Balloon Estimators for Improving and Scaling the Nonparametric Off-Policy Policy Gradient

Project Manager
Joao Carvalho

Principal Investigator
Prof. Jan Peters (PhD)

Project Term
2020 - 2020

Clusters
Lichtenberg Cluster Darmstadt

Software
PyTorch

Institute
Intelligent Autonomous Systems

University
Technische Universität Darmstadt

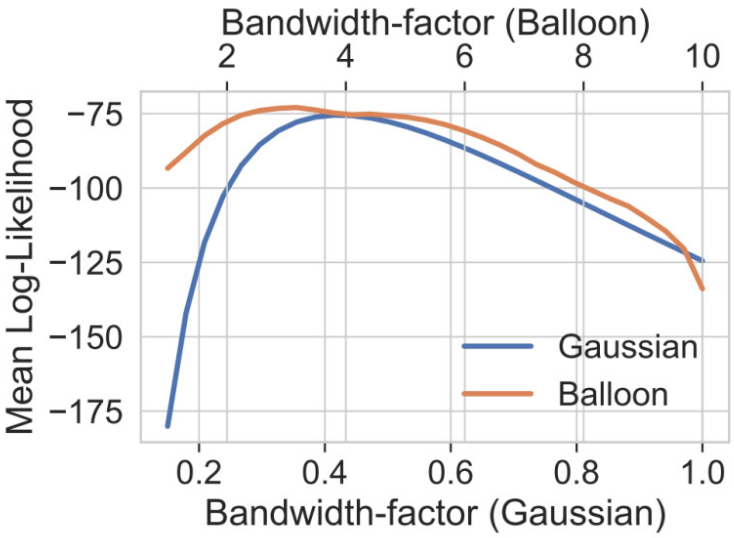


Figure 1: Analysis of mean log-likelihood for different bandwidth factors for Balloon estimator and Gaussian (bandwidth). The plot shows that the Balloon estimator achieves higher likelihood across a range of bandwidths.

Introduction

The Nonparametric Off-Policy Policy Gradient (NOPG) is a policy gradient algorithm to solve reinforcement learning tasks in continuous state and action spaces focusing on low sample complexity. Using a small number of samples is important in robotics because of the difficult to obtain realworld data due to time constraints and hardware wear and tear. NOPG solves the sample efficiency problem by using nonparametric regression and kernel density estimation methods to model the reward and state transition functions. In previous work, only fixed bandwidth Gaussian kernels were used. In this work, we investigate using an adaptive bandwidth estimator - the balloon estimator. This estimator has higher performance in estimating sparse and multi-modal data in comparison to Gaussian kernel density estimation. We provide a proof of concept for using Balloon Estimators with NOPG and compare the results and performance to Gaussian KDE in the mountain car task.

Methods

In our research we used different nonparametric kernel density estimation techniques to improve the NOPG algorithm. NOPG is a method that optimizes a parametrized policy using offline and off-policy data, either collected with an expert agent or through

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN
TECHNISCHE UNIVERSITÄT DARMSTADT
UNIKASSEL VERSITÄT
Philipps Universität Marburg

human
demonstrations. In the original formulation, to estimate the transition and reward functions with nonparametric regression, gaussian kernel density estimators with a fixed bandwidth were used. This bandwidth is an hyperparameter that can be optimized by maximizing the model loglikelihood with respect to the bandwidth, e.g. using gradient ascent. Balloon estimators instead consider a variable bandwidth that is dependent on the point at test time, i.e. not known before hand during training and simply inferred at test time for each single desired point.

Results

Our experiments show that the log-likelihood of the Balloon estimator is higher than that of the Gaussian estimator, and therefore we chose this method to query points of the transition and reward function not present in the training dataset. Due to a better model (measured by the loglikelihood), we obtain a better sum of discounted rewards, which is the original objective of a reinforcement learning agent. Therefore, our experiment results confirm that having a better model is crucial for the performance of the NOPG algorithm.

Discussion

We implemented the adaptive bandwidth Balloon kernel density estimation variant, which is able to optimize sub-optimal trajectories for the Nonparametric Off-Policy Policy Gradient. We compared the results to a baseline using the fixed bandwidth Gaussian kernel density estimation, used in the original NOPG algorithm. We showed that the Balloon estimator is able solve the mountain car environment on par with the Gaussian estimator and even achieving a better performance result, measured with the sum of discounted rewards. Choosing a good bandwidth factor is crucial to a good density estimation, both for Balloon and Gaussian estimators, and thus should be treated as an hyperparameter and optimized, for instance, by maximizing the loglikelihood of the transition and rewards models (e.g. via gradient ascent).

Outlook

As a future work, an interesting direction would be to further analyze the influence of the bandwidth selection process in the whole optimization of the reinforcement learning agent policy.

Last Update: 2022-06-24 11:24