Interleaved Q-Learning with Partially Coupled Training Process*

Main Track

Min He
University of Electronic Science and
Technology of China
Chengdu, China
uestchemin@gmail.com

Hongliang Guo
University of Electronic Science and
Technology of China
Chengdu, China
guohl1983@uestc.edu.cn

ABSTRACT

This paper studies estimating the *maximum expected value* (MEV) of several independent random variables (RVs). No unbiased estimator exists without knowing the distributions of those RVs a priori. Two of the most famous estimators, maximum estimator (ME) and double estimator (DE), yield positive bias and negative bias respectively. We propose a coupled estimator (CE) which subsumes ME and DE as special cases and yields a bias between that of ME and DE, while maintaining the same variance bound. Furthermore, a simple yet effective variance reduction technique is proposed and verified in the experiments. The instantiated algorithm in the Markov decision process (MDP) setting, called interleaved Q-learning, outperforms Q-learning and double Q-learning in some highly stochastic environments. Insights on how to adapt the coupling ratio in CE and hence make interleaved Q-learning automatically shift between Q-learning and double Q-learning are provided and verified in the experimental section.

KEYWORDS

Maximum Estimator; Double Estimator; Coupled Estimator; Estimation Bias; Q-Learning; Double Q-Learning; Interleaved Q-learning

ACM Reference Format:
Min He and Hongliang Guo. 2019. Interleaved Q-Learning with Partially Coupled Training Process. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 9 pages.

1 INTRODUCTION

Many stochastic sequential decision making problems require estimating the maximum expected value of several RVs[1], given samples collected from each variable [22]. Two of the most famous estimators are ME and DE respectively, and both of them have been applied to various problem settings. For instance, in reinforcement learning (RL), Q-learning employs ME to estimate the optimal value of an action in a state; while double Q-learning uses DE for action-value estimation [8]. It is widely accepted that in some highly stochastic environments, Q-learning may perform poorly largely due to the overestimation of action values from ME. On the other hand, double Q-learning steadily underestimates the action values because of using DE, and performs better than Q-learning in some environmental settings [22].

This paper takes a fresh perspective to investigate the overestimation/underestimation problem of ME/DE, and thereby delivers an in-depth insight on which type of stochastic scenarios favours ME and which type favours DE respectively. Inspired from the insight, we propose a unifying estimator, namely *coupled estimator* (CE) which takes advantage of both ME and DE, and subsumes them as special cases for MEV estimation. Instead of having a single set for MEV estimation in ME or two completely disjoint sets in DE, CE partitions the samples of RVs into two *partially coupled* sets (Set A and Set B), in which the DE logic (using one set to evaluate another one) is used to estimate MEV.

We further propose a coupling ratio adaptation algorithm, which makes CE's estimation bias find a balance between ME's overestimation and DE's underestimation as verified through simulation results. Moreover, we further show that through interleaving the two evaluation terms from $\hat{\mu}_j(S_1)$ and $\hat{\mu}_j(S_2)$, CE's estimation variance is greatly decreased (up to a factor of 1/2) as validated in the experiments.

The CE-instantiated RL algorithm, namely *interleaved Q-learning*, absorbs the merits of both Q-learning and double Q-learning, and subsumes them as special cases in two extreme conditions as what CE does for ME and DE. Interleaved Q-learning embraces (1) a partially coupled training process to mimic the partially overlapped set partition process in CE, and (2) an interleaved evaluation process similar to CE's variance reduction technique.

The original contributions of the paper include: (1) a unifying estimator (CE) which subsumes ME and DE as special cases and a self-adaptive strategy for the coupling ratio adjustment; (2) a simple yet effective interleaving approach decreasing the estimator's variance up to a factor of 1/2; (3) a CE-instantiated off-policy reinforcement learning algorithm, namely interleaved Q-learning, which steadily outperforms Q-learning and double Q-learning in various stochastic environment and also subsumes Q-learning and double Q-learning as special cases.

2 MAXIMUM EXPECTED VALUE ESTIMATION

In this section, we first present the problem formulation of MEV estimation for several independent RVs, i.e. $\max_i \mathbb{E}(X_i)$, then introduce two of the most famous estimators (ME and DE) in the literature and end the section with a brief literature review on miscellaneous MEV estimators.

*Titlenote — equal contribution by the first two authors.

[1] Without loss of generality, we assume that the target is to find the *maximum* expected value over several RVs, the *negative* case applies in the same way.

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

2.1 Notations and Problem Formulation

The problem is to estimate the maximum expected value over a finite set of $N \geq 2$ random variables $X = \{X_1, X_2, \cdots, X_N\}$. The probability density function (PDF) and cumulative density function (CDF) of X_i are denoted as $f_i(x)$ and $F_i(x)$, respectively, with μ_i and σ_i^2 denoting the mean and variance of X_i. The *true* maximum expected value $\mu_*(X)$ is defined as:

$$\mu_*(X) = \max_i \mu_i = \max_i \int_{-\infty}^{+\infty} x f_i(x)\, dx. \tag{1}$$

In most application scenarios, the PDFs (f_i) are unknown beforehand and therefore $\mu_* = \max_i \mu_i$ cannot be found analytically. In this paper, we use $S = \{S_1, S_2, \cdots, S_N\}$ to denote the samples of X, and denote $\hat{\mu}_i(S)$ as an estimator for μ_i based on the sampled set S. Similarly, $\hat{\mu}_*(S)$ is used as an estimator for μ_*[2].

We define $A_j(S)$ as the event that the RV X_j's sample average, out of the given sample set S, is the largest out of all the sample averages. In this case, $\mathrm{P}(A_j)$ refers to the probability that event A_j happens over the sampling space Ω. The term $\mathrm{P}(A_j)$ is of great importance when analysing the overestimation/underestimation of ME/DE in later subsections.

Definition 2.1. The bias of an estimator ($\hat{\mu}(S)$) to the true value (μ_*) is defined as $\mathrm{Bias}(\hat{\mu}) = \mathbb{E}(\hat{\mu}) - \mu_*$.

It has been proved in [22] that no general unbiased estimator exists even if all $\hat{\mu}_i$ are unbiased. Two of the most famous estimators, ME and DE, are described and analysed in the following two subsections, respectively.

2.2 The Maximum Estimator

The maximum estimator for $\mu_*(S)$ is

$$\hat{\mu}_*^{\mathrm{ME}}(S) \equiv \max_i \hat{\mu}_i(S), \tag{2}$$

where $\hat{\mu}_i(S)$ is an estimator of μ_i, which can be unbiased. ME is widely used in practice because it is conceptually simple and easy to implement. There have been many works proving ME's overestimation of true MEV, i.e. [17] and [22]. However, most of the existing proofs rely on some logical explanation to justify ME's overestimation, this paper presents a concise mathematical derivation process for proving ME's overestimation and the derivation process inspires the bias reduction technique used in CE.

Before laying down the overestimation fact of ME and its corresponding proof, we first present the following two lemmas.

Lemma 2.2. *Let φ be a convex function, and X be an integrable real-valued random variable, we have:*

$$\varphi\big(\mathbb{E}(X)\big) \leq \mathbb{E}\left(\varphi(X)\right) \tag{3}$$

Lemma 2.3. *The* max *operator is a convex function.*

Both of the two lemmas are well known results. Lemma 2.2 is a direct application of Jensen's inequality, and the proof can be found in [14]. The proof of Lemma 2.3 is in [4].

Theorem 2.4. *For any given sampled set S out of the sampling space Ω and unbiased estimators $\hat{\mu}_i(S)$, i.e., $\mathbb{E}\left(\hat{\mu}_i(S)\right) = \mu_i$, we have $\mathbb{E}\left(\hat{\mu}_*^{ME}(S)\right) \geq \mu_*$.*

Proof.

$$\begin{aligned}
\mathbb{E}\left(\hat{\mu}_*^{\mathrm{ME}}(S)\right) &= \mathbb{E}\left(\max_i \hat{\mu}_i(S)\right) \\
&\geq \max_i \mathbb{E}\left(\hat{\mu}_i(S)\right) \\
&= \max_i \mu_i \\
&= \mu_*
\end{aligned}$$

The inequality is from Jensen's inequality in Lemma 2.2 and the fact that the max operator is a convex function as stated in Remark 2.3. We have assumed that $\hat{\mu}_i(S)$ is an unbiased estimator of μ_i, thus $\mathbb{E}\left(\hat{\mu}_i(S)\right) = \mu_i$. □

Insights on ME's bias: from Theorem 2.4's proving process, we can see that the inequality comes into play when we are exchanging the max operator with the expectation operator. It means that, if the arg max operator returns a constant RV index recommendation which makes max operate on a single fixed RV, the inequality will become equality, and ME will be unbiased. In MEV use cases, it means that out of the set of RVs, there is one RV dominating others (meaning that the RV's expected value is much larger than the rest of the RVs), ME tends to be unbiased.

Theorem 2.4 states that ME is always having a positive bias and thus provides a lower bound of ME as zero. The following two theorems provide upper bounds of ME's bias and variance, respectively.

Theorem 2.5. *The upper bound of ME's bias is expressed as:*

$$\mathrm{Bias}\big(\hat{\mu}_*^{ME}(S)\big) = \mathbb{E}\left(\hat{\mu}_*^{ME}(S)\right) - \mu_* \leq \sqrt{\frac{N-1}{N}\sum_{i=1}^{N}\mathrm{Var}(\hat{\mu}_i)}. \tag{4}$$

Theorem 2.6. *The upper bound of ME's variance is provided in Eq. 5.*

$$\mathrm{Var}\big(\hat{\mu}_*^{ME}(S)\big) \leq \sum_{i=1}^{N}\mathrm{Var}(\hat{\mu}_i) \tag{5}$$

The proofs of Theorem 2.5 and Theorem 2.6 are provided in [2] and [22] respectively.

2.3 The Double Estimator

The double estimator $\hat{\mu}_*^{\mathrm{DE}}(S)$ first partitions the sampling set S into two disjoint sets, namely S_1 and S_2. DE uses one of the sets, say Set S_1 to determine which RV to choose, and then uses Set S_2 to determine the value of the estimation. The calculation process is:

$$\hat{\mu}_*^{\mathrm{DE}}(S) = \hat{\mu}_j(S_2), \tag{6}$$

where $j = \arg\max_j \hat{\mu}_j(S_1)$. Note that the RV index selection process and the value estimation process are *separated* and functioning over two disjoint and independent sets of samples. In this way, the overestimation bias is eliminated, however, the underestimation bias is introduced as stated in the following theorem.

Theorem 2.7. *For any given sampled set S out of the sampling space Ω and unbiased estimators $\hat{\mu}_i(S)$, we have $\mathbb{E}\left(\hat{\mu}_*^{DE}(S)\right) \leq \mu_*$.*

Proof.

$$\mathbb{E}\left(\hat{\mu}_*^{\mathrm{DE}}(S)\right) = \mathbb{E}\left(\hat{\mu}_j(S_2)\right)$$

[2] Note that we write $\hat{\mu}_i$ and $\hat{\mu}_*$ when X and/or S is clear within the context.

$$
\begin{aligned}
&= \mathbb{E}\left(\sum_j \mathrm{P}(A_j)\hat{\mu}_j(S_2)\right)\\
&= \sum_j \mathrm{P}(A_j)\,\mathbb{E}\left(\hat{\mu}_j(S_2)\right)\\
&= \sum_j \mathrm{P}(A_j)\mu_j\\
&\le \sum_j \mathrm{P}(A_j)\max_j \mu_j\\
&= \mu_*.
\end{aligned}
$$

The key derivation point is that index j is also a random variable, and when calculating the expected value of DE, we need to represent the probability of selecting each index explicitly. In our deduction process, the probability of selecting index j as the largest RV's index is equal to the probability that event A_j happens out of the sampling space Ω, which is denoted as $\mathrm{P}(A_j)$[3]. □

Insights on DE's bias: from Theorem 2.7's proving process, we can see that the inequality comes into play due to the fact that $\mu_j \le \mu_*$. It means that, if the arg max operator only returns the set of RVs with the expectation value of μ_*, the inequality become equality and DE is unbiased. In real MEV use cases, it means that when there is a subset of RVs having the same (or similar) expectation values, and all the other RVs' expectation values are dominated by the subset, DE tends to be unbiased.

The following two theorems give the lower bound (upper bound is zero) and variance of DE respectively and the corresponding proofs are provided in [22].

Theorem 2.8. *The lower bound of DE's bias is expressed as:*

$$\mathrm{Bias}\left(\hat{\mu}_*^{DE}(S)\right) = \mathbb{E}\left(\hat{\mu}_*^{DE}(S)\right) - \mu_* \ge -\frac{1}{2}\sqrt{\sum_{i=1}^{N}\mathrm{Var}(\hat{\mu}_i(S_2))}. \quad (7)$$

Theorem 2.9. *The upper bound of DE's variance is*

$$\mathrm{Var}\left(\hat{\mu}_*^{DE}(S)\right) \le \sum_{i=1}^{N}\mathrm{Var}(\hat{\mu}_i(S_2)). \quad (8)$$

2.4 Miscellaneous MEV Estimators

Several other estimators targeting at alleviating the overestimation of ME and/or underestimation of DE exist. Zhang et.al. propose a weighted double estimator (WDE) for MEV estimation, and the weight is decided by the difference between the largest estimated value over the set of RVs and the smallest one [26]. A Gaussian estimator (GE) is proposed by [6], and the authors assume a Gaussian distribution of the sample averages, with this assumption, GE is able to reach an estimation with the bias bound smaller than that of both DE and ME. The extension of GE to an *infinite* number of RVs is reported in [5].

3 COUPLED ESTIMATOR

In the previous section, we have introduced ME and DE respectively. On the one hand, ME does not partition the sample set, and uses the same one (S) for both index selection and value estimation; in this case, the strong correlation (between index selection and value estimation) makes ME overestimate the MEV. On the other hand, DE partitions the sample set S into two completely *disjoint* sets, S_1 and S_2; in this case, S_1 and S_2 are independent with each other, and as a consequence, DE underestimates the MEV. We conjecture that if we partition the sample set S into two *partially coupled* sets, say S_1 and S_2, and tune the coupling ratio from 0 to 1, the corresponding estimator, namely coupled estimator (CE), will shift continuously from underestimation of DE to overestimation of ME. In the following subsections, we will formally present the coupled estimator, analyse its bias bound, and corresponding variance bound.

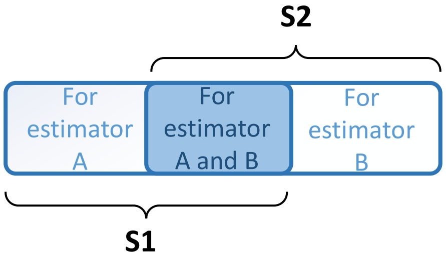


Figure 1: Illustration of the partially coupled partition(estimators A and B can be any kind of estimators)

3.1 The Coupled Estimator

Before introducing the coupled estimator, we lay down the definition of coupling ratio and illustrate its influence in the sampling set partition process.

Definition 3.1. The coupling ratio $\eta \in [0, 1]$ of a sampling set S is defined as the percentage of the shared data between the two partitioned sets, S_1 and S_2, over the total data set.

Fig. 1 illustrates the meaning of coupling ratio of the sampling set. Consider the sampling set S with only one variable for ease of understanding, and there are N samples. After the partially coupled/overlapped partition into S_1 and S_2, with N_1 and N_2 samples for S_1 and S_2 respectively. The coupling ratio η is calculated as $\eta = (N_1 + N_2 - N)/N$. The coupled estimator $\hat{\mu}_*^{\mathrm{CE}}(S)$ first partitions S into two partially coupled sets, namely $S1$ and $S2$. CE uses one of the sets, say Set S_1 to determine which index to choose, and then uses the other set to determine the value of the estimation. The calculation process is depicted as:

$$\hat{\mu}_*^{\mathrm{CE}}(S) = \hat{\mu}_j(S_1), \quad (9)$$

where $j = \arg\max_j \hat{\mu}_j(S_2)$. Note that CE's calculation process is similar to that of DE; the key difference lies in the partition process. In CE, the partition process is partially coupled by the coupling ratio (η), which is essential for the estimator's bias reduction.

3.2 CE's Bias, Bound and Variance

Theorem 3.2. *For any given sampling set S out of the sampling space Ω and unbiased estimators $\hat{\mu}_i(S)$, i.e., $\mathbb{E}\left(\hat{\mu}_i(S)\right) = \mu_i$, we have $\mathbb{E}\left(\hat{\mu}_*^{DE}(S)\right) \le \mathbb{E}\left(\hat{\mu}_*^{CE}(S)\right) \le \mathbb{E}\left(\hat{\mu}_*^{ME}(S)\right)$.*

Proof.

$$\mathbb{E}\left(\hat{\mu}_*^{\mathrm{CE}}(S)\right) = \mathbb{E}\left(\hat{\mu}_j(S_2)\right)$$

[3] Note that $\sum_j \mathrm{P}(A_j) = 1$

$$
\begin{aligned}
&= \mathbb{E}\Big(\sum_j \mathrm{P}(A_j)\hat{\mu}_j(S_2)\Big)\\
&= \sum_j \mathrm{P}(A_j)\,\mathbb{E}\left(\hat{\mu}_j(S_2)\right)\\
&= \sum_j \mathrm{P}(A_j)\,\mathbb{E}\left(\theta\hat{\mu}_j(S_1 \cap S_2) + (1-\theta)\hat{\mu}_j(\bar{S}_1 \cap S_2)\right)\\
&= \sum_j \mathrm{P}(A_j)\Big(\theta\,\mathbb{E}\left(\hat{\mu}_j(S_1 \cap S_2)\right) + (1-\theta)\,\mathbb{E}\left(\hat{\mu}_j(\bar{S}_1 \cap S_2)\right)\Big)\\
&= \sum_j \mathrm{P}(A_j)\Big(\theta\,\mathbb{E}\big(\max_j \hat{\mu}_j\big) + (1-\theta)\,\mathbb{E}\left(\hat{\mu}_j\right)\Big).
\end{aligned}
$$

The derivation process till now is straightforward; the expectation of the RV $\hat{\mu}_j(S_2)$ is a convex combination of two RVs, namely $\hat{\mu}_j(S_1 \cap S_2)$, and $\hat{\mu}_j(\bar{S}_1 \cap S_2)$[4]. Continuing with the derivation process, we will obtain the upper bound and lower bound of $\mathrm{E}\big(\hat{\mu}_j(S)\big)$. For upper bound:

$$
\begin{aligned}
\mathbb{E}\left(\hat{\mu}_j(S)\right) &= \sum_j \mathrm{P}(A_j)\Big(\theta\,\mathbb{E}\big(\max_j \hat{\mu}_j\big) + (1-\theta)\,\mathbb{E}\left(\hat{\mu}_j\right)\Big)\\
&\leq \sum_j \mathrm{P}(A_j)\Big(\theta\,\mathbb{E}\big(\max_j \hat{\mu}_j\big) + (1-\theta)\,\mathbb{E}\big(\max_j \hat{\mu}_j\big)\Big)\\
&= \sum_j \mathrm{P}(A_j)\,\mathbb{E}\big(\max_j \hat{\mu}_j\big)\\
&= \mathbb{E}\big(\max_j \hat{\mu}_j\big)\\
&= \mathbb{E}\left(\hat{\mu}_*^{\mathrm{ME}}(S)\right).
\end{aligned}
$$

For lower bound:

$$
\begin{aligned}
\mathbb{E}\left(\hat{\mu}_j(S)\right) &= \sum_j \mathrm{P}(A_j)\Big(\theta\,\mathbb{E}\big(\max_j \hat{\mu}_j\big) + (1-\theta)\,\mathbb{E}\left(\hat{\mu}_j\right)\Big)\\
&\geq \sum_j \mathrm{P}(A_j)\Big(\theta\,\mathbb{E}\left(\hat{\mu}_j\right) + (1-\theta)\,\mathbb{E}\left(\hat{\mu}_j\right)\Big)\\
&= \sum_j \mathrm{P}(A_j)\,\mathbb{E}\left(\hat{\mu}_j\right)\\
&= \sum_j \mathrm{P}(A_j)\mu_j\\
&= \mathbb{E}\left(\hat{\mu}_*^{\mathrm{DE}}(S)\right).
\end{aligned}
$$

□

Theorem 3.3. *For a large enough given sampled set S out of the sampling space Ω and unbiased estimators $\hat{\mu}_i(S)$, i.e., $\mathbb{E}\left(\hat{\mu}_i(S)\right) = \mu_i$, there exists a coupling ratio $\eta \in [0,1]$, which makes CE unbiased, i.e., $\mathbb{E}\left(\hat{\mu}_*^{CE}(S)\right) = \mu_*$.*

Theorem 3.4. *The lower bound and upper bound of CE's bias is expressed as:*

$$
-\frac{1}{2}\sqrt{\sum_{i=1}^{N}\mathrm{Var}(\hat{\mu}_i(S_2))} \leq \mathrm{Bias}\big(\hat{\mu}_*^{CE}(S)\big) \leq \sqrt{\frac{N-1}{N}\sum_{i=1}^{N}\mathrm{Var}(\hat{\mu}_i)}. \tag{10}
$$

The proof for Theorem 3.3 is a direct application of the mean value theorem introduced in [9]. The proof for Theorem 3.4 is trivial given Theorem 3.2, and is omitted here. CE's variance bound is stated in Theorem 3.5.

Theorem 3.5. *The upper bound of CE's variance is expressed as:*

$$
\mathrm{Var}\big(\hat{\mu}_*^{CE}(S)\big) \leq \sum_{i=1}^{N}\mathrm{Var}\big(\hat{\mu}_i(S_2)\big). \tag{11}
$$

The proof process is trivial and omitted here due to page limitation.

3.3 Coupling Ratio Adaptation

Theorem 3.3 indicates that, for any MEV estimation problem, there must exist an optimal coupling ratio $\eta^* \in [0,1]$, which makes CE unbiased. However, there is no analytical expression for the optimal coupling ratio. In this paper, we give out an intuitive/heuristic expression for setting η. The calculation process starts with setting $\eta = 0$, in which case S_1 and S_2 become two completely disjoint sets, and then follows the MEV index selection procedure:

$$
j_1 = \arg\max_j \hat{\mu}_j(S_1), \tag{12}
$$

$$
j_2 = \arg\max_j \hat{\mu}_j(S_2), \tag{13}
$$

and then, calculate η as:

$$
\eta = \frac{1 - \exp\Big(-\sigma^2\big(\hat{\mu}_{j_1}(S_1) - \hat{\mu}_{j_2}(S_1)\big)\Big)}{1 + \exp\Big(-\sigma^2\big(\hat{\mu}_{j_1}(S_1) - \hat{\mu}_{j_2}(S_1)\big)\Big)}, \tag{14}
$$

where $\sigma^2 > 0$ is the meta parameter. The underlying rationale of Eq. 14 is as follows: (1) the distribution disparity of those RVs influences the performance of DE and ME greatly (large distribution disparity favours ME and small disparity favours DE) [22]; (2) but it is impossible to calculate exactly the distribution disparity between RVs without knowing their underlying distributions; (3) $\max_i\left(\hat{\mu}_i(S)\right) - \min_i\left(\hat{\mu}_i(S)\right)$ is normally used as a rough approximation to the distribution disparity, i.e. [26], however, we believe that the approximation is too conservative in that the minimum expected value of a set of RVs is seldom selected for *maximum* expected value estimation; (4) we propose to use the difference between two selected RVs (out of two sampling sets) as the distribution disparity approximator, which is $\hat{\mu}_{j_1}(S_1) - \hat{\mu}_{j_2}(S_1)$, and perform a non-linear mapping to map the difference to the range of [0,1] so as to set the value of η.

3.4 Interleaved CE for Variance Reduction

CE unifies two seemingly disparate MEV estimators (ME and DE), and subsumes them as special cases in two extreme conditions. The bias of CE is between that of ME who overestimates MEV and that of DE who underestimates MEV. However, we cannot evaluate the goodness of an estimator only based on its bias, the variance of CE is also important. Theorem 3.5 provides CE's variance bound, which is the same as that of ME and DE. In this subsection, we propose a simple yet effective approach to reduce CE's variance, namely interleaved CE.

In interleaved CE, the two (partial coupled) sampling sets (S_1 and S_2) are treated equally important when they are used in either MEV

[4] Note that the combination coefficient θ has a non-linear relationship with the coupling ratio η, which cannot be analytically derived.

index selection or value estimation, and the final estimated value is calculated as the average of the estimation from both sampling sets. Specifically, we use Set S_1 to select the MEV index, and use Set S_2 to estimate the value, and we perform the procedure with swapped order (S_2 for index selection and S_1 for value estimation), then the final estimation is the average of the two estimated value. The calculation process for interleaved CE is:

$$\hat{\mu}_*^{\text{CE}}(S) = \frac{1}{2}\big(\hat{\mu}_{j_1}(S_2) + \hat{\mu}_{j_2}(S_1)\big), \tag{15}$$

where $j_1 = \arg\max_j \hat{\mu}_j(S_1)$ and $j_2 = \arg\max_j \hat{\mu}_j(S_2)$.

In this way,the variance of the interleaved CE in Eq. 15 is no lager than that of normal CE in Eq. 9. The equality is strict when $\eta = 1$, and the smaller the η is, the smaller the variance of the interleaved CE than that of normal CE.(roughly by a factor of $1/2$ in the expectation when η is small enough). The proof process is straightforward and hence omitted here.

4 INTERLEAVED Q-LEARNING WITH PARTIALLY COUPLED TRAINING PROCESS

In this section, we introduce one of CE's most popular instantiated algorithms in Markov decision process (MDP) settings, namely interleaved Q-learning with partially coupled training process[5]. In this section, we will lay down the backgrounds of interleaved Q-learning, followed by the algorithm introduction with the pseudo code depicting its flow process, inter-Q-learning algorithm's convergence proof, and the variance reduction technique.

4.1 Backgrounds

A Markov decision process (MDP) consists of a tuple of five elements $(S, A, P_{sa}^{s'}, \gamma, R)$, where S is a set of state space, A is the action and $P_{sa}^{s'}$ is the state transition probabilities, R is the mean expected reward after executing action a in state s and $\gamma \in (0, 1)$ is the discount factor [15]. The reinforcement learning problem is to find a policy maximizing the expected discounted cumulative reward, i.e. $\mathbb{E}(\sum_{t=0}^{\infty} \gamma^t R_{t+1})$ [19]. For any policy π, the state value function is defined as:

$$V^{\pi}(s) = \mathbb{E}\Big(\sum_{t=0}^{\infty} \gamma^t R_{t+1}\Big), \tag{16}$$

An optimal policy π^* maximizes the return for each state.

Q-learning is a popular reinforcement learning algorithm that was proposed by Watkins and can be used to optimally solve MDP problems [24, 25]. It is an instantiation of the maximum estimator to estimate the maximum expected values of subsequent state values. The update of Q-learning is:

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha\big(r_{t+1} + \gamma \max_a Q_t(s_{t+1}, a) - Q_t(s_t, a_t)\big), \tag{17}$$

where s_t is the current state, a_t is the action chosen at s_t, r_{t+1} is the reward and s_{t+1} is the next state. Many research works have proven the overestimation of the Q-learning algorithm in stochastic environment [8], and the original Q-learning algorithm inspires several improvements, such as delayed Q-learning [18], phased Q-learning [10], fitted Q-iteration [7], bias-corrected Q-learning [12] and deep Q-networks (DQN) [13]. It has been proven that Q-learning reaches the optimal value function Q^* with probability one in the limit under some mild conditions on the learning rates and an exploration policy [21].

Double Q-learning [8] is an instantiation of DE for MEV estimation. It stores two independent state-action value tables/functions (Q^A and Q^B), and each action value is updated with a value from the other action value table/function for the next state. The original double Q-learning also inspires several improvements, such as deep double DQN [23], double delayed Q-learning [1] and weighted multi-agent deep double Q-learning [27].

4.2 Interleaved Q-learning with Partially Coupled Training Process

It is not straightforward to instantiate CE for MEV estimation in MDP settings, as one does not have the sampled data set at hand to perform the partially coupled partition. In the reinforcement learning context, data streams are provided on-line as the reinforcement learning agent interacts with the environment. Therefore, one needs an on-line data partition process to mimic the MEV estimation in CE. Before introducing the inter-Q-learning algorithm, we first define a crucial concept within the algorithm.

Definition 4.1. The coupled co-training rate(η) is defined as the probability that a new data tuple is used in the training process of both two action value estimators.

The coupled co-training rate (η) in inter-Q-learning, functions similarly to the overlapping ratio (η) in the coupled estimator. Inter-Q-learning initializes two independent action value estimators, namely Q^A and Q^B respectively, similar to what double Q-learning does. When a new data tuple $\langle s_t, a_t, r_{t+1}, s_{t+1}\rangle$ arrives, one can (a) update Q^A with Q^B's estimated value according to Eq. 18 with probability $(1-\eta)/2$, or (b) update Q^B with Q^A's estimated value according to Eq. 19 with probability $(1-\eta)/2$ or (c) perform both (a) and (b) with probability η.

$$\begin{aligned} Q^A(s_t, a_t) =& Q^A(s_t, a_t) + \alpha\Big(r_{t+1} \\ &+ \gamma(Q^B(s_{t+1}, \arg\max_a Q^A(s_{t+1}, a))) - Q^A(s_t, a_t)\Big), \end{aligned} \tag{18}$$

$$\begin{aligned} Q^B(s_t, a_t) =& Q^B(s_t, a_t) + \alpha\Big(r_{t+1} \\ &+ \gamma(Q^A(s_{t+1}, \arg\max_a Q^B(s_{t+1}, a))) - Q^B(s_t, a_t)\Big), \end{aligned} \tag{19}$$

Note that on one extreme condition, if the inter-Q-learning algorithm only perform (a) and (b) alternatively or probabilistically, it is equivalent to double Q-learning. On the other extreme condition, if the inter-Q-learning algorithm always performs (c), it is equivalent to Q-learning. Therefore, inter-Q-learning is a generalization of Q-learning and double Q-learning, and subsumes them as special cases. Performing (a),(b) and (c) in a probabilistic way enables a partial co-training process of Q^A and Q^B, therefore, inter-Q-learning mimics the partially coupled partition process in CE and can be deemed as a CE-instantiated reinforcement learning algorithm. Compared to double Q-learning, inter-Q-learning increases its estimate by

[5] In the following part of the paper, we name the algorithm as 'inter-Q-learning' or interleaved Q-learning for short.

performing the (partially) co-training process; on the other hand, inter-Q-learning decreases its estimate by performing independent updates in (a) or (b) when compared to Q-learning. We can instantiate interleaved CE in the RL context through 'interleaving' Q^A's evaluation over Q^B's action selection and vice versa, in this case, the update of Q^A and Q^B become:

$$Q^A(s_t,a_t) = Q^A(s_t,a_t) + \alpha\Big(r_{t+1} + \gamma\big(Q^B(s_{t+1}, \arg\max_a Q^A(s_{t+1},a)) + Q^A(s_{t+1}, \arg\max_a Q^B(s_{t+1},a))\big)/2 - Q^A(s_t,a_t)\Big), \tag{20}$$

$$Q^B(s_t,a_t) = Q^B(s_t,a_t) + \alpha\Big(r_{t+1} + \gamma\big(Q^A(s_{t+1}, \arg\max_a Q^B(s_{t+1},a)) + Q^B(s_{t+1}, \arg\max_a Q^A(s_{t+1},a))\big)/2 - Q^B(s_t,a_t)\Big). \tag{21}$$

The algorithm flow process of inter-Q-learning is depicted in Algorithm 1.

Algorithm 1 Interleaved Q-learning

```
Initialize $Q^A, Q^B, s, \alpha, \gamma, \eta$
Define $a_1^* = \arg\max_a Q^A(s', a)$
Define $a_2^* = \arg\max_a Q^B(s', a)$
repeat
    Choose $a$ for $s$ according to $\epsilon$-greedy policy based on the value of $(Q^A + Q^B)/2$
    Execute action $a$, obtain $r, s'$
    Sample $p$ from $(0, 0.5)$ according to uniform distribution
    if $p < (1-\eta)/2$ then
        Choose to update $Q^A$
        $\Delta Q^A(s,a) \leftarrow \alpha\Big(r+\gamma\frac{1}{2}\big(Q^B(s',a_1^*)+Q^A(s',a_2^*)\big)-Q^A(s,a)\Big)$
    else if $p > (1+\eta)/2$ then
        Choose to update $Q^B$
        $\Delta Q^B(s,a) \leftarrow \alpha\big(r+\gamma\frac{1}{2}\big(Q^A(s',a_2^*)+Q^B(s',a_1^*)\big)-Q^B(s,a)\big)$
    else
        Choose to update both $Q^A$ and $Q^B$
        $\Delta Q^A(s,a) \leftarrow \alpha\big(r+\gamma\frac{1}{2}\big(Q^B(s',a_1^*)+Q^A(s',a_2^*)\big)-Q^A(s,a)\big)$
        $\Delta Q^B(s,a) \leftarrow \alpha\big(r+\gamma\frac{1}{2}\big(Q^A(s',a_2^*)+Q^B(s',a_1^*)\big)-Q^B(s,a)\big)$
    end if
    $s \leftarrow s'$
until END
```

4.3 Convergence Proof

In this subsection, we show that inter-Q-learning converges asymptotically to the optimal action values. Before the theorem proving process, we first give out an intuitive explanation. As inter-Q-learning is a generalization of both Q-learning and double Q-learning and subsumes them as special cases, in the meanwhile, both Q-learning and double Q-learning converge in the limit, inter-Q-learning also converges in the limit to the optimal action values. Before posing the theorem and its proof, we first lay down the following lemma, whose proof is provided in [16].

LEMMA 4.2. *Consider a stochastic process $(\alpha_t, \Delta_t, F_t)$, $t \geq 0$, where $\alpha_t, \Delta_t, F_t : X \rightarrow \mathcal{R}$ satisfy the equations:*

$$\Delta_{t+1}(x) = \big(1-\alpha_t(x)\big)\Delta_t(x) + \alpha_t(x)F_t(x),\ x \in X,\ t = 0, 1, 2, \ldots \tag{22}$$

Let P_t be a sequence of increasing $\sigma-$fields such that α_0 and Δ_0 are P_0-measurable, and α_t, Δ_t and F_{t-1} are P_t-measurable, $t = 1, 2, \ldots$. Assume that the following conditions hold:(1) the set X is finite; (2) $0 \leq \alpha_t(x) \leq 1$, $\sum_t \alpha_t(x) = \infty$, $\sum_t \alpha_t^2(x) < \infty$ w.p.1; (3) $\|\mathbb{E}\big(F_t(\cdot)|P_t\big)\| \leq \kappa\|\Delta_t\| + c_t$, where $\kappa \in [0,1)$ and c_t converges to zero w.p.1;(4) $\mathrm{Var}\big(F_t(x)|P_t\big) \leq K\big(1+\|\Delta_t\|\big)^2$, where K is some constant. Then Δ_t converges to zero with probability one (w.p.1).

THEOREM 4.3. *Both Q^A and Q^B as updated by inter-Q-learning in Algorithm 1 will converge to the optimal value Q^* w.p.1 if an infinite number of experiences for each state action pair are presented to the learning algorithm. The additional conditions are: 1) The MDP is finite, i.e. $|S \times A| < \infty$; 2) $\gamma \in [0,1)$; 3) Q^A and Q^B are stored in lookup tables; 4) both Q^A and Q^B are updated an infinite number of times; 5) $\alpha_t(s,a) \in [0,1]$, $\sum_t \alpha_t(s,a) = \infty$, $\sum_t \big(\alpha_t(s,a)\big)^2 < \infty$ w.p.1; 6) $\mathrm{Var}(R(s,a)) < \infty$.*

In the proving process, we apply Lemma 4.2 with $P_t = \{Q_0^A, Q_0^B, s_0, a_0, \alpha_0, r_1, s_1, \ldots, s_t, a_t\}$, $X = S \times A$, $\Delta_t = Q_t^A - Q^*$,$\zeta = \alpha$ and $F_t(s_t,a_t) = r_t + \gamma Q_t^B(s_{t+1},a^*) - Q^*(s_t,a_t)$ to prove Theorem 4.3. Requirements (1),(2) and (4) in Lemma 4.2 are straightforward to verify, and omitted here.

PROOF.

$$\begin{aligned} F_t(s_t,a_t) &= r_t + \gamma Q_t^B(s_{t+1},a^*) - Q_t^*(s_t,a_t) \\ &= r_t + \gamma Q_t^A(s_{t+1},a^*) - Q_t^*(s_t,a_t) \\ &\quad + \gamma\big(Q_t^B(s_{t+1},a^*) - Q_t^A(s_{t+1},a^*)\big). \end{aligned}$$

It has been proved in [25] that $\mathbb{E}\left(r_t + \gamma Q_t^A(s_{t+1},a^*) - Q_t^*(s_t,a_t)\right) \leq \gamma\|\Delta_t\|$. Therefore, we need to verify that $c_t = \gamma\big(Q_t^B(s_{t+1},a^*) - Q_t^A(s_{t+1},a^*)\big)$ converges to zero w.p.1. Let $\Delta_t^{BA} = Q_t^B - Q_t^A$, it suffices to prove that Δ_t^{BA} converges to zero. Defining the following two terms,

$$F_t^B(s_t,a_t) \equiv \big(r_t + \gamma Q_t^A(s_{t+1},b^*) - Q_t^B(s_t,a_t)\big), \tag{23}$$

$$F_t^A(s_t,a_t) \equiv \big(r_t + \gamma Q_t^B(s_{t+1},a^*) - Q_t^A(s_t,a_t)\big), \tag{24}$$

and depending on whether Q^B or Q^A or both Q^B and Q^A are updated, the update of Δ_t^{BA} at time $t+1$ is represented as:

$$\Delta_{t+1}^{BA} = \Delta_t^{BA} + \alpha F_t^B \quad w.p.(1-\eta)/2, \tag{25}$$

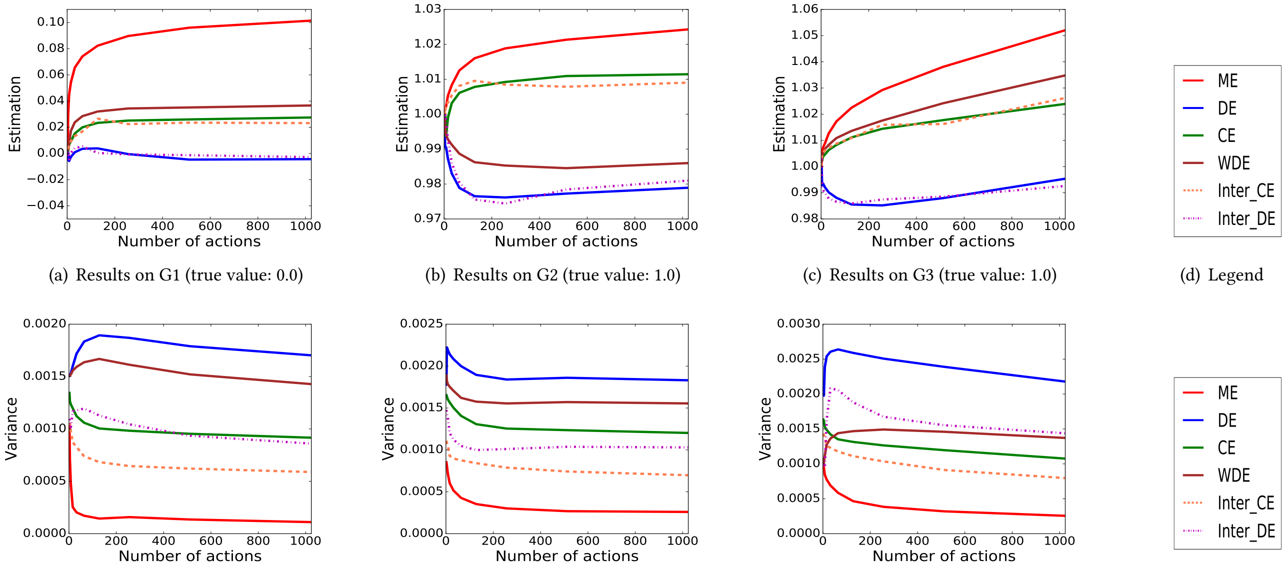
$$\Delta_{t+1}^{BA} = \Delta_t^{BA} - \alpha F_t^A \quad w.p.(1-\eta)/2, \tag{26}$$

$$\Delta_{t+1}^{BA} = \Delta_t^{BA} + \alpha(F_t^B - F_t^A) \qquad w.p.\eta. \tag{27}$$

Reapply Lemma 4.2 to analyse the stochastic process for Δ_t^{BA}, and perform corresponding expectation operations[6], we conclude that Δ_t^{BA} converges to zero w.p.1. With $\mathbb{E}\left(r_t + \gamma Q_t^A(s_{t+1},a^*) - Q_t^*(s_t,a_t)\right) \leq \gamma\|\Delta_t\|$ and $c_t = \gamma\big(Q_t^B(s_{t+1},a^*) - Q_t^A(s_{t+1},a^*)\big)$ converges to zero w.p.1, we conclude that condition (3) in Lemma 4.2 is satisfied, hence completes the proof for Theorem 4.3.

□

[6] Detailed derivation process proving Δ_t^{BA}'s convergence to zero is sketched in [8]



(a) Results on G1 (true value: 0.0)

(b) Results on G2 (true value: 1.0)

(c) Results on G3 (true value: 1.0)

(d) Legend

(e) Results on G1

(f) Results on G2

(g) Results on G3

(h) Legend

Figure 2: (a-d) Estimated values and true values and (e-h) Variance comparison for different MEV estimators on three groups of multi-arm bandit problems, G1, G2, and G3. All the figures are best viewed in color.

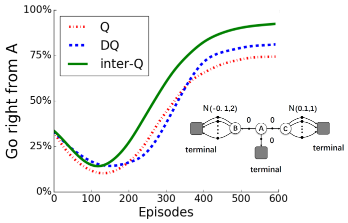


Figure 3: Comparison of Q-learning (Q), Double Q-learning (DQ) and interleaved Q-learning (inter-Q) on a simple episodic MDP (shown inset). The parameters are set as α = 0.1, ϵ = 0.1 and γ = 1.

5 AN ILLUSTRATIVE EXAMPLE

The small MDP (twisted from Fig. 6.7 in [20]) shown inset in Fig. 3 provides a simple example of how CE's low estimation bias benefits TD control algorithms compared to ME and DE. The MDP has three non-terminal states A, B and C. Episodes always start in A with a choice among three actions, left, right and down. The down action transitions immediately to the terminal state with a reward of zero. The right action transitions deterministically to C with a reward of zero, from which there are many possible actions all of which cause immediate termination with a reward drawn from a normal distribution with mean 0.1 and variance 1. Thus, the expected return for any trajectory starting with C is 0.1. The left action transitions deterministically to B with a reward of zero, from which there are many possible actions all of which cause immediate termination with a reward drawn from a normal distribution with mean -0.1 and variance 2. Thus, the expected return for any trajectory starting with B is -0.1. In this case, the optimal action from A is to go right as the expected return is 0.1 which is the largest. However, since ME overestimates the return and the larger the variance is, the larger the overestimation is, it has some probability of take left as the optimal action from A. Similarly, DE underestimates the return, and it has some probability of choosing the down action as the optimum. The simulation results are shown in Fig. 3.

6 SIMULATION RESULTS AND ANALYSIS

In this section, we mainly perform experiments over two types of scenarios, namely multi-armed bandits and grid world[7]. For multi-armed bandits, we compare the proposed MEV estimator (CE and interleaved CE) with state of the arts including ME [22], DE [22], WDE [26], and interleaved DE[8]. For grid world, we compare the instantiated interleaved Q-learning (inter-Q) with canonical Q-learning (Q) [25], double Q-learning (DQ) [8] and weighted double Q-learning (WDQ) [26].

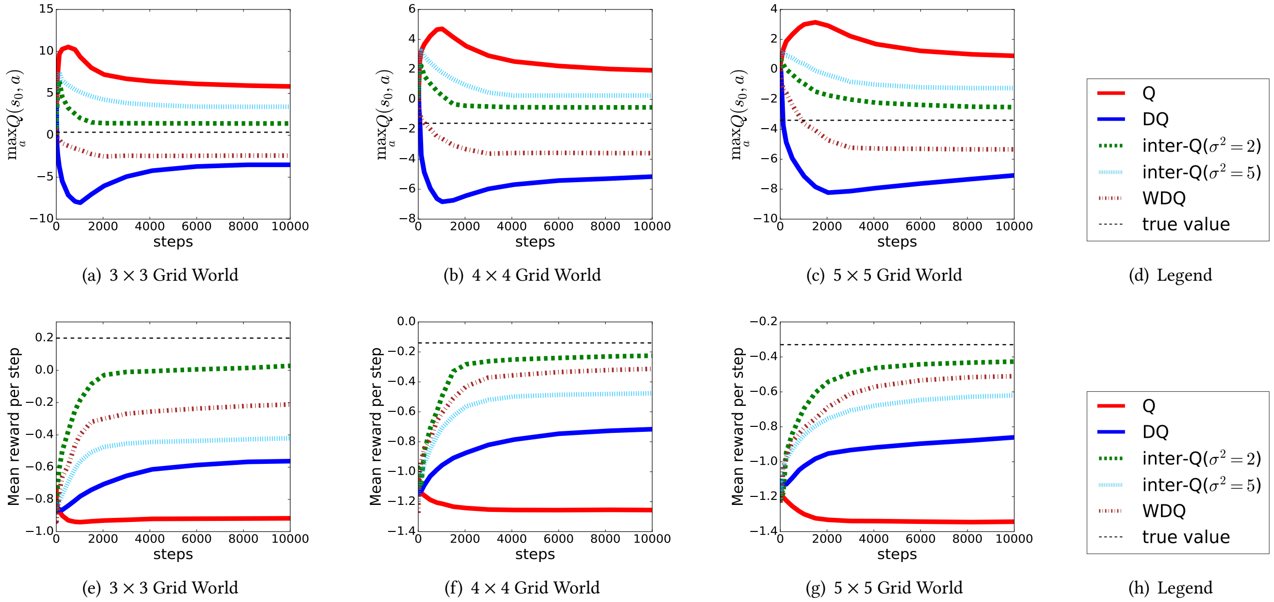
6.1 Multi-Armed Bandits

Multi-armed bandit problem is a classical scenario for MEV estimation [19]. Our experiments are conducted on three groups of multi-armed bandit problems[9]: (G1) $\mathbb{E}\{X_i\} = 0$, for $i \in \{1, 2, \cdots, N\}$; (G2) $\mathbb{E}\{X_1\} = 1$ and $\mathbb{E}\{X_i\} = 0$ for $i \in \{2, 3, \cdots, N\}$; and (G3) $\mathbb{E}\{X_i\} = i/N$, for $i \in \{1, 2, \cdots, N\}$. Here N refers to the number of actions in the multi-armed bandit context. In the scenario set up, $\max_i \mathbb{E}\{X_i\} = 0$ in G1, and $\max_i \mathbb{E}\{X_i\} = 1$ in G2 and G3. In

[7] It is worth noting that these two types of scenarios are also selected as benchmark scenarios for the evaluation of double Q-learning [8], and weighted double Q-learning [26]. Maintaining the same benchmark scenarios ensures clear algorithm comparison.

[8] Note that the interleaving procedure can also be applied to DE for variance reduction

[9] Note that the scenario setup is the same as what is described in Section 5.1 in [26].



(a) 3 × 3 Grid World (b) 4 × 4 Grid World (c) 5 × 5 Grid World (d) Legend

(e) 3 × 3 Grid World (f) 4 × 4 Grid World (g) 5 × 5 Grid World (h) Legend

Figure 4: The maximum action value in the initial state $s0$ and mean reward per step comparison on the $n \times n$ grid world problems, where n ranges from 3 to 5. Results are averaged over 1000 runs

the experiment, we run the simulation for 100 independent times, and each time, the sample values (x_i) are draw from $\mathcal{N}(\mathbb{E}\{X_i\}, 1)$, and the sample size is 1000. We follow [26] to set β (with $c = 1$) for WDE, and set σ^2's default value as 2.

Fig. 2(a-c) shows the empirical results for MEV estimation with different estimators. In the figure, we can see that in terms of estimation bias: (1) DE performs the best in G1, because this type of scenario favours DE 'exactly'; CE's performance is quite close to DE; ME and WDE overestimate the MEV; (2) CE's estimated value is the closest to the true value, which is 1.0, in G2; (3) in G3, CE performs the best. The results well justify the underlying rationale of when ME overestimates, DE underestimates and how to set the optimal η in CE, as described in Section 3.3. It is worth noting that the interleaved CE and interleaved DE yield similar estimations to those of CE and DE. These results show that the interleaving process does not change the bias of the corresponding estimator, instead, the variance is greatly decreased as shown in Fig. 2(e-g). In Fig. 2(e-g), we can see that the interleaved estimators (both interleaved CE and interleaved DE) yields much smaller variance than their 'non-interleaved' counterparts[10].

6.2 Grid World

We compare interleaved Q-learning, Q-learning, double Q-learning and weighted double Q-learning in the grid world scenario which scales form 3 × 3 to 5 × 5. In an $n \times n$ grid world, the starting state is in the bottom left position and goal state is in the top right. Each state has four actions: up,down,right,left to go to the adjacent state. If the agent chooses an action that walks off the grid, the agent stays in the same state. The agent receives a random reward of $\mathcal{N}(-1, 1)$ for actions to non-goal states, and receive a reward of $\mathcal{N}(5, 1)$ for actions to goal state. The optimal mean reward per step is $\frac{5-2(n-1)}{2(n-1)+1}$. With a discount factor $\gamma = 0.95$, the optimal value of maximally valued action in the starting state is $5\gamma^{2(n-1)} - \sum_{i=0}^{2n-3} \gamma^i$. Adopting the grid world scenario set up in [26], we set the reward of non-goal state action and goal state action subject to a normal distribution with a variance 1, which makes the situation more stochastic.

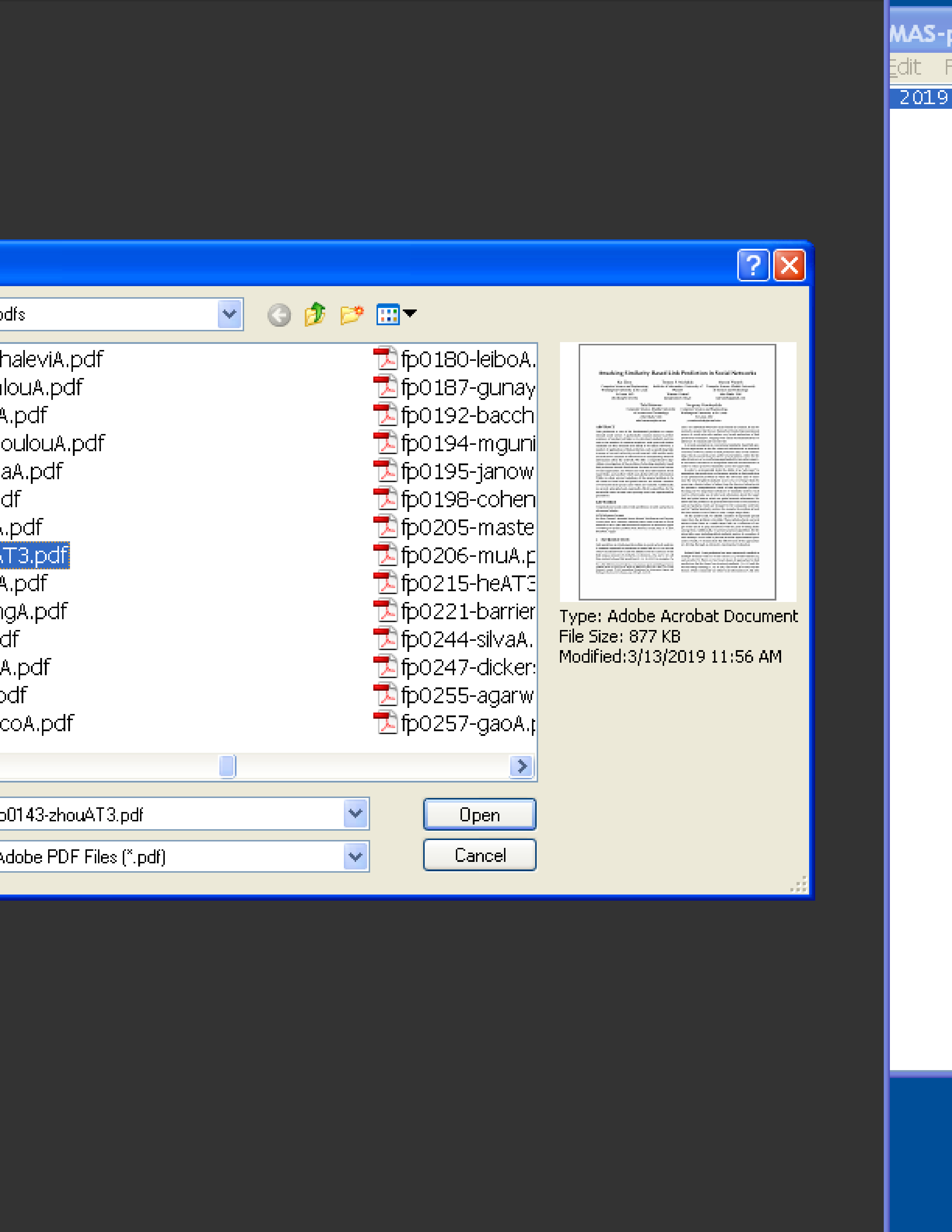
Fig. 4(a-c) and Fig. 4(e-g) show the performance comparison of inter-Q-learning with state of the arts. In terms of the optimal state value estimation of the starting state (s_0) and the average reward per action, we can see that the proposed inter-Q-learning algorithm (with $\sigma^2 = 2$) performs better than other state of the arts.

7 CONCLUSION AND FUTURE WORKS

This paper presents a coupled estimator for MEV estimation to alleviate the overestimation of ME as well as the underestimation of DE, and subsumes ME and DE as special cases. A simple yet effective interleaving approach is proposed to reduce CE's variance while maintaining the estimation bias. The instantiated RL algorithm, namely inter-Q-learning, in MDP settings inherits the merits of CE and performs better than state of the arts.

An interesting future work direction is to expand the inter-Q-learning to the continuous feature application domain such as Atari video games [3] with function approximation techniques such as deep convolution neural networks [11]. We are also keen on extending the inter-Q-learning algorithm to the continuous action RL problems as what the authors in [5] do for Gaussian estimator. In the meanwhile, extending interleaved Q-learning for multi-step TD learning algorithms is also a promising direction.

[10] Note that the variance of ME is the smallest for all the three settings, because it uses the total data set for estimation, and the variance is undoubtedly the smallest.



dfs

haleviA.pdf
louA.pdf
A.pdf
oulouA.pdf
aA.pdf
df
.pdf
T3.pdf
A.pdf
ngA.pdf
df
A.pdf
pdf
coA.pdf

fp0180-leiboA.
fp0187-gunay
fp0192-bacch
fp0194-mguni
fp0195-janow
fp0198-cohen
fp0205-maste
fp0206-muA.p
fp0215-heAT3
fp0221-barrier
fp0244-silvaA.
fp0247-dicker
fp0255-agarw
fp0257-gaoA.p

Type: Adobe Acrobat Document
File Size: 877 KB
Modified:3/13/2019 11:56 AM

o0143-zhouAT3.pdf

Adobe PDF Files (*.pdf)