View Reviews

Paper ID	2770
Paper Title	Optimistic Policy Optimization via Multiple Importance Sampling

Reviewer #1

Questions

1. Please enter a detailed review describing the strengths and weaknesses of the submission.

The paper looks at the problem of policy search in MDPs. Policy search problems usually suffer with the problem of off-policy evaluation and have therefore been studied mostly in local search or gradient search type methods. Instead of that, this paper proposes a more bandit like framework for the same while using truncated multiple importance sampling for off-policy estimation. Since, the basic framework is that of MAB, they show regret bounds of order \sqrt{T}. Now, I'll go in detail about each step.

For policy search, the authors don't consider the usual case where one optimizes for the best policy in a class. Rather, the authors consider parameter based policy optimization which draws the policy's parameters from a distribution which is parameterized by the learnt variable \theta. Therefore, this removes the exponential variance problem induced due to long horizon for importance weighting methods. This was recently also studied by Metelli et. al. who also proposed an importance sampling based approach but in a gradient descent style algorithm. Augmenting their results, the authors in this paper, study multiple importance sampling which works with multiple distributions by taking the empirical sum by distribution dependent mixture weights. Choosing the right set of mixture weights can have an impact on the variance of the resulting estimator. A bound on the variance of this estimator is provided by using the Renyi-divergence between the evaluation distribution and the mixture of the proposals. The authors write the analysis and text in a way which gives the impression that, the variance of the weighted estimator is the same as the variance of an IS estimator where the proposal itself is the mixture. It should be noted that this is not true and it is just an upper bound which the authors can show and it can turn out to be a crude upper bound even when ignoring any coverage constraints on the set of distributions. This bound on the variance is shown in Lemma 1 and is a straightforward modification of the result in Metelli et. al. Now, as the IS weights in this estimator can still get very large, the authors show the bias and variance bounds for the case where importance weights are truncated at a value M. To do that, the assumption needed is over the (1+\epsilon)-Renyi divergence between the distributions. Showing the bias-variance results again uses the same arguments as before with the added truncation step. A nice result following this is a finite sample concentration result for the estimator which uses an adaptively chosen truncation threshold depending on the sample size. Again, proving this uses simple tools like the Bernstein concentration bound. However, this is a nice result showing the way in which one can control the confidence width around the estimator. Till now, the presented results are nice and have good empirical appeal too. However, the policy search spin-up with this asks for additional scrutiny.

The authors consider, using a MAB framework by considering the set of all possible \theta values as arms. The UCB indices are built using the result from Theorem 1 in the paper. This is quite important as the estimators now use the same data and therefore, have to utilize any correlation present between the arms. Using a common estimator provides one way to do this.
Mistakes: In equation 17, the Renyi divergence terms use p_{x_t} in the second term which should be p_x. Similarly, assumption 1 bounds the Renyi divergence for the sequence of hyperpolicies chosen by the learner. Controlling this means one has to control the divergence for all policies in the set which seems quite difficult to achieve. Also, the authors have mistakenly not removed the TODO where a way to control this is to be proposed. This is a strong desiderate for me. Further, it is important to see for the experiments whether the UCB indices actually turn out to be non-vacuous numbers.
Now, there are various caveats with this:
1. Evaluating the \arg max over \Theta for the UCB index cannot be solved for a non-finite set as the problem is

non-convex and non-differentiable which the authors also mention.
2. Even when one uses discretization or has a finite set \Theta, evaluating this requires evaluating the Renyi-divergence between the distributions. This divergence is calculated over the trajectory distribution induced by the hyper-policy and reduces to Renyi-divergence between the parameter distributions. This is not tractable for non Gaussian examples to my knowledge. For example, evaluating the divergence for a NN parametrized is difficult as the distribution is sensitive to small changes in the parameters. Further, it is extremely important because the Renyi divergence has to be controlled in some way as shown in Assumption 1. Similarly, this is hard to show for a large class of methods in Lemma 3.

Given these assumptions and the concentration results, the regret bounds follow through with the usual machinery. As expected, for the discretization case, one needs to show the Lipschitzness property of the reward wrt \theta. Also, the computation time for discretization is exponential in the dimension of the space.

Related work: I think it can use a little addition about work in parameter based policy search algorithms and work done in statistics and importance sampling as it forms the basis of their work.

Experiments: Firstly, the paragraph in section 8.1 is repeated. The part of the experiment where different MAB approaches are compared is not really reflective of the papers contribution and probably should be compared with traditional policy search approaches to see how they compare in regret evaluation. Formulating this as a MAB problem is something which I'm apprehensive about in the first place. Since, PGPE gives better results, I think, even for regret they will give better values. As GPUCB performs better than OPTIMIST, I'm further intrigued if this shows the inadequacy of the PBPO approach. Similar arguments follow for the mountain car experiments.

As such, I find the initial set of results about truncated MIS interesting but at the same time find the MAB formulation to be interesting as a theoretical exercise. I conclude with two questions:
1. How can the authors justify assumption 1 made for the regret analysis. How bad can the regret bounds be without that assumption?
2. Can any of this be extended to non-Gaussian hyperpolicies wrt to computing the Renyi-divergences, solving the optimization problem in line 5, discretization and maintaining assumption 1.

2. Please provide an overall score for the submission.
Weak Reject: Borderline, tending to reject

3. Please enter a 2-3 sentence summary of your review explaining your overall score.
I find the initial set of results about truncated MIS interesting but at the same time find the MAB formulation to be interesting as a theoretical exercise only. It is difficult to justify the assumptions made in this section and the results don't seem particularly useful for non-Gaussian hyper-policies. Further justification for the queries in the review might be helpful.

5. Please rate your confidence in the score assigned.
High: Reviewer has understood the main arguments in the paper, and has made high level checks of the proofs.

Reviewer #2

Questions

1. Please enter a detailed review describing the strengths and weaknesses of the submission.
Major comments:
1. This paper casts policy search problem as multi armed bandit problem. With multiple importance sampling method and weight truncation, upper confidence bound and regret bound are proved.
2. The order regret bound seems to be the same as previous work.
3. OPTMIST and OPTMIST 2 need to draw a whole trajectory. The optimization problem to select arm has high complexity, and is usually non-convex and non-differentiable.
4. In the experiment, OPTIMIST 2 does not beat the policy gradient methods in the long run. Improvement on the

experiment would make the paper more convincing.
5. In Figure 2, it would be better to conduct more trials to avoid randomness.

Minor comments:
1. Redundant text in Line 408-416.

2. Please provide an overall score for the submission.
Weak Accept: Borderline, tending to accept

3. Please enter a 2-3 sentence summary of your review explaining your overall score.
Overall, this paper is clearly written and technically sound. But the experiments can be improved.

5. Please rate your confidence in the score assigned.
Medium: Reviewer has understood the main points in the paper, but skipped the proofs and technical details.

Reviewer #3

Questions

1. Please enter a detailed review describing the strengths and weaknesses of the submission.

The authors applied advanced-modern importance sampling schemes to policy search applications. I think that the paper contains interesting material and is well-written in general. However, some parts are quite obscure (the explanation should be improved).

2. Please provide an overall score for the submission.
Weak Accept: Borderline, tending to accept

3. Please enter a 2-3 sentence summary of your review explaining your overall score.
I think that the paper contains interesting material and is well-written in general. I have some suggestions to improve the quality of the manuscript.

- There are some typos; for instance, see sentence at page 4, lines 178-182, starting with “This models…”

- Section 2.1 should improved is not completely clear and well-written, in my opinion.

I have also a couple of observation regarding the state-of-the-art discussion:

- Please, take into account the clipping technique describes in (Ionides, 2008) has been also proposed and used in Population Monte Carlo context, in

E. Koblents, and J. Miguez. A population Monte Carlo scheme with transformed weights and its application to stochastic kinetic models, Statistics and Computing, 25 (2), 407-425, 2015.

- Moreover, The Multiple Importance Sampling becomes costly when the number of proposal (N) grows; efficiently alternatives are studied in

V. Elvira, L. Martino, D. Luengo, M. Bugallo, "Efficient Multiple Importance Sampling Estimators", IEEE Signal Processing Letters, Volume 22, Issue 10, Pages: 1757-1761, 2015

5. Please rate your confidence in the score assigned.
Medium: Reviewer has understood the main points in the paper, but skipped the proofs and technical details.

Reviewer #4

Questions

1. Please enter a detailed review describing the strengths and weaknesses of the submission.

SUMMARY

This paper tackles the problem of policy optimization in reinforcement learning using an approach based on multi-armed bandits and optimism in the face of uncertainty. The goal is to identify the best policy from a parameterized class, in terms of the expected return of the policy over an episode. The idea is to treat each possible value of the policy parameters as an arm in a multi-armed bandit problem, with the reward being the evaluated return of the policy over one episode. Moreover, the sampled reward of one arm (policy) can be used to estimate the rewards of other arms using importance weights; in this sense the arms' rewards are mutually informative. Finally, by establishing upper confidence bounds on these estimates of different policies' values, an optimistic (UCB-like) algorithm can be used to optimize the policy parameters.

The authors combines a few different ideas. Firstly, after a number of episodes, the algorithm has access to the sampled return under several different policies. Each of these samples can be used with an importance weight to estimate the return of each potential new policy. Furthermore, a weighted combination of these importance weighted estimates is also an unbiased estimate of the return of the new policy; the weights can even depend on each sample. As a heuristic to minimize variance; each sample is weighted proportionally to the importance ratio of that sample.

The second idea is to truncate the importance weights to control the variance of this "multiple importance sampling" estimator. This, of course, introduces bias, but the truncation is done adaptively to optimize the bias-variance tradeoff. Finally, the concentration properties of the estimator are exploited to derive high-probability confidence bounds on the estimates. The adaptive truncation threshold as well as the confidence interval require knowledge of the Renyi divergence between the target policy and the mixture of all previously evaluated policies; the tractability of computing this Renyi divergence largely dictates the implementability of the algorithm in any particular setting.

The algorithms have sublinear regret bounds on the order of T^(1/(1+epsilon)), as long as the policies have finite Renyi divergences of order (1+epsilon); the authors focus on the epsilon=2 case which gives sqrt(T) regret. The paper also includes variants of the algorithm that operate on discretizations of the policy parameter space, along with corresponding regret bounds. Finally, there are evaluations of this algorithm on illustrative toy domains (LQG and Mountain Car).

CLARITY

I found the paper well motivated, clearly laid out and easy to follow.

SIGNIFICANCE & NOVELTY

This is the first work that I'm aware of that directly applies a UCB-type algorithm to the MAB formulation of the policy optimization problem while taking advantage of the shared information between the policies (arms). This, of course, required the development of an appropriate estimator for the return of each policy: having low variance while taking advantage of all the available off-policy returns. I think both of these components are significant contributions.

CORRECTNESS

I checked the regret analysis of the discrete case as well as the other ancillary proofs and found no errors. I only lightly checked the proofs for the other regret bounds.

There were a few minor typos:

* line 178, column 1: "mining" should be "minimizing"?

* line 426, column 2: "empirica" should be "empirical"

* lines 641 and 713: the supremum norm of f should be squared

2. Please provide an overall score for the submission.

Accept: Good paper

3. Please enter a 2-3 sentence summary of your review explaining your overall score.

I believe this work is interesting and novel and should be accepted for the conference. Furthermore, it is possible that the individual techniques developed (e.g. using ideas from heavy-tailed bandits to provide confidence bounds for importance-weighted estimators) could find use in other settings. There are, of course, various open questions but this work provides a useful starting point.

5. Please rate your confidence in the score assigned.

High: Reviewer has understood the main arguments in the paper, and has made high level checks of the proofs.