Markov Decision Process Framework for Control-Based Reinforcement Learning

Yingdong Lu, Mark S. Squillante, Chai Wah Wu
Mathematical Sciences Department, IBM Research
Thomas J. Watson Research Center
Yorktown Heights, NY 10598, USA

yingdong@us.ibm.com, mss@us.ibm.com, cwwu@us.ibm.com

1. INTRODUCTION

For many years, reinforcement learning (RL) has proven to be very successful in solving a wide variety of learning and decision making under uncertainty (DMuU) problems, including those related to game playing and robotic control. Many different RL approaches, with varying levels of success, have been developed to address these problems.

Among these different approaches, model-free RL has been successful in solving various DMuU problems without any prior knowledge. Such model-free approaches, however, often suffer from high sample complexity that can require an inordinate amount of samples for some problems which can be prohibitive in practice, especially for problems limited by time or other constraints. Model-based RL has been successful in demonstrating significantly reduced sample complexity and in outperforming model-free approaches for various DMuU problems. Such model-based approaches, however, can often suffer from the difficulty of learning an appropriate system model and from worse asymptotic performance than model-free approaches due to model bias from inherently assuming that the learned system dynamics model accurately represents the true system environment; in addition, an approximate solution of the optimal control policy is often obtained based on the learned system dynamics model [4].

We propose herein a novel form of RL for seeking to directly learn an optimal control policy of a general underlying (unknown) dynamical system and to directly apply the corresponding learned optimal control policy within the dynamical system. This general approach is in strong contrast to many traditional model-based RL methods that, after learning the system dynamics model often of high complexity and dimensionality, then use this system dynamics model to compute an optimal solution of a corresponding dynamic programming problem, often applying model predictive control [4]. Our control-based RL approach instead learns the optimal parameters that derive an optimal policy function from a family of control policy functions, often of much lower complexity and dimensionality, from which the optimal control policy is directly obtained. Furthermore, we establish that our general approach converges to an optimal solution analogous to model-free RL approaches while eliminating the problems of model bias in traditional model-based RL approaches.

The theoretical foundation and analysis of our control-based RL approach is introduced within the context of a general Markov decision process (MDP) framework that extends the policy associated with the classical Bellman operator to a family of control policy functions derived from a corresponding parameter set, expands the domain of these policies from a single state to span across states, and extends the associated optimality criteria through these generalizations of the definition and scope of a control policy, all providing theoretical support for our general control-based RL approach. Within this MDP framework, we establish results on convergence w.r.t. both a contraction operator and a corresponding form of Q-learning, establish results on various aspects of optimality and optimal control policies, and introduce a new form of policy-parameter gradient ascent. To the best of our knowledge, this is the first proposal and analysis of such a general control-based RL approach based on theoretical support from an underlying extended MDP framework.

Generally speaking, the basic idea of learning a parameterized policy within an MDP framework to reduce sample complexity is not a new idea. One such popular approach concerns policy gradient methods [1], where gradient ascent of the value function in a space of policies is used together with projection to obtained an optimal policy. These ideas have been further refined in neural network based policy optimization approaches such as TRPO and PPO [1]. In strong contrast, our proposed approach derives the optimal policy through control-policy functions that map estimates of a few global (and local) parameters to optimal control policies in an iterative manner based on observations from applying the control policy of the current estimate of parameters.

We next present the MDP framework supporting our general approach that directly learns the parameters of the optimal control policy, together with the corresponding theoretical results on convergence and optimality as well as a new form of policy-parameter gradient ascent. We refer to [3] for all proofs and additional details and references.

2. MDP FRAMEWORK

Consider a discrete-space (which can be relaxed within our framework and results), discrete-time (DSDT) discounted MDP framework defined over a set of states $\mathbb{X}$, a set of actions $\mathbb{A}$, a transition probability kernel $\mathbb{P}$, a reward function r mapping state-action pairs to a bounded subset of $\mathbb{R}$, and a discount factor $\gamma \in [0, 1)$. Let $\mathbb{E}_{\mathbb{P}}$ denote expectation w.r.t. the probability kernel $\mathbb{P}$. Then the discounted infinite-horizon stochastic dynamic programming formulation associated with the DMuU problems of interest can be expressed as

$$\max_{a_1,a_2,\dots} \mathbb{E}_{\mathbb{P}}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t, x_{t+1})\right] \text{ s.t. } x_{t+1} = f(x_t, a_t), \quad (1)$$

MAMA 2023 Orlando, FL, USA.

where $x_t \in \mathbb{X}$ represents the state of the system, $a_t \in \mathbb{A}$ represents the control action decision variable, $f(\cdot,\cdot)$ represents the evolution function of the stochastic dynamical system characterizing the system state given the previous state and the taken action, and $r(\cdot,\cdot,\cdot)$ represents a reward-based objective function of both the system states and control action.

We note that (1) can represent a wide variety of stochastic dynamic programs associated with DMuU problems based on the different forms taken by the evolution function $f(\cdot,\cdot)$, together with the transition probability kernel $\mathbb{P}$; $f(\cdot,\cdot)$ can also characterize the discretized evolutionary system dynamics governed by (stochastic) differential equations or (stochastic) partial differential equations; and $r(\cdot,\cdot,\cdot)$ is also allowed to take on various general forms, and thus can represent any combination of cumulative and terminal rewards.

We first present a mathematical framework for our general approach to control-based RL, turn to establish corresponding theoretical results on convergence and optimality, and close with a new form of policy-parameter gradient ascent.

Mathematical Framework

Consider a DSDT discounted MDP denoted by $(\mathbb{X}, \mathbb{A}, \mathbb{P}, r, \gamma)$, where $\mathbb{X}$, $\mathbb{A}$, $\mathbb{P}$, r and $\gamma \in [0, 1)$ are as defined above. Let $\mathbb{Q}(\mathbb{X} \times \mathbb{A})$ denote the space of bounded real-valued functions over $\mathbb{X} \times \mathbb{A}$ with supremum norm. For the state x_t at time $t \in \mathbb{Z}$ in which action a is taken, i.e., $(x_t, a) \in \mathbb{X}\times\mathbb{A}$, denote by $\mathbb{P}(\cdot|x_t, a)$ the conditional transition probability for the next state x_{t+1} and precisely define $\mathbb{E}_{\mathbb{P}} := \mathbb{E}_{x_{t+1}\sim\mathbb{P}(\cdot|x_t,a)}$ to be the expectation w.r.t. $\mathbb{P}(\cdot|x_t, a)$. A stationary policy $\pi(\cdot|x) : \mathbb{X} \to \mathbb{A}$ defines a distribution of available control actions given the current state x, which reduces to a deterministic policy when the conditional distribution renders a constant action for state x; with slight abuse of notation, we always write policy $\pi(x)$. Let Π denote the set of all policies $\mathbb{X} \to \mathbb{A}$, and define the Bellman operator $\mathcal{T}_B : \mathbb{Q} \times \mathbb{Q}$ as

$$\mathcal{T}_B Q(x, a) := r(x, a) + \gamma \mathbb{E}_{\mathbb{P}} \max_{b\in A} Q(x', b),$$

with x' denoting the next state upon transition from x. Let $Q^*(x, a)$ denote the optimal action-value function, $V^*(x) = \max_a Q^*(x, a)$ the optimal value function, and $\pi^*(x) = \arg\max_a Q^*(x, a)$ the optimal action; $Q^*(x, a)$ is the unique fixed point of $\mathcal{T}_B$, a contraction in supremum norm.

We next introduce two key ideas to this standard MDP framework, one extending the policy $\pi : \mathbb{X} \to \mathbb{A}$ associated with the Bellman operator to a family of control-policy functions that map a parameter vector from a parameter set to a control policy that is optimal (or approximately optimal) under the given parameter vector; and the other extending the domain of these control policies from a single state to span across all (or a large subset of) states in $\mathbb{X}$. Let $\mathcal{P}$ be a subset of a metric space (e.g., Euclidean spaces) that serves as a parameter set. A control policy mapping $\mathcal{G} : \mathcal{P} \to \Pi$ identifies a family $\mathcal{G}(\mathcal{P}) \subseteq \Pi$ of control policies derived from vectors in the parameter set, where for any $p \in \mathcal{P}$ the control-policy function $\mathcal{G}(\cdot)$ identifies a particular control policy $\mathcal{F}_p$ in Π. Then, $\mathbb{F}$ represents a family of control-policy functions $\mathcal{G} : p \mapsto \mathcal{F}_p$ that yield control policies $\mathcal{F}_p : \mathbb{X} \to \mathbb{A}$ between the set of states and the set of actions. The family $\mathbb{F}$ includes control-policy functions $\mathcal{G}$ that provide the best rewards in expectation across all (or a large subset of) states $x \in \mathbb{X}$ of the MDP from among all control-policy functions in $\mathbb{F}$, derived w.r.t. the parameter vector $p \in \mathcal{P}$ that encodes both global and local information but is unknown and needs to be learned. It is evident, in the case $\mathbb{F}$ contains the policy functions which recover the control policies $\mathcal{F}_p(x) = \arg\max_{a\in\mathbb{A}} q(x, a)$ for all $q(x, a) \in \mathbb{Q}(\mathbb{X}\times\mathbb{A})$ and for $p \in \mathcal{P}$, that the introduction of the family $\mathbb{F}$ in our framework achieves the same outcome as the standard MDP framework, but with great reductions in the sample complexity over the Bellman equation and operator. Another very important difference is that our search is across all (or a large subset of) states $x \in \mathbb{X}$ to find a single (or small collection of) optimal parameter vector(s) $p^* \in \mathcal{P}$ that derives a single (or small collection of) optimal control-policy function(s) $\mathcal{G} \in \mathbb{F}$ which coincides with the Bellman equation for each state.

For each Q-function $q(x, a) \in \mathbb{Q}(\mathbb{X} \times \mathbb{A})$, we define the generic function $\tilde{q} : \mathbb{X} \times \mathcal{G}(\mathcal{P}) \to \mathbb{R}$ as $\tilde{q}(x, \mathcal{F}_p) = q(x, \mathcal{F}_p(x))$ where the control policy $\mathcal{F}_p$ is obtained from the control-policy function $\mathcal{G}$ derived from the parameter vector $p \in \mathcal{P}$; thus, $\tilde{q}(x, \mathcal{F}_p) \in \mathbb{Q}(\mathbb{X} \times \mathcal{G}(\mathcal{P}))$ is readily apparent. Iterations w.r.t. the operator of our mathematical framework (defined precisely below) then consist of improving the estimates of the parameter vector p while applying the optimal control-policy function $\mathcal{G}$ derived from the current estimate of p.

Convergence and Global Optimality

Define the operator $\mathbf{T}$ on $\mathbb{Q}(\mathbb{X} \times \mathcal{G}(\mathcal{P}))$ as

$$(\mathbf{T}\tilde{q})(x, \mathcal{F}_p) = \sum_{y\in\mathbb{X}} \mathbb{P}_{\mathcal{F}_p(x)}(x, y) \Big[r(x, \mathcal{F}_p(x), y) + \gamma \sup_{p'\in\mathcal{P}} \tilde{q}(y, \mathcal{F}_{p'}) \Big]. \quad (2)$$

The operator $\mathbf{T}$ is an analog of the Bellman operator within our MDP framework, for which we have the following result.

LEMMA 2.1. *For any $\gamma \in (0, 1)$, the operator $\mathbf{T}$ in (2) is a contraction in the supremum norm.*

Therefore, for any function $\tilde{q} \in \mathbb{X} \times \mathcal{G}(\mathcal{P})$, the iterations $\mathbf{T}^t(\tilde{q})$ converge as $t \to \infty$ to $\tilde{q}^*(x, \mathcal{F}_p)$, the unique fixed point of the contraction operator $\mathbf{T}$, and $\tilde{q}^*(x, \mathcal{F}_p)$ equals

$$\sum_{y\in\mathbb{X}} \mathbb{P}_{\mathcal{F}_p(x)}(x, y) \Big[r(x, \mathcal{F}_p(x), y) + \gamma \sup_{p'\in\mathcal{P}} \tilde{q}^*(y, \mathcal{F}_{p'}) \Big]. \quad (3)$$

Next consider convergence of the Q-learning algorithm within the context of our general MDP framework. In particular, we focus on the classical Q-learning update rule:

$$\begin{aligned} \tilde{q}_{t+1}(x_t, \mathcal{F}_{p,t}) = \tilde{q}_t(x_t, \mathcal{F}_{p,t}) + \alpha_t(x_t, \mathcal{F}_{p,t}) \Big[r_t \quad & (4) \\ + \gamma \sup_{p'\in\mathcal{P}} \tilde{q}_t(x_{t+1}, \mathcal{F}_{p'}) - \tilde{q}_t(x_t, \mathcal{F}_{p,t}) \Big], & \end{aligned}$$

for $0 < \gamma < 1$, $0 \le \alpha_t(x_t, \mathcal{F}_{p,t}) \le 1$ and iterations t. Let $\mathcal{F}_{p,t}$ be a sequence of control policies that covers all state-action pairs and r_t the corresponding reward of applying $\mathcal{F}_{p,t}$ to state x_t. We then have the following convergence result.

THEOREM 2.1. *Suppose $\mathbf{T}$ is a contraction operator as defined in (2). If $\sum_t \alpha_t = \infty$, $\sum_t \alpha_t^2 < \infty$, and r_t are bounded, then $\tilde{q}_t$ converges to $\tilde{q}^*$ as $t \to \infty$.*

Lastly, we introduce an important assumption followed by the corresponding globally optimal convergence result.

ASSUMPTION 2.1. *There exist a policy function $\mathcal{G}$ in the family $\mathbb{F}$ and a unique parameter vector p^* in the parameter set $\mathcal{P}$ such that, for any state $x \in \mathbb{X}$, $\pi^*(x) = \mathcal{F}_{p^*}(x)$= $\mathcal{G}(p^*)(x)$.*

Intuitively, this says $\mathbb{F}$ is rich enough to include a global policy that coincides with the Bellman operator for each state. We then have the following global convergence result.

THEOREM 2.2. *Suppose Assumption 2.1 holds for family of policy functions $\mathbb{F}$ and its parameter set $\mathcal{P}$ with contraction operator $\mathbf{T}$ in (2). If $\sum_t \alpha_t = \infty$, $\sum_t \alpha_t^2 < \infty$, and r_t are bounded, then $\tilde{q}_t$ converges to $\tilde{q}^*$ as $t \to \infty$ and the optimal policy function is derived from a unique parameter vector p^*.*

Convergence and Approximate Optimality
Let $\mathbb{F}$ be sufficiently rich to satisfy Assumption 2.1. We then consider our general MDP framework under a less rich family $\mathbb{F}_1 \subset \mathbb{F}$ of policy functions $\mathcal{G}^{(1)} \in \mathbb{F}_1$ derived from parameter vectors p of the parameter set $\mathcal{P}_1$. Define the operator $\mathbf{T}_1 : \mathbb{Q}(\mathbb{X}\times\mathcal{G}^{(1)}(\mathcal{P}_1)) \to \mathbb{Q}(\mathbb{X}\times\mathcal{G}^{(1)}(\mathcal{P}_1))$ as in (2) for any function $\tilde{q}_1(x, \mathcal{F}_p^{(1)}) \in \mathbb{Q}(\mathbb{X}\times\mathcal{G}^{(1)}(\mathcal{P}_1))$, namely $(\mathbf{T}_1\tilde{q}_1)(x, \mathcal{F}_p^{(1)}) = \sum_{y\in\mathbb{X}} \mathbb{P}_{\mathcal{F}_p^{(1)}(x)}(x, y)\big[r(x, \mathcal{F}_p^{(1)}(x), y) + \gamma \sup_{p'\in\mathcal{P}_1} \tilde{q}_1(y, \mathcal{F}_{p'}^{(1)})\big]$. From Lemma 2.1, operator $\mathbf{T}_1$ is a contraction in supremum norm and therefore, in the limit as $t \to \infty$, $\mathbf{T}_1^t(\tilde{q}_1)$ converges to the unique fixed point $\tilde{q}_1^*(x, \mathcal{F}_p^{(1)})$ of the contraction operator, for any function $\tilde{q}_1 \in \mathbb{Q}(\mathbb{X} \times \mathcal{G}^{(1)}(\mathcal{P}_1))$. Theorem 2.1 implies that the corresponding Q-learning iterates $\tilde{q}_{1,t}$ converge to the corresponding $\tilde{q}_1^*$ satisfying (3) as $t \to \infty$.

Now consider two families $\mathbb{F}_1$ and $\mathbb{F}_2$ of policy functions, where $\mathbb{F}_1 \subset \mathbb{F}_2 \subset \mathbb{F}$ and the members of $\mathbb{F}_i$ are derived from the parameter vectors of the corresponding parameter sets $\mathcal{P}_i$, $i = 1, 2$. From Lemma 2.1 and Theorem 2.1, for $i = 1, 2$, the contraction operators $\mathbf{T}_i : \mathbb{Q}(\mathbb{X}\times\mathcal{G}^{(i)}(\mathcal{P}_i)) \to \mathbb{Q}(\mathbb{X}\times\mathcal{G}^{(i)}(\mathcal{P}_i))$ under the parameter sets $\mathcal{P}_i$ converge to the unique fixed points $\tilde{q}_i^*(x, \mathcal{F}_p^{(i)})$ that equals, for all $x, p \in \mathcal{P}_i$:

$$\sum_{y\in\mathbb{X}} \mathbb{P}_{\mathcal{F}_p^{(i)}(x)}(x, y)\Big[r(x, \mathcal{F}_p^{(i)}(x), y) + \gamma \sup_{p'\in\mathcal{P}_i} \tilde{q}_i^*(y, \mathcal{F}_{p'}^{(i)})\Big]. \quad (5)$$

LEMMA 2.2. *Assume the state and action spaces are compact and $\mathcal{F}_p$ is uniformly continuous for each p. For the two parameter sets $\mathcal{P}_1$ and $\mathcal{P}_2$ above and any two parameter vectors $p_1 \in \mathcal{P}_1$ and $p_2 \in \mathcal{P}_2$, let $d(\cdot,\cdot)$ be a* sup*-norm distance function defined over the action space $\mathbb{A}$, i.e., $d(\mathcal{F}_{p_1}^{(1)}, \mathcal{F}_{p_2}^{(2)}) = \sup_{x\in\mathbb{X}} \|\mathcal{F}_{p_1}^{(1)}(x) - \mathcal{F}_{p_2}^{(2)}(x)\|$. Then, for all $\epsilon > 0$ there exists $\delta > 0$ such that, if $\forall p_1 \in \mathcal{P}_1\ \exists p_2 \in \mathcal{P}_2$ with $d(\mathcal{F}_{p_1}^{(1)}, \mathcal{F}_{p_2}^{(2)}) < \delta$ and if $\forall p_2 \in \mathcal{P}_2\ \exists p_1 \in \mathcal{P}_1$ with $d(\mathcal{F}_{p_1}^{(1)}, \mathcal{F}_{p_2}^{(2)}) < \delta$, we have $\sup_{x\in\mathbb{X}} \|\tilde{q}_1^* - \tilde{q}_2^*\| < \epsilon$.*

Intuitively, Lemma 2.2 shows that, for any policy families $\mathbb{F}_1 \subset \mathbb{F}_2$ sufficiently close to each other, the fixed points $\tilde{q}_1$, $\tilde{q}_2$ of the corresponding operators $\mathbf{T}_1$ and $\mathbf{T}_2$ are also close to each other. When the policy families $\mathbb{F}_1 \subset \mathbb{F}_2$ are sufficiently rich and approach $\mathbb{F}$, then the fixed points of the corresponding operators $\mathbf{T}_1$,$\mathbf{T}_2$ approach the unique fixed point of $\mathbb{F}$ satisfying (3), and therefore they approach the optimal q-value as promised by Bellman. We formally characterize this asymptotic convergence of approximate optimality to global optimality in the following result.

THEOREM 2.3. *Assume the state and action spaces are compact and $\mathcal{F}_p$ is uniformly continuous for each p. Consider a sequence of families of policy functions $\mathbb{F}_1 \subset \mathbb{F}_2 \cdots \subset \mathbb{F}_{k-1} \subset \mathbb{F}_k$ such that $\bigcup_{i=1}^k \mathbb{F}_i \to \mathbb{F}$ as $k \to \infty$, with $\mathcal{P}$ and $\mathcal{P}_i$ respectively denoting the parameter sets corresponding to $\mathbb{F}$ and $\mathbb{F}_i$, $i = 1, \ldots, k$. Then, $\sup_{x\in\mathbb{X}} \|\tilde{q}_k^* - \tilde{q}^*\| \to 0$ as $k \to \infty$.*

One specific instance of a sequence of the families of policy functions $\mathbb{F}_1 \subset \mathbb{F}_2 \cdots \subset \mathbb{F}_{k-1} \subset \mathbb{F}_k$ in Theorem 2.3 consists of piecewise-linear control policies of increasing richness (e.g., the class of CPWL functions in [2]) w.r.t. finer and finer granularity of the policy function space converging towards $\mathbb{F}$.

Control Policy Parameter-Vector Gradient Ascent
Building on the above results, we seek to determine the components of the unknown parameter vector p from which to derive the optimal control policy $\mathcal{F}_{p^*}$. One such approach consists of a corresponding gradient ascent method where the policy $\mathcal{F}_p$ is chosen according to an optimal control objective in terms of the value function V w.r.t. the parameter vector p. More formally, for stepsize η, the gradient ascent formulation of our general policy gradient method is given by

$$p_{t+1} = p_t + \eta \frac{\partial V}{\partial \mathcal{F}_{p_t}} \frac{\partial \mathcal{F}_{p_t}}{\partial p_t}. \quad (6)$$

Note that standard policy gradient methods are a special case of (6) where the parameter vector p is directly replaced by the policy π. In particular, the special case of $\mathcal{G}$ being an identity map with $\frac{\partial V}{\partial \mathcal{F}_{p_t}} \frac{\partial \mathcal{F}_{p_t}}{\partial p_t}$ replaced by $\frac{\partial V}{\partial \pi_t}$ corresponds to the direct policy gradient parameterization case in [1].

One important instance of the map $\mathcal{G} : p \mapsto \mathcal{F}_p$ concerns policies $\mathcal{F}_p$ that are derived from an optimal control formulation when the dynamics of the system are defined in terms of the (unknown) parameter vector p. For example, consider a general LQR optimal control problem where $A(p), B(p)$ denote the coefficients of a linear dynamical system, i.e., $\dot{x} = A(p)x + B(p)u$, with unknown p. At each step t of the policy gradient method in (6), we compute the solution $K(p_t)$ of the Riccati (algebraic or differential depending on the horizon) equations for the current parameter estimate p_t, and derive the optimal linear feedback control $\mathcal{F}_{p_t} = u = K(p_t)x$; the parameter estimate p_{t+1} is then updated according to (6).

As an illustrative example, consider the problem of landing a lunar module (LLM) to maximize the cumulative reward comprising positive points for successful degrees of soft landing and negative points for fuel usage and crashing. At each step t of the policy gradient method in (6), we compute the solution for the soft LLM problem w.r.t. a Riccati-like matrix differential equation [5] in terms of p_t, then derive a nonlinear optimal feedback control from this solution of the corresponding Riccati equation, and lastly update the parameter estimate p_{t+1} according to (6). Here, the unknown parameter vector p_t consists of the mass of the spacecraft, the thrust of the engines, and the gravity of the moon.

3. REFERENCES

[1] A. Agarwal, et al. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *JMLR*, 22(98):1–76, 2021.

[2] J.-N. Lin, R. Unbehauen. Canonical piecewise-linear approximations. *IEEE Trans Circuits and Systems: Fund Theory Appl*, 39(8):697–699, 1992.

[3] Y. Lu, M.S. Squillante, C.W. Wu. Markov Decision Process Framework for Control-Based Reinforcement Learning. *Preprint*, May 2023.

[4] A. Nagabandi, et al. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. *arXiv:1708.02596v2*, 2017.

[5] J. Zhou, K.L. Teo, D. Zhou, G. Zhao. Nonlinear optimal feedback control for lunar module soft landing. *J Global Opt*, 52:211–227, 2012.