ICML 2018
International Conference on Machine Learning 2018
July 10 to July 15 2018, Stockholm, Sweden

Reviews For Paper

Paper ID	925
Title	Stochastic Variance-Reduced Policy Gradient

Masked Reviewer ID: Assigned_Reviewer_1
Review:

Question	
Summary of the paper (Summarize the main claims/contributions of the paper.)	This paper proposes a natural, but non-trivial, application of ideas from the optimization field to reinforcement learning. In particular, it is investigated wh be sped up. The theoretical contributions seem sound. It is unclear how important these results are in practice, but I welcome the effort to investigate these issues. The empirical part of the paper is weak. The new algorithm is extended with multiple heuristics, which are not covered by the theory, which is apparently then no ablation or parameter studies are shown to highlight the importance of these choices. Also, this severely limits any interpretation on the value of th Additionally, the results are weak and barely seem to beat the baseline, which itself seems much weaker than earlier published results. I would encourage the authors to first show that the algorithm works empirically, in the sense that it improves over vanilla policy-gradient algorithms, and heuristics and tricks to make it work better, while showing the reader how much each addition changes the performance (e.g., through an ablation study). be done on simple problems, although the chosen problems (e.g., Cart-Pole) arguably are already quite toy, so these might be quite suitable for such exper As it stands, the paper is interesting, but it is completely unclear whether the proposed algorithmic contributions have any real practical benefits. That said, I think the paper would be of sufficient interest to the community, and therefore I rate it as a "weak accept".
Clarity (Assess the clarity of the presentation and reproducibility of the results.)	Above Average
Significance (Does the paper contribute a major breakthrough or an incremental advance?)	Below Average
Correctness (Is the paper technically correct?)	Paper is technically correct
Overall Rating	Weak accept
Detailed comments. (Explain the basis for your ratings while providing constructive feedback.)	*** Major comments/question: Lemma 3.1 is stated, but not proven, also not in the supplement? It also seems unlikely: why wouldn’t the variance depend on the mini-batch size? ** On the experiments: Performance on the Swimmer task seems quite poor, compared to that reported in, say, “Benchmarking Deep Reinforcement Learning for Continuous Con the authors prefer to call G(PO)MDP) is listed as obtaining returns >90. This makes me doubt the quality of the used baseline. Why are no learning curves shown for the Half-Cheetah? The newly proposed method does not seem to fare that well when compared to TRPO, and other algorithms. Why not report your own values AND Duan et al. values in all cases? There is now a weird mismatch, where Duan et al.'s results are only use in one place place where the new algorithm beats the results listed there. There is some ambiguity about which version of Cart-Pole is used, but it seems to be the one higher scores? *** Minor comments: P2, L070, C1: “in Section 3 propose” -> “in Section 3 we propose” P2, L096, C2: It’s helpful to define REINFORCE and G(PO)MDP. P3, L111, C1: “even them” -> “even they” “Reducing their effectiveness” -> this is unclear. Baselines can really help, even if estimated from data. Also, besides priors, data is all we really have — the data. L265, C1: “what done” -> “what is done”
Reviewer confidence	Reviewer is an expert

Masked Reviewer ID: Assigned_Reviewer_2
Review:

Question	
Summary of the paper (Summarize the main claims/contributions of the paper.)	This paper considers the problem of on-policy control of MDPs using policy gradient methods. A popular line of research in supervised learning is to opti smooth functions and SVRG is an algorithm that comes with stronger convergence guarantees as compared to regular SGD for this setting. The authors at problem of control in MDPs and derive a non-asymptotic rate result for SVRG+PG by following the technique of [Reddi et al. 2016]. In contrast to the S additional challenges in applying an idea like SVRG, as the authors note. The work combines a few existing ideas in the RL toolkit, such as policy gradien estimates from REINFORCE/GPMDP), importance weighting together with the SVRG principle, which is to thrown in a zero-mean term that involves the point and employ epochs for computational efficiency.
Clarity (Assess the clarity of the presentation and reproducibility of the results.)	Above Average
Significance (Does the paper contribute a major breakthrough or an incremental advance?)	Above Average
Correctness (Is the paper technically correct?)	Paper is technically correct
Overall Rating	Weak accept
Detailed comments. (Explain the basis for your ratings while providing constructive feedback.)	The authors tackle an important problem and take a shot at improving the rate of policy gradient type algorithms using SVRG. I appreciate the effort auth technical result, but the bound in Thm 4.4 remains opaque (at least to me). In particular, I cannot infer about the optimal choices of a host of parameters th comments below justify this view and I will look forward to authors response before deciding to vouch for acceptance or not. On Thm 4.4: For a non-asymptotic rate result to be useful, in my opinion, there should be sufficient hints about the various parameter choices and unfortu the version stated in the main paper) doesn't provide much information to infer optimal parameter choice. To elaborate, 1) How should m and S be chosen? Assumption 4.3 would constrain "m", but Im not sure if there aren't other constraints. 2) How to choose step-size \alpha? A set of complicated constraints involving c_t and beta_t are available in Lemma B.8, however, the choice for alpha_t

	on L is believable, since it common in smooth optimization in SL settings. The rest of the constraints on \beta_t,c_t seem to mirror those in [Reddi et al. 2 implementation advice. 3) Is \theta^* a global minimizer of (1)? If yes, it should be assumed/stated upfront. 4) Is theta_0 the same as \tilde\theta^0? 5) Under Assumptions 4.1 and 4.2, if one employs a non-SVRG plain PG, is the rate weaker? Thm 4.4 would be bolstered if one a regular PG variant, ma 6) The gradient estimates assume a horizon of H and the effect of a particular choice for H on the bound in Thm 4.4 ain't clear to me. On experiments: 1) Was the comparison with regular PG fair in terms the # of sample trajectories used? SVRG+PG obtains N trajectories at the beginning of each epoch in minibatch. 2) Keeping the minibatch size constant and not increasing as the iterations progress: as Thm 4.4 suggest, this may be suboptimal 3) The average reward in the Cart-pole task is of the order 10^3 and so, running upto 100 steps only with a discount that is very close to 1 is surprising. 4) How was \alpha chosen? Minor: Doesn't the policy gradient theorem require the existence of stationary distribution equivalents for the discounted setting?
Reviewer confidence	Reviewer is an expert

Masked Reviewer ID: Assigned_Reviewer_3

Review:

Question	
Summary of the paper (Summarize the main claims/contributions of the paper.)	The paper presents an application of the finite-sum variance reduction technique of Johnson & Zhang 2013 (SVRG) to a considerably harder problem of policy gradient estimation. - lifting the technique from supervised learning given finite training sets to policy-gradient learning over infinite sequences - dropping the strong convexity assumption for a weaker Lipschitz smoothness assumption - dealing with non-stationary of the optimization problem
Clarity (Assess the clarity of the presentation and reproducibility of the results.)	Excellent (Easy to follow)
Significance (Does the paper contribute a major breakthrough or an incremental advance?)	Excellent (substantial, novel contribution)
Correctness (Is the paper technically correct?)	Paper is technically correct
Overall Rating	Strong accept
Detailed comments. (Explain the basis for your ratings while providing constructive feedback.)	The paper is very well written and presents excellent solutions to the mentioned problems. I especially like the idea of approximating the full gradient that is computed over a finite SAG/SAGA/SVRG approaches by a Monte Carlo estimation of an expectation, and showing how the size of this sample plays into the convergence rate. The same goes for the mir convergence rate. I have to admit that I did not have the time to go over all details of the proof, but I have a few remarks on related work that might be integrated to make the work even better. For ex adaption of the minibatch size has been presented by Byrd, Chin, Nocedal, Wu. 2012. Sample size selection in optimization methods for machine learning (https://link.springer.con 0572-5). A technique to bound the variance of the importance weight has been introduced as "clipping" by Ionides. 2008. Truncated Importance Sampling (https://www.google.de/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwjhyMya7P3ZAhWMiKYKHUq6AMcQFggpMAA&url=http%3A%2F%2Fwww.stat.lsa.umich.edu revised.pdf&usg=AOvVaw3L3scwHDutltwhVTKZlsbT) A related paper that lifts MISO to objectives that are expectations has been presented by Bietti & Mairal. 2016. Stochastic Optimization with Variance Reduction for Infinite Datase (https://arxiv.org/abs/1610.00970).
Reviewer confidence	Reviewer is knowledgeable