Seminar Paper Outline: Variance Reduction For Reinforcement Learning In Input-Driven Environments

Stefan Werner

Summer term 2020

Abstract: The seminar paper's abstract should give a broad overview of input-driven environment, why difficulties specific to such environments may arise for common policy gradient methods and summarize the main contributions of the considered paper.

- Include network function virtualization as an example for an input-driven environments in the scope of resource scheduling (particularly relevant to this project group compared to e.g. traffic flow)
- Mention that common baselines for policy gradient methods do not successfully reduce variance in input-driven environments (they can be improved upon)
- List the main contributions of the seminar paper:
  - they formalize input-driven environments as a partially-observed MDPs
  - they derive an optimal (input-dependent) baseline to reduce the variance of policy gradient methods in input-driven environments
  - they propose two heuristics to approximate such an optimal baseline, where particularly the Meta-Learning based baseline demonstrates strong performance
  - (mention that such input-dependent baselines do not introduce further bias)?

1 Introduction

The introduction should elaborate on some aspects mentioned in the abstract and particularly discuss the following points:

- List relevant fields from practice or the literature where input process naturally arise under the met assumptions (i.e. the stochastic input process should be independent from taken actions)
- Specifically elaborate on the NFV resource scheduling problem discussed in the first three seminar papers. This should involve:
  - A descriptive figure that visualizes the considered NFV setting, the influence of the (stochastic) input process on the resource scheduling problem and the decoupling of the input process and the problem's state
  - An example that illustrates why the general formulation of an MDP does not suffice to describe input-driven environments, i.e. give an example that shows that an inferior policy can accumulate more reward than an otherwise optimal policy when confronted with an stochastic input process. More precisely, compare a well performing baseline from the literature against a naive approach in a simplified NFV setting and argue that the naive approach may accumulate more reward if the stochastic input process favors

it by chance.

  - Argue that the former phenomenon may introduce high variance in an intuitive manner. Specifically argue that attributing rewards to 'good actions' is harder in such a complex setting.

- Elaborate on why naive policy gradient methods commonly are subject to high variance, hence emphasize the necessity of relating estimates to a baseline by arguing how suitable baselines may reduce variance:
  - Demonstrate high variance of updates using the *REINFORCE* example. REINFORCE consideres Monte Carlo updates which, in turn, dependent on a long sequence of random components. Hence, such returns are subject to high variance, since a larger number of random variables implies higher variance (exacerbated by the fact that that rewards are not independent from taken actions, visited states, etc.)
  - The former point could be illustrated analogous to the example given in the DRL lecture Berkeley (CS294). Therefore, refer to the course of sampled trajectories in order to illustrate the variance concept for policy gradient methods.
  - State the effects of high variance samples (e.g. computationally expensive → many samples are required for convergence, which is infeasible for large-scale problems)
  - Argue that baselines may diminish this problem in an intuitive manner (e.g. if we take the state-value function as our baseline, we successfully distinguish the cases 1. highly rewarding state where any action generates high reward and 2. the rewards was due to advantageous action)
  - Argue why common baselines fall short on reducing variance in input-driven environments. Consequently imply the need for an adaptation of such baselines and elaborate on some necessary properties of baselines in such settings (i.e. must respect input-behavior, should be bias-free)

2 Problem Formulation and Setting

While this seminar paper should give an overview to primary concepts of the paper, the considered paper is based on some key mathematical derivations and thus should be included. Therefore, I intend to include the formal representation of the considered setting and its respective extension to include a stochastic input-process, as well as the derivation of the optimal input-dependent baseline.

QUESTION: how should we deal with mathematical formalisms in general? My assigned paper involves some key derivations and to some extend it would be easier to point out specific properties if I may adopt the paper's notation.

This section should elaborate on the following aspects:

- Formal definition of the problem setting (I will skip case 1 of the definition for whom the authors conclude that the problem can be restates as a fully-observable MDP). Refer back to the introductory example and explain the met assumptions and processes while giving an intuitive interpretation (if possible).

- Restate theorem 2 of the paper (optimal input-dependent baseline). The paper does not provide an insightful, while intuitive interpretation of the derived optimal baseline. Contributing an intuitive explanation (if possible) would therefore be highly beneficial to the reader.

- Discuss why learning such an optimal baseline is usually not practically feasible (not entirely clear to me yet). Consequently imply that we must estimate an input-dependent baseline.

3 Heuristic Estimation of Input-Dependent Baselines

This section should describe what methodologies the authors follow to estimate beneficial input-dependent baselines. Specifically, I intend to mostly omit the paper's discussion regarding a sequence learning approach to input behavior.

1. Describe the multi-value-network approach to estimating variance reduction baselines. The paper mostly neglects an intuitive argumentation in favor of this approach and how replaying input sequences may diminish problems arising for input-oblivious baselines.

2. Argue that this simple heuristic is particularly useful due to its simplicity and effectiveness.

3. **QUESTION:** should I formally introduce the paper's Multi-value-network approach?

4 Improving Efficiency via Meta-Learning

This section should introduce Meta-Learning and how it relates to this topic, i.e. Meta-Learning allows for an effective adaptation of previously beneficial input-dependent baselines to the current task. Hence, Meta-Learning rapidly provides a beneficial instantiation of the input-dependent baseline through an adaptation from previous experience (learning about learning!). Specifically elaborate on the methodology of MAML which is an integral part of the paper's approach to rapidly infer beneficial input-dependent baselines for a new task.

- Description of key insights from Meta-Learning

- Methodology of MAML and how their approach relates to the paper's algorithm

- Explanation of the paper's Meta-Learning approach to learn input-dependent baselines.

- **QUESTION:** should I formally introduce the paper's Meta-Learning based approach?

5 Conclusion & Future Work

Since restating the experimental evaluation and empirical analysis of the paper enables no additional insights to key concepts, this section should *briefly* summarize the results (e.g. the authors determine through an empirical evaluation that their Meta-Learning based approach outperforms their Multi-value-network approach, however, nevertheless find that accounting for input-dependent behavior is beneficial in all experiments). Therefore, this section should accomplish the following:

- Summarize key results of the paper's empirical evaluation

- Recap key insights of the author's approach to accounting for input-driven aspects of the environment

- If possible give an outlook on how to improve / adapt the paper's results

- **QUESTION:** should I include a dedicated section which discusses potential future work, adaptations, shortcomings of the paper (if I can think of such an improvement)?