Model Selection for Off-Policy Policy Evaluation

Yao Liu
Carnegie Mellon University
Pittsburgh PA, 15213
yaoliu@cs.cmu.edu

Philip S. Thomas
Carnegie Mellon University
Pittsburgh PA, 15213
philipt@cs.cmu.edu

Emma Brunskill
Stanford University
Stanford CA, 94305
ebrun@cs.stanford.edu

Abstract

In this work we study the off-policy policy evaluation problem, which is about how to predict the value of a policy by data from other policies. This is crucial for many applications where we can not deploy new policy directly due to safety or cost. We consider the model selection problems for a better off-policy estimators, when we have models from different sources. Traditional off-policy policy evaluation method can be divided into importance sampling estimators or model based estimators, and they respectively suffer from high variance and bias. Recent work such as doubly robust [1] and MAGIC [5], shows that we can get benefit from combining importance sampling method with model value. However they all assume that they have only one model. In case we have several different models, which is common in some complex domains, it may be hard to select the best one from them, and may lose the potential benefit from all models. we present a evidence example to show that select model by simply minimizing the notation of error in previous estimator (MAGIC) can fall into a wrong model, which suggest that selecting a best model for off-policy policy evaluation is non-trivial and worth of further exploration. We propose two new estimators of model bias and a cross validation way to help to choose a model, and shows the preliminary result.

Keywords: reinforcement learning, off-policy evaluation, model selection

Acknowledgements

1 Introduction

Off-policy evaluation is an important problem for reinforcement learning [4], where one aims to evaluate the performance of one policy with data collected by other policies. It is closely related to counterfactual reasoning problem which is important and of independent interest in statistics. Off-policy policy evaluation is also useful for many applications, where we can not evaluate our new learned policy by directly running it, because of its cost, risk, or legal concern. For example, in digital marketing domains [6], personalized curriculum recommendation [2], and customer modeling domains [7], it is necessary to evaluate the potential risk of new policy, before deploy the untested policy. Off-policy evaluation can also help online/on-policy learning such as policy gradient, since it allow us to make use of more data from historical policies to compute the gradient.

Current off-policy evaluation method can be roughly divided into two groups: importance sampling (IS) [3] and model based method. Importance sampling methods is unbiased but suffer from large variance, and model based methods have relatively small variance, but have a potential bias due to function approximation or state abstraction. Recently, some methods combining those two group of estimators have been proposed, such as doubly robust [1], weighted doubly robust, and MAGIC (Model and Guided Importance Sampling Combined) estimator [5]. Those methods try to balance the bias, and show that combining those two group of estimators can benefit off-policy evaluation.

However these methods use only one model and did not consider the case where we have several models from different source, for example different domain knowledge, experts, or mathematical modeling method. In case that we have several different models, it may be hard to select the best one from them, or may lose the potential benefit from combining multiple models. We aim to find the best way to make use of multiple models to give a better off-policy estimator. It can be choosing the best model for off-policy evaluation, or combining the models. In some domains, it also can be helpful to give interpretation of different models' ability to predict policy value.

In this paper, we present a evidence example to show that select model by simply minimizing the notation of error in the estimator (MAGIC) can fall into a wrong model. We propose several new estimators of model bias, and using cross validation to help to choose a model.

2 Approach and Preliminary Result

We firstly introduce the MAGIC estimator, since our approach is based on this estimator. Consider that we have two estimators, one (IS) is from importance sampling group defined in [5] which has importance ratio ρ and another (AM) is pure model based estimators which does not have importance ratio at all. Then we build a vector of estimators where the ith element is:

$$g^{(j)} = IS^{[0:j]} + AM^{[j+1:\infty]}$$

, which means we use IS to estimate the first j time step return and use model to estimate the others. The aim of MAGIC is find a weight vector $\mathbf{x}$ to minimize the mean square error of $\mathbf{x}^T\mathbf{g}$. Since:

$$MSE(\mathbf{x}^T\mathbf{g}) = \mathbf{x}^T cov(\mathbf{g},\mathbf{g})\mathbf{x} + \mathbf{x}^T \left(bias(\mathbf{g})^T bias(\mathbf{g})\right)\mathbf{x}$$

, they use an empirical estimation to estimate covariance matrix. As for bias part, since IS is unbiased, they use bootstrapping confidence interval as an conservative estimate of the true value, to estimate the model bias by its closest distance to the confidence interval. In MAGIC, they use weighted doubly robust as IS estimator and treat model as an oracle.

Before introduce the empirical result of our approach, we firstly introduce an toy MDP, MultiModel, as an evidence example to show the necessity of careful model selection for off-policy evaluation. This MDP is a 5-state, stochastic MDP with horizon $= 20$. Figure 1 shows reward and transition probabilities in detail. Here we start at s_0 and we have 3 action, each of them will lead to a corresponding state with probability 0.6 and lead to others with probability 0.2. If we are not in state s_0, we will be back to s_0 and receive different reward according to the current state. The behavior policy will take these action with probability softmax$(1, 0, 0)$, and the evaluation policy will take action with probability softmax$(-1, 1, 0)$. Here we assume we have 3 kind of models: The first model can observe all the state build the model. The second one will abstract s_3 and s_1 as one state. The third one will abstract s_3 and s_2 as one state. Clearly here the first one will give the unbiased estimate to the value and other models will be biased.

Consider in case that we do not know which model is better, and we build MAGIC estimators on each of them. Then the most straight forward way is choosing the one with smallest mean square error estimate, since MAGIC will also provide an error estimate of its evaluation value. In figure 2, curve Minimize MSE corresponds to this naive method minimizing empirical MSE. The plot shows that this naive model selection method will lead to a much higher error, compared with always using the correct model, which suggests we need to carefully select the model we use for MAGIC.

One way to control the bias of model selection is cross validation. We split the data into two part, the first one is used to build MAGIC in the normal way, then use the other part to estimate the empirical mean square error of the first part.

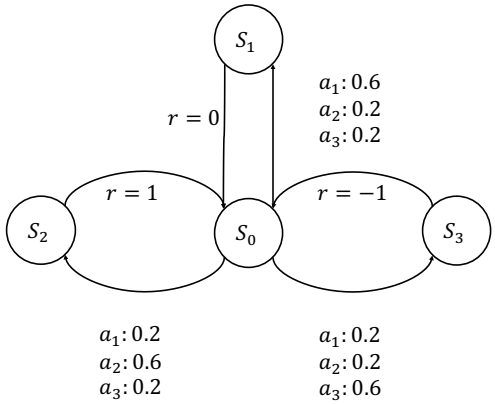

Figure 1: MultiModel MDP

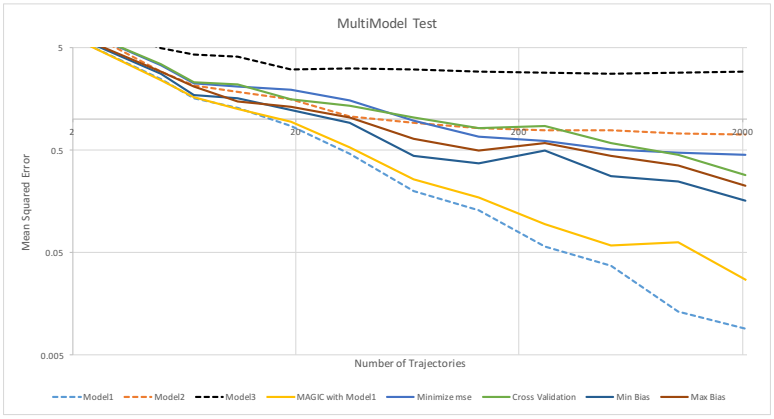

Figure 2: Mean square error of different off-policy estimators with multiple models, averaged over 128 trials.

Another change comes from the observation that the original bias estimator in MAGIC is conservative and rough, which is fine to tune the balance between IS and model estimators. However, model selection problem is more sensitive to the bias part. Recall that in MAGIC, we use bootstrapping confidence interval distance of weighted doubly robust to estimate the bias, but the weighted doubly robust itself also need a model. In MAGIC, they use the same model as AM part, which will tend to underestimate the model bias. Here we proposed two kind of bias estimators:

1. Min Bias: Use other models to build bootstrapping confidence interval, and compute the shortest distances with respect to each model. Then choose the smallest distance as bias estimator.

2. Max Bias: Build confidence intervals and compute distances in the same way. Then choose the largest distance as bias estimator.

Figure 2 shows that with cross validation and those new bias estimators, the average mean square error is better than just minimizing the mean square error estimate. However it is still not as good as always using the best model.

3 Conclusion

We have shared a evidence example to show that how models can benefit the off-policy evaluation and current approach can not make use of it, since it can not efficiently and accurately choose the best model to do off-policy evaluation. It shows that model selection for off-policy evaluation is an interesting and non-trivial problem. Our preliminary result suggest that a more careful estimator of model bias may help to choose a better model.

References

[1] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 652–661, 2016.

[2] Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. Offline policy evaluation across representations with applications to educational games. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1077–1084. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[3] Doina Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000.

[4] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[5] Philip S Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, 2016.

[6] Philip S Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *AAAI*, pages 3000–3006, 2015.

[7] Li Zhou and Emma Brunskill. Latent contextual bandits and their application to personalized recommendations for new users. *arXiv preprint arXiv:1604.06743*, 2016.