Estimating the Maximum Expected Value through Gaussian Approximation

Carlo D'Eramo CARLO.DERAMO@POLIMI.IT
Alessandro Nuara ALESSANDRO.NUARA@MAIL.POLIMI.IT
Marcello Restelli MARCELLO.RESTELLI@POLIMI.IT
Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano

Abstract

This paper is about the estimation of the maximum expected value of a set of independent random variables. The performance of several learning algorithms (e.g., Q-learning) is affected by the accuracy of such estimation. Unfortunately, no unbiased estimator exists. The usual approach of taking the maximum of the sample means leads to large overestimates that may significantly harm the performance of the learning algorithm. Recent works have shown that the cross validation estimator—which is negatively biased—outperforms the maximum estimator in many sequential decision-making scenarios. On the other hand, the relative performance of the two estimators is highly problem-dependent. In this paper, we propose a new estimator for the maximum expected value, based on a weighted average of the sample means, where the weights are computed using Gaussian approximations for the distributions of the sample means. We compare the proposed estimator with the other state-of-the-art methods both theoretically, by deriving upper bounds to the bias and the variance of the estimator, and empirically, by testing the performance on different sequential learning problems.

1. Introduction

In many machine learning problems it is necessary to estimate the *maximum expected value* (MEV) of a set of random variables, given samples collected from each variable (van Hasselt, 2013). For instance, in reinforcement learning, the optimal policy can be found by taking, in each state, the action that attains the maximum expected cumulative reward. The optimal value of an action in a state, on its turn, depends on the maximum expected values of the actions available in the reached states. Since errors propagate through all the state-action pairs, a bad estimator for the maximum expected value negatively affects the speed of learning (van Hasselt, 2010). Another example of the importance of producing accurate estimates for the maximum expected value is provided by sponsored search auctions, where the search engine needs to select which ad to show from a pool of candidates. To select the best ad, usually, one needs to estimate the probability that a random user will click on it. Since, each time an ad is clicked, the search engine charges to the advertiser a fee that depends on the click probabilities of the top two ads, a good estimate of the maximum expected value is essential to maximize the revenue of the search engine (Xu et al., 2013).

The most common estimator is the *Maximum Estimator* (ME), which consists of taking the maximum of the sample means. It is well known (Smith & Winkler, 2006; Van den Steen, 2004; van Hasselt, 2010) that ME overestimates the maximum expected value. To avoid such positive bias, a common approach is the *Double Estimator* (DE), that consists in a *cross-validatory* approach where the sample set is split into two sample sets A and B (Stone, 1974). DE results from the average of two estimates. For the first estimate, the sample set A is used to determine which is the variable with the largest mean, while sample set B is used to estimate the value of the variable. The second estimate is obtained by switching the roles of A and B. Although the absolute bias of CV can be larger than the one of ME (van Hasselt, 2013), CV is negatively biased and this can be an advantage in some applications (van Hasselt, 2010; Xu et al., 2013; van Hasselt et al., 2015). Unfortunately, an unbiased estimator for the maximum expected value *does not exist* for many common distributions (e.g., Gaussian, Binomial, and Beta) (Blumenthal & Cohen, 1968; Dhariyal et al., 1985). On the other hand, having an unbiased estimator does not entail a small expected Mean Squared Error (MSE), since also the variance of the estimator has to be considered.

In this paper we propose to estimate the maximum expected value using the *Weighted Estimator* (WE), that con-

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

sists of a weighted average of the sample means, where the weights are obtained by estimating the probability that each variable has the largest expected value. To compute such probabilities we would need to know the distributions of the sample means. Relying on the central limit theorem, we approximate the distributions of the samples means with Gaussian distributions parameterized by the sample mean and sample variance. Such weighting mechanism reduces the bias w.r.t. ME without increasing the variance as DE does. The *original contributions* are: (1) the introduction of the WE estimator, (2) a theoretical analysis of the bias and the variance of the estimation error, (3) an extensive empirical analysis that compares ME, DE, WE when used for learning in sequential decision problems.

The rest of the paper is organized as follows. In the next section we introduce the basic notation and discuss the most related works. Section 3 contains the description of the proposed approach, whose theoretical properties are presented in Section 4. In Section 5, we show the empirical results where WE is compared with the state of the art. Section 6 draws conclusions and discusses future work.

2. Preliminaries

Problem definition We consider the problem of finding the maximum expected value of a finite set of $M \geq 2$ independent random variables $X = \{X_1, ..., X_M\}$. We denote with $f_i : \mathbb{R} \to \mathbb{R}$ the probability density function (PDF), with $F_i : \mathbb{R} \to \mathbb{R}$ the cumulative density function (CDF), with μ_i the mean, and with σ_i^2 the variance of variable X_i. The maximum expected value $\mu_*(X)$ is defined as

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

Assuming that the PDFs are unknown, $\mu_*(X)$ cannot be found analytically. Given a set of noisy samples $S = \{S_1, ..., S_N\}$ retrieved by the unknown distributions of each X_i, we are interested in finding an accurate estimator $\hat{\mu}_*(S) \approx \mu_*(X)$. These random samples have means $\hat{\mu}_1, ..., \hat{\mu}_N$ that are unbiased estimators of the true mean μ_i. The PDF and CDF of $\hat{\mu}_i(S)$ are denoted by $\hat{f}_i^S$ and $\hat{F}_i^S$.

Related works The maximum expected value can be approximated with the maximum of the sample means:

$$\hat{\mu}_*^{ME}(S) = \max_i \hat{\mu}_i(S) \approx \mu_*(X). \quad (2)$$

This method is called *Maximum Estimator* (ME) and it is used, for instance, by Q-Learning to approximate the value of the following state by maximizing over the estimated action values in that state. Unfortunately, as proved in (Smith & Winkler, 2006), this estimator is positively biased and this is critical in Q-Learning where the approximation error can increase step by step due to the overestimation of the state-action values. To understand the presence of this positive bias, consider the CDF $\hat{F}_{\max}(x)$ of the maximal estimator $\max_i \hat{\mu}_i$ that is the probability that ME is less than or equal to x. This probability is equal to the probability that all other estimates are less than or equal to x: $\hat{F}_{\max}(x) = P(\max_i \hat{\mu}_i \leq x) = \prod_{i=1}^M P(\hat{\mu}_i \leq x) = \prod_{i=1}^M \hat{F}_i(x)$. Considering the PDF $\hat{f}_{\max}$, the expected value of the maximum estimator is $E\left[\hat{\mu}_*^{ME}\right] = E[\max_i \hat{\mu}_i] = \int_{-\infty}^{\infty} x \hat{f}_{\max}(x) dx$. This is equal to

$$\begin{aligned} E\left[\hat{\mu}_*^{ME}\right] &= \int_{-\infty}^{\infty} x \frac{d}{dx} \prod_{j=1}^{M} \hat{F}_j(x)\, \mathrm{d}x \\ &= \sum_i^M \int_{-\infty}^{\infty} x \hat{f}_i(x) \prod_{i \neq j}^{M} \hat{F}_j(x)\, \mathrm{d}x. \end{aligned} \quad (3)$$

However, this is the expected value of the ME, not the maximum expected value in (1). The positive bias can be explained by the presence of x in the integral which correlates with the monotonically increasing product $\prod_{i \neq j}^M \hat{F}_j(x)$.

In order to avoid this issue, an alternative method, called *Double Estimator* (DE), has been proposed in (van Hasselt, 2010) and theoretically analyzed in (van Hasselt, 2013). In this technique, like in the case of the maximum estimator, a sample set S retrieved by the true unknown distribution is used, but in this case it is divided in two disjoint subsets $S^A = \{S_1^A, ..., S_N^A\}$ and $S^B = \{S_1^B, ..., S_N^B\}$. If the sets are split in a proper way, for instance randomly, the sample means $\hat{\mu}_i^A$ and $\hat{\mu}_i^B$ are unbiased, like the means $\hat{\mu}_i$ in the case of the single estimator. An estimator a^*, such that $\hat{\mu}_{a^*}^A(X) = \max_i \hat{\mu}_i^A(X)$, is used to pick an estimator $\hat{\mu}_{a^*}^B$ that is an estimate for $\max_i E[\hat{\mu}_i^B]$ and for $\max_i E[X_i]$. Obviously, this can be done the opposite way, using an estimator b^* to retrieve the estimator value $\hat{\mu}_{b^*}^A$. DE takes the average of these two estimators. The bias of DE can be found in the same way as for ME with

$$E\left[\hat{\mu}_*^{DE}\right] = \sum_i^M E\left[\hat{\mu}_i^B\right] \int_{-\infty}^{\infty} \hat{f}_i^A(x) \prod_{j \neq i}^{M} \hat{F}_j^A(x)\, \mathrm{d}x \quad (4)$$

when using an estimator a^* (the same holds by swapping A and B). This formula can be seen as a weighted sum of the expected values of the random variables where the weights are the probabilities of each variable to be the maximum. Since these probabilities sum to one, the approximation given by DE results in a value that is lower than or equal to the maximal expected value. Even if the underestimation does not guarantee better estimation than the ME, it can be helpful to avoid an incremental approximation error in some learning problems. For instance, Double Q-Learning (van Hasselt, 2010) is a variation of Q-Learning that exploits this technique to avoid the previously described issues due to overestimation. Double Q-Learning has been tested in some very noisy environments and succeeded to find better policies than Q-Learning. Double Q-Learning has been applied also in the field of Deep Rein-

forcement Learning as a modification of the widely known DQN model (van Hasselt et al., 2015). Another remarkable application of DE is presented in (Xu et al., 2013) where it achieves better results than ME in a sponsored search auction problem.

3. The Proposed Method

Differently from ME and DE that output the sample average of the variable that is estimated to be the one with the largest mean, we propose to estimate the maximum expected value $\mu_*(X)$ with the *Weighted Estimator* (WE) that computes a weighted mean of all the sample averages:

$$\hat{\mu}_*^{WE}(S) = \sum_{i=1}^{M} \hat{\mu}_i(S) w_i^S.$$

Ideally, each weight w_i^S should be the probability of $\hat{\mu}_i(S)$ being larger than all other samples means:

$$w_i^S = P\left(\hat{\mu}_i(S) = \max_j \hat{\mu}_j(S)\right).$$

If we knew the PDFs $\hat{f}_i^S$ for each $\hat{\mu}_i(S)$ we could compute the *Distribution-aware Weighted Estimator* (DWE):

$$\hat{\mu}_*^{DWE}(S) = \sum_{i=1}^{M} \hat{\mu}_i(S) \int_{-\infty}^{+\infty} \hat{f}_i^S(x) \prod_{j \neq i} \hat{F}_j^S(x)\, \mathrm{d}x. \quad (5)$$

We know that the sample mean $\hat{\mu}_i(S)$ is a random variable whose expected value is μ_i and whose variance is $\frac{\sigma_i^2}{|S_i|}$. Unfortunately, its PDF $\hat{f}_i^S$ depends on the PDF f_i of variable X_i that is assumed to be unknown. In particular, if X_i is normally distributed, then, independently of the sample size, the sampling distribution of its sample mean is normal too: $\hat{\mu}_i(S) \sim \mathcal{N}\left(\mu_i, \frac{\sigma_i^2}{|S_i|}\right)$. On the other hand, by the central limit theorem, the sampling distribution $\hat{f}_i^S$ of the sample mean $\hat{\mu}_i(S)$ approaches the normal distribution as the number of samples $|S_i|$ increases, independently of the distribution of X_i. Leveraging on these considerations, we propose to approximate the distribution of the sample mean $\hat{\mu}_i(S)$ with a normal distribution, where we replace the (unknown) population mean and variance of variable X_i with their (unbiased) sample estimates $\hat{\mu}_i(S)$ and $\hat{\sigma}_i(S)$:

$$\hat{f}_i^S \approx \tilde{f}_i^S = \mathcal{N}\left(\hat{\mu}_i(S), \frac{\hat{\sigma}_i^2(S)}{|S_i|}\right),$$

so that WE is computed as:

$$\hat{\mu}_*^{WE}(S) = \sum_{i=1}^{M} \hat{\mu}_i(S) \int_{-\infty}^{+\infty} \tilde{f}_i^S(x) \prod_{j \neq i} \tilde{F}_j^S(x)\, \mathrm{d}x. \quad (6)$$

It is worth noting that WE is consistent with $\mu_*(X)$. In fact, as the number of samples grows to infinity, each sample mean $\hat{\mu}_i$ converges to the related population mean μ_i, and the variance of the normal distribution $\tilde{f}_i$ tends to zero, so that the weights of the variables with expected value less than $\mu_*(X)$ go to zero, so that $\hat{\mu}_*^{WE} \to \mu_*(X)$.

4. Estimation Error

In this section, we theoretically analyze the estimation error of $\hat{\mu}_*^{WE}(S)$ in terms of bias and variance, comparing it with the results available for ME and DE. Although DWE cannot be used in practice, we include it in the following analysis since it provides an upper limit to the accuracy of WE.

4.1. Bias

We start with summarizing the main results about the bias of ME and DE reported in (van Hasselt, 2013). For what concerns the direction of the bias, ME is positively biased, while DE is negatively biased. If we look at the absolute bias, there is no clear winner. For instance, when all the random variables are identically distributed, DE is unbiased, while the same setting represents a worst case for ME . On the other hand, when the maximum expected value is sufficiently larger than the expected values of the other variables, the absolute bias of ME can be significantly smaller than the one of DE (see Section 5). The bias of ME is bounded by:

$$\mathrm{Bias}\left(\hat{\mu}_*^{ME}\right) \leq \sqrt{\frac{M-1}{M} \sum_{i=1}^{M} \frac{\sigma_i^2}{|S_i|}}.$$

For the bias of DE, van Hasselt (2013) conjectures the following bound (which is proved for two variables):

$$\mathrm{Bias}\left(\hat{\mu}_*^{DE}\right) \geq -\frac{1}{2}\left(\sqrt{\sum_{i=1}^{M} \frac{\sigma_i^2}{|S_i^A|}} + \sqrt{\sum_{i=1}^{M} \frac{\sigma_i^2}{|S_i^B|}}\right).$$

In the next theorem we provide a relationship between the bias of WE and the one of ME.

Theorem 1. *For any given set X of M random variables:*

$$\mathrm{Bias}(\hat{\mu}_*^{WE}) \leq \mathrm{Bias}(\hat{\mu}_*^{ME}) \leq \sqrt{\frac{M-1}{M} \sum_{i=1}^{M} \frac{\sigma_i^2}{|S_i|}}.$$

As we will see in Section 5, this does not mean that the absolute bias of WE is necessarily smaller than the one of ME, since (as we will see later) the bias of WE can be also negative. In order to better characterize the bias of WE, we put it in relation with the bias of DE.

Theorem 2. *For any given set X of M random variables:*

$$\mathrm{Bias}(\hat{\mu}_*^{WE}) \geq \mathrm{Bias}(\hat{\mu}_*^{DE}).$$

Example In Figures 1 and 2 we visualize the bias of the different MEV estimators in a setting with two normally distributed random variables ($X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$) as a function of the difference of their expected values. Both variables have variance equal to 10 ($\sigma_1^2 = \sigma_2^2 = 10$) and we assume to have 100 samples for each variable ($|S_1| = |S_2| = 100$). Figure 1 confirms the previous theoretical analysis: the bias of ME is always positive, while the biases of DWE and DE are always negative,

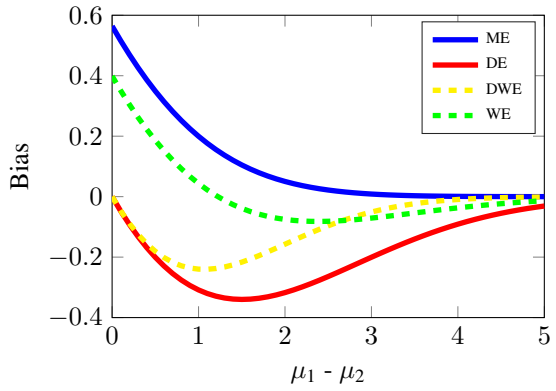


Figure 1. Comparison of the bias of the different estimators.

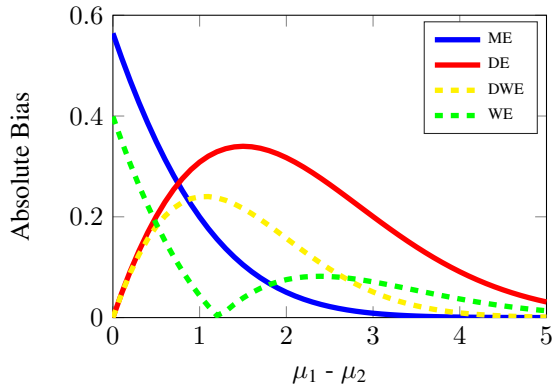


Figure 2. Comparison of the absolute bias of the different estimators.

with the latter always worse than the former. The bias of WE can be positive or negative according to the situation, but it always falls in the range identified by the biases of ME and DE. Looking at the absolute biases shown in Figure 2, we can notice that there is not a clear winner. As previously mentioned, when the variables have the same mean, both DE and DWE are unbiased, while it represents a worst case for the bias of ME and WE. It follows that, when the difference of the two means is small (less than 0.5 in the example), DE suffers less absolute bias than ME and WE. For moderate differences of the means (between 0.5 and 1.8 in the example), WE has the minimum absolute bias, while ME is preferable for larger differences. Such results can be generalized as follows: DE suffers a small bias when there are several variables that have expected values close (w.r.t. their variances) to the maximum one, while ME provides the best estimate when there is one variable whose expected value is significantly larger (w.r.t. the variances) than all the expected values of all the other variables. In all the other cases, WE is less biased.

4.2. Variance

We cannot evaluate the goodness of an estimator by analyzing only its bias. In fact, since the MSE of an estimator is the sum of its squared bias and its variance, we need to take into consideration also the latter.

van Hasselt (2013) proved that both the variance of ME and the one of DE can be upper bounded with the sum of the variances of the sample means: $\mathrm{Var}\left(\hat{\mu}_*^{ME}\right) \leq \sum_{i=1}^{M} \frac{\sigma_i^2}{|S_i|}$, $\mathrm{Var}\left(\hat{\mu}_*^{DE}\right) \leq \sum_{i=1}^{M} \frac{\sigma_i^2}{|S_i|}$. The next theorem shows that the same upper bound holds also for the variance of WE.

Theorem 3. *The variance of WE is upper bounded by*

$$\mathrm{Var}\left(\hat{\mu}_*^{WE}\right) \leq \sum_{i=1}^{M} \frac{\sigma_i^2}{|S_i|}.$$

The bound in Theorem 3 is overly pessimistic; in fact, even if each weight w_i^S is correlated to the other weights and to the sample mean $\hat{\mu}_i(S)$, their sum is equal to one. For sake of comparison, we upper bound the variance of DWE.

Theorem 4. *The variance of DWE is upper bounded by*

$$\mathrm{Var}\left(\hat{\mu}_*^{DWE}\right) \leq \max_{i \in 1,\ldots,M} \frac{\sigma_i^2}{|S_i|}.$$

Example As done for the bias, in Figure 3 we show the variance of the different estimators under the same settings described above. As the difference of the means of the two variables grows, the variance of all the estimators converges to the variance of the sample mean of the variable with the maximum expected value. DE is the estimator with the largest variance since its sample means are computed using half the number of samples w.r.t. the other estimators. WE exhibits a variance slightly larger than the one of ME, while, as expected, the variance of DWE is always the smallest.

Finally, in Figure 4 we show the MSE (variance + bias^2) of the different estimators. When the difference between the two means is less than one, WE suffers from a lower MSE than the other two estimators. On the other hand, ME is preferable when there is a variable with an expected value that is significantly larger than the other ones.

5. Experiments

In this section we empirically compare the performance of WE, ME, and DE on four sequential decision-making problems: two multi-armed bandit domains and two MDPs.

5.1. Multi-Armed Bandits

In the multi-armed bandit problem, a learner wants to identify the action (arm) associated to the largest mean reward. In some domains, we are not interested only in knowing which is the best action, but we want also an accurate estimate of its expected reward.

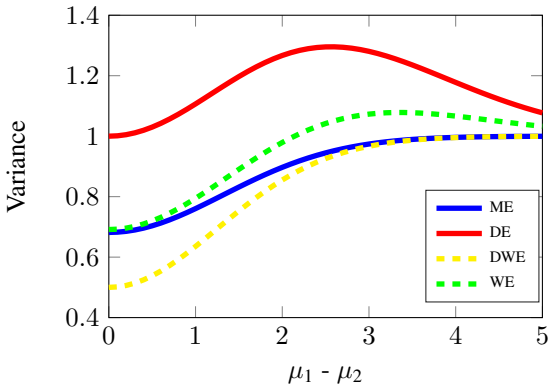


Figure 3. Comparison of the variance of the different estimators.

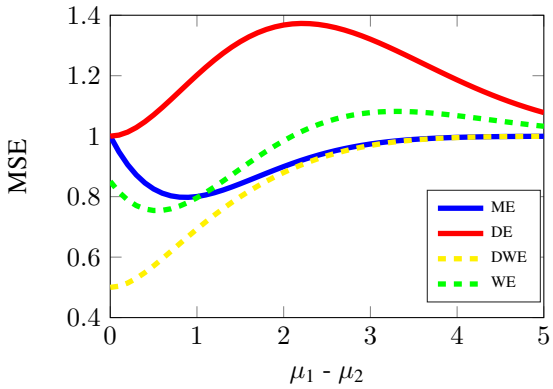
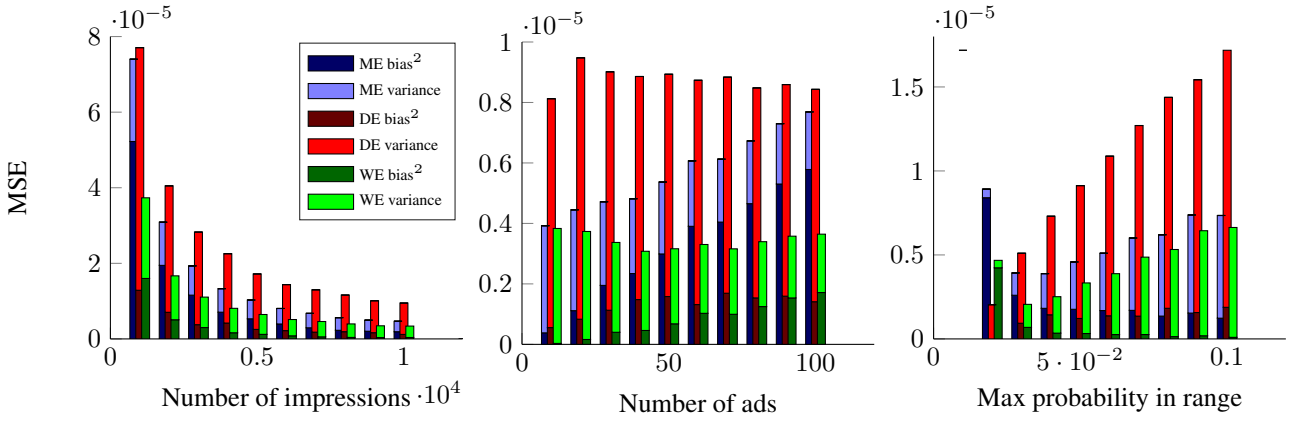


Figure 4. Comparison of the MSE of the different estimators.



(a) First experiment (b) Second experiment (c) Third experiment

Figure 5. MSE for each setting. Results are averaged over 2,000 experiments.

5.1.1. Internet Ads

We consider the problem as formulated in (van Hasselt, 2013). The goal of this problem is to select the best ad to show on a website among a set of M possible ads, each one with an unknown expected return per visitor. Assuming that each ad has the same return per click, the best ad is the one with the maximum click rate. Since the click rates are unknown, we need to estimate them from data. In our setting, given N visitors, each ad is shown the same number of times, so that we have N/M samples to compute the sample click rate. A quick and accurate estimate of the maximum click rate may be relevant to determine future investment strategies. We compared the results of ME, DE, and WE in three different settings. We consider a default configuration where we have $N = 300,000$ visitors, $M = 30$ ads and mean click rates uniformly sampled from the interval $[0.02, 0.05]$. In the first experiment, we vary the number of visitors $N = \{30,000, 60,000, ..., 270,000, 300,000\}$, so that the number of impressions per ad ranges from $1,000$ to $10,000$. In the second experiment, we vary the number of ads $M = \{10, 20, ..., 90, 100\}$ and the number of visitors is set to $N = 10,000 \cdot M$. In the last experiment, we modify the interval of the mean click rates by changing the value of the upper limit with values in $\{0.02, 0.03, ..., 0.09, 0.1\}$, with the lower fixed at 0.02.

In Figure 5, we show the $MSE = bias^2 + variance$ for the three experiments comparing the results obtained by each estimator. In the first experiment (Figure 5(a)), as expected, the MSE decreases for all estimators as the number of impressions increases and WE has the lowest MSE in all cases. It is interesting to see how the ME estimator has a very large bias in the leftmost case, which shows (accordingly to Figure 2) that ME estimator suffers large bias when only a few samples are available and, therefore, the variances of the sample means are large. From Figure 5(b) we can notice that an increasing number of actions has a negative effect on ME and a positive effect on the DE due to the fact that a larger number of ads implies a larger number of variables with a mean close to the maximum expected value, that represents a worst case for ME and a best case

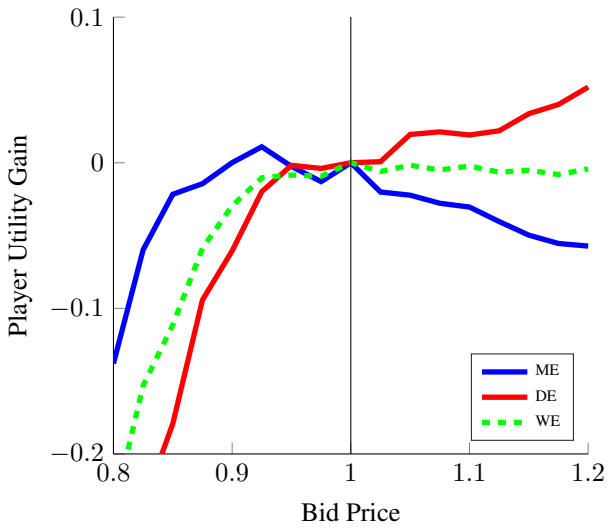


Figure 6. Relative player 1 utility gain for different value of the bid defined as $\frac{utility(b)}{utility(v)} - 1$. Results are averaged over 2,000 experiments.

Algorithm 1 Weighted Q-Learning

initialize $\forall(s,a) : Q(s,a) = \mu(s,a) = 0, \sigma(s,a) = \infty$
repeat
  Initialize s.
  repeat
    Choose action a from state s using policy derived from Q (e.g. ϵ-greedy).
    Take action a, observe reward r and next state s'.
    for all $a_i \in \mathbf{a}$ **do**
      $w_i \leftarrow \int_{-\infty}^{+\infty} \tilde{f}_i^S(s', a_i) \prod_{j \neq i} \tilde{F}_j^S(s', a_j)\, \mathrm{d}x.$
    end for
    $W \leftarrow w^T * Q(s', \mathbf{a}).$
    $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma W - Q(s,a)].$
    Update $\mu(s,a)$ and $\sigma(s,a)$.
    $s \leftarrow s', a \leftarrow a'.$
  until s is terminal
until

for DE. The MSE of WE is the lowest in all cases and does not seem to suffer the increasing number of actions. The same happens in Figure 5(c) when all the ads share the same click rate (0.02), where DE is the best. However, it starts to have large variance as soon as the range of probabilities increases (Figure 3). WE has the lowest MSE, but, as the range increases, it gets similar to the MSE of ME.

5.1.2. Sponsored Search Auctions

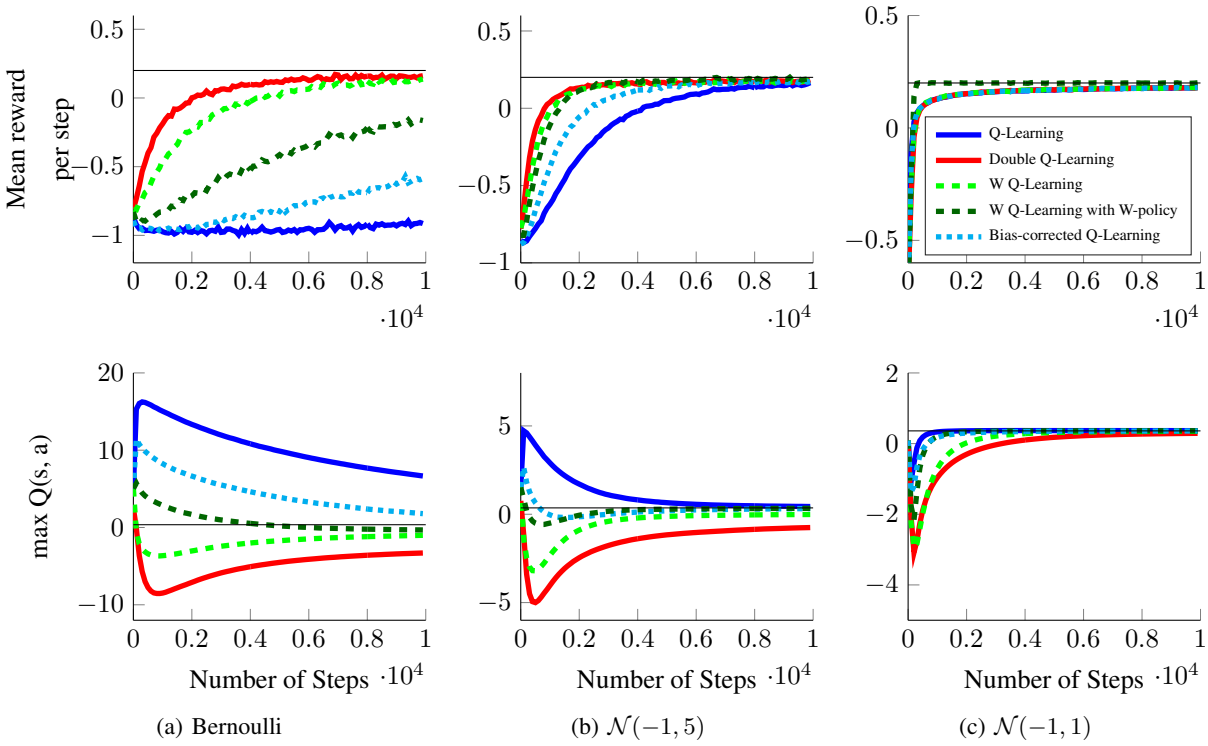
We considered the domain described in (Xu et al., 2013), where a search engine runs an auction to select the best ad to show from a pool of candidates with the goal of maximizing over a value that depends on the bid of each advertiser and its click probability. Each time an ad is clicked, the search engine charges the advertiser a fee that depends on the bids b of the advertisers and the click through rates (CTRs) ρ of the ads. Since in general the CTRs are unknown, it is crucial for the search engine to estimate from the data which is the best ad (i.e., the one that maximizes $b \cdot \rho$) and the payment in case of click. Wrong estimations may significantly harm the revenue. On the other hand, the advertisers have to decide the value of their bid b_i according to the true values v_i of a click. A desirable condition in auctions, called *incentive compatibility*, requires that the advertisers maximize their utility by truthfully bidding $b_i = v_i$. Incentive incompatibility may occur if the estimate of the click probabilities are not accurate, therefore it is interesting to evaluate how the estimators perform on this task. We measured the utility gain of advertiser 1, whose true per click value is $v_1 = 1$, for different bid b_1 values and competing with four other advertisers whose bids are $b_{-1} = \{0.9, 1, 2, 1\}$. The CTRs are: $\rho = \{0.15, 0.11, 0.1, 0.05, 0.01\}$. Following the approach suggested in (Xu et al., 2013), CTRs are estimated from data collected using the UCB1 algorithm (Auer et al., 2002) in a learning phase consisting of $10,000$ rounds of exploration (i.e., impressions).

Figure 6 shows the utility gain of advertiser 1 when using ME, DE, and WE.[1] It can be seen that ME does not achieve incentive compatibility because utility has positive values before the true bid price (which is highlighted with a black vertical bar). On the other hand, with DE the advertiser has no incentive to underbid, but there is an incentive to overbid using DE. With WE, there is no significant incentive to underbid or overbid showing that it succeeds to achieve incentive compatibility.

5.2. Markov Decision Process

In the following experiments we compare Q-Learning, Double Q-Learning and a modified version of Q-Learning, that we call *Weighted Q-Learning* (see Algorithm 1), which uses WE to estimate the maximum action values. Since WE computes the probability of each action to be the optimal one, it is quite natural to exploit them to define a policy, that we call *Weighted policy*. At each step, this policy selects an action using the probability estimated by WE. As we show in the following experiments, the exploration induced by this policy is effective in reducing the estimation error of the value function.

[1]The debiasing algorithm proposed in (Xu et al., 2013) is a cross validation approach, but differs from the estimators considered in this paper. It averages the values used for selection and the values used for estimation, thus being a hybrid of DE and ME.

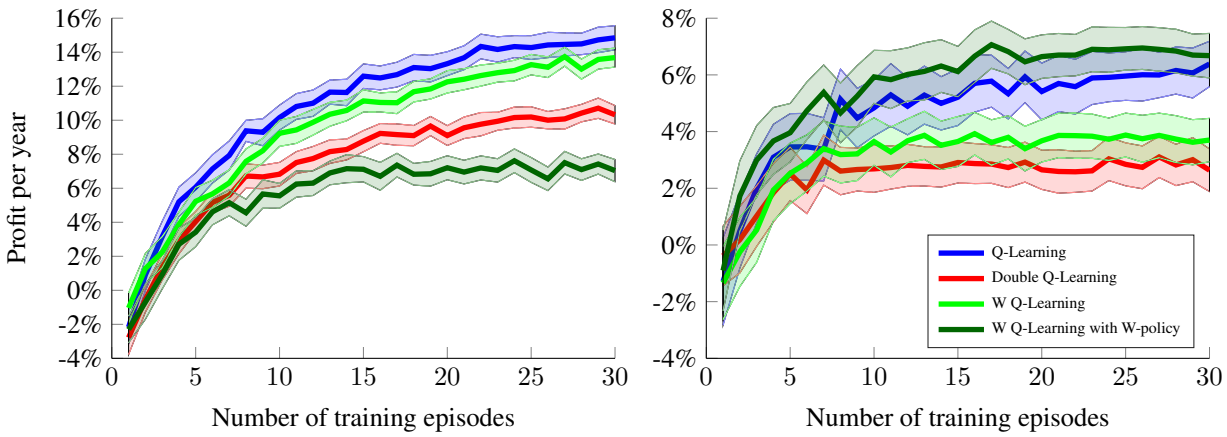


(a) Bernoulli (b) $\mathcal{N}(-1, 5)$ (c) $\mathcal{N}(-1, 1)$

Figure 7. Grid world results with the three reward functions averaged over 10,000 experiments. Optimal policy is the black line.

5.2.1. Grid World

We analyzed the performance of the different estimators on a 3×3 grid world where the start state in the lower-left cell and the goal state in the upper-right cell (van Hasselt, 2010). In this domain, we also compared the performance of the Bias-corrected Q-Learning algorithm, a modified version of Q-learning that, assuming Gaussian rewards, corrects the positive bias of ME by subtracting to each Q-value a quantity that depends on the standard deviation of the reward and on the number of actions (Lee & Powell, 2012; Lee et al., 2013). Learning rate is $\alpha_t(s,a) = \frac{1}{n_t(s,a)^{0.8}}$ where $n_t(s,a)$ is the current number of updates of that action value and the discount factor is $\gamma = 0.95$. In Double Q-Learning we use two learning rates $\alpha_t^A(s,a) = \frac{1}{n_t^A(s,a)^{0.8}}$ and $\alpha_t^B(s,a) = \frac{1}{n_t^B(s,a)^{0.8}}$ where $n_t^A(s,a)$ and $n_t^B(s,a)$ are respectively the number of times when table A and table B are updated. We use an ϵ-greedy policy with $\epsilon = \frac{1}{\sqrt{n(s)}}$ where $n(s)$ is the number of times the state s has been visited. For the reward function we consider three different settings: (1) Bernoulli, -12 or 10 randomly at each step, (2) Gaussian with mean $\mu = -1$ and standard deviation $\sigma = 5$, (3) Gaussian with mean $\mu = -1$ and standard deviation $\sigma = 1$. Once in the goal state, each action ends the episode and returns a reward of 5. Since the optimal policy ends the episode in five actions, the optimal average reward per step is 0.2. Moreover, the optimal value of the action maximizing the Q-value is $5\gamma^4 - \sum_{k=0}^{3}\gamma^k \approx 0.36$. In Figure 7, the top plots show the average reward per step obtained by each algorithm and the plots at the bottom show the estimate of the maximum state-action value at the starting state for each algorithm. Figures 7(a) and 7(b) show that when the variance is large with respect to the differences between the means, the underestimation of Double Q-Learning allows to achieve the best policy faster than other algorithms, even if the approximation of the Q-function is not accurate as for Weighted Q-Learning. Bias-corrected Q-Learning performs better than Q-Learning, but worse than the other algorithms (at least when the reward variance is large), even if the approximation of the Q-function is quite accurate. In all settings, Weighted Q-Learning shows much less bias than the other estimators (in particular using the weighted policy). Using the weighted policy, it achieves the best performance in the case with $\sigma = 1$ (see Figure 7(c)). This happens because the weighted policy exploits the good approximation of the Q-function computed by Weighted Q-Learning and reduces the exploration faster than ϵ greedy. It is worth to point out that the Gaussian approximation of Q-values used by Weighted Q-Learning works well for both Gaussian and Bernoullian rewards, showing that WE is effective even with non-Gaussian distributions.



(a) Train (b) Test

Figure 8. Profit per year during training (left) and test (right) phase. Results are averaged over 100 experiments.

5.2.2. FOREX

Finally, we want to evaluate the performance of the three estimators in a more challenging learning problem. Foreign Exchange Market (Forex) is known to be an environment with hardly predictable dynamics. For this reason, it is very difficult to estimate the Q-values and, therefore, the expected profit. In real cases it becomes a crucial issue, especially in terms of risk management, to avoid overestimation or underestimation of the Q-value of a state-action pair. To better evaluate the results, we defined a more basic environment compared to the real market. In our Forex MDP the agent enters in the market always with 1$ and each time the agent enters on long or short position a fixed spread value of 0.0002$ is paid. The possible actions taken from the agent can be -1, 0 or 1, which mean respectively 'enter on a short position', 'close a position' and 'enter on long position'. As state space we consider the suggestions (in terms of actions) provided by 7 common Forex indicators and the action chosen by the agent at the previous time step. The state space is $S = \{-1, 0, 1\}^8$ with $s_{i=1...7}(t) = \{-1, 0, 1\}$ and $s_8(t) = a(t-1)$. The action taken by the agent is $a(t) = \{-1, 0, 1\}$. The reward $r(t)$ is a function of the previous and current action chosen and of the difference between the current closing price $c(t)$ and the previous closing price $c(t-1)$:
$r(t) = a(t-1)(c(t)-c(t-1))+0.5*spread|a(t)-a(t-1)|$.
The same four algorithms used in the grid world domain were trained using historical daily data of GBP/USD exchange rate from 09/22/1997 to 01/10/2005 and tested on data from 01/11/2005 to 05/27/08. During the training phase, we set learning rate $\alpha(s,a) = \frac{1}{n(s,a)}$, discount factor $\gamma = 0.8$ and $\epsilon = \frac{1}{\sqrt{n(s)}}$.

In Figure 8 is shown the profit per year of the four algorithms on the training set (Figure 8(a)) and test set (Figure 8(b)) in relation to the number of training episodes. In this training phase an ϵ-greedy policy is used for Q-learning and Double Q-Learning, while Weighted Q-Learning has been done with both the ϵ-greedy policy and the Weighted policy. During training, Q-learning performs better than Double Q-learning and also than Weighted Q-learning. Weighted Q-learning with the Weighted policy has the worst performance on the training set, but it performs significantly better on the test set. This is because the Weighted policy is more explorative than the ϵ-greedy policy, so the performance can be worse during the training phase, but the estimation of the Q-values is more accurate. Double Q-learning performs worse than Q-learning and Weighted Q-learning both on training set and test set. The reason is that in many states there is an action that is significantly better than the others, that represents the case where ME gives the best results, while DE suffers.

6. Conclusions

We have presented a new estimator for the maximum expected value of a set of random variables, based on a weighted average of the sample means. An extensive empirical analysis confirms what has been observed in the theoretical analysis, showing that no estimator is always the best. Nevertheless, our estimator has a robust behavior performing well in all conditions and obtaining the best performance in most of the "non-trivial" cases.
As future work, in order to apply the proposed estimator to continuous reinforcement learning problems, it will be interesting to extend WE to the case where Q-values are not stored in a table, but are represented through regression models such as trees or neural networks.

References

Auer, Peter, Cesa-Bianchi, Nicolo, and Fischer, Paul. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Blumenthal, Saul and Cohen, Arthur. Estimation of the larger of two normal means. *Journal of the American Statistical Association*, 63(323):861–876, 1968.

Dhariyal, I., Sharma, D., and Krishnamoorthy, K. Nonexistence of unbiased estimators of ordered parameters. *Statistics*, 1985.

Lee, Daewoo, Defourny, Boris, and Powell, Warren B. Bias-corrected q-learning to control max-operator bias in q-learning. In *Adaptive Dynamic Programming And Reinforcement Learning (ADPRL), 2013 IEEE Symposium on*, pp. 93–99. IEEE, 2013.

Lee, Donghun and Powell, Warren B. An intelligent battery controller using bias-corrected q-learning. In *AAAI*. Citeseer, 2012.

Smith, James E and Winkler, Robert L. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006.

Stone, Mervyn. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society. Series B (Methodological)*, pp. 111–147, 1974.

Van den Steen, Eric. Rational overoptimism (and other biases). *American Economic Review*, pp. 1141–1151, 2004.

van Hasselt, Hado. Double q-learning. In *Advances in Neural Information Processing Systems*, pp. 2613–2621, 2010.

van Hasselt, Hado. Estimating the maximum expected value: an analysis of (nested) cross-validation and the maximum sample average. *arXiv preprint arXiv:1302.7175*, 2013.

van Hasselt, Hado, Arthur, Guez, and David, Silver. Deep reinforcement learning with double q-learning. *CoRR*, abs/1509.06461, 2015. URL http://arxiv.org/abs/1509.06461.

Xu, Min, Qin, Tao, and Liu, Tie-Yan. Estimation bias in multi-armed bandit algorithms for search advertising. In *Advances in Neural Information Processing Systems*, pp. 2400–2408, 2013.

A. Proofs

Theorem 1. The proof of Theorem 1 follows directly from observing that $\hat{\mu}_*^{WE}$ is always smaller than $\hat{\mu}_*^{ME}$. In fact, the ME estimator can be seen as a weighted estimator that gives probability one to the variable associated to the largest sample mean $\hat{\mu}_i$, so that any other weighting cannot produce a larger value. □

Theorem 2. If we compare the expected value of DE reported in Equation (4) with the value of the estimator WE in Equation (3), we can notice strong similarities. The main difference is that in DE the sample mean of variable X_i and its probability of being the maximum are computed w.r.t. two independent set of samples, while in WE these two quantities are positively correlated. It follows that WE has a positive bias w.r.t. DE. □

Theorem 3. Starting from the definition of WE (3), we can derive the bound to the variance as follows

$$\begin{aligned}\mathrm{Var}\left(\hat{\mu}_*^{WE}\right) &= \mathrm{Var}\left(\sum_{i=1}^{M}\hat{\mu}_i(S)w_i^S\right)\\ &\leq \mathrm{Var}\left(\sum_{i=1}^{M}\hat{\mu}_i(S)\right)\\ &= \sum_{i=1}^{M}\mathrm{Var}\left(\hat{\mu}_i(S)\right),\end{aligned}$$

where the inequality is a consequence of the maximization of each weight w_i^S with one and the last equality comes from the independence of the sample means. □

Theorem 4. Since the weights w_i computed by DWE are not random variables, it follows

$$\begin{aligned}\mathrm{Var}\left(\hat{\mu}_*^{DWE}\right) &= \mathrm{Var}\left(\sum_{i=1}^{M}\hat{\mu}_i(S)w_i\right)\\ &= \sum_{i=1}^{M}w_i^2\mathrm{Var}\left(\hat{\mu}_i(S)\right)\\ &\leq \max_{i\in 1,\ldots,M}\frac{\sigma_i^2}{|S_i|},\end{aligned}$$

where the inequality is motivated by $w_i^2 \leq 1,\ \forall i$. □

B. Forex

The indicators chosen to define the states are: Moving Average Convergence/Divergence indicator, Relative Strength Index, Momentum, Channel Commodity Index, Stochastic Oscillator, Bollinger Bands, Moving Average Cross-Over The actions suggested by the indicators are computed by setting the parameters and the entry-exit conditions following the most used and common rules for these indicators. In particular all the signals are defined using implemented

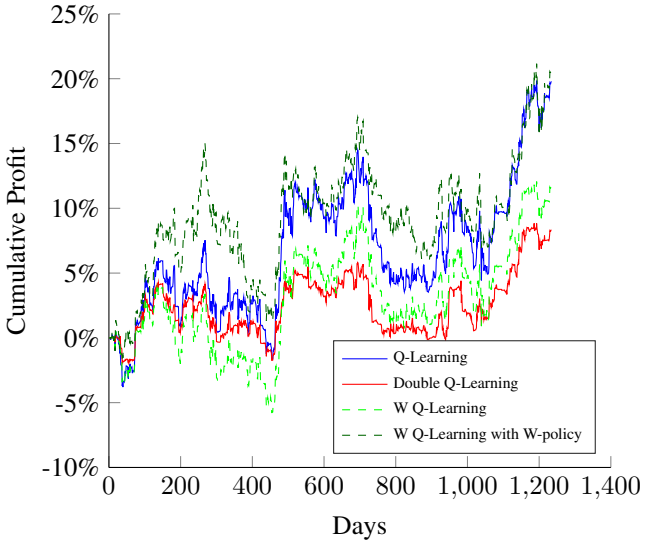


Figure 9. Cumulative profit in test set after 30 training episodes. Results are averaged over 100 experiments.

Matlab Financial Toolbox functions and setting parameters and conditions like below.

- Moving Average Covergence/Divergence indicator: the MACD is calculated by subtracting the 26-period exponential moving average from the 12-period moving average. A 9 period exponential moving average is used as signal line. When the MACD falls below the signal line a long position signal is produced. Otherwise if the MACD surpasses the signal line a short position signal is produced.

- Relative Strength Index: the period chosen for the RSI is 20 days. When the RSI is under the value of 30 an oversold market condition occurs, so a long position signal is produced. Similarly, when the RSI is over 70, a short position signal is generated. When the value is between these two bounds a close position signal is produced.

- Momentum: the momentum uses a period of 14 days. The strategy is to be always in the market. In particular if the value of the momentum is less than zero a long position signal is generated. Otherwise a short position signal is generated.

- Channel Commodity Index: the CCI is used with a 20 days period. The long position signal is produced when the CCI cross and surpasses the lower bound of -100. The short position is taken when the CCI falls behind the value of 100. Otherwise the close position signal is generated.

- Stochastic Oscillator: the Stochastic oscillator uses the high prices, low prices, and close prices with a 14 days period for the %K line and 3 for the %D line. If both lines are above the value of 80 and the %K line falls behind %D line, then a short position signal is generated. If both lines

are below 20 and the %K line surpasses the %D line, a long position signal is generated. In other conditions a close position signal is generated.

- Bollinger Bands: the Bollinger Bands signal is produced with a window size of 20. Long position signal is produced when the closing price surpasses the upper Bollinger Band. When the closing price falls down the lower band, then a short position signal is taken. Otherwise a close position signal is generated.

- Moving Average Cross-Over: we used two moving average of 20 and 200 periods. A long position signal is produced when the 20 period moving average falls behind the 200 periods one. Otherwise a short position signal is generated.

In Figure 9, we can compare the cumulative rewards of the algorithms on the test set after 30 training episodes. Observing the ranges of the values, we can see that the Double Q-learning agent gains and loses less than the other agents, so we can deduce that it entries in the market less often.