On Adaptivity and Safety in Sequential Decision Making

Sapana Chaudhary
Texas A&M University, College Station, TX, USA
sapanac@tamu.edu

Abstract

Sequential decision making is an important field in machine learning, encompassing techniques such as online optimization, structured bandits, and reinforcement learning, which have numerous applications such as recommendation systems, online advertising, conversational agents, and robot learning. However, two key challenges face real-world sequential decision making: the need for adaptable models and the need for safety during both learning and execution. Adaptability refers to the ability of a model to quickly adapt to new and diverse environments, which is especially challenging in environments where feedback is sparse. To address this challenge, we propose using meta reinforcement learning with sub-optimal demonstration data. Safety is also critical in real-world sequential decision making. A model that adheres to safety requirements can avoid dangerous outcomes and ensure the safety of humans and other agents in the environment. We propose an approach based on online convex optimization that ensures safety at every time step. Addressing these challenges can lead to the development of more robust, safe, and adaptable AI systems that can perform a wide range of tasks and operate in a variety of environments.

1 Introduction

Sequential decision making systems are all around us, from recommendation systems used by Amazon, Netflix, *etc* for personalized recommendations to online advertising done on platforms like Google or Bing. Even robot learning through environment or simulator access can also be cast as sequential decision making problem. More recently, conversational agents like ChatGPT employ sequential decision making to understand the context and the history of a particular conversation to provide more accurate responses.

Particular instances of decision making algorithms are employed to solve these practical systems. Contextual bandits, for example, are used in recommendation systems to select the most relevant content or products to recommend to users based on their previous behavior and preferences. Depending on the robotic application at hand, model predictive control or reinforcement learning are used to optimize for robot control policies. And GPT-3 is adapted to dialogue format, giving ChatGPT, using reinforcement learning from human feedback.

We focus on two important challenges that arise in these practical systems, namely need for *safety* and *adaptability*. We consider adaptability in the further challenging setting of *sparse reward* feedback. We study these challenges through algorithmic paradigms of reinforcement learning (RL), and online convex optimization (OCO).

Adaptability In the real-world applications, it is not enough to train an agent to perform a single task or operate in a specific environment. Instead, agents must be able to generalize their knowledge and adapt to new situations. For example, in robotics, a robot may need to navigate to new environments or perform new tasks.

Sparse reward or feedback Most formulations of sequential decision making rely on learning a meaningful agent model/policy through means of a feedback signal called cost/reward function. Most real-world applications, however, do not provide a feedback at every time instance in the decision making process. Such sparsity in feedback is detrimental to the learning of agent model.

Safety In many real-world applications, the actions selected by the decision maker must satisfy some necessary safety constraints over the decision set. For example, in robotics applications, the control actions should maintain the closed-loop stability of the system. Typically, such constraints are represented using a safe decision set definition. The control action then must lie inside this safe decision set for safe operation of the system.

2 Background

Reinforcement Learning and Meta Reinforcement Learning Reinforcement learning (RL) is an intuitive way to model learning through interaction. Traditional RL algorithms require a significant amount of training data to learn a single task or operate in a specific environment. On the contrary, in many real-world applications, the environment is inherently uncertain and dynamic, and agents need to be able to adapt *quickly* to new situations. Adapting to such environment changes quickly may require agents to learn from a small amount of data, which can be extremely challenging.

Meta reinforcement learning (meta-RL), a rapidly growing area of research, addresses this challenge of quick and efficient adaptability by enabling agents to *learn how to learn*. Meta-RL allows agents to develop a set of meta-policies or meta-knowledge that can be used to guide their learning process and adapt to new situations quickly.

Online Convex Optimization Online convex optimization (OCO) models sequential decision making only through the action and the respective cost function sequences. Here, an agent repeatedly makes decisions based on the feedback received from the environment. Agent's goal is to minimize cumulative cost with respect to the best action in hindsight, a quantity commonly referred to as 'Regret'. Simply put, OCO can be considered as a stateless version of RL, and is studied to understand a problem with its probable solution approaches in a simpler setting.

3 Contributions

Inspired from the need for adaptability and safety mentioned in Sec. 1, we have devised algorithms (1) that perform model adaptation using sub-optimal demonstration data in sparse-reward environments with meta reinforcement learning , and (2) that respect safety constraints at every time step in online convex optimization.

Meta-RL in sparse reward environments Meta-RL is a powerful approach for solving real-world problems [Finn *et al.*, 2017], but it faces a significant challenge when reward functions are sparsely specified, meaning that feedback values are only available for certain decisions. To address this issue, we investigate the use of sub-optimal demonstration data that is available for each task. We propose an algorithm called Enhanced Meta-RL using Demonstrations (EMRLD) [Rengarajan *et al.*, 2022] that leverages the demonstration data to provide a proxy for missing reward feedback during training. EMRLD combines reinforcement learning and supervised learning over demonstration data to generate a meta-policy that improves performance in a monotonic fashion. We conducted experiments on various sparse reward environments, including a mobile robot, and found that our EMRLD algorithm significantly outperforms existing approaches.

OCO with unknown linear constraints We consider the problem of safe online convex optimization, where we need to choose an action at each time step that satisfies a set of linear safety constraints, even though we don't know the specific parameters that define these constraints. We can only observe noisy feedback about the constraints for the actions we choose. Our proposed algorithm, called SO-PGD [Chaudhary and Kalathil, 2022], achieves a regret of $O(T^{2/3}\sqrt{log(T)})$ if we have access to a safe baseline action. This means that we can optimize our actions without violating the safety constraints. We have also developed algorithms to ensure safety in multi-agent settings with unknown linear safety constraints using distributed-OCO [Chang *et al.*, 2023]. For convex loss functions, our algorithm achieves a dynamic regret of $O(T^{2/3}\sqrt{log(T)} + T^{1/3}C_T^*)$, where C_T^* is the length of the best minimizer sequence. For certain non-convex problems, we achieve a dynamic regret of $O(T^{2/3}\sqrt{log(T)}+T^{1/3}C_T^*)$. In addition, we have explored a relaxed form of safety that involves smooth policies in imitation learning [Chaudhary and Ravindran, 2022]. This is an important step towards extending our results to high-dimensional problems.

4 Conclusion and Future Work

Our ultimate goal is to develop safe and efficient RL algorithms that can learn from experience and adapt to changing environments while ensuring stability and avoiding dangerous or unintended behaviors. In this regard, we are further exploring new ways to address these challenges using the context-based meta-RL approaches. Context-based meta-RL broadly encompasses probabilistic meta-RL [Rakelly *et al.*, 2019] and Bayesian meta-RL [Zintgraf *et al.*, 2019; Dorfman *et al.*, 2020], methods that reason about uncertainty in the RL system and make more informed decisions.

References

[Chang *et al.*, 2023] Ting-Jui Chang, Sapana Chaudhary, Dileep Kalathil, and Shahin Shahrampour. Dynamic regret analysis of safe distributed online optimization for convex and non-convex problems. *arXiv preprint arXiv:2302.12320*, 2023.

[Chaudhary and Kalathil, 2022] Sapana Chaudhary and Dileep Kalathil. Safe online convex optimization with unknown linear safety constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6175–6182, 2022.

[Chaudhary and Ravindran, 2022] Sapana Chaudhary and Balaraman Ravindran. Smooth imitation learning via smooth costs and smooth policies. In *5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD)*, pages 63–71, 2022.

[Dorfman *et al.*, 2020] Ron Dorfman, Idan Shenfeld, and Aviv Tamar. Offline meta learning of exploration. *arXiv preprint arXiv:2008.02598*, 2020.

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[Rakelly *et al.*, 2019] Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *International conference on machine learning*, pages 5331–5340. PMLR, 2019.

[Rengarajan *et al.*, 2022] Desik Rengarajan, Sapana Chaudhary, Jaewon Kim, Dileep Kalathil, and Srinivas Shakkottai. Enhanced meta reinforcement learning using demonstrations in sparse reward environments. *arXiv preprint arXiv:2209.13048*, 2022.

[Zintgraf *et al.*, 2019] Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. Varibad: A very good method for bayes-adaptive deep rl via meta-learning. *arXiv preprint arXiv:1910.08348*, 2019.