“Which Off-Policy Evaluation (OPE) Method, and When?”

Jonathan Huml

Abstract

Offline, off-policy reinforcement learning is of critical importance in use cases where online algorithms might be unsafe, costly, or unethical. A host of off-policy methods exist for the *evaluation problem*, or estimating the value of a state for a proposed “evaluation” policy π_e, but with data collected under a behavior policy π_b. However, benchmarking the performance of these various OPE methods is an ongoing effort. This paper will evaluate the evaluators: given data generated by a behavior policy in a set of environments, which method should we choose, and for which environments? We test importance sampling (IS), weighted (or normalized) importance sampling (WIS), fitted Q-evaluation (FQE), and doubly robust FQE methods on two carefully chosen environments meant to expose the underlying properties of each method and their respective differences in performance.

1. Introduction

We first consider an analogy: if an example of reinforcement learning is the vehicle control problem, then offline, off-policy learning is like driving with one hand tied behind your back. In this case, we have batches of data generated by a behavior policy π_b and a target or evaluation policy π_e that is not (or almost always not) equivalent to the data-generating policy. The evaluation problem, most generally, concerns the case where we wish to estimate the true value of a state, $V(s)$, under an given policy, which itself a distribution.

Consider, however, that $\pi_e \neq \pi_b$: this *distributional shift* is the central off-policy evaluation problem. First, what happens if the behavior policy is far from optimal? In the offline setting, we have no way to interact with the environment, and thus a policy that never explores, or only exploits poor states, or otherwise acts suboptimally, has already tied our hand to the car seat. Perhaps we should procure some scissors? Yet, herein lies the out-of-distribution problem: the further we stray from the behavior policy, the more error we accumulate and a faster rate (quadratically with horizon H) even *with* optimal action labels (1). To analogize once more, this time with a supervised learning problem, we would be hard-pressed to use a neural network to identify cat breeds if we only trained only on dog breeds.

A host of methods exist to overcome this challenge. However, standard benchmarking tasks to compare these OPE methods in different contexts are an ongoing effort. Given an environment and a known (not estimated) behavior policy, which method should we choose for a given target policy?

1.1. Contributions

This paper builds upon the findings of the Caltech OPE Benchmarking Suite (2). Much like (2), our ultimate goal is to have a checklist or decision tree that, after enumerating the fundamental, high-level properties of an environment, we can choose the best method for that environment. We show the proposed tree of (2) below in Figure 1. This paper will only consider the tabular setting, and we focus our experiments on the first two branches under the “Properly Specified” node: how do horizon and policy mismatch inform one another? The order of these nodes is of critical importance. While errors accumulate quadratically in horizon length, and thus may dominate all other decision factors, we consider off-policy learning in its greater context. We *evaluate* a policy, but then we typically wish to *iterate* on this policy; that is, we have some control over the evaluation policy and its subsequent updates, but not the horizon

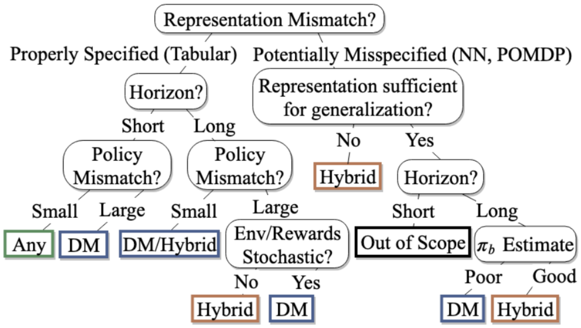


Figure 1. This paper presents experiments that re-evaluate the horizon and policy mismatch nodes, and we briefly present results that further explore the stochasticity of the environment. More details on stochasticity can be found in the appendix, but the focus is on whether or not we should re-order horizon and policy mismatch in this tree.

(or other properties of the data). If we can constrain π_e to a neighborhood of π_b, perhaps this horizon scaling effect is more nuanced. The experiments presented here show exactly this: even over longer horizons, the more "underpowered methods" (like IS or WIS) may be competitive performers.

2. Related Work

This work explores four methods that can be categorized into three buckets: inverse propensity scoring (IPS), direct methods, and hybrid methods. Importance sampling and weighted importance sampling (3) fall into the IPS bucket, and far predate reinforcement learning applications. These were originally used for Monte Carlo sampling methods, being particularly useful for numerical integration methods. Direct methods are a broad and diverse class, encompassing regression-based techniques that model almost any relevant aspect of the reinforcement learning problem, whether estimating the transition dynamics or the reward/value function. In this paper, we examine the fitted Q-evaluation algorithm (10), which is a flexible, model-free (i.e. no transition dynamics considered) method that has gained much recent traction in the RL community. Hybrid methods attempt to combine the best of both direct and IPS methods, and the doubly robust method (11) is a plug-and-play estimator that can use any direct method estimator, guiding the value estimates with the importance weights of IS.

2.1. Importance Sampling (IS)

Importance sampling is a remarkably simple and straightforward method for off-policy evaluation (3). We will show the short derivation for any distributions $p(x)$ and $q(x)$, where we have samples drawn from q (and hence an approximation of its expectation) and want the expectation of the random variable $x \in \Omega$ under p:

$$\begin{aligned}\mathbb{E}_p[x] &= \int_{x\in\Omega} xp(x)dx \\ &= \int_\Omega xp(x)\frac{q(x)}{q(x)}dx \\ &= \int_\Omega xq(x)\frac{p(x)}{q(x)}dx \\ &= \mathbb{E}_q\left[x\frac{p(x)}{q(x)}\right]\end{aligned}$$

Recall that a value function of state s is nothing more than an expectation itself: namely, it is the state-action value at s weighted by the probability of taking action $a \in \mathcal{A}$ at s. Taking this intuition and simple derivation, we introduce the importance sampling estimator over N trajectories:

$$\hat{V}_{IS}(s) = \frac{1}{N}\sum_{i=1}^{N} \underbrace{\left(\prod_{t=1}^{\tau}\frac{\pi_e(a_t|s_t)}{\pi_b(a_t|s_t)}\right)}_{w_i} \underbrace{\left(\sum_{t=1}^{\tau}\gamma^t r_t^i\right)}_{\text{discounted rewards}}$$

Several remarks are of note. First, in the event that $\pi_e = \pi_b$, the importance weights cancel and we obtain our "dumb" estimate of the value at a state, or just the average over all discounted rewards obtained from that state across the trajectories. The sample mean, of course, is a trivially unbiased estimator. Second, there are longer proofs that start from the full Bellman equation (or at least our estimate of it), but the (estimated) transition probabilities cancel out regardless of where we start. In this sense, importance sampling is "model-free" in that only information used in the weights are *how* the actions are selected (i.e. the policy), which motivates its benefits as a computationally inexpensive, baseline model. As such, this estimator retains two well-known properties: it *unbiased*, but *high variance* (3). The estimator is not even defined at certain behavior policies: we must make the *coverage assumption* where $\pi_e > 0$ implies $\pi_b > 0$. Letting $\pi_b = \varepsilon > 0$ for all $a \in \mathcal{A}$ and $s \in \mathcal{S}$, a rather extreme but informative example is fixing $k \in [0, 1]$ on the same state-action space for π_e, and taking $\lim_{\varepsilon\to 0}\frac{k}{\varepsilon}$. The variance of any *one* importance weight is hence unbounded, and the multiplicative weighting of *several* importance weights across horizon length τ can exacerbate the issue, even for the non-limiting case. Thus, there are two ways that IS can collapse, and collapse fast: ε and τ.

2.2. Weighted Importance Sampling (WIS)

Importance sampling presents several problems, but the method is simple, both fast to implement and easy to understand and troubleshoot. The variance, however, is often too high to be useful in any practical setting, especially considering that we would appreciate some safety guarantees in the mission-critical scenarios that often motivate off-policy learning in the first place (this motivates the idea of the high-confidence off-policy evaluation framework (4)).

A simple modification to lower the variance of the IS estimator is to divide by the importance weights, thus normalizing the estimator. While the variance is still unbounded in a limiting sense (consider $\lim_{\varepsilon\to 0}\frac{\sum_i k_i}{\varepsilon}$ where $k_j = \varepsilon$ for all but one $i \neq j$, and we see the same problem even though we have mostly patched the problem), in practice, this often makes IS more usable. WIS attacks the variance problem through the limiting ε problem, but we can also take the approach through τ via a *per-decision importance sampler* by considering only (FINISH...). This is not considered in our experiments for the sake of clarity and focus, but does motivate the idea that importance sampling can indeed be extended and usable.

The weight normalization induces bias into the originally-unbiased IS estimator, but greatly reduces the variance and is a strongly consistent estimator. The estimator is given by:

$$\hat{V}_{WIS}(s) = \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i \left(\sum_{t=1}^{\tau} \gamma^t R_t^i \right)$$

2.3. Fitted Q-Evaluation (FQE)

Fitted Q evaluation (10) is a popular, model-free direct method (in the sense that it does not rely on estimates of Markov transition matrices) that estimates a $Q-$function from data given a class of function approximators $\mathcal{F}$. The typical program looks something like, for k iterations and model parameters θ:

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

where

$$y_t^i = R_t^i + \gamma \mathbb{E}_{\pi_e} \hat{Q}_{k-1}(s_{t+1}^i; \theta)$$

Note that the $Q-$function estimate $\hat{Q}_{k-1}$ is often denoted as f in the literature, and the arguments of the minima are taken over $f \in \mathcal{F}$ instead of the actual parameters θ of f. As such, this is a very general framework that can incorporate deep learning, tree-based, or simple linear regression techniques. In this paper, we opt for the latter with $|\theta| = |\mathcal{S}|$. With generality, of course, comes the potential for model misspecification. While importance sampling estimates are of high variance, they are also of low bias. Because we make assumptions about the number of parameters or the actual model class, the bias for direct methods is typically much larger while the variance is much smaller.

The model tuning and practical considerations are thus more complex for FQE or other direct methods than simple IPS methods. On the number of k iterations, for example, we are guaranteed to asymptotically converge as $k \to \infty$, but of course cannot run infinite iterations in practice. We found in our preliminary experiments that the number of required iterations to meet a small threshold can vary greatly with the number of trajectories and horizon length. While this is not a particular problem for smaller MDPs using simple regression as in this paper, this could become more computationally expensive in real-world settings.

2.4. Doubly Robust FQE (DR FQE)

Like importance sampling estimators, doubly robust estimation techniques (5) also predate the application to the off-policy evaluation problem (11). This hybrid method combines the low variance and (potentially) high bias of regression-based techniques with the high variance and low bias of importance sampling estimators. Doubly robust estimators have been used, for example, in dynamic treatment regimes (6) and the contextual bandit setting (7). We will first present the estimator in this contextual bandit setting since the recursive translation to the sequential setting is motivated in (11) by solving a bandit problem at each horizon $t \in [\tau]$. This is given by:

$$V_{DR} := \hat{V}(s) + \underbrace{\frac{\pi_e(a|s)}{\pi_b(a|s)}}_{\rho} \left(r - \hat{R}(s,a) \right)$$

where $\hat{R}(s,a)$ is estimated in some fashion (typically by performing regression), and $\hat{V}(s)$ is simply the expectation of this reward estimate over the randomness of the policy. This form requires several remarks. First, the specification of $\hat{R}(s,a)$ is intentionally vague: the user can specify this estimate in any desired way. Thus, while this paper explores a doubly-robust version of FQE for the sequential setting, we could have just as easily specified any other reward or state-action value estimation method here as well. Second, we can interpret this form as a sort of "gradient descent" on the value function with learning rate ρ and "gradient" on error $\left(r - \hat{R}(s,a) \right)$. In this sense, the doubly robust estimator is using the importance weights to guide the direction of update, while anchoring on the actual value estimate we obtain through whatever pre-specified method we choose. If $\hat{R}$ is well-specified, then the error term goes to zero, and if the importance weight is close to 1 (i.e. little distributional shift), then the only update is done on the error of the estimate of the reward function. We thus obtain "two shots on goal," where if either of ρ or the reward function class is properly specified, the estimator is asymptotically unbiased. In this paper, we assume that the importance weights are given to us, and therefore this particular instance of DR-FQE is guaranteed to be asymptotically unbiased. The question then becomes at what *rate* we achieve this unbiasedness.

The translation to the sequential setting only requires applying the bandit doubly robust estimator for each t. This form is given by:

$$V_{DR}^{\tau+1-t} := \hat{V}(s_t) + \rho_t \left(r_t + \gamma V_{DR}^{\tau-t} - \hat{Q}(s_t, a_t) \right)$$

Note that the estimation is now done on the $Q-$function, which is where we simply plug in our FQE estimates from the previous section.

While the specifics of the result are dependent on the underlying Markov Decision Process, (11) also shows that the doubly robust estimator achieves the Cramer-Rao lower bound on the variance of an unbiased estimator; that is, of all unbiased estimators, this is the absolute best that we can achieve. We do note, however, that a biased estimator could have a lower mean-squared error.

2.5. Experiment Hypotheses and Intuition

The ordering by which the methods were presented was intentional. From IS to DR-FQE, the methods become more sophisticated and, theoretically, better performers. Starting with importance sampling, we observed a method with unbounded variance and no guarantees on the "best-case scenario," to a doubly robust method that achieves the Cramer-Rao lower bound and combines the sophistication of direct methods with the simplicity of inverse propensity scoring methods. As such, we expect this ordering *on average*. There is, of course, no free lunch. While we expect and roughly validate the findings presented in Figure 1, we also note that there are a host of intricacies presented by the environment "knobs" that are not fully captured in previous work. By knowing the importance weights, for example, the quadratic horizon scaling may become less relevant for the IPS methods, especially if the behavior policy is "good" in the sense that it balances exploration and exploitation.

3. Environments

In this paper, we examine two environments for the benchmarking task. Each were initially chosen to separate various environment-specific aspects of the off-policy evaluation task. However, as we explore in this section, some of these initially-posited differences were less relevant than others. We also add that the methodologies presented here are model-free: while they may not necessarily transport across environments smoothly, they do not *explicitly* rely on a "good" model of the underlying environment dynamics. Still, policy exploration and environment dimensionality are highly relevant, environment-specific dependencies that can shift the ordering of method performance.

3.1. The Chain Environment

The chain environment (9) encapsulates the policy exploration problem. We receive some small reward by starting (and remaining) in the first state, but there is a much larger reward at the end of the chain in the final state. We receive no rewards by traversing the middle states, and we ensure that the discount factor γ is large enough that there is indeed a difference between the first and last states.

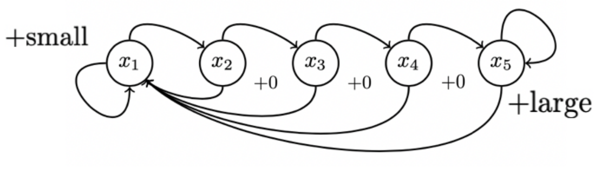


Figure 2.

The action space $\mathcal{A}$ has two elements: left or right. At the beginning and end of the chain, a left and right move, respectively, translates to staying in that first or final state. Thus, the optimal policy is quite simple and known by examination: we should always move right, but we have to make this decision at least n times in a row to obtain the large reward, so the behavior or collection policy must have explored most of the state space. If we choose the wrong action even once, we go back to the starting state. The rewards are deterministic, but there is potentially stochasticity in the transition dynamics. With probability ε, we "slip" to the opposite action chosen.

3.2. The Mixing Environment

The mixing environment, in the form presented here, more explicitly accounts for the dimensionality problem. Like the chain environment, the action space here also has two elements: "stay" or "leave," where leaving places us at state s_{i+1} from state s_i. However, there are now far more states and edges to learn than in the chain environment, and they interact in a less straightforward way. The stochasticity in this environment can also be quantified by a slippage factor ε, but we now slip to any other possible state with probability $\frac{\varepsilon}{n-1}$. There is a deterministic reward hidden at one state s_i, and all other states give no reward, always.

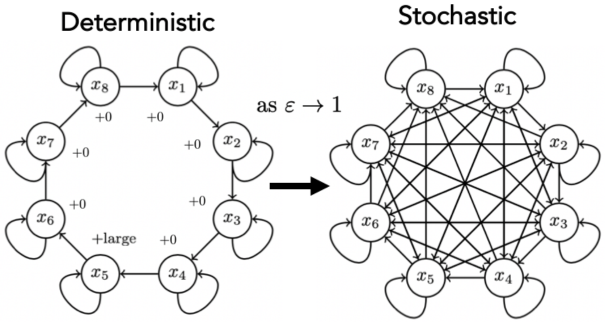


Figure 3. The mixing environment slowly becomes a fully connected graph as the slippage factor increases.

3.3. Environmental Factors

There exist many possible dimensions along which an environment can be classified. Stochasticity in the transition dynamics or reward function, dimensionality of the state space, number of trajectories, or horizon length are all highly relevant in off-policy learning. We quickly mention a preliminary experiment here with the slippage factor since we initially posited this as a measure of "difficulty" as we built our environments. However, in contrast to our initial supposition, we find almost no discernible effect in the mean-squared error of the estimated value with respect to this slippage factor ε in certain cases. We use the mean-squared error as the measure of choice for comparing performances, which we define across n experiments as:

$$\text{MSE}_{\hat{V}(s)} = \frac{1}{n}\sum_{j=1}^{n}\Big(V(s) - \hat{V}(s)\Big)^2$$

We examine this in the context of the mixing environment, since the number of edges here were hypothesized as the central difficulty. Here, we fix a small horizon $\tau = 10$ (which, by later experiments, all methods were found to be of similar performance), small policy mismatch (where we add and subtract some small positive constant $\delta \approx 0$ from a fixed behavior policy to obtain a fixed evaluation policy), and fixed number of trajectories $N = 100$. By the figure below, we see that the methods are mostly invariant under increased ε in this setting.

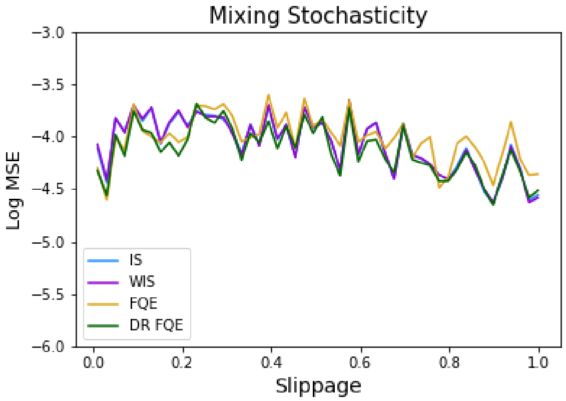


Figure 4. The slippage factor seems to have no discernible effect on method performance when the policy mismatch is small. Note that $k = 50$ experiments (or seeds) were ran for all figures reported in this paper unless otherwise stated, and we do not report the 95% confidence intervals for the sake of readability since the error is roughly negligible for this many experiments. The only sources of stochasticity across experiments are the transition probabilities and the policy selection, and we must be careful not to confuse the variance of the *experiments* with the variance of the underlying *methods*. Importance sampling, for example, has a mean square error that will be all variance (or high variance) since it is unbiased, but the variance across many experiments could be quite low.

With the benefit of hindsight, this result is expected and illuminates the sense in which these methods are "model-free." We provide a proposition in the appendix to further explore this finding in a deeper sense. For importance sampling methods, at least, this can be explained as follows. Given a starting state s_0, the probability of a certain state-action trajectory can be written as:

$$\begin{aligned} &\Pr\{a_0, s_1, a_1, \ldots, s_\tau | s_0, a_{0:\tau-1} \sim \pi\} \\ &= \pi(a_0|s_0)p(s_1|s_0, a_0)\pi(a_1|s_1) \cdot \ldots \cdot p(s_\tau|s_{\tau-1}, a_{\tau-1}) \\ &= \prod_{t=0}^{\tau-1} \pi(a_t|s_t)p(s_{t+1}|s_t, a_t) \end{aligned}$$

The importance weights for π_e and π_b can then be written as:

$$\rho = \frac{\prod_{t=0}^{\tau-1} \pi_e(a_t|s_t)p(s_{t+1}|s_t, a_t)}{\prod_{t=0}^{\tau-1} \pi_b(a_t|s_t)p(s_{t+1}|s_t, a_t)}$$

The transition probabilities therefore cancel and the importance sampling ratio is not explicitly a function of this slippage factor ε.

The experimental result also motivates the difficulty in choosing environments that actually illuminate the differences as a function of the underlying environment; that is, the chain environment and the mixing environment were not quite as different as we had initially supposed.

4. Experiments

The main focus of this paper is attempting to solve a variant of the "chicken and the egg problem": does the policy mismatch or the horizon primarily determine the method that we should choose? As we have seen in the introduction, error accumulates quadratically in the horizon regardless of how "good" the behavior policy is. However, if the policy mismatch is small, this quadratic accumulation can be manageable. How "small" must this distributional shift be for each method? Does this change the rankings of the method performances?

We first present two main experiments that later motivate a more comprehensive third experiment to explore how horizon and policy mismatch interplay for each of the four methods.

4.1. Methodology: First Experiment

In the first experiment, we fix a known (not estimated) behavior policy π_b for each of the chain and mixing environments with $\pi_b = [0.5, 0.5]$ since both environments have an action space where $|\mathcal{A}| = 2$. Note that, in general, we will often drop the state dependence in our notation for π since all policies are state-independent and completely uniform for all $s \in \mathcal{S}$ for both π_e and π_b. While not particularly realistic, this does provide a way to study the effect of a (clearly) suboptimal behavior policy, and we emphasize that the main problem of off-policy learning, in general, is the distributional shift between π_e and π_b as opposed to the actual optimality of π_b. While we would need to stray less from a close-to-optimal behavior policy, that we must stray at all is the most pressing problem.

The motivation behind this specific choice for π_b is that the behavior policy has been chosen in a completely random fashion with maximum entropy and no insight, but also explores enough of the state space to be usable.

Next, we fix the slippage factor $\varepsilon = 0.2$ for both the chain and mixing environments. As discussed in the previous section, this variable has little effect on the outcome and thus we largely ignore this in our analysis. We also fix the number of trajectories at $N = 100$ for each environment, with horizon equal to the number of states for the mixing environment and the $2N$ for the chain environment, which are $|\mathcal{S}_{mixing}| = 10$ and $|\mathcal{S}_{chain}| = 5$, respectively.

Next, to make the plots easier to read, we will define a general, two action policy as $\pi = [\delta, 1-\delta]$ for $\delta \in [0,1]$. The plots below vary this δ for π_e when we are in state s_i and choose either "leave" or "stay" for the mixing or chain environments. Since the policy only maps to two actions, we could choose either of the two actions in the plots and obtain the same performance orderings.

We measure the MSE across $n = 50$ experiments, where we measure the difference of the true value of the starting state obtained via dynamic programming (with specified, known dynamics and reward function) and the estimated value of the starting state.

4.2. Results and Discussion: First Experiment

The results in Figures 5 and 6 show that we achieve a global optimum at $\delta = 0.5$ as expected. This is where the policy mismatch is exactly zero, however we wish to choose to measure this mismatch. At $\delta = 0.5$, we could choose any of the methods as they would all perform equally well.

We also note that importance sampling estimators, outside of this small neighborhood around $\delta = 0.5$, seem to perform the worst on average as we expected. However, the other methods do seem to shift their performance ordering based in the specific environment and fixed environment parameters.

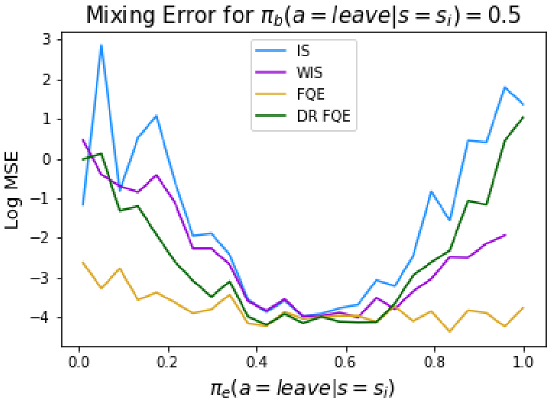


Figure 5. With $\tau = 10$, FQE seems to far outperform all other methods. The variance of the importance weights seems to shift the doubly robust method toward higher error than FQE alone.

With so many different environment parameters, the exact cause of this ordering cannot be rigorously quantified without more experiments that vary other aspects of the environment. Weighted importance sampling seems to mostly outperform fitted Q evaluation for the chain environment, but fitted Q evaluation seems to vastly outperform all other methods for the mixing environment. Knowing that error accumulates quadratically with policy mismatch over horizon τ, we hypothesize that these specific orderings are a result of the fixed horizons in addition to the actual policy mismatch. This briefly motivates our second experiment, which is a bridge to our more complete third experiment.

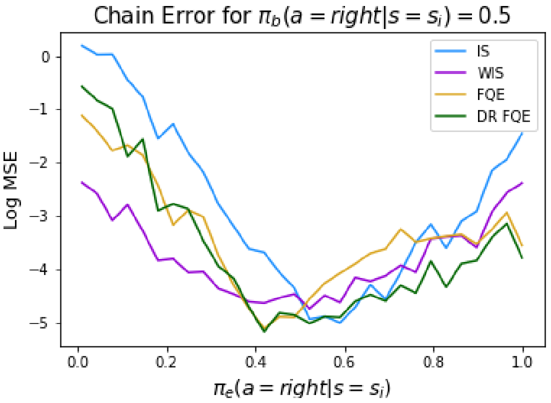


Figure 6. With $\tau = 10$ for the chain environment, FQE is less competitive. While Figure 5 appears more symmetric in its policy mismatch, Figure 6 appears less so. This is likely an artifact of the scale of the rewards at the first and last state.

4.3. Results and Discussion: Second Experiment

The second experiment shows a counterexample to a portion of the first experiment. Here, we fix a small policy mismatch for all methods, which corresponds to the $\delta \approx 0.5$ regime in Figures 5 and 6. Note that because the policies are state-independent and stationary, WIS and IS would be the same if the policy mismatch was exactly zero. We then vary the horizon to test how this ordering would scale if the horizon was longer than $\tau = 10$.

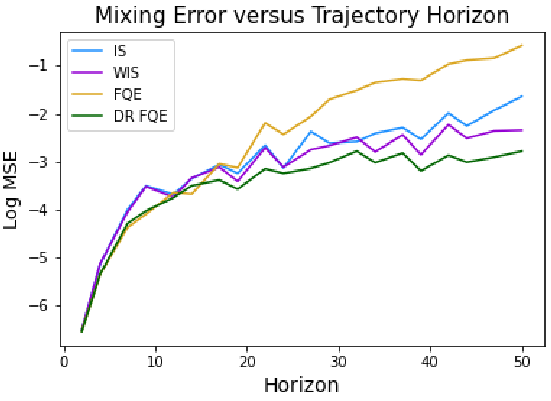
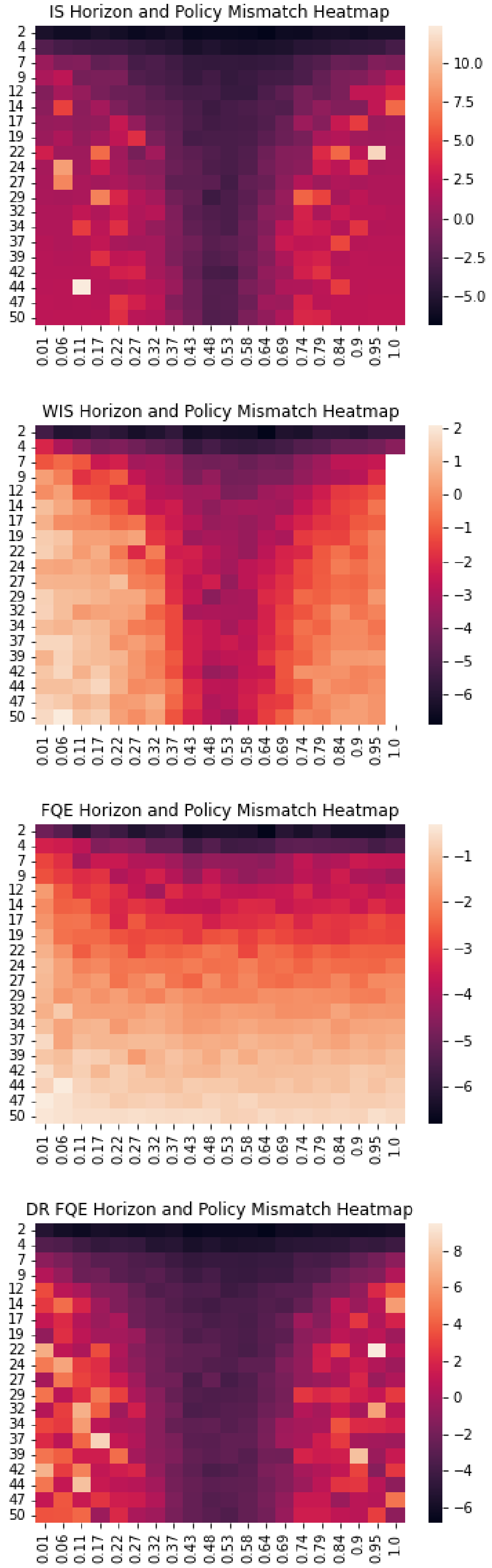


Figure 7. If the policy mismatch is about zero, then IPS weights are a number close to 1 raised to the τ^{th} power by our policy construction. In log scale, we observe the quadratic scaling in horizon we have mentioned so many times throughout this paper.

This figure shows a different story for the mixing environment. While FQE seemed to blow away the other methods at $\tau = 10$ for varying policy mismatches, Figure 7 shows that if the policy mismatch is essentially zero, then IPS methods scale more favorably with horizon. This interplay motivates our final experiment.


IS Horizon and Policy Mismatch Heatmap

WIS Horizon and Policy Mismatch Heatmap

FQE Horizon and Policy Mismatch Heatmap

DR FQE Horizon and Policy Mismatch Heatmap

Figure 8. The $y-$axis of the heatmap specifies the horizon, while the $x-$axis specifies the value of π_e for the first action (or δ as in the first experiment). The results are shown here for the mixing environment.

4.4. Results and Discussion: Third Experiment

In the first experiment, we varied the policy mismatch but not the horizon. In the second experiment, we varied the horizon but fixed a small policy mismatch. This allowed us to observe the global minima of error where the policy mismatch is zero, and to see the quadratic scaling of error with respect to horizon. However, neither of these experiments alone can inform us about the original question, which is "Should horizon or policy mismatch come first in our decision tree?"

The third experiment combines the first two experiments by varying both horizon (on the $y-$axis) and the first action behaviour (δ on the $x-$axis), fixing the behavior policy at $\pi_b = [0.5, 0.5]$ again, and then plotting the log mean-squared error as the "heat" in the heatmap. These results are shown in Figure 8, and they are highly informative.

We first note the ranges of the log MSE scales for each method. IS has an error range of 15; WIS has an error range of 8; FQE has an error range of 5; DR-FQE has an error range of 14. Across all combinations of horizon and policy mismatch, FQE is highly dependable. However, we also see that the combination of the importance weights with FQE can lead to substantially improved performance within a specific range of policy mismatches.

Perhaps the most compelling part of these heatmaps is the "width of feasibility" for the IS and DR-FQE methods. We see that for longer horizons and more pronounced mismatches, these methods quickly become untenable. However, the hybrid model performs well in the neighborhood $\delta \in (0.37, 0.64)$, while the IS estimator performs well only in the neighborhood $\delta \in (0.48, 0.53)$. By combining the IPS and direct methods, we can allow greater policy mismatch and still observe competitive performance even with long horizons. While FQE at least does not degenerate to the degree that IPS methods do with respect to horizon and policy mismatch, the bias induced by our linear model limits its ability to take advantage of small mismatches and achieve lower errors.

By these findings, we question why the policy mismatch should not dominate the horizon in the ordering of the decision tree. We typically have some degree of control over the target policy in the greater off-policy learning context. We can constrain the mismatch in some way during value iteration, but the horizon is fixed and given. Thus, we can choose which regime we exist in during our decision process. However, the width of feasibility for the policy mismatch is quite small: we can only move so far away from the behavior policy. In contrast, especially for methods like FQE, the horizon seems to have less of an effect in the limit. [IS THIS AN ARTIFACT OF THE LOG SCALE???]

5. Conclusion and Future Directions

In this paper, we explored importance sampling, fitted Q-evaluation, and variants thereof for the problem of off-policy evaluation benchmarking. The presented experimental findings challenge the

6. Appendix

Proposition 1: *Given slippage or stochasticity factor $\varepsilon > 0$ and state-independent, stationary policies π_e and π_b on environment $\mathcal{M}$, $\mathcal{M}$ can always be re-interpreted as a deterministic environment with a fixed policy mismatch defined as* $\sup_{s \in \mathcal{S}} \frac{\pi_e(a_t|s_t)}{\pi_b(a_t|s_t)}$.

Remarks: This requires not so much of a "proof" as it requires confirmation that the policies will still lie on the probability simplex regardless of interpretation. The policy mismatch is already constant, and every point attains the supremum since the policies are themselves constant and independent of state. We simply collapse the stochasticity factor into the policies, turning state-action pairs into state-action-epsilon tuples. We can then treat $\mathcal{M}$ as functionally deterministic since our policies are now just the probabilities of actually going right or left. For the case $|\mathcal{A}| = 2$ we let $\vec{\varepsilon} = [1 - \varepsilon, \varepsilon]$, for example, giving:

$$\pi_e^*(\text{left}|s_t) = (1-\varepsilon)\pi_e(\text{left}|s_t) + (\varepsilon)\pi_e(\text{right}|s_t)$$
$$\pi_e^*(\text{right}|s_t) = (1-\varepsilon)\pi_e(\text{right}|s_t) + (\varepsilon)\pi_e(\text{left}|s_t)$$

where $\mathbb{P}_e[\text{go left}] + \mathbb{P}_e[\text{go right}] = 1$. Of course, $||\vec{\varepsilon}||_1 = 1$ and $||\vec{\pi}||_1 = 1$ implies $2\vec{\varepsilon} \cdot \vec{\pi} = 1$, so we are still on the probability simplex.

$\blacksquare$

References

[1] Ross, Gordon, Bagnell. “A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning.” 2011.

[2] Voloshin, Le, Jiang, Yue. ”Empirical Study of Off-Policy Policy Evaluation for Reinforcement Learning.” 2019.https://arxiv.org/abs/1911.06854

[3]

[4] Thomas, Theocharous, Ghavamzadeh.“High Confidence Off-Policy Evaluation.” 2015.

[5]

[6] 2001 paper

[7] 2011 paper

[8] Levine, Kumar, Tucker, Fu. ”Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems.” 2020. https://arxiv.org/pdf/2005.01643.pdf

[9] Strens. “A Bayesian Framework for Reinforcement Learning.” 2000. http://ceit.aut.ac.ir/~shiry/lecture/machine-learning/papers/BRL-2000.pdf

[10] Le, Voloshin, Yue. “Batch Learning Under Constraints.” 2019. https://arxiv.org/pdf/1903.08738.pdf

[11] Jiang, Le. “Doubly Robust Off-policy Value Evaluation for Reinforcement Learning.” 2015. https://arxiv.org/pdf/1511.03722.pdf

[12] Sutton, Barto. “Reinforcement Learning: An Introduction.” 1992.

[13] Fu, Norouzi, Nachum, Tucker, Wang, Novikov, Yang, Zhang, Chen, Kumar, Paduraru, Levine, Paine. “Benchmarks for deep off-policy evaluation.”