Review-33

Variational Policy Gradient Method for Reinforcement Learning with General Utilities

Application of general utilities in Reinforcement Learning (RL) has gained attention as a result of the increasing complexity of tasks. This generality often invalidates the Bellman equation resulting in a setup where task rewards and value functions fail. This leaves one to rethink policy search in light of general utility functions of state-action occupancy measure. To this end, the work presents Variational Policy Gradients which extend the Policy Gradient Theorem towards well-defined general concave utility measures. The proposed framework models parameterized policy gradient as the solution to a stochastic saddle point problem involving Fenchel dual of utility measures. Variational policy gradients globally converge to the optimal policy of general objective with the imprroved order of $O(1/t)$ as a result of hidden convexity.

Utilization of general utilities often invalidates the usage of Dynamic Programming which renders Bellman equations and value functions of little significance. The policy gradient theorem allows tractable policy search in the case of task rewards. Motivated by this insight, the work generalises policy gradient theorem for concave utility functions in the case of parameterized policies. Based on gradient of Fenchel dual, Variational Policy Gradient theorem formulates the problem of optimal occupancy measure as a stochastic saddle point problem. Estimation of gradient is carried out using truncating trajectories which gives rise to sample-average approximations of policy gradient. The scheme further derives a bound on the policy gradient estimates which is an improvement over previous bounds in literature. Variational policy gradients globally converge to the optimal policy when applied in the policy gradient ascent setting. This is proved by showing that the optimization problem has no spurious extrema despite its nonconvexity in the case of policy parameterization. Generalization of this result towards stationary policies is carried out by leveraging hidden convexity which demonstrates a bijection between parameterized policies and occupancy measures.

Suitability of variational policy gradient scheme is demonstrated by the increasing cosine similarity between variational and true policy gradients. The scheme is additionally found suitable on maximum entropy exploration utility. On the other hand, theoretical assumptions required for generalization of policy gradients may be further refined. Convergence properties of variational policy gradients heavily rely on the Lipschitz assumption which do not necessarily translate to the practical setting of policy parameterization. A more rigorous analysis would have to do away with bounding the policy gradient and seek alternatives to bound this intractable expression. Moreover, the bound on convergence is proved theoretically and empirical evaluations of its accuracy are found lacking. Validating this result in practice could be an interesting direction for future work.

The work generalizes policy gradients in the case of general utilities using the theoretical framework of saddle point problem. The scheme presents two new directions for future work. Firstly, doind away with the Lipschitz assumption for bounding the intractable policy gradient motivates alternatives such as the usage of likelihood ratio trick. And secondly, empirical validation of convergence properties could be an interesting avenue from the optimization perspective.

Utilization of general utility measures often hampers the usage of dynamic programming and value functions in RL. To this end, the work presents variational policy gradient theorem for general utilities which is a generalization of the policy gradient result. The scheme models policy search as a saddle point problem making use of state-action occupancy measures. Variational policy gradients are found to converge to the stationary optimal policy even in the case of nonconvex environment settings. Furthermore, the algorithm has a bounded $O(1/t)$ convergence rate.