Meta Learning the Step Size in Policy Gradient Methods

L. SABBIONI, F. CORDA AND M. RESTELLI
(luca.sabbioni, marcello.restelli)@polimi.it, francesco.corda@mail.polimi.it

Motivation
- We want to apply Hyperparameter Optimization for Policy Gradient Learning.
- A dynamic learning rate may allow a fast and stable learning process.
- If the process depends on external variables, hyperparameters can be adapted to the context.

Contributions
1. Meta-MDP: Definition of an abstract class of hyperparameter decision processes called Meta-MDP, a framework for modeling meta-learning;
2. Bounds for Lipschitz Contextual-MDPs: Derivation of general guarantees on the return with respect to the parameterization of the context under smoothness assumptions;
3. FQI on Meta-MDP: Fitted Q-Iteration algorithm [Ernst et al., 2005] on the meta-MDP, using Natural Gradient Ascent as update rule, used to choose an adaptive step size throughout the learning process.

Properties
Assumptions: Lipschitz MDPs
\[ K(P_s, \omega) \leq L \] for all \( s \) and \( \omega \)
- These assumptions allow the Value Function to be Lipschitz continuous with constant \( L \) [Pirotta et al., 2015].

Assumptions: Lipschitz Policy
\[ K(P_s, \omega) \leq L \] for all \( s \) and \( \omega \)

Results: Lipschitz Value functions/return
\[ Q_2(s, a) = W_2(s, a) \]

References