = 0.10$-Greedy(DDQN), $\pi_e = 1.00$-Greedy(DDQN).

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>9.5E-1</td>
<td>8.5E-1</td>
<td>7.1E-1</td>
<td>7.7E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>9.2E-1</td>
<td>7.5E-1</td>
<td>4.3E-1</td>
<td>4.8E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.0E0</td>
<td>8.5E-1</td>
<td>4.2E-1</td>
<td>4.5E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>$\bf{4.3E-2}$</td>
<td>3.3E-2</td>
<td>1.6E-1</td>
<td>$\bf{3.2E-2}$</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>5.7E-1</td>
<td>4.6E-1</td>
<td>2.0E-1</td>
<td>2.2E-1</td>
</tr>
<tr>
<td>$Q^*$($\lambda$)</td>
<td>2.1E1</td>
<td>2.0E1</td>
<td>1.8E1</td>
<td>2.7E1</td>
</tr>
<tr>
<td>Tree</td>
<td>4.5E-1</td>
<td>3.4E-1</td>
<td>1.3E-1</td>
<td>1.7E-1</td>
</tr>
<tr>
<td>IH</td>
<td>4.0E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
**Table 415**: MC, relative MSE. Model Type: linear. $T = 250, N = 128, \pi_b = 1.00$-Greedy(DDQN), $\pi_e = 0.00$-Greedy(DDQN).

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.4E-1</td>
<td><strong>3.2E-1</strong></td>
</tr>
<tr>
<td>Q-REG</td>
<td>6.6E-1</td>
<td>5.1E2</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.4E-1</td>
<td>9.7E0</td>
</tr>
<tr>
<td>FQE</td>
<td>2.2E-1</td>
<td>1.8E0</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.6E-2</td>
<td>3.4E0</td>
</tr>
<tr>
<td>Q$^\theta$($\lambda$)</td>
<td>9.1E-2</td>
<td>3.6E0</td>
</tr>
<tr>
<td>TREE</td>
<td><strong>1.5E-2</strong></td>
<td>3.4E0</td>
</tr>
<tr>
<td>IH</td>
<td>5.2E0</td>
<td>-</td>
</tr>
</tbody>
</table>

**IPS**

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>WIS</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.2E0</td>
<td>-</td>
</tr>
</tbody>
</table>

**Table 416**: MC, relative MSE. Model Type: linear. $T = 250, N = 256, \pi_b = 1.00$-Greedy(DDQN), $\pi_e = 0.00$-Greedy(DDQN).

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.2E-1</td>
<td>1.7E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>7.0E-1</td>
<td>2.5E23</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.6E-1</td>
<td>6.7E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>2.1E-1</td>
<td>2.3E-1</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.6E-2</td>
<td><strong>8.0E-3</strong></td>
</tr>
<tr>
<td>Q$^\theta$($\lambda$)</td>
<td>8.2E-2</td>
<td>2.2E-1</td>
</tr>
<tr>
<td>TREE</td>
<td><strong>1.5E-2</strong></td>
<td>8.4E-3</td>
</tr>
<tr>
<td>IH</td>
<td>5.1E0</td>
<td>-</td>
</tr>
</tbody>
</table>

**IPS**

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E0</td>
<td><strong>6.7E-1</strong></td>
</tr>
<tr>
<td>WIS</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.1E0</td>
<td>-</td>
</tr>
</tbody>
</table>

**Table 417**: MC, relative MSE. Model Type: NN. $T = 250, N = 128, \pi_b = 1.00$-Greedy(DDQN), $\pi_e = 0.00$-Greedy(DDQN).

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>8.6E-1</td>
<td>7.3E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>8.4E-1</td>
<td>5.5E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.4E-1</td>
<td>5.0E-1</td>
</tr>
<tr>
<td>FQE</td>
<td><strong>1.2E-1</strong></td>
<td><strong>3.9E-1</strong></td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.5E0</td>
<td>1.7E0</td>
</tr>
<tr>
<td>Q$^\theta$($\lambda$)</td>
<td>5.4E0</td>
<td>5.6E0</td>
</tr>
<tr>
<td>TREE</td>
<td>2.0E0</td>
<td>2.3E0</td>
</tr>
<tr>
<td>IH</td>
<td>5.2E0</td>
<td>-</td>
</tr>
</tbody>
</table>

**IPS**

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E0</td>
<td><strong>5.4E-1</strong></td>
</tr>
<tr>
<td>WIS</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.2E0</td>
<td>-</td>
</tr>
</tbody>
</table>

**Table 418**: MC, relative MSE. Model Type: NN. $T = 250, N = 256, \pi_b = 1.00$-Greedy(DDQN), $\pi_e = 0.00$-Greedy(DDQN).

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>8.6E-1</td>
<td>4.7E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>7.7E-1</td>
<td>4.8E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.0E-1</td>
<td>4.7E-1</td>
</tr>
<tr>
<td>FQE</td>
<td><strong>2.7E-2</strong></td>
<td><strong>1.3E-1</strong></td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.0E0</td>
<td>1.2E0</td>
</tr>
<tr>
<td>Q$^\theta$($\lambda$)</td>
<td>5.6E0</td>
<td>5.6E0</td>
</tr>
<tr>
<td>TREE</td>
<td>2.1E0</td>
<td>2.5E0</td>
</tr>
<tr>
<td>IH</td>
<td>5.1E0</td>
<td>-</td>
</tr>
</tbody>
</table>

**IPS**

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E0</td>
<td><strong>5.3E-1</strong></td>
</tr>
<tr>
<td>WIS</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NAIVE</td>
<td>5.1E0</td>
<td>-</td>
</tr>
</tbody>
</table>
### Table 419: MC, relative MSE. Model Type: linear. $T = 250, N = 128, \pi_b = 1.00$-Greedy(DDQN), $\pi_e = 0.10$-Greedy(DDQN).

<table>
<thead>
<tr>
<th>DM Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>9.3E-2</td>
<td>3.2E-1</td>
<td>2.9E0</td>
<td>1.0E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>6.3E-1</td>
<td>4.6E-1</td>
<td>4.7E0</td>
<td>1.6E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>8.5E-1</td>
<td>6.0E-1</td>
<td>3.5E-1</td>
<td>3.5E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>2.7E-1</td>
<td>1.4E-1</td>
<td>7.1E-1</td>
<td>6.1E-1</td>
</tr>
<tr>
<td>R(λ)</td>
<td>6.5E-2</td>
<td>1.8E-2</td>
<td>3.5E-1</td>
<td>2.9E-1</td>
</tr>
<tr>
<td>Q(λ)</td>
<td>9.4E-3</td>
<td>1.1E-1</td>
<td>4.9E-1</td>
<td>3.9E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>6.3E-2</td>
<td>1.8E-2</td>
<td>3.7E-1</td>
<td>3.1E-1</td>
</tr>
<tr>
<td>IH</td>
<td>3.0E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 420: MC, relative MSE. Model Type: linear. $T = 250, N = 256, \pi_b = 1.00$-Greedy(DDQN), $\pi_e = 0.10$-Greedy(DDQN).

<table>
<thead>
<tr>
<th>DM Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>8.7E-1</td>
<td>6.1E-1</td>
<td>7.7E-1</td>
<td>7.5E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>8.1E-1</td>
<td>5.7E-1</td>
<td>3.7E-1</td>
<td>4.1E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.6E-1</td>
<td>6.9E-1</td>
<td>6.5E-1</td>
<td>6.6E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>1.7E-2</td>
<td>3.9E-3</td>
<td>3.9E-1</td>
<td>2.0E-1</td>
</tr>
<tr>
<td>R(λ)</td>
<td>8.1E-1</td>
<td>9.0E-1</td>
<td>1.5E0</td>
<td>1.4E0</td>
</tr>
<tr>
<td>Q(λ)</td>
<td>3.7E0</td>
<td>3.8E0</td>
<td>3.2E0</td>
<td>3.2E0</td>
</tr>
<tr>
<td>TREE</td>
<td>8.9E-1</td>
<td>9.8E-1</td>
<td>1.6E0</td>
<td>1.4E0</td>
</tr>
<tr>
<td>IH</td>
<td>3.1E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 421: MC, relative MSE. Model Type: linear. $T = 250, N = 128, \pi_b = 1.00$-Greedy(DDQN), $\pi_e = 0.10$-Greedy(DDQN).

<table>
<thead>
<tr>
<th>DM Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>8.7E-1</td>
<td>6.1E-1</td>
<td>7.7E-1</td>
<td>7.5E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>8.1E-1</td>
<td>5.7E-1</td>
<td>3.7E-1</td>
<td>4.1E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.6E-1</td>
<td>6.9E-1</td>
<td>6.5E-1</td>
<td>6.6E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>1.7E-2</td>
<td>3.9E-3</td>
<td>3.9E-1</td>
<td>2.0E-1</td>
</tr>
<tr>
<td>R(λ)</td>
<td>8.1E-1</td>
<td>9.0E-1</td>
<td>1.5E0</td>
<td>1.4E0</td>
</tr>
<tr>
<td>Q(λ)</td>
<td>3.7E0</td>
<td>3.8E0</td>
<td>3.2E0</td>
<td>3.2E0</td>
</tr>
<tr>
<td>TREE</td>
<td>8.9E-1</td>
<td>9.8E-1</td>
<td>1.6E0</td>
<td>1.4E0</td>
</tr>
<tr>
<td>IH</td>
<td>3.1E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 422: MC, relative MSE. Model Type: linear. $T = 250, N = 256, \pi_b = 1.00$-Greedy(DDQN), $\pi_e = 0.10$-Greedy(DDQN).

<table>
<thead>
<tr>
<th>DM Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>8.8E-1</td>
<td>4.0E-1</td>
<td>5.8E-1</td>
<td>4.0E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>7.6E-1</td>
<td>4.1E-1</td>
<td>1.9E-1</td>
<td>2.1E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.7E-1</td>
<td>5.3E-1</td>
<td>3.1E-1</td>
<td>3.2E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>5.0E-3</td>
<td>1.4E-1</td>
<td>1.3E-1</td>
<td>1.5E-2</td>
</tr>
<tr>
<td>R(λ)</td>
<td>1.1E0</td>
<td>1.2E0</td>
<td>2.1E0</td>
<td>1.6E0</td>
</tr>
<tr>
<td>Q(λ)</td>
<td>2.7E0</td>
<td>2.8E0</td>
<td>2.4E0</td>
<td>2.4E0</td>
</tr>
<tr>
<td>TREE</td>
<td>8.8E-1</td>
<td>1.0E0</td>
<td>1.6E0</td>
<td>1.3E0</td>
</tr>
<tr>
<td>IH</td>
<td>3.1E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 423: MC, relative MSE. Model Type: linear. $T = 250, N = 256, \pi_b = 1.00$-Greedy(DDQN), $\pi_e = 0.10$-Greedy(DDQN).

<table>
<thead>
<tr>
<th>DM Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.0E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Q-REG</td>
<td>0.0E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MRDR</td>
<td>0.0E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FQE</td>
<td>0.0E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>R(λ)</td>
<td>0.0E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Q(λ)</td>
<td>0.0E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TREE</td>
<td>0.0E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IH</td>
<td>0.0E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
### 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).

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.6E5</td>
<td>2.1E5</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>6.8E-3</td>
<td>8.8E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.7E-3</td>
<td>3.0E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>3.2E-3</td>
<td>1.1E-3</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Q($\pi$($\lambda$))</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tree</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IH</td>
<td>4.4E-3</td>
<td>-</td>
</tr>
</tbody>
</table>

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).

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.4E5</td>
<td>8.2E4</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.0E-1</td>
<td>3.6E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.3E-1</td>
<td>9.3E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>2.6E-3</td>
<td>7.1E-4</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Q($\pi$($\lambda$))</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tree</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IH</td>
<td>1.3E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

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).

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>5.1E2</td>
<td>1.2E3</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>3.8E-3</td>
<td>3.8E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>3.6E-2</td>
<td>4.5E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>1.5E-3</td>
<td>8.0E-4</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Q($\pi$($\lambda$))</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tree</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IH</td>
<td>3.8E-3</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>IS</td>
<td>WIS</td>
</tr>
<tr>
<td>Per-Decision</td>
<td>2.3E-1</td>
<td>3.4E-4</td>
</tr>
<tr>
<td>IPS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard</td>
<td>IS</td>
<td>WIS</td>
</tr>
<tr>
<td>Per-Decision</td>
<td>2.3E-1</td>
<td>3.4E-4</td>
</tr>
</tbody>
</table>
### 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.).

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>4.2E-2</td>
<td>4.3E-2</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.0E-1</td>
<td>4.7E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.6E-1</td>
<td>4.5E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>3.7E-2</td>
<td>3.6E-2</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.4E0</td>
<td>7.1E-2</td>
</tr>
<tr>
<td>Q(\lambda)</td>
<td>2.3E0</td>
<td>7.5E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>1.1E0</td>
<td>6.6E-2</td>
</tr>
<tr>
<td>IH</td>
<td><strong>2.1E-2</strong></td>
<td>-</td>
</tr>
</tbody>
</table>

#### IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>4.7E-2</td>
<td>6.4E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>1.6E-2</td>
<td><strong>6.6E-3</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>9.6E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 427: Gridworld, relative MSE. \( T = 25, N = 128, \pi_b = 0.20\)-Greedy(V iter.), \( \pi_e = 0.10\)-Greedy(V iter.).

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td><strong>3.1E-3</strong></td>
<td>1.9E-3</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>3.9E-2</td>
<td>1.5E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>7.8E-2</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>1.2E-2</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.4E0</td>
<td>1.7E-2</td>
</tr>
<tr>
<td>Q(\lambda)</td>
<td>1.6E0</td>
<td>1.5E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>9.1E-1</td>
<td>1.7E-2</td>
</tr>
<tr>
<td>IH</td>
<td>1.1E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

#### IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.3E-2</td>
<td>1.9E-2</td>
</tr>
<tr>
<td>WIS</td>
<td>4.1E-3</td>
<td><strong>1.1E-3</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>9.6E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 428: Gridworld, relative MSE. \( T = 25, N = 256, \pi_b = 0.20\)-Greedy(V iter.), \( \pi_e = 0.10\)-Greedy(V iter.).

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td><strong>1.0E-3</strong></td>
<td>1.5E-3</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>6.6E-3</td>
<td>3.4E-3</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.9E-2</td>
<td>3.0E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>3.5E-3</td>
<td>2.3E-3</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.7E-1</td>
<td>5.7E-3</td>
</tr>
<tr>
<td>Q(\lambda)</td>
<td>2.3E-1</td>
<td>4.0E-3</td>
</tr>
<tr>
<td>Tree</td>
<td>4.3E-1</td>
<td>4.8E-3</td>
</tr>
<tr>
<td>IH</td>
<td>4.8E-3</td>
<td>-</td>
</tr>
</tbody>
</table>

#### IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>4.8E-3</td>
<td>6.3E-3</td>
</tr>
<tr>
<td>WIS</td>
<td>1.0E-3</td>
<td><strong>5.1E-4</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>1.1E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 429: Gridworld, relative MSE. \( T = 25, N = 512, \pi_b = 0.20\)-Greedy(V iter.), \( \pi_e = 0.10\)-Greedy(V iter.).

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>5.9E-4</td>
<td>1.2E-3</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>5.0E-3</td>
<td>4.1E-4</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.8E-2</td>
<td>4.1E-4</td>
</tr>
<tr>
<td>FQE</td>
<td>5.6E-4</td>
<td>2.4E-4</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>2.8E-3</td>
<td>4.4E-4</td>
</tr>
<tr>
<td>Q(\lambda)</td>
<td><strong>2.4E-4</strong></td>
<td>2.4E-4</td>
</tr>
<tr>
<td>Tree</td>
<td>3.4E-1</td>
<td>4.2E-4</td>
</tr>
<tr>
<td>IH</td>
<td>4.1E-3</td>
<td>-</td>
</tr>
</tbody>
</table>

#### IPS

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>9.9E-4</td>
<td>1.3E-3</td>
</tr>
<tr>
<td>WIS</td>
<td>4.0E-4</td>
<td><strong>3.7E-4</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>9.0E-2</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 430: Gridworld, relative MSE. $T = 25, N = 1024, \pi_b = 0.20$-Greedy(V iter.), $\pi_e = 0.10$-Greedy(V iter.).

<table>
<thead>
<tr>
<th>Method</th>
<th>AM</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM Hybrid</td>
<td>3.6E-4</td>
<td>1.7E-4</td>
<td>1.7E-4</td>
<td>1.1E-4</td>
</tr>
<tr>
<td>Direct</td>
<td>3.6E-4</td>
<td>1.7E-4</td>
<td>1.7E-4</td>
<td>1.1E-4</td>
</tr>
<tr>
<td>DR</td>
<td>2.0E-3</td>
<td>3.2E-4</td>
<td>3.2E-4</td>
<td>5.4E-4</td>
</tr>
<tr>
<td>WDR</td>
<td>2.3E-2</td>
<td>4.0E-4</td>
<td>3.9E-4</td>
<td>7.9E-4</td>
</tr>
<tr>
<td>MAGIC</td>
<td>4.2E-4</td>
<td>7.3E-4</td>
<td>6.3E-4</td>
<td>9.8E-4</td>
</tr>
<tr>
<td>FQE</td>
<td>8.2E-4</td>
<td>3.3E-4</td>
<td>3.3E-4</td>
<td>4.7E-4</td>
</tr>
<tr>
<td>R(λ)</td>
<td>3.0E-3</td>
<td>4.6E-4</td>
<td>4.6E-4</td>
<td>1.5E-3</td>
</tr>
<tr>
<td>Q(λ)</td>
<td>3.6E-4</td>
<td>3.4E-4</td>
<td>3.4E-4</td>
<td>3.6E-4</td>
</tr>
<tr>
<td>Tree</td>
<td>3.4E-1</td>
<td>3.9E-4</td>
<td>3.6E-4</td>
<td>8.2E-4</td>
</tr>
<tr>
<td>IPS</td>
<td>1.8E-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 432: Gridworld, relative MSE. $T = 25, N = 128, \pi_b = 0.40$-Greedy(V iter.), $\pi_e = 0.10$-Greedy(V iter.).

<table>
<thead>
<tr>
<th>Method</th>
<th>AM</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM Hybrid</td>
<td>3.4E-2</td>
<td>1.2E-2</td>
<td>1.2E-2</td>
<td>1.2E-2</td>
</tr>
<tr>
<td>Direct</td>
<td>3.4E-2</td>
<td>1.2E-2</td>
<td>1.2E-2</td>
<td>1.2E-2</td>
</tr>
<tr>
<td>DR</td>
<td>2.0E-1</td>
<td>3.4E-2</td>
<td>2.7E-2</td>
<td>3.2E-2</td>
</tr>
<tr>
<td>WDR</td>
<td>2.1E-1</td>
<td>5.8E-2</td>
<td>2.0E-2</td>
<td>2.1E-2</td>
</tr>
<tr>
<td>MAGIC</td>
<td>2.8E-2</td>
<td>2.1E-3</td>
<td>9.8E-4</td>
<td>1.5E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>5.1E-1</td>
<td>6.1E-2</td>
<td>8.3E-3</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>R(λ)</td>
<td>1.4E-3</td>
<td>1.3E-3</td>
<td>1.1E-3</td>
<td>1.5E-3</td>
</tr>
<tr>
<td>Q(λ)</td>
<td>3.6E-2</td>
<td>3.4E-2</td>
<td>3.4E-2</td>
<td>3.6E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>8.9E-1</td>
<td>1.5E-1</td>
<td>1.3E-2</td>
<td>3.8E-2</td>
</tr>
<tr>
<td>IPS</td>
<td>3.5E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 431: Gridworld, relative MSE. $T = 25, N = 64, \pi_b = 0.40$-Greedy(V iter.), $\pi_e = 0.10$-Greedy(V iter.).

<table>
<thead>
<tr>
<th>Method</th>
<th>AM</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM Hybrid</td>
<td>2.0E-1</td>
<td>1.4E-1</td>
<td>1.2E-1</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>Direct</td>
<td>2.0E-1</td>
<td>1.4E-1</td>
<td>1.2E-1</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>DR</td>
<td>2.4E-1</td>
<td>1.1E-1</td>
<td>1.5E-1</td>
<td>3.8E-2</td>
</tr>
<tr>
<td>WDR</td>
<td>2.9E-1</td>
<td>8.6E-2</td>
<td>6.5E-2</td>
<td>3.1E-2</td>
</tr>
<tr>
<td>MAGIC</td>
<td>7.8E-3</td>
<td>2.5E-3</td>
<td>6.6E-3</td>
<td>9.8E-4</td>
</tr>
<tr>
<td>FQE</td>
<td>7.6E-3</td>
<td>2.8E-2</td>
<td>5.0E-2</td>
<td>6.6E-3</td>
</tr>
<tr>
<td>R(λ)</td>
<td>5.6E-2</td>
<td>5.2E-3</td>
<td>5.9E-3</td>
<td>9.8E-4</td>
</tr>
<tr>
<td>Q(λ)</td>
<td>1.0E0</td>
<td>1.7E-1</td>
<td>2.9E-2</td>
<td>1.6E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>4.8E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IPS</td>
<td>7.4E-4</td>
<td>8.4E-4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 433: Gridworld, relative MSE. $T = 25, N = 256, \pi_b = 0.40$-Greedy(V iter.), $\pi_e = 0.10$-Greedy(V iter.).

<table>
<thead>
<tr>
<th>Method</th>
<th>AM</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM Hybrid</td>
<td>1.8E-2</td>
<td>2.1E-2</td>
<td>2.0E-2</td>
<td>1.8E-2</td>
</tr>
<tr>
<td>Direct</td>
<td>1.8E-2</td>
<td>2.1E-2</td>
<td>2.0E-2</td>
<td>1.8E-2</td>
</tr>
<tr>
<td>DR</td>
<td>7.9E-2</td>
<td>3.5E-3</td>
<td>2.9E-3</td>
<td>4.6E-3</td>
</tr>
<tr>
<td>WDR</td>
<td>8.1E-2</td>
<td>1.1E-2</td>
<td>6.7E-3</td>
<td>4.1E-3</td>
</tr>
<tr>
<td>MAGIC</td>
<td>2.1E-2</td>
<td>4.8E-4</td>
<td>2.7E-4</td>
<td>4.9E-4</td>
</tr>
<tr>
<td>FQE</td>
<td>4.7E-1</td>
<td>1.5E-2</td>
<td>3.4E-3</td>
<td>4.4E-3</td>
</tr>
<tr>
<td>R(λ)</td>
<td>4.2E-4</td>
<td>3.1E-4</td>
<td>2.9E-4</td>
<td>4.9E-4</td>
</tr>
<tr>
<td>Q(λ)</td>
<td>8.7E-1</td>
<td>4.2E-2</td>
<td>4.9E-3</td>
<td>8.9E-3</td>
</tr>
<tr>
<td>Tree</td>
<td>2.8E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IPS</td>
<td>6.6E-2</td>
<td>8.5E-2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 432: Gridworld, relative MSE. $T = 25, N = 128, \pi_b = 0.40$-Greedy(V iter.), $\pi_e = 0.10$-Greedy(V iter.).

<table>
<thead>
<tr>
<th>Method</th>
<th>AM</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM Hybrid</td>
<td>3.4E-4</td>
<td>1.7E-4</td>
<td>1.7E-4</td>
<td>1.1E-4</td>
</tr>
<tr>
<td>Direct</td>
<td>3.4E-4</td>
<td>1.7E-4</td>
<td>1.7E-4</td>
<td>1.1E-4</td>
</tr>
<tr>
<td>DR</td>
<td>2.0E-3</td>
<td>3.2E-4</td>
<td>3.2E-4</td>
<td>5.4E-4</td>
</tr>
<tr>
<td>WDR</td>
<td>2.3E-2</td>
<td>4.0E-4</td>
<td>3.9E-4</td>
<td>7.9E-4</td>
</tr>
<tr>
<td>MAGIC</td>
<td>4.2E-4</td>
<td>7.3E-4</td>
<td>6.3E-4</td>
<td>9.8E-4</td>
</tr>
<tr>
<td>FQE</td>
<td>8.2E-4</td>
<td>3.3E-4</td>
<td>3.3E-4</td>
<td>4.7E-4</td>
</tr>
<tr>
<td>R(λ)</td>
<td>3.0E-3</td>
<td>4.6E-4</td>
<td>4.6E-4</td>
<td>1.5E-3</td>
</tr>
<tr>
<td>Q(λ)</td>
<td>3.6E-4</td>
<td>3.4E-4</td>
<td>3.4E-4</td>
<td>3.6E-4</td>
</tr>
<tr>
<td>Tree</td>
<td>3.4E-1</td>
<td>3.9E-4</td>
<td>3.6E-4</td>
<td>8.2E-4</td>
</tr>
<tr>
<td>IPS</td>
<td>1.8E-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 433: Gridworld, relative MSE. $T = 25, N = 256, \pi_b = 0.40$-Greedy(V iter.), $\pi_e = 0.10$-Greedy(V iter.).

<table>
<thead>
<tr>
<th>Method</th>
<th>AM</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM Hybrid</td>
<td>3.4E-4</td>
<td>1.7E-4</td>
<td>1.7E-4</td>
<td>1.1E-4</td>
</tr>
<tr>
<td>Direct</td>
<td>3.4E-4</td>
<td>1.7E-4</td>
<td>1.7E-4</td>
<td>1.1E-4</td>
</tr>
<tr>
<td>DR</td>
<td>2.0E-3</td>
<td>3.2E-4</td>
<td>3.2E-4</td>
<td>5.4E-4</td>
</tr>
<tr>
<td>WDR</td>
<td>2.3E-2</td>
<td>4.0E-4</td>
<td>3.9E-4</td>
<td>7.9E-4</td>
</tr>
<tr>
<td>MAGIC</td>
<td>4.2E-4</td>
<td>7.3E-4</td>
<td>6.3E-4</td>
<td>9.8E-4</td>
</tr>
<tr>
<td>FQE</td>
<td>8.2E-4</td>
<td>3.3E-4</td>
<td>3.3E-4</td>
<td>4.7E-4</td>
</tr>
<tr>
<td>R(λ)</td>
<td>3.0E-3</td>
<td>4.6E-4</td>
<td>4.6E-4</td>
<td>1.5E-3</td>
</tr>
<tr>
<td>Q(λ)</td>
<td>3.6E-4</td>
<td>3.4E-4</td>
<td>3.4E-4</td>
<td>3.6E-4</td>
</tr>
<tr>
<td>Tree</td>
<td>3.4E-1</td>
<td>3.9E-4</td>
<td>3.6E-4</td>
<td>8.2E-4</td>
</tr>
<tr>
<td>IPS</td>
<td>1.8E-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
### Table 434: Gridworld, relative MSE. $T = 25, N = 512, \pi_b = 0.40$-Greedy(V iter.), $\pi_e = 0.10$-Greedy(V iter.).

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.1E-2</td>
<td>7.2E-3</td>
<td>6.9E-3</td>
<td>5.6E-3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.0E-2</td>
<td>6.2E-4</td>
<td>6.2E-4</td>
<td>9.3E-4</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.3E-2</td>
<td>1.0E-3</td>
<td>3.5E-4</td>
<td>6.8E-4</td>
</tr>
<tr>
<td>FQE</td>
<td>1.8E-2</td>
<td>1.1E-4</td>
<td>4.6E-5</td>
<td>1.4E-4</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>4.6E-1</td>
<td>6.7E-3</td>
<td>1.8E-3</td>
<td>3.1E-3</td>
</tr>
<tr>
<td>Q$^\pi$(\lambda)</td>
<td><strong>5.7E-5</strong></td>
<td>3.6E-5</td>
<td><strong>2.9E-5</strong></td>
<td>9.1E-5</td>
</tr>
<tr>
<td>TREE</td>
<td>8.8E-1</td>
<td>1.3E-2</td>
<td>3.5E-3</td>
<td>6.1E-3</td>
</tr>
<tr>
<td>IH</td>
<td>2.7E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 435: Gridworld, relative MSE. $T = 25, N = 1024, \pi_b = 0.40$-Greedy(V iter.), $\pi_e = 0.10$-Greedy(V iter.).

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.9E-1</td>
<td>2.9E-1</td>
<td>2.5E-1</td>
<td>2.6E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>5.6E-1</td>
<td>1.8E0</td>
<td>1.0E0</td>
<td>3.1E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.0E-1</td>
<td>3.7E0</td>
<td>1.7E0</td>
<td>4.6E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>1.1E-1</td>
<td>2.0E-2</td>
<td>1.3E-2</td>
<td>2.9E-3</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.2E0</td>
<td>7.4E-1</td>
<td>9.3E-2</td>
<td>6.9E-1</td>
</tr>
<tr>
<td>Q$^\pi$(\lambda)</td>
<td><strong>5.0E-1</strong></td>
<td>3.4E-3</td>
<td><strong>2.7E-3</strong></td>
<td>2.9E-3</td>
</tr>
<tr>
<td>TREE</td>
<td>9.0E-1</td>
<td>6.1E-3</td>
<td>5.6E-4</td>
<td>2.0E-3</td>
</tr>
<tr>
<td>IH</td>
<td>8.7E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 436: Gridworld, relative MSE. $T = 25, N = 64, \pi_b = 0.60$-Greedy(V iter.), $\pi_e = 0.10$-Greedy(V iter.).

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>6.0E-1</td>
<td>3.6E-1</td>
<td>3.2E-1</td>
<td>5.3E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.7E0</td>
<td>2.7E0</td>
<td>2.4E0</td>
<td>1.3E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.3E0</td>
<td>8.1E0</td>
<td>2.7E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>FQE</td>
<td>1.2E-1</td>
<td>1.1E-1</td>
<td>1.9E-2</td>
<td>1.4E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.2E0</td>
<td>1.2E0</td>
<td>2.4E-1</td>
<td>1.1E0</td>
</tr>
<tr>
<td>Q$^\pi$(\lambda)</td>
<td><strong>1.8E-2</strong></td>
<td>2.3E-2</td>
<td><strong>1.1E-2</strong></td>
<td>1.5E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>1.2E0</td>
<td>1.5E0</td>
<td>3.4E-1</td>
<td>1.3E0</td>
</tr>
<tr>
<td>IH</td>
<td>1.0E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 437: Gridworld, relative MSE. $T = 25, N = 128, \pi_b = 0.60$-Greedy(V iter.), $\pi_e = 0.10$-Greedy(V iter.).

<table>
<thead>
<tr>
<th></th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.4E-2</td>
<td>3.3E-3</td>
<td>3.3E-3</td>
<td>2.7E-3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>6.6E-3</td>
<td>1.5E-4</td>
<td>1.5E-4</td>
<td>2.3E-4</td>
</tr>
<tr>
<td>MRDR</td>
<td>6.0E-2</td>
<td>6.6E-4</td>
<td>2.7E-4</td>
<td>4.9E-4</td>
</tr>
<tr>
<td>FQE</td>
<td>1.8E-2</td>
<td>5.9E-5</td>
<td>5.4E-5</td>
<td>9.0E-5</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>4.8E-1</td>
<td>2.6E-3</td>
<td>2.2E-4</td>
<td>5.8E-4</td>
</tr>
<tr>
<td>Q$^\pi$(\lambda)</td>
<td><strong>2.7E-5</strong></td>
<td>3.3E-5</td>
<td><strong>3.5E-5</strong></td>
<td>5.7E-5</td>
</tr>
<tr>
<td>TREE</td>
<td>9.0E-1</td>
<td>6.1E-3</td>
<td>5.6E-4</td>
<td>2.0E-3</td>
</tr>
<tr>
<td>IH</td>
<td>2.8E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 438: Gridworld, relative MSE. $T = 25, N = 128, \pi_b = 0.60$-Greedy(V iter.), $\pi_e = 0.10$-Greedy(V iter.).

<table>
<thead>
<tr>
<th></th>
<th>IPS Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.1E-2</td>
<td>1.7E-2</td>
</tr>
<tr>
<td>WIS</td>
<td><strong>1.1E-3</strong></td>
<td>4.2E-3</td>
</tr>
<tr>
<td>NAIVE</td>
<td>1.2E0</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 439: Gridworld, relative MSE. $T = 25, N = 128, \pi_b = 0.60$-Greedy(V iter.), $\pi_e = 0.10$-Greedy(V iter.).

<table>
<thead>
<tr>
<th></th>
<th>IPS Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.1E0</td>
<td>1.6E0</td>
</tr>
<tr>
<td>WIS</td>
<td><strong>2.9E-1</strong></td>
<td>4.5E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.9E0</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 440: Gridworld, relative MSE. $T = 25, N = 128, \pi_b = 0.60$-Greedy(V iter.), $\pi_e = 0.10$-Greedy(V iter.).

<table>
<thead>
<tr>
<th></th>
<th>IPS Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>5.5E-1</td>
<td>1.4E0</td>
</tr>
<tr>
<td>WIS</td>
<td>2.9E-1</td>
<td><strong>1.6E-1</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>
### Table 438: Gridworld, relative MSE. $T = 25, N = 256, \pi_b = 0.60$-Greedy(V iter.), $\pi_e = 0.10$-Greedy(V iter.).

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>2.5E-1</td>
<td>2.1E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.0E-1</td>
<td>1.6E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.8E-1</td>
<td>3.9E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>1.2E-1</td>
<td>5.1E-3</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.2E0</td>
<td>4.0E-1</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>1.4E-3</td>
<td>1.1E-3</td>
</tr>
<tr>
<td>Tree</td>
<td>1.2E0</td>
<td>4.7E-1</td>
</tr>
<tr>
<td>IH</td>
<td>8.0E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 439: Gridworld, relative MSE. $T = 25, N = 512, \pi_b = 0.60$-Greedy(V iter.), $\pi_e = 0.10$-Greedy(V iter.).

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.5E-1</td>
<td>6.4E-2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>6.2E-2</td>
<td>1.5E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.7E-1</td>
<td>7.2E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>1.3E-1</td>
<td>2.8E-3</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.3E0</td>
<td>2.4E-1</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>3.0E-4</td>
<td>3.0E-4</td>
</tr>
<tr>
<td>Tree</td>
<td>1.2E0</td>
<td>2.8E-1</td>
</tr>
<tr>
<td>IH</td>
<td>6.7E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 440: Gridworld, relative MSE. $T = 25, N = 1024, \pi_b = 0.60$-Greedy(V iter.), $\pi_e = 0.10$-Greedy(V iter.).

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.1E0</td>
<td>1.4E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.1E0</td>
<td>3.2E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.1E0</td>
<td>8.4E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>3.6E-1</td>
<td>2.0E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.2E0</td>
<td>1.2E0</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>2.1E-1</td>
<td>2.2E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>1.2E0</td>
<td>1.3E0</td>
</tr>
<tr>
<td>IH</td>
<td>2.1E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 441: Gridworld, relative MSE. $T = 25, N = 64, \pi_b = 0.80$-Greedy(V iter.), $\pi_e = 0.10$-Greedy(V iter.).

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.9E-1</td>
<td>3.0E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>8.5E-3</td>
<td>4.2E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.1E0</td>
<td>1.8E0</td>
</tr>
<tr>
<td>FQE</td>
<td>1.8E0</td>
<td>1.6E0</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.2E0</td>
<td>1.3E0</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>2.1E-1</td>
<td>2.2E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>1.2E0</td>
<td>1.3E0</td>
</tr>
<tr>
<td>IH</td>
<td>2.1E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 442: Gridworld, relative MSE. $T = 25, N = 262, \pi_b = 0.60$-Greedy(V iter.), $\pi_e = 0.10$-Greedy(V iter.).

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>2.5E-1</td>
<td>2.1E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.0E-1</td>
<td>1.6E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.8E-1</td>
<td>3.9E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>1.2E-1</td>
<td>5.1E-3</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.2E0</td>
<td>4.0E-1</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>1.4E-3</td>
<td>1.1E-3</td>
</tr>
<tr>
<td>Tree</td>
<td>1.2E0</td>
<td>4.7E-1</td>
</tr>
<tr>
<td>IH</td>
<td>8.0E-2</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 442: Gridworld, relative MSE. $T = 25, N = 128, \pi_b = 0.80\text{-Greedy(V iter.)}, \pi_e = 0.10\text{-Greedy(V iter.)}$.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.0E0</td>
<td>2.3E0</td>
<td>1.2E0</td>
<td>7.2E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.3E0</td>
<td>3.9E1</td>
<td>2.5E1</td>
<td>4.0E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.2E0</td>
<td>7.2E1</td>
<td>1.6E1</td>
<td>4.3E0</td>
</tr>
<tr>
<td>FQE</td>
<td>3.1E-1</td>
<td>9.5E-2</td>
<td>8.6E-2</td>
<td>5.9E-2</td>
</tr>
<tr>
<td>R(λ)</td>
<td>1.3E0</td>
<td>7.8E0</td>
<td>1.4E0</td>
<td>2.4E0</td>
</tr>
<tr>
<td>Q^6(λ)</td>
<td>8.0E-2</td>
<td>6.9E-2</td>
<td>7.3E-2</td>
<td>7.5E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>1.2E0</td>
<td>1.1E1</td>
<td>1.6E0</td>
<td>2.7E0</td>
</tr>
<tr>
<td>IH</td>
<td>7.9E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 443: Gridworld, relative MSE. $T = 25, N = 256, \pi_b = 0.80\text{-Greedy(V iter.)}, \pi_e = 0.10\text{-Greedy(V iter.)}$.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>7.0E-1</td>
<td>3.5E0</td>
<td>5.8E-1</td>
<td>3.7E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>9.8E0</td>
<td>1.3E1</td>
<td>2.7E0</td>
<td>3.9E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.3E0</td>
<td>1.4E2</td>
<td>1.5E1</td>
<td>2.2E1</td>
</tr>
<tr>
<td>FQE</td>
<td>2.9E-1</td>
<td>5.0E-1</td>
<td>1.5E-2</td>
<td>1.5E-2</td>
</tr>
<tr>
<td>R(λ)</td>
<td>1.3E0</td>
<td>3.8E1</td>
<td>3.6E-1</td>
<td>1.4E0</td>
</tr>
<tr>
<td>Q^6(λ)</td>
<td>1.9E-2</td>
<td>1.7E-1</td>
<td>8.8E-3</td>
<td>1.6E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>1.2E0</td>
<td>4.2E1</td>
<td>3.8E-1</td>
<td>1.5E0</td>
</tr>
<tr>
<td>IH</td>
<td>9.5E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 444: Gridworld, relative MSE. $T = 25, N = 512, \pi_b = 0.80\text{-Greedy(V iter.)}, \pi_e = 0.10\text{-Greedy(V iter.)}$.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>7.0E-1</td>
<td>3.5E0</td>
<td>5.8E-1</td>
<td>3.7E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>9.8E0</td>
<td>1.3E1</td>
<td>2.7E0</td>
<td>3.9E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.3E0</td>
<td>1.4E2</td>
<td>1.5E1</td>
<td>2.2E1</td>
</tr>
<tr>
<td>FQE</td>
<td>2.9E-1</td>
<td>5.0E-1</td>
<td>1.5E-2</td>
<td>1.5E-2</td>
</tr>
<tr>
<td>R(λ)</td>
<td>1.3E0</td>
<td>3.8E1</td>
<td>3.6E-1</td>
<td>1.4E0</td>
</tr>
<tr>
<td>Q^6(λ)</td>
<td>1.9E-2</td>
<td>1.7E-1</td>
<td>8.8E-3</td>
<td>1.6E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>1.2E0</td>
<td>4.2E1</td>
<td>3.8E-1</td>
<td>1.5E0</td>
</tr>
<tr>
<td>IH</td>
<td>9.5E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 445: Gridworld, relative MSE. $T = 25, N = 1024, \pi_b = 0.80\text{-Greedy(V iter.)}, \pi_e = 0.10\text{-Greedy(V iter.)}$.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>5.7E-1</td>
<td>4.1E-1</td>
<td>1.3E-1</td>
<td>1.9E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.1E0</td>
<td>1.2E0</td>
<td>1.0E0</td>
<td>4.6E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.0E0</td>
<td>1.3E0</td>
<td>1.5E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>FQE</td>
<td>2.8E-1</td>
<td>4.7E-2</td>
<td>1.7E-2</td>
<td>9.6E-3</td>
</tr>
<tr>
<td>R(λ)</td>
<td>1.3E0</td>
<td>1.1E0</td>
<td>7.8E-1</td>
<td>1.5E0</td>
</tr>
<tr>
<td>Q^6(λ)</td>
<td>8.1E-3</td>
<td>8.7E-3</td>
<td>3.7E-3</td>
<td>3.2E-3</td>
</tr>
<tr>
<td>TREE</td>
<td>1.2E0</td>
<td>1.2E0</td>
<td>8.3E-1</td>
<td>1.5E0</td>
</tr>
<tr>
<td>IH</td>
<td>1.1E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 446: Gridworld, relative MSE. $T = 25, N = 1024, \pi_b = 0.80\text{-Greedy(V iter.)}, \pi_e = 0.10\text{-Greedy(V iter.)}$.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>7.0E-1</td>
<td>1.2E0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.1E-1</td>
<td>8.5E-1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.0E0</td>
<td>1.2E0</td>
<td>8.3E-1</td>
<td>1.5E0</td>
</tr>
<tr>
<td>FQE</td>
<td>2.8E-1</td>
<td>4.7E-2</td>
<td>1.7E-2</td>
<td>9.6E-3</td>
</tr>
<tr>
<td>R(λ)</td>
<td>1.3E0</td>
<td>1.1E0</td>
<td>7.8E-1</td>
<td>1.5E0</td>
</tr>
<tr>
<td>Q^6(λ)</td>
<td>8.1E-3</td>
<td>8.7E-3</td>
<td>3.7E-3</td>
<td>3.2E-3</td>
</tr>
<tr>
<td>TREE</td>
<td>1.2E0</td>
<td>1.2E0</td>
<td>8.3E-1</td>
<td>1.5E0</td>
</tr>
<tr>
<td>IH</td>
<td>1.1E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 446: Gridworld, relative MSE. $T = 25, N = 64, \pi_b = 1.00$-Greedy(V iter.), $\pi_e = 0.10$-Greedy(V iter.).

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.1E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.5E0</td>
<td>2.4E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.2E0</td>
<td>2.3E0</td>
</tr>
<tr>
<td>FQE</td>
<td>1.2E0</td>
<td>1.2E0</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.1E0</td>
<td>1.2E0</td>
</tr>
<tr>
<td>Q$^\dagger$(\lambda)</td>
<td>9.9E0</td>
<td>1.3E1</td>
</tr>
<tr>
<td>Tree</td>
<td>1.1E0</td>
<td>1.2E0</td>
</tr>
<tr>
<td>IH</td>
<td>1.3E0</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 458: Gridworld, relative MSE. $T = 25, N = 512, \pi_b = 1.00$-Greedy(V iter.), $\pi_e = 0.10$-Greedy(V iter.).

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.2E0</td>
<td>4.3E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.8E0</td>
<td>5.7E2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.6E0</td>
<td>1.5E3</td>
</tr>
<tr>
<td>FQE</td>
<td>3.9E-1</td>
<td>8.3E0</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.2E0</td>
<td>2.2E0</td>
</tr>
<tr>
<td>Q$^\dagger$(\lambda)</td>
<td>9.6E-2</td>
<td>4.3E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>1.2E0</td>
<td>4.2E0</td>
</tr>
<tr>
<td>IH</td>
<td>2.8E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 448: Gridworld, relative MSE. $T = 25, N = 128, \pi_b = 1.00$-Greedy(V iter.), $\pi_e = 0.10$-Greedy(V iter.).

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.2E0</td>
<td>1.2E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.7E0</td>
<td>3.2E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.7E0</td>
<td>2.8E0</td>
</tr>
<tr>
<td>FQE</td>
<td>4.4E-1</td>
<td>3.4E-1</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.2E0</td>
<td>1.5E0</td>
</tr>
<tr>
<td>Q$^\dagger$(\lambda)</td>
<td>3.7E-1</td>
<td>2.9E0</td>
</tr>
<tr>
<td>Tree</td>
<td>1.1E0</td>
<td>1.6E0</td>
</tr>
<tr>
<td>IH</td>
<td>7.2E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 447: Gridworld, relative MSE. $T = 25, N = 256, \pi_b = 1.00$-Greedy(V iter.), $\pi_e = 0.10$-Greedy(V iter.).

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.2E0</td>
<td>1.2E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.7E0</td>
<td>3.2E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.7E0</td>
<td>2.8E0</td>
</tr>
<tr>
<td>FQE</td>
<td>4.4E-1</td>
<td>3.4E-1</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.2E0</td>
<td>1.5E0</td>
</tr>
<tr>
<td>Q$^\dagger$(\lambda)</td>
<td>3.7E-1</td>
<td>2.9E0</td>
</tr>
<tr>
<td>Tree</td>
<td>1.1E0</td>
<td>1.6E0</td>
</tr>
<tr>
<td>IH</td>
<td>7.2E-1</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 449: Gridworld, relative MSE. $T = 25, N = 512, \pi_b = 1.00$-Greedy(V iter.), $\pi_e = 0.10$-Greedy(V iter.).

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.2E0</td>
<td>4.3E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.8E0</td>
<td>5.7E2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.6E0</td>
<td>1.5E3</td>
</tr>
<tr>
<td>FQE</td>
<td>3.9E-1</td>
<td>8.3E0</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.2E0</td>
<td>2.2E0</td>
</tr>
<tr>
<td>Q$^\dagger$(\lambda)</td>
<td>9.6E-2</td>
<td>4.3E-1</td>
</tr>
<tr>
<td>Tree</td>
<td>1.2E0</td>
<td>4.2E0</td>
</tr>
<tr>
<td>IH</td>
<td>2.8E-1</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 450: Gridworld, relative MSE. $T = 25, N = 1024, \pi_b = 1.00\text{-Greedy(V iter.)}, \pi_e = 0.10\text{-Greedy(V iter.)}.$

<table>
<thead>
<tr>
<th>Method</th>
<th>DM Direct</th>
<th>DM DR</th>
<th>DM WDR</th>
<th>DM MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.1E0</td>
<td>1.1E0</td>
<td>2.1E0</td>
<td>9.2E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.5E0</td>
<td>8.0E1</td>
<td>8.4E1</td>
<td>2.2E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.0E0</td>
<td>8.5E0</td>
<td>3.1E1</td>
<td>1.9E1</td>
</tr>
<tr>
<td>FQE</td>
<td>3.5E-1</td>
<td>1.6E-1</td>
<td>5.1E-2</td>
<td>7.5E-2</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.2E0</td>
<td>6.2E0</td>
<td>1.6E0</td>
<td>2.7E0</td>
</tr>
<tr>
<td>$Q^\pi(\lambda)$</td>
<td><strong>3.6E-2</strong></td>
<td>3.0E-2</td>
<td><strong>1.9E-2</strong></td>
<td>3.6E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>1.1E0</td>
<td>9.4E0</td>
<td>1.6E0</td>
<td>2.7E0</td>
</tr>
<tr>
<td>IH</td>
<td>2.4E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IPS Standard</th>
<th>IPS Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>8.9E0</td>
</tr>
<tr>
<td>WIS</td>
<td>1.7E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>1.1E1</td>
</tr>
</tbody>
</table>
F.7 Detailed Results for Pixel Gridworld

Table 451: Pixel-Gridworld, relative MSE. $T = 25, N = 64, \pi_b = 0.20\text{-Greedy(V iter.)}, \pi_e = 0.10\text{-Greedy(V iter.)}$. Note: we use the same policy as in Gridworld. $\pi_b$ known. Stochastic environment.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.3E0</td>
<td>2.3E1</td>
<td>2.3E1</td>
<td>7.3E-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.1E-1</td>
<td>4.3E-3</td>
<td>4.1E-3</td>
<td>4.5E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>1.5E-1</td>
<td>1.5E-2</td>
<td>9.7E-3</td>
<td>1.4E-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>1.8E-2</td>
<td>1.9E-3</td>
<td>1.8E-3</td>
<td>3.6E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.3E-3</td>
<td>8.3E-4</td>
<td>8.1E-4</td>
<td>6.9E-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q'(\lambda)</td>
<td>2.0E-3</td>
<td>2.0E-3</td>
<td>2.0E-3</td>
<td>2.1E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>2.9E-3</td>
<td>1.6E-3</td>
<td>1.6E-3</td>
<td>2.0E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>2.1E-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 452: Pixel-Gridworld, relative MSE. $T = 25, N = 128, \pi_b = 0.20\text{-Greedy(V iter.)}, \pi_e = 0.10\text{-Greedy(V iter.)}$. Note: we use the same policy as in Gridworld. $\pi_b$ known. Stochastic environment.

<table>
<thead>
<tr>
<th>Method</th>
<th>DM</th>
<th>Hybrid</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.4E2</td>
<td>8.0E1</td>
<td>7.7E1</td>
<td>6.0E1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-Reg</td>
<td>4.7E-2</td>
<td>2.3E-3</td>
<td>2.1E-3</td>
<td>1.9E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>1.9E-1</td>
<td>6.4E-3</td>
<td>4.6E-3</td>
<td>4.1E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>8.9E-3</td>
<td>9.6E-4</td>
<td>1.1E-3</td>
<td>2.0E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.2E-3</td>
<td>6.5E-4</td>
<td>6.3E-4</td>
<td>5.8E-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q'(\lambda)</td>
<td>3.0E-3</td>
<td>1.4E-3</td>
<td>1.4E-3</td>
<td>1.3E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>3.1E-3</td>
<td>6.0E-4</td>
<td>6.2E-4</td>
<td>1.1E-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>1.1E-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 453: Pixel-Gridworld, relative MSE. $T = 25, N = 256, \pi_b = 0.20\text{-Greedy(V iter.)}, \pi_e = 0.10\text{-Greedy(V iter.)}$. Note: we use the same policy as in Gridworld. $\pi_b$ known. Stochastic environment.

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>Hybrid</th>
<th>Magic</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.0E3</td>
<td>1.2E1</td>
<td>9.5E1</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>8.7E-3</td>
<td>5.5E-4</td>
<td>6.7E-4</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.2E-1</td>
<td>3.0E-3</td>
<td>3.2E-3</td>
</tr>
<tr>
<td>FQE</td>
<td>2.9E-3</td>
<td>2.7E-4</td>
<td>2.7E-4</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>6.4E-4</td>
<td>2.3E-4</td>
<td>2.3E-4</td>
</tr>
<tr>
<td>Q'(\lambda)</td>
<td>3.5E-3</td>
<td>2.6E-4</td>
<td>2.6E-4</td>
</tr>
<tr>
<td>Tree</td>
<td>7.5E-4</td>
<td>3.8E-4</td>
<td>3.9E-4</td>
</tr>
<tr>
<td>IH</td>
<td>1.6E-3</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 454: Pixel-Gridworld, relative MSE. $T = 25, N = 512, \pi_b = 0.20\text{-Greedy(V iter.)}, \pi_e = 0.10\text{-Greedy(V iter.)}$. Note: we use the same policy as in Gridworld. $\pi_b$ known. Stochastic environment.

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>Hybrid</th>
<th>Magic</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.4E2</td>
<td>6.8E0</td>
<td>6.8E0</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>5.2E-3</td>
<td>3.2E-4</td>
<td>2.8E-4</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.2E-1</td>
<td>1.2E-3</td>
<td>6.1E-4</td>
</tr>
<tr>
<td>FQE</td>
<td>9.7E-4</td>
<td>1.7E-4</td>
<td>1.7E-4</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>2.0E-3</td>
<td>2.2E-4</td>
<td>2.0E-4</td>
</tr>
<tr>
<td>Q'(\lambda)</td>
<td>9.5E-4</td>
<td>2.0E-4</td>
<td>2.0E-4</td>
</tr>
<tr>
<td>Tree</td>
<td>1.9E-3</td>
<td>1.8E-4</td>
<td>2.0E-4</td>
</tr>
<tr>
<td>IH</td>
<td>2.6E-4</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
<td>WDR</td>
</tr>
<tr>
<td>AM</td>
<td>2.3E0</td>
<td>7.7E2</td>
<td>1.4E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.5E0</td>
<td>7.3E2</td>
<td>1.4E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.0E0</td>
<td>7.9E2</td>
<td>1.5E0</td>
</tr>
<tr>
<td>FQE</td>
<td>1.6E-2</td>
<td>7.7E0</td>
<td>2.3E-2</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>3.2E-3</td>
<td>8.4E0</td>
<td>4.8E-3</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>2.5E-3</td>
<td>1.6E1</td>
<td>3.0E-3</td>
</tr>
<tr>
<td>Tree</td>
<td>1.3E-3</td>
<td>9.8E0</td>
<td>7.8E-3</td>
</tr>
<tr>
<td>IH</td>
<td>3.2E-2</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
<td>WDR</td>
</tr>
<tr>
<td>AM</td>
<td>2.8E1</td>
<td>8.4E0</td>
<td>4.9E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.6E-1</td>
<td>2.3E-2</td>
<td>1.6E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>7.6E-1</td>
<td>1.0E-1</td>
<td>2.1E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>1.9E-3</td>
<td>2.5E-4</td>
<td>2.9E-4</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>2.2E-3</td>
<td>1.6E-3</td>
<td>1.4E-3</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>2.3E-3</td>
<td>4.8E-3</td>
<td>2.3E-3</td>
</tr>
<tr>
<td>Tree</td>
<td>4.8E-4</td>
<td>2.3E-3</td>
<td>1.3E-3</td>
</tr>
<tr>
<td>IH</td>
<td>4.2E-2</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
<td>WDR</td>
</tr>
<tr>
<td>AM</td>
<td>3.2E-3</td>
<td>8.4E0</td>
<td>4.9E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.6E-1</td>
<td>2.3E-2</td>
<td>1.6E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>7.6E-1</td>
<td>1.0E-1</td>
<td>2.1E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>1.9E-3</td>
<td>2.5E-4</td>
<td>2.9E-4</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>2.2E-3</td>
<td>1.6E-3</td>
<td>1.4E-3</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>2.3E-3</td>
<td>4.8E-3</td>
<td>2.3E-3</td>
</tr>
<tr>
<td>Tree</td>
<td>4.8E-4</td>
<td>2.3E-3</td>
<td>1.3E-3</td>
</tr>
<tr>
<td>IH</td>
<td>4.2E-2</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
<td>WDR</td>
</tr>
<tr>
<td>AM</td>
<td>3.2E-3</td>
<td>8.4E0</td>
<td>4.9E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.6E-1</td>
<td>2.3E-2</td>
<td>1.6E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>7.6E-1</td>
<td>1.0E-1</td>
<td>2.1E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>1.9E-3</td>
<td>2.5E-4</td>
<td>2.9E-4</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>2.2E-3</td>
<td>1.6E-3</td>
<td>1.4E-3</td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>2.3E-3</td>
<td>4.8E-3</td>
<td>2.3E-3</td>
</tr>
<tr>
<td>Tree</td>
<td>4.8E-4</td>
<td>2.3E-3</td>
<td>1.3E-3</td>
</tr>
<tr>
<td>IH</td>
<td>4.2E-2</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 459: Pixel-Gridworld, relative MSE. \( T = 25, N = 128, \pi_b = 0.20\-\text{Greedy(V iter.)}, \pi_e = 0.10\-\text{Greedy(V iter.)}. \) Note: we use the same policy as in Gridworld. \( \pi_b \) unknown. Stochastic environment.

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid</th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>2.2E2</td>
<td>7.4E17</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.4E2</td>
<td>1.1E14</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.2E0</td>
<td>6.9E7</td>
</tr>
<tr>
<td>FQE</td>
<td>1.4E-1</td>
<td>2.4E13</td>
</tr>
<tr>
<td>R((\lambda)))</td>
<td>\textbf{4.6E-2}</td>
<td>6.7E12</td>
</tr>
<tr>
<td>Q^b((\lambda)))</td>
<td>5.4E-2</td>
<td>3.6E12</td>
</tr>
<tr>
<td>Tree</td>
<td>4.7E-2</td>
<td>6.2E12</td>
</tr>
<tr>
<td>IH</td>
<td>9.1E0</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 460: Pixel-Gridworld, relative MSE. \( T = 25, N = 256, \pi_b = 0.20\-\text{Greedy(V iter.)}, \pi_e = 0.10\-\text{Greedy(V iter.)}. \) Note: we use the same policy as in Gridworld. \( \pi_b \) unknown. Stochastic environment.

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid</th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>3.5E3</td>
<td>3.5E3</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.6E-1</td>
<td>6.3E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.4E-1</td>
<td>1.3E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>1.5E-1</td>
<td>6.2E-3</td>
</tr>
<tr>
<td>R((\lambda)))</td>
<td>5.9E-3</td>
<td>3.1E-3</td>
</tr>
<tr>
<td>Q^b((\lambda)))</td>
<td>8.3E-3</td>
<td>\textbf{1.2E-3}</td>
</tr>
<tr>
<td>Tree</td>
<td>4.2E-3</td>
<td>4.9E-3</td>
</tr>
<tr>
<td>IH</td>
<td>5.3E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 461: Pixel-Gridworld, relative MSE. \( T = 25, N = 64, \pi_b = 0.40\-\text{Greedy(V iter.)}, \pi_e = 0.10\-\text{Greedy(V iter.)}. \) Note: we use the same policy as in Gridworld. \( \pi_b \) known. Stochastic environment.

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid</th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>3.3E0</td>
<td>1.7E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.6E-1</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>5.8E-1</td>
<td>1.6E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>7.3E-2</td>
<td>5.5E-3</td>
</tr>
<tr>
<td>R((\lambda)))</td>
<td>6.6E-4</td>
<td>6.3E-4</td>
</tr>
<tr>
<td>Q^b((\lambda)))</td>
<td>9.2E-3</td>
<td>1.2E-3</td>
</tr>
<tr>
<td>Tree</td>
<td>4.1E-3</td>
<td>1.2E-3</td>
</tr>
<tr>
<td>IH</td>
<td>5.0E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 462: Pixel-Gridworld, relative MSE. \( T = 25, N = 128, \pi_b = 0.40\-\text{Greedy(V iter.)}, \pi_e = 0.10\-\text{Greedy(V iter.)}. \) Note: we use the same policy as in Gridworld. \( \pi_b \) known. Stochastic environment.

<table>
<thead>
<tr>
<th>DM</th>
<th>Hybrid</th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>2.7E-1</td>
<td>3.2E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.2E-2</td>
<td>1.9E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.2E0</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIRECT</td>
<td>2.8E0</td>
<td>1.2E0</td>
<td>1.1E0</td>
<td>5.1E-1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.1E-1</td>
<td>6.0E-3</td>
<td>5.9E-3</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>6.0E-1</td>
<td>6.5E-3</td>
<td>4.9E-3</td>
<td>1.4E-2</td>
</tr>
<tr>
<td>FQE</td>
<td>1.6E-1</td>
<td>1.1E-3</td>
<td>6.7E-4</td>
<td>1.1E-3</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>3.2E-3</td>
<td>4.1E-4</td>
<td>3.8E-4</td>
<td>1.3E-3</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>8.8E-3</td>
<td>5.6E-4</td>
<td>2.4E-4</td>
<td>1.4E-3</td>
</tr>
<tr>
<td>TREE</td>
<td>2.7E-3</td>
<td>3.8E-4</td>
<td>1.8E-4</td>
<td>5.3E-4</td>
</tr>
<tr>
<td>IH</td>
<td>5.5E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIRECT</td>
<td>5.9E1</td>
<td>4.6E8</td>
<td>5.0E1</td>
<td>1.5E1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.1E0</td>
<td>1.1E14</td>
<td>5.1E-1</td>
<td>5.0E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.3E0</td>
<td>4.2E19</td>
<td>4.3E0</td>
<td>1.5E0</td>
</tr>
<tr>
<td>FQE</td>
<td>1.5E-1</td>
<td>5.6E15</td>
<td>7.8E-2</td>
<td>3.8E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>2.0E-3</td>
<td>1.2E16</td>
<td>1.8E-3</td>
<td>1.1E-3</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>6.6E-3</td>
<td>6.2E15</td>
<td>3.4E-2</td>
<td>1.3E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>4.5E-3</td>
<td>2.7E15</td>
<td>3.6E-2</td>
<td>8.7E-3</td>
</tr>
<tr>
<td>IH</td>
<td>7.0E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIRECT</td>
<td>3.0E19</td>
<td>1.9E8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.5E-1</td>
<td>2.1E-1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.1E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FQE</td>
<td>6.6E-2</td>
<td>3.4E-3</td>
<td>2.4E-3</td>
<td>6.6E-3</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.6E-1</td>
<td>2.5E0</td>
<td>1.9E-2</td>
<td>1.8E-3</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>7.0E-2</td>
<td>6.7E0</td>
<td>3.7E-3</td>
<td>1.7E-3</td>
</tr>
<tr>
<td>TREE</td>
<td>6.6E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IH</td>
<td>1.1E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIRECT</td>
<td>3.0E19</td>
<td>1.9E8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.5E-1</td>
<td>2.1E-1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.1E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FQE</td>
<td>6.6E-2</td>
<td>3.4E-3</td>
<td>2.4E-3</td>
<td>6.6E-3</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.6E-1</td>
<td>2.5E0</td>
<td>1.9E-2</td>
<td>1.8E-3</td>
</tr>
<tr>
<td>Q($\lambda$)</td>
<td>7.0E-2</td>
<td>6.7E0</td>
<td>3.7E-3</td>
<td>1.7E-3</td>
</tr>
<tr>
<td>TREE</td>
<td>6.6E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IH</td>
<td>1.1E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 463: 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.
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
<td>DR</td>
<td>WDR</td>
</tr>
<tr>
<td>AM</td>
<td>5.5E1</td>
<td>1.8E2</td>
<td>5.5E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>8.3E-1</td>
<td>2.2E1</td>
<td>3.1E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>8.5E-1</td>
<td>1.8E1</td>
<td>1.7E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>3.1E-2</td>
<td>8.9E-2</td>
<td>3.1E-3</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>$1.7E-3$</td>
<td>4.1E-2</td>
<td>2.9E-3</td>
</tr>
<tr>
<td>$Q^*(\lambda)$</td>
<td>2.1E-3</td>
<td>5.9E-2</td>
<td>3.0E-3</td>
</tr>
<tr>
<td>TREE</td>
<td>5.9E-3</td>
<td>6.5E-2</td>
<td>6.0E-3</td>
</tr>
<tr>
<td>IH</td>
<td>1.0E-1</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>4.2E2</td>
<td>3.3E10</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.9E0</td>
<td>1.2E6</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.3E0</td>
<td>6.1E8</td>
</tr>
<tr>
<td>FQE</td>
<td>5.2E-1</td>
<td>1.1E6</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>$7.0E-2$</td>
<td>1.3E9</td>
</tr>
<tr>
<td>$Q^*(\lambda)$</td>
<td>1.6E-1</td>
<td>4.6E8</td>
</tr>
<tr>
<td>TREE</td>
<td>7.1E-2</td>
<td>5.3E8</td>
</tr>
<tr>
<td>IH</td>
<td>1.4E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>3.3E3</td>
<td>1.4E6</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.1E1</td>
<td>2.6E3</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.1E0</td>
<td>1.2E9</td>
</tr>
<tr>
<td>FQE</td>
<td>3.9E-1</td>
<td>2.1E5</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>$5.2E-2$</td>
<td>6.0E7</td>
</tr>
<tr>
<td>$Q^*(\lambda)$</td>
<td>1.1E-1</td>
<td>1.8E7</td>
</tr>
<tr>
<td>TREE</td>
<td>6.2E-2</td>
<td>1.9E7</td>
</tr>
<tr>
<td>IH</td>
<td>1.3E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.3E3</td>
<td>4.9E2</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.5E0</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.6E0</td>
<td>1.9E0</td>
</tr>
<tr>
<td>FQE</td>
<td>2.3E-1</td>
<td>7.0E-2</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>$3.5E-2$</td>
<td>9.0E-2</td>
</tr>
<tr>
<td>$Q^*(\lambda)$</td>
<td>6.5E-2</td>
<td>5.3E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>6.2E-2</td>
<td>7.8E-2</td>
</tr>
<tr>
<td>IH</td>
<td>1.0E0</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>9.1E0</td>
<td>1.1E1</td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>$4.1E-2$</td>
<td>5.1E-2</td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>1.2E0</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.9E10</td>
<td>3.1E2</td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>2.0E0</td>
<td>$1.2E0$</td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.0E0</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>2.2E11</td>
<td>1.9E5</td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>$2.3E0$</td>
<td>6.9E0</td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.3E0</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.3E0</td>
<td>1.1E0</td>
<td></td>
</tr>
<tr>
<td>WIS</td>
<td>3.1E-1</td>
<td>$1.9E-1$</td>
<td></td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.9E0</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning

Table 471: Pixel-Gridworld, relative MSE. \( T = 25, N = 64, \pi_b = 0.60\text{-Greedy(V iter.)}, \pi_e = 0.10\text{-Greedy(V iter.)}. \) Note: we use the same policy as in Gridworld. \( \pi_b \) known. Stochastic environment.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
<td>DR</td>
<td>WDR</td>
</tr>
<tr>
<td>AM</td>
<td>6.0E1</td>
<td>2.6E2</td>
<td>1.6E1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.9E0</td>
<td>3.2E0</td>
<td>1.6E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>7.7E-1</td>
<td>2.5E0</td>
<td>1.6E0</td>
</tr>
<tr>
<td>FQE</td>
<td>2.5E-1</td>
<td>4.1E-1</td>
<td>3.4E-2</td>
</tr>
<tr>
<td>(( R(\lambda) ))</td>
<td>7.2E-3</td>
<td>3.6E-2</td>
<td>2.7E-2</td>
</tr>
<tr>
<td>Q(^\pi)(( \lambda ))</td>
<td>3.0E-2</td>
<td>4.0E-2</td>
<td>1.6E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>9.0E-3</td>
<td>3.4E-1</td>
<td>2.7E-2</td>
</tr>
<tr>
<td>IH</td>
<td>4.7E-1</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 472: Pixel-Gridworld, relative MSE. \( T = 25, N = 128, \pi_b = 0.60\text{-Greedy(V iter.)}, \pi_e = 0.10\text{-Greedy(V iter.)}. \) Note: we use the same policy as in Gridworld. \( \pi_b \) known. Stochastic environment.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
<td>DR</td>
<td>WDR</td>
</tr>
<tr>
<td>AM</td>
<td>5.4E1</td>
<td>1.5E1</td>
<td>1.9E1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>3.1E-1</td>
<td>1.5E-1</td>
<td>3.8E-2</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.1E0</td>
<td>6.3E-1</td>
<td>3.7E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>7.4E-2</td>
<td>5.9E-3</td>
<td>6.5E-3</td>
</tr>
<tr>
<td>(( R(\lambda) ))</td>
<td>8.1E-3</td>
<td>5.5E-3</td>
<td>4.7E-3</td>
</tr>
<tr>
<td>Q(^\pi)(( \lambda ))</td>
<td>4.1E-2</td>
<td>1.1E-2</td>
<td>2.1E-3</td>
</tr>
<tr>
<td>TREE</td>
<td>3.2E-3</td>
<td>3.3E-3</td>
<td>2.4E-3</td>
</tr>
<tr>
<td>IH</td>
<td>5.2E-1</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 473: Pixel-Gridworld, relative MSE. \( T = 25, N = 256, \pi_b = 0.60\text{-Greedy(V iter.)}, \pi_e = 0.10\text{-Greedy(V iter.)}. \) Note: we use the same policy as in Gridworld. \( \pi_b \) known. Stochastic environment.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
<td>DR</td>
<td>WDR</td>
</tr>
<tr>
<td>AM</td>
<td>3.5E0</td>
<td>8.7E0</td>
<td>7.0E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.5E0</td>
<td>1.5E-1</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.7E0</td>
<td>3.8E0</td>
<td>1.0E0</td>
</tr>
<tr>
<td>FQE</td>
<td>3.4E-2</td>
<td>3.2E-3</td>
<td>1.3E-3</td>
</tr>
<tr>
<td>(( R(\lambda) ))</td>
<td>1.1E-2</td>
<td>1.6E-2</td>
<td>4.9E-3</td>
</tr>
<tr>
<td>Q(^\pi)(( \lambda ))</td>
<td>4.6E-2</td>
<td>2.0E-2</td>
<td>7.1E-3</td>
</tr>
<tr>
<td>TREE</td>
<td>8.3E-3</td>
<td>9.8E-3</td>
<td>5.0E-3</td>
</tr>
<tr>
<td>IH</td>
<td>5.2E-1</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 474: Pixel-Gridworld, relative MSE. \( T = 25, N = 512, \pi_b = 0.60\text{-Greedy(V iter.)}, \pi_e = 0.10\text{-Greedy(V iter.)}. \) Note: we use the same policy as in Gridworld. \( \pi_b \) known. Stochastic environment.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
<td>DR</td>
<td>WDR</td>
</tr>
<tr>
<td>AM</td>
<td>1.2E0</td>
<td>1.7E0</td>
<td>-</td>
</tr>
<tr>
<td>Q-REG</td>
<td>4.6E-2</td>
<td>1.5E-1</td>
<td>-</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.3E0</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th></th>
<th>STANDARD</th>
<th>PER-DECISION</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>2.8E0</td>
<td>3.9E0</td>
</tr>
<tr>
<td>WIS</td>
<td>\textbf{2.0E-1}</td>
<td>5.6E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.9E0</td>
<td>-</td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th></th>
<th>STANDARD</th>
<th>PER-DECISION</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>3.1E-1</td>
<td>5.1E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>\textbf{6.3E-2}</td>
<td>1.4E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.3E0</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>9.8E0</td>
<td>1.6E20</td>
<td>9.4E1</td>
<td>1.1E1</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>1.0E1</td>
<td>7.3E10</td>
<td>1.3E1</td>
<td>2.3E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.2E0</td>
<td>2.9E20</td>
<td>2.8E0</td>
<td>3.5E0</td>
</tr>
<tr>
<td>FQE</td>
<td>3.9E-1</td>
<td>1.5E17</td>
<td>6.8E-2</td>
<td>1.7E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>2.7E-2</td>
<td>2.5E18</td>
<td><strong>1.2E-2</strong></td>
<td>2.2E-2</td>
</tr>
<tr>
<td>Q(\lambda)</td>
<td>5.1E-2</td>
<td>6.0E18</td>
<td>3.5E-2</td>
<td>2.2E-2</td>
</tr>
<tr>
<td>Tree</td>
<td><strong>1.9E-2</strong></td>
<td>2.5E18</td>
<td>2.5E-2</td>
<td>2.3E-2</td>
</tr>
<tr>
<td>IH</td>
<td>4.8E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.5E21</td>
<td>2.4E5</td>
</tr>
<tr>
<td>WIS</td>
<td><strong>2.1E-1</strong></td>
<td>9.3E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.5E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>3.6E1</td>
<td>1.7E2</td>
<td>5.0E1</td>
<td>8.0E1</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>5.7E-1</td>
<td>6.4E-1</td>
<td>1.2E-1</td>
<td>4.9E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.3E-1</td>
<td>1.6E1</td>
<td>5.3E0</td>
<td>3.5E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>4.1E-2</td>
<td>9.9E-2</td>
<td>3.3E-3</td>
<td><strong>1.8E-3</strong></td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>2.3E-2</td>
<td>9.6E-2</td>
<td>7.3E-3</td>
<td>3.6E-3</td>
</tr>
<tr>
<td>Q(\lambda)</td>
<td>3.5E-2</td>
<td>1.0E-1</td>
<td>1.8E-2</td>
<td>1.1E-2</td>
</tr>
<tr>
<td>Tree</td>
<td><strong>7.9E-3</strong></td>
<td>4.2E-2</td>
<td>2.6E-3</td>
<td>3.6E-3</td>
</tr>
<tr>
<td>IH</td>
<td>5.7E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>7.5E-1</td>
<td>2.3E0</td>
</tr>
<tr>
<td>WIS</td>
<td>1.0E-1</td>
<td><strong>1.0E-1</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.4E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>1.4E1</td>
<td>5.5E14</td>
<td>8.8E1</td>
<td>1.2E1</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>6.2E-1</td>
<td>3.0E15</td>
<td>1.3E-1</td>
<td>6.5E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.1E-1</td>
<td>1.6E18</td>
<td>3.4E-1</td>
<td>5.6E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>1.4E-1</td>
<td>3.4E15</td>
<td>2.5E-2</td>
<td>3.9E-2</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td><strong>9.0E-3</strong></td>
<td>3.0E15</td>
<td>1.1E-2</td>
<td><strong>3.3E-3</strong></td>
</tr>
<tr>
<td>Q(\lambda)</td>
<td>2.2E-2</td>
<td>3.3E16</td>
<td>1.3E-2</td>
<td>2.2E-2</td>
</tr>
<tr>
<td>Tree</td>
<td>1.2E-2</td>
<td>9.1E16</td>
<td>1.9E-2</td>
<td>7.9E-3</td>
</tr>
<tr>
<td>IH</td>
<td>5.9E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>6.1E18</td>
<td>4.4E5</td>
</tr>
<tr>
<td>WIS</td>
<td><strong>1.7E-1</strong></td>
<td>5.0E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>4.5E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.0E2</td>
<td>4.8E16</td>
<td>3.0E2</td>
<td>2.2E2</td>
</tr>
<tr>
<td>Q-Reg</td>
<td>9.9E0</td>
<td>8.3E11</td>
<td>2.0E1</td>
<td>1.6E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.5E0</td>
<td>6.5E18</td>
<td>4.0E1</td>
<td>2.7E1</td>
</tr>
<tr>
<td>FQE</td>
<td>9.0E-1</td>
<td>2.3E13</td>
<td>5.2E-1</td>
<td>5.9E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>2.9E-1</td>
<td>1.4E16</td>
<td>3.1E-1</td>
<td>2.2E-1</td>
</tr>
<tr>
<td>Q(\lambda)</td>
<td>5.0E-1</td>
<td>1.3E16</td>
<td>2.8E-1</td>
<td>2.8E-1</td>
</tr>
<tr>
<td>Tree</td>
<td><strong>1.8E-1</strong></td>
<td>1.3E15</td>
<td>3.5E-1</td>
<td><strong>1.6E-1</strong></td>
</tr>
<tr>
<td>IH</td>
<td>2.9E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

IPS

<table>
<thead>
<tr>
<th>Method</th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>2.6E18</td>
<td>2.2E5</td>
</tr>
<tr>
<td>WIS</td>
<td>2.2E0</td>
<td><strong>2.2E0</strong></td>
</tr>
<tr>
<td>NAIVE</td>
<td>1.2E1</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>7.4E1</td>
<td>2.9E12</td>
<td>1.1E4</td>
<td>9.7E1</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.6E0</td>
<td>7.0E7</td>
<td>1.6E0</td>
<td>2.4E0</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>3.8E0</td>
<td>6.6E14</td>
<td>4.1E0</td>
<td>4.3E0</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>1.0E0</td>
<td>8.8E12</td>
<td>3.4E-1</td>
<td>4.5E-1</td>
<td></td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>4.0E-1</td>
<td>2.5E14</td>
<td>2.3E-1</td>
<td>1.7E-1</td>
<td></td>
</tr>
<tr>
<td>$Q^*(\lambda)$</td>
<td>5.9E-1</td>
<td>8.7E13</td>
<td>3.2E-1</td>
<td>3.0E-1</td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>3.9E-1</td>
<td>9.4E13</td>
<td>3.1E-1</td>
<td>2.5E-1</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>3.1E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.6E0</td>
<td>8.2E1</td>
<td>2.3E1</td>
<td>4.1E0</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.2E0</td>
<td>1.8E0</td>
<td>1.1E0</td>
<td>2.5E0</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>1.2E0</td>
<td>8.8E-1</td>
<td>9.6E-1</td>
<td>2.1E0</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>2.8E0</td>
<td>1.5E0</td>
<td>4.1E-1</td>
<td>1.7E0</td>
<td></td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.4E-1</td>
<td>1.7E-1</td>
<td>9.5E-2</td>
<td>7.1E-2</td>
<td></td>
</tr>
<tr>
<td>$Q^*(\lambda)$</td>
<td>6.2E0</td>
<td>4.3E0</td>
<td>7.5E-1</td>
<td>3.7E0</td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>4.2E-1</td>
<td>3.7E-1</td>
<td>2.6E-1</td>
<td>3.7E-1</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>1.6E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>9.2E3</td>
<td>3.3E5</td>
<td>6.5E4</td>
<td>9.4E3</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.4E0</td>
<td>3.6E2</td>
<td>1.4E0</td>
<td>1.7E0</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>9.9E-1</td>
<td>2.3E11</td>
<td>1.4E0</td>
<td>1.5E0</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>3.9E-1</td>
<td>4.7E8</td>
<td>7.3E-2</td>
<td>7.2E-2</td>
<td></td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.2E-1</td>
<td>7.3E10</td>
<td>1.1E-1</td>
<td>6.9E-2</td>
<td></td>
</tr>
<tr>
<td>$Q^*(\lambda)$</td>
<td>3.8E-1</td>
<td>3.0E10</td>
<td>3.7E-1</td>
<td>2.5E-1</td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>1.7E-1</td>
<td>2.9E10</td>
<td>1.2E-1</td>
<td>9.9E-2</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>3.2E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>9.8E1</td>
<td>1.5E3</td>
<td>4.9E2</td>
<td>9.6E1</td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.2E0</td>
<td>3.3E0</td>
<td>1.5E0</td>
<td>2.3E0</td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>8.4E-1</td>
<td>9.7E0</td>
<td>3.0E1</td>
<td>1.3E0</td>
<td></td>
</tr>
<tr>
<td>FQE</td>
<td>4.7E-1</td>
<td>2.0E-1</td>
<td>9.9E-2</td>
<td>1.7E-1</td>
<td></td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>7.3E-2</td>
<td>1.7E-1</td>
<td>7.9E-2</td>
<td>3.9E-2</td>
<td></td>
</tr>
<tr>
<td>$Q^*(\lambda)$</td>
<td>1.8E-1</td>
<td>3.9E-1</td>
<td>6.2E-2</td>
<td>3.9E-2</td>
<td></td>
</tr>
<tr>
<td>Tree</td>
<td>2.2E-2</td>
<td>7.2E-1</td>
<td>6.4E-2</td>
<td>4.9E-2</td>
<td></td>
</tr>
<tr>
<td>IH</td>
<td>1.4E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 483: 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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>2.2E13</td>
<td>3.1E0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.6E0</td>
<td>1.8E0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRDR</td>
<td>1.2E1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning

Table 483: Pixel-Gridworld, relative MSE. $T = 25, N = 256, \pi_b = 0.80\text{-Greedy(V iter.)}, \pi_e = 0.10\text{-Greedy(V iter.)}. Note: we use the same policy as in Gridworld. $\pi_b$ known. Stochastic environment.

<table>
<thead>
<tr>
<th>DM</th>
<th>DIRECT</th>
<th>HYBRID</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>7.4E0</td>
<td>1.2E2</td>
<td>1.4E2</td>
<td>5.2E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.4E0</td>
<td>3.1E0</td>
<td>3.0E0</td>
<td>2.0E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.4E-1</td>
<td>3.1E0</td>
<td>3.2E0</td>
<td>1.1E0</td>
</tr>
<tr>
<td>FQE</td>
<td>3.6E-1</td>
<td>6.2E-2</td>
<td>4.6E-2</td>
<td>4.5E-2</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>9.0E-2</td>
<td>1.9E-1</td>
<td>3.5E-2</td>
<td><strong>2.8E-2</strong></td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td><strong>8.3E-2</strong></td>
<td>1.8E-1</td>
<td>7.4E-2</td>
<td>5.4E-2</td>
</tr>
<tr>
<td>TREE</td>
<td>2.5E-1</td>
<td>1.1E-1</td>
<td>7.8E-2</td>
<td>1.3E-1</td>
</tr>
<tr>
<td>IH</td>
<td>1.5E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 484: Pixel-Gridworld, relative MSE. $T = 25, N = 512, \pi_b = 0.80\text{-Greedy(V iter.)}, \pi_e = 0.10\text{-Greedy(V iter.)}. Note: we use the same policy as in Gridworld. $\pi_b$ known. Stochastic environment.

<table>
<thead>
<tr>
<th>DM</th>
<th>DIRECT</th>
<th>HYBRID</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>4.8E0</td>
<td>1.4E15</td>
<td>7.4E0</td>
<td>5.3E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.2E0</td>
<td>4.7E11</td>
<td>2.7E0</td>
<td>3.9E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.3E0</td>
<td>2.5E16</td>
<td>1.6E0</td>
<td>3.6E0</td>
</tr>
<tr>
<td>FQE</td>
<td>1.3E0</td>
<td>1.1E12</td>
<td>2.2E-1</td>
<td>7.5E-1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>6.5E-2</td>
<td>4.7E14</td>
<td>8.6E-2</td>
<td><strong>4.9E-2</strong></td>
</tr>
<tr>
<td>$Q^b(\lambda)$</td>
<td>6.2E-1</td>
<td>1.1E16</td>
<td>4.7E-1</td>
<td>4.3E-1</td>
</tr>
<tr>
<td>TREE</td>
<td>2.7E-1</td>
<td>1.4E15</td>
<td>1.4E-1</td>
<td>1.8E-1</td>
</tr>
<tr>
<td>IH</td>
<td>1.3E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 485: Pixel-Gridworld, relative MSE. $T = 25, N = 64, \pi_b = 0.80\text{-Greedy(V iter.)}, \pi_e = 0.10\text{-Greedy(V iter.)}. Note: we use the same policy as in Gridworld. $\pi_b$ unknown. Stochastic environment.

<table>
<thead>
<tr>
<th>IPS</th>
<th>STANDARD</th>
<th>PER-DECISION</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.4E0</td>
<td>1.9E0</td>
</tr>
<tr>
<td>WIS</td>
<td><strong>4.7E-1</strong></td>
<td>6.6E-1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>8.1E0</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 486: Pixel-Gridworld, relative MSE. $T = 25, N = 128, \pi_b = 0.80\text{-Greedy(V iter.)}, \pi_e = 0.10\text{-Greedy(V iter.)}. Note: we use the same policy as in Gridworld. $\pi_b$ unknown. Stochastic environment.

<table>
<thead>
<tr>
<th>IPS</th>
<th>STANDARD</th>
<th>PER-DECISION</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>2.6E17</td>
<td>7.6E1</td>
</tr>
<tr>
<td>WIS</td>
<td><strong>1.1E0</strong></td>
<td>3.0E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>8.3E0</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>7.2E1</td>
<td>1.1E15</td>
</tr>
<tr>
<td>Q-REG</td>
<td>5.1E0</td>
<td>5.6E11</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.7E0</td>
<td>4.5E18</td>
</tr>
<tr>
<td>FQE</td>
<td>7.8E-2</td>
<td>2.5E13</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>5.7E-2</td>
<td>2.6E17</td>
</tr>
<tr>
<td>Q$^\theta$($\lambda$)</td>
<td>2.2E-1</td>
<td>2.3E17</td>
</tr>
<tr>
<td>Tree</td>
<td>2.5E-1</td>
<td>2.4E17</td>
</tr>
<tr>
<td>IH</td>
<td>1.4E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>7.7E0</td>
<td>2.7E17</td>
</tr>
<tr>
<td>Q-REG</td>
<td>5.4E0</td>
<td>5.8E14</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.1E0</td>
<td>8.4E21</td>
</tr>
<tr>
<td>FQE</td>
<td>1.0E1</td>
<td>1.9E19</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.7E1</td>
<td>1.6E20</td>
</tr>
<tr>
<td>Q$^\theta$($\lambda$)</td>
<td>2.0E1</td>
<td>1.3E21</td>
</tr>
<tr>
<td>Tree</td>
<td>1.7E1</td>
<td>1.4E20</td>
</tr>
<tr>
<td>IH</td>
<td>5.9E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>4.8E1</td>
<td>1.3E13</td>
</tr>
<tr>
<td>Q-REG</td>
<td>2.3E0</td>
<td>1.9E9</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.4E1</td>
<td>5.5E14</td>
</tr>
<tr>
<td>FQE</td>
<td>1.1E0</td>
<td>2.9E13</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>1.2E0</td>
<td>7.5E13</td>
</tr>
<tr>
<td>Q$^\theta$($\lambda$)</td>
<td>7.8E0</td>
<td>3.0E14</td>
</tr>
<tr>
<td>Tree</td>
<td>1.3E0</td>
<td>2.1E14</td>
</tr>
<tr>
<td>IH</td>
<td>5.7E0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>1.9E16</td>
<td>1.1E1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>8.9E0</td>
<td>3.6E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>2.4E1</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct</td>
<td>DR</td>
</tr>
<tr>
<td>AM</td>
<td>7.2E1</td>
<td>1.1E15</td>
</tr>
<tr>
<td>Q-REG</td>
<td>5.1E0</td>
<td>5.6E11</td>
</tr>
<tr>
<td>MRDR</td>
<td>4.7E0</td>
<td>4.5E18</td>
</tr>
<tr>
<td>FQE</td>
<td>7.8E-2</td>
<td>2.5E13</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>5.7E-2</td>
<td>2.6E17</td>
</tr>
<tr>
<td>Q$^\theta$($\lambda$)</td>
<td>2.2E-1</td>
<td>2.3E17</td>
</tr>
<tr>
<td>Tree</td>
<td>2.5E-1</td>
<td>2.4E17</td>
</tr>
<tr>
<td>IH</td>
<td>1.4E0</td>
<td>-</td>
</tr>
</tbody>
</table>
**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.

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AM</td>
<td>2.1E0</td>
<td>3.7E1</td>
<td>1.4E1</td>
<td>5.7E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.1E0</td>
<td>1.8E0</td>
<td>3.2E0</td>
<td>4.0E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.1E0</td>
<td>8.8E-1</td>
<td>3.6E0</td>
<td>4.5E0</td>
</tr>
<tr>
<td>FQE</td>
<td>2.3E1</td>
<td>2.1E1</td>
<td>6.1E0</td>
<td>1.4E1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>9.1E2</td>
<td>2.0E3</td>
<td>7.6E4</td>
<td>9.1E2</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>5.7E1</td>
<td>5.7E1</td>
<td>7.3E1</td>
<td>4.8E1</td>
</tr>
<tr>
<td>Tree</td>
<td>2.0E1</td>
<td>2.1E2</td>
<td>4.2E2</td>
<td>6.9E1</td>
</tr>
<tr>
<td>IH</td>
<td>2.0E0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td>1.0E0</td>
<td>1.7E0</td>
</tr>
<tr>
<td>WIS</td>
<td>1.6E1</td>
<td>3.3E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>1.1E1</td>
<td>-</td>
</tr>
</tbody>
</table>

---

**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.

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AM</td>
<td>3.2E0</td>
<td>2.1E1</td>
<td>3.8E1</td>
<td>7.6E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.3E0</td>
<td>1.6E0</td>
<td>5.9E0</td>
<td>6.1E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.1E0</td>
<td>9.1E-1</td>
<td>8.0E0</td>
<td>8.1E0</td>
</tr>
<tr>
<td>FQE</td>
<td>1.1E1</td>
<td>1.1E1</td>
<td>2.8E0</td>
<td>6.5E0</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>5.1E0</td>
<td>4.8E0</td>
<td>2.7E0</td>
<td>2.8E0</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>3.1E1</td>
<td>2.7E1</td>
<td>9.8E0</td>
<td>1.2E1</td>
</tr>
<tr>
<td>Tree</td>
<td>3.2E0</td>
<td>3.2E0</td>
<td>1.4E0</td>
<td>2.0E0</td>
</tr>
<tr>
<td>IH</td>
<td>2.1E0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td>1.0E0</td>
<td>1.7E0</td>
</tr>
<tr>
<td>WIS</td>
<td>1.6E1</td>
<td>3.3E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>1.1E1</td>
<td>-</td>
</tr>
</tbody>
</table>

---

**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.

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AM</td>
<td>3.1E0</td>
<td>6.2E1</td>
<td>3.2E1</td>
<td>3.5E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.2E0</td>
<td>2.0E0</td>
<td>3.8E0</td>
<td>4.3E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.1E0</td>
<td>1.2E0</td>
<td>2.5E0</td>
<td>2.9E0</td>
</tr>
<tr>
<td>FQE</td>
<td>3.9E-1</td>
<td>3.7E-1</td>
<td>3.3E-1</td>
<td>2.9E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>7.7E0</td>
<td>1.1E1</td>
<td>5.9E0</td>
<td>5.8E0</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>3.9E1</td>
<td>3.8E1</td>
<td>9.5E0</td>
<td>5.8E0</td>
</tr>
<tr>
<td>Tree</td>
<td>5.8E-1</td>
<td>4.0E0</td>
<td>3.0E-1</td>
<td>3.1E-1</td>
</tr>
<tr>
<td>IH</td>
<td>2.1E0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td>1.0E0</td>
<td>1.7E0</td>
</tr>
<tr>
<td>WIS</td>
<td>4.1E0</td>
<td>4.2E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>1.1E1</td>
<td>-</td>
</tr>
</tbody>
</table>

---

**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.

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AM</td>
<td>4.9E0</td>
<td>1.1E3</td>
<td>3.7E2</td>
<td>6.3E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.9E0</td>
<td>2.4E0</td>
<td>3.5E0</td>
<td>4.4E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.7E-1</td>
<td>2.5E1</td>
<td>4.7E0</td>
<td>2.7E0</td>
</tr>
<tr>
<td>FQE</td>
<td>5.7E-1</td>
<td>7.3E0</td>
<td>6.6E-2</td>
<td>1.3E-1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>5.4E-1</td>
<td>5.6E0</td>
<td>1.1E-1</td>
<td>2.2E-1</td>
</tr>
<tr>
<td>Q^*(\lambda)</td>
<td>3.6E1</td>
<td>3.2E1</td>
<td>6.4E0</td>
<td>1.2E1</td>
</tr>
<tr>
<td>Tree</td>
<td>1.3E0</td>
<td>3.5E1</td>
<td>3.5E-1</td>
<td>6.5E-1</td>
</tr>
<tr>
<td>IH</td>
<td>2.1E0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td>9.8E-1</td>
<td>8.7E0</td>
</tr>
<tr>
<td>WIS</td>
<td>2.4E0</td>
<td>4.2E0</td>
</tr>
<tr>
<td>NAIVE</td>
<td>1.1E1</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AM</td>
<td>2.1E0</td>
<td>7.8E0</td>
<td>9.7E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.2E0</td>
<td>1.6E0</td>
<td>6.5E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>$1.1E0$</td>
<td>$9.6E-1$</td>
<td>8.1E0</td>
</tr>
<tr>
<td>FQE</td>
<td>2.8E1</td>
<td>2.3E1</td>
<td>1.2E1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.5E1</td>
<td>1.6E1</td>
<td>8.7E0</td>
</tr>
<tr>
<td>Q$^\theta$(\lambda)</td>
<td>5.2E1</td>
<td>3.4E1</td>
<td>1.4E1</td>
</tr>
<tr>
<td>Tree</td>
<td>4.8E1</td>
<td>4.9E1</td>
<td>1.1E1</td>
</tr>
<tr>
<td>IH</td>
<td>2.1E0</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AM</td>
<td>5.3E0</td>
<td>2.5E19</td>
<td>4.1E1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.1E0</td>
<td>1.4E16</td>
<td>4.5E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>$8.8E-1$</td>
<td>5.1E21</td>
<td>5.2E0</td>
</tr>
<tr>
<td>FQE</td>
<td>2.4E1</td>
<td>2.2E19</td>
<td>8.0E0</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>3.1E1</td>
<td>1.6E20</td>
<td>4.6E0</td>
</tr>
<tr>
<td>Q$^\theta$(\lambda)</td>
<td>2.7E1</td>
<td>1.7E20</td>
<td>8.7E0</td>
</tr>
<tr>
<td>Tree</td>
<td>1.4E2</td>
<td>2.1E19</td>
<td>7.1E0</td>
</tr>
<tr>
<td>IH</td>
<td>2.0E0</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AM</td>
<td>3.4E0</td>
<td>5.4E1</td>
<td>1.7E1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.7E0</td>
<td>3.6E0</td>
<td>1.9E1</td>
</tr>
<tr>
<td>MRDR</td>
<td>$1.2E0$</td>
<td>$8.5E-1$</td>
<td>3.0E1</td>
</tr>
<tr>
<td>FQE</td>
<td>1.2E2</td>
<td>1.1E2</td>
<td>3.4E1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.1E3</td>
<td>2.5E3</td>
<td>1.4E4</td>
</tr>
<tr>
<td>Q$^\theta$(\lambda)</td>
<td>1.0E2</td>
<td>1.8E2</td>
<td>1.9E1</td>
</tr>
<tr>
<td>Tree</td>
<td>2.8E1</td>
<td>4.5E1</td>
<td>3.2E1</td>
</tr>
<tr>
<td>IH</td>
<td>7.8E0</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>Hybrid</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AM</td>
<td>2.6E0</td>
<td>6.8E15</td>
<td>3.5E1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.2E0</td>
<td>2.0E8</td>
<td>3.8E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>$1.2E0$</td>
<td>6.1E16</td>
<td>4.5E0</td>
</tr>
<tr>
<td>FQE</td>
<td>2.5E0</td>
<td>1.4E13</td>
<td>5.8E-1</td>
</tr>
<tr>
<td>$R(\lambda)$</td>
<td>1.4E0</td>
<td>3.0E16</td>
<td>$4.2E-1$</td>
</tr>
<tr>
<td>Q$^\theta$(\lambda)</td>
<td>2.8E1</td>
<td>6.0E16</td>
<td>9.4E0</td>
</tr>
<tr>
<td>Tree</td>
<td>2.1E0</td>
<td>1.9E16</td>
<td>5.1E-1</td>
</tr>
<tr>
<td>IH</td>
<td>2.0E0</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>HYBRID</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
<td>DR</td>
<td>WDR</td>
<td>MAGIC</td>
</tr>
<tr>
<td>AM</td>
<td>5.9E0</td>
<td>8.7E1</td>
<td>7.0E1</td>
<td>1.1E1</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.5E0</td>
<td>2.1E0</td>
<td>8.5E0</td>
<td>9.3E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.1E0</td>
<td>6.7E0</td>
<td>1.2E1</td>
<td>1.4E1</td>
</tr>
<tr>
<td>FQE</td>
<td>6.1E1</td>
<td>6.0E1</td>
<td>1.6E1</td>
<td>2.4E1</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>2.3E1</td>
<td>2.4E2</td>
<td>1.0E1</td>
<td>1.0E1</td>
</tr>
<tr>
<td>Q^+(\lambda)</td>
<td>3.9E1</td>
<td>8.4E1</td>
<td>1.6E1</td>
<td>2.5E1</td>
</tr>
<tr>
<td>TREE</td>
<td>1.8E1</td>
<td>2.3E1</td>
<td>9.1E0</td>
<td>1.1E1</td>
</tr>
<tr>
<td>IH</td>
<td>8.4E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>HYBRID</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIRECT</td>
<td>DR</td>
<td>WDR</td>
<td>MAGIC</td>
</tr>
<tr>
<td>AM</td>
<td>4.4E0</td>
<td>1.8E3</td>
<td>2.4E2</td>
<td>7.3E0</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.7E0</td>
<td>3.4E0</td>
<td>6.7E0</td>
<td>8.5E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.5E0</td>
<td>8.8E2</td>
<td>8.6E0</td>
<td>8.9E0</td>
</tr>
<tr>
<td>FQE</td>
<td>6.5E0</td>
<td>5.0E1</td>
<td>2.3E0</td>
<td>2.7E0</td>
</tr>
<tr>
<td>R(\lambda)</td>
<td>1.4E1</td>
<td>3.8E4</td>
<td>4.5E0</td>
<td>5.7E0</td>
</tr>
<tr>
<td>Q^+(\lambda)</td>
<td>6.6E1</td>
<td>5.0E4</td>
<td>9.6E0</td>
<td>3.2E1</td>
</tr>
<tr>
<td>TREE</td>
<td>1.6E1</td>
<td>1.5E4</td>
<td>3.9E0</td>
<td>6.4E0</td>
</tr>
<tr>
<td>IH</td>
<td>7.8E0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard</td>
</tr>
<tr>
<td>IS</td>
<td>2.6E3</td>
</tr>
<tr>
<td>WIS</td>
<td>3.6E1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.8E1</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>IPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard</td>
</tr>
<tr>
<td>IS</td>
<td>3.7E5</td>
</tr>
<tr>
<td>WIS</td>
<td>2.2E1</td>
</tr>
<tr>
<td>NAIVE</td>
<td>3.5E1</td>
</tr>
</tbody>
</table>
F.8 Detailed Results for Enduro

Table 501: Enduro, relative MSE. Model Type: conv. $T = 500, N = 512, \pi_b = 0.10$-Greedy(DDQN), $\pi_e = 0.00$-Greedy(DDQN).

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Q-REG</td>
<td>8.9E-1</td>
<td>8.3E-1</td>
<td>9.8E-1</td>
<td>2.9E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>9.3E-1</td>
<td>1.5E0</td>
<td>1.5E0</td>
<td>3.1E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>3.2E-1</td>
<td>8.5E-2</td>
<td>1.4E-1</td>
<td>3.8E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Q($\pi$($\lambda$))</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TREE</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IH</td>
<td>1.0E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 502: Enduro, relative MSE. Model Type: conv. $T = 500, N = 512, \pi_b = 0.25$-Greedy(DDQN), $\pi_e = 0.00$-Greedy(DDQN).

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Q-REG</td>
<td>1.0E0</td>
<td>1.1E0</td>
<td>4.7E0</td>
<td>6.6E0</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.0E0</td>
<td>1.0E0</td>
<td>1.3E-1</td>
<td>1.5E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>7.1E-1</td>
<td>7.1E-1</td>
<td>7.4E-2</td>
<td>8.9E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Q($\pi$($\lambda$))</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TREE</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IH</td>
<td>1.5E-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 503: Enduro, relative MSE. Model Type: conv. $T = 500, N = 512, \pi_b = 0.25$-Greedy(DDQN), $\pi_e = 0.10$-Greedy(DDQN).

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>DR</th>
<th>WDR</th>
<th>MAGIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Q-REG</td>
<td>9.0E-1</td>
<td>6.8E-1</td>
<td>7.5E-1</td>
<td>5.1E-1</td>
</tr>
<tr>
<td>MRDR</td>
<td>1.0E0</td>
<td>4.1E0</td>
<td>2.2E1</td>
<td>3.4E-1</td>
</tr>
<tr>
<td>FQE</td>
<td>6.5E-1</td>
<td>3.5E-1</td>
<td>8.6E-2</td>
<td>4.5E-2</td>
</tr>
<tr>
<td>R($\lambda$)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Q($\pi$($\lambda$))</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TREE</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IH</td>
<td>9.5E-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 504: Enduro, relative MSE. Model Type: conv. $T = 500, N = 512, \pi_b = 0.25$-Greedy(DDQN), $\pi_e = 0.00$-Greedy(DDQN).

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E0</td>
<td>8.8E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>1.5E-1</td>
<td>8.9E-2</td>
</tr>
<tr>
<td>NAIVE</td>
<td>9.9E-2</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 505: Enduro, relative MSE. Model Type: conv. $T = 500, N = 512, \pi_b = 0.10$-Greedy(DDQN), $\pi_e = 0.00$-Greedy(DDQN).

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Per-Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>1.0E0</td>
<td>8.5E-1</td>
</tr>
<tr>
<td>WIS</td>
<td>8.9E-2</td>
<td>8.2E-2</td>
</tr>
<tr>
<td>NAIVE</td>
<td>1.1E-2</td>
<td>-</td>
</tr>
</tbody>
</table>