Policy Gradient for Reinforcement Learning with General Utilities

Navdeep Kumar
Technion
navdeepkumar@alum.iisc.ac.in

Kaixin Wang
Technion
kaixin.wang@u.nus.edu

Utkarsh Pratiush
University of Tennessee, Knoxville
utkarshp@alum.iisc.ac.in

Kfir Levy
Technion
kfirylevy@technion.ac.il

Shie Mannor
Technion
shie@ee.technion.ac.il

Abstract

We derive policy gradient theorem for reinforcement learning (RL) with the objective which is a general (non-linear and non-convex) function of the occupancy measure of the policy. This setting incorporates many problems in literature such as apprenticeship learning, pure exploration and variational intrinsic control, etc. Our proposed policy gradient theorem shares the same elegance and ease of implementability as the standard policy gradient theorem, can be generalized easily to model-free settings suitable for large scale problems.

1 Introduction

Reinforcement Learning (RL) is a sequential decision problem where an agent interacts with an environment and learns to behave optimally Sutton & Barto (2018). Most of the work, is dedicated to Linear RL, where the goal is to learn the policy that maximizes the objective that is linear in occupancy measure of the policy Puterman (1994). However, many supervised and unsupervised RL problems are not covered in the Linear RL framework, such as apprenticeship learning, pure exploration, skill discovery and variational intrinsic control Abbeel & Ng (2004); Hazan et al. (2019); Bagaria et al. (2020); Zahavy et al. (2021), where the objectives are non-linear functions of the occupancy measures (non-linear RL).

Non-linear RL is tremendously challenging as the standard methods such as dynamic programming, value iterations, policy gradients, fail to trivially generalize to this setting. To the best of our knowledge, non prior works exist for general non-linear RL. However, there are few works for convex RL where the objective function is convex in the occupancy measure Zhang et al. (2020); Geist et al. (2022); Zhang et al. (2020); Zahavy et al. (2021); Mutti et al. (2022). Zhang et al. (2020) proposed the variational policy gradient for convex RL, Zahavy et al. (2021) reformulated the convex MDP as a min-max game involving policy-player and the cost player. Both the works use Fenchel duality that relies heavily on the convexity of the objective function.

This is the first work that directly derives the policy gradient thoerem for non-linear RL. Further, it proposes the model-free algorithm for the same that can be used in the large scale problems. Further, this policy gradient converges to global optimal solution, under some condition, in the convex RL Zhang et al. (2020). This makes this method an attractive methods for large scale problems.

2 Main

A Markov Decision Process (MDP) is defined by a sextuple $(\mathcal{S}, \mathcal{A}, \gamma, R, P, q)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\gamma \in [0,1)$ is the discount factor, $R \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ is the reward vector, $P \in (\Delta_{\mathcal{S}})^{\mathcal{S}\times\mathcal{A}}$ is the transition kernel, $\Delta_{\mathcal{A}}$ represents the space of probability distribution over the set $\mathcal{A}$, and $q \in \Delta_{\mathcal{S}}$ is the initial distribution over the state space $\mathcal{S}$ Sutton & Barto (2018). A stationary policy $\pi \in (\Delta_{\mathcal{A}})^{\mathcal{S}}$ maps states to probability distributions over actions. Moreover, $P(s'|s,a), \pi(a|s)$ represent the probability of transition from state s under action a to state s' and probability of taking action a in state s by policy π respectively. Let Π be the set of all stationary policies. Let $\mu^\pi \in (\Delta_{\mathcal{S}})^{\mathcal{S}\times\mathcal{A}}$ be the occupancy measure of policy π, defined as Puterman (1994)

$$\mu^\pi(s,a) := \sum_{n=0}^{\infty} \gamma^n \mathbb{E}[\mathbf{1}(s_n = s, a_n = a)|s_0 \sim q, a_m \sim \pi(\cdot|s_m), s_m \sim P(\cdot|s_{m-1}, a_{m-1})]. \quad (1)$$

Our objective in non-linear RL is

$$\min_{\mu^\pi \in \mathcal{K}} f(\mu^\pi) \quad (2)$$

where $f : \mathcal{K} \to \mathbb{R}$ is a differentiable function, and $\mathcal{K} := \{\mu^\pi | \pi \in \Pi\}$ is set of occupancy measure of all stationary policies Puterman (1994). In case of linear RL, apprentice learning Abbeel & Ng (2004) and pure exploration Hazan et al. (2019), objective functions are $f(\mu^\pi) = -\langle R, \mu^\pi \rangle$, $f(\mu^\pi) = ||\mu^\pi - \mu_{\text{expert}}||^2$ and $f(\mu^\pi) = \mu^\pi \cdot \log(\mu^\pi)$ respectively (see Table 1 in Zahavy et al. (2021) for more).

Unfortunately value iterations, dynamic programming methods can't be directly be applied here. Fortunately, the following gradient method

$$\theta_{k+1} = \theta_k + \eta_k \frac{\mathrm{d}f(\mu^{\pi_{\theta_k}})}{\mathrm{d}\theta},$$

is proven to converge to global optima, under mild condition Zhang et al. (2020). However, computing the gradient remains an open question, which is the key contribution of this work, discussed next.

Let Q^π_R be the Q-value function Sutton & Barto (2018) for the policy π and reward vector R, that be defined as $Q^\pi_R(s,a) := \sum_{n=0}^{\infty} \gamma^n \mathbb{E}[R(s_n, a_n)|s_0 = s, a_0 = a, a_t \sim \pi(\cdot|s_t), s_t \sim P(\cdot|s_{t-1}, a_{t-1})]$. Further, let $\mu^\pi(s) = \sum_a \pi(a|s)\mu^\pi(s,a)$ as a shorthand, then $\mu^\pi(s,a) = \mu^\pi(s)\pi(a|s)$ by definition. We assume the policy π_θ is parameterize by the parameter $\theta \in \Theta$. The following results derives the gradient of $f(\mu^{\pi_\theta})$ w.r.t. θ.

Theorem 1. *(Policy Gradient Theorem for non-linear RL)*

$$\frac{\mathrm{d}f(\mu^{\pi_\theta})}{\mathrm{d}\theta} = \sum_{s,a} \mu^{\pi_\theta}(s) Q^{\pi_\theta}_{R_\theta}(s,a) \frac{\mathrm{d}\pi_\theta(s,a)}{\mathrm{d}\theta}, \quad \textit{where } R_\theta = \frac{df(x)}{dx}|_{x=\mu^{\pi_\theta}}. \quad (3)$$

To compute the above policy gradient, we need to compute the Q-value w.r.t reward function R_θ, which is the gradient of the objective function $f(\mu^{\pi_\theta})$ w.r.t. occupation measure μ^{π_θ}. The result below states that the occupation measure can be efficiently computed iteratively.

Proposition 1. *For all policy π and kernel P, the iterative sequence given by*

$$\mu_{n+1}(s) := q(s) + \gamma \sum_a \pi(a|s) \sum_{s'} P(s'|s,a)\mu_n(s'), \quad \forall n \in \mathbb{N},$$

converges linearly to μ^π, more precisely, $\|\mu^\pi - \mu^\pi_n\|_1 \leq \gamma^n \|\mu^\pi - \mu^\pi_0\|_1, \quad \forall n \in \mathbb{N}.$

The above results can be combined to design model-free algorithm for non-linear RL, as illustrated in the algorithm 1. The algorithm 1 can be shown to converge using the techniques of two-time scale algorithm Borkar (2008). Further, global convergence of the algorithm, can be established for convex function f, under the similar mild conditions as in Zhang et al. (2020).

We presented policy gradient method for RL with non-linear objective function. The method is versatile and easily implementable which can used to tackle large scale problems.

Algorithm 1 Policy Gradient Algorithm for RL with general utilities

```
while not converged do
    Play action $a_t$ according to the policy $\pi_{\theta_t}$ and sample the next state $s_{t+1}$.
    Update the occupancy measure as $\mu(s_t) = \mu(s_t) + \eta_t \left[ \mu_0(s_t) + \gamma\mu(s_{t+1}) - \mu(s_t) \right]$
    Get reward vector $R = f'(\mu)$ and update Q-value as
        $$Q(s_t, a_t) = Q(s_t, a_t) + \epsilon_t[R(s_t, a_t) + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)].$$
    Update policy parameter as $\theta = \theta - \eta_t \frac{d \log(\pi_\theta(s_t,a_t))}{d\theta} Q(s_t, a_t)$.
end while
```

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

References

Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, pp. 1, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138385. doi: 10.1145/1015330.1015430. URL https://doi.org/10.1145/1015330.1015430.

Akhil Bagaria, Jason Crowley, Jing Wei Nicholas Lim, and George Konidaris. Skill discovery for exploration and planning using deep skill graphs. In *4th Lifelong Machine Learning Workshop at ICML 2020*, 2020. URL https://openreview.net/forum?id=-mvAo5hWNp.

V.S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008. ISBN 9780521515924. URL https://books.google.co.il/books?id=QLxIvgAACAAJ.

Matthieu Geist, Julien Pérolat, Mathieu Laurière, Romuald Elie, Sarah Perrin, Olivier Bachem, Rémi Munos, and Olivier Pietquin. Concave utility reinforcement learning: the mean-field game viewpoint, 2022.

Elad Hazan, Sham M. Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration, 2019.

Mirco Mutti, Riccardo De Santi, Piersilvio De Bartolomeis, and Marcello Restelli. Challenging common assumptions in convex reinforcement learning, 2022.

Martin L. Puterman. Markov decision processes: Discrete stochastic dynamic programming. In *Wiley Series in Probability and Statistics*, 1994.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. URL http://incompleteideas.net/book/the-book-2nd.html.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. Solla, T. Leen, and K. Müller (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 2000. URL https://proceedings.neurips.cc/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf.

Tom Zahavy, Brendan O'Donoghue, Guillaume Desjardins, and Satinder Singh. Reward is enough for convex mdps, 2021.

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities, 2020.

A Appendix

A.1 Related Work

There have been many works that aim to solve special cases of General RL such as Apprentenship Learning, Pure Exploration, etc. Junyu et al. Zhang et al. (2020) were among the first, who identified these problems as Convex MDP. They showed that Policy Gradient for the Convex MDP would converge to global minima under certain conditions. However, they were unable to obtain policy gradient in Convex RL analogous to policy gradient in Linear RL. But instead, they derived Variational Policy Gradient method using the solution of a stochastic saddle point problem involving the Fenchel dual of the convex RL objective. In Zahavy et al. (2021), the authors used Frenchel duality to convert the convex RL problem into a two-player zero-sum game between the agent (policy player) and an adversary that produces rewards (cost player) that agent must maximize.

We considered a more general objective and derived Policy Gradient Theorem for RL with general utilities that is as elegant and easily implementable as the Policy Gradient Theorem for Linear RL by Sutton et al. (2000). We recover the Policy Gradient Theorem for Linear RL by having Linear RL objective in Policy Gradient Theorem for RL with general utilities. We believe that it will play the same role in convex RL as Policy Gradient Theorem by Sutton et al. (2000) played in Linear RL. It gave rise to a very simple algorithm for RL with general utilities.

A.2 Proof of Policy Gradient Theorem 1

To apply gradient descent on policy space (or parameter space), we need the gradient w.r.t policy (or policy parameter):

$$\nabla_\theta f := \frac{\mathrm{d} f(\mu^\pi)}{\mathrm{d}\theta}. \tag{4}$$

By the chain rule, we have

$$\nabla_\theta f(\mu^\pi) = (\nabla_{\mu^\pi} f(\mu^\pi))^\top \frac{\mathrm{d}\mu^\pi}{\mathrm{d}\theta} = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \frac{\partial f(\mu^\pi)}{\partial \mu^\pi(s,a)} \nabla_\theta \mu^\pi(s,a). \tag{5}$$

Let $\beta_k^\pi(s,a,s',a') = \mathbb{E}[\mathbf{1}(s_k = s', a_k = a') \mid s_0 = s, a_0 = a, a_t \sim \pi(\cdot|s_t), s_t \sim P(\cdot|s_{t-1}, a_{t-1}), \forall t \leq k]$ is the probability of transition from state-action (s,a) to state-action (s',a') in exactly k steps following the policy π. In addition, we denote $\beta^\pi(s',a',s,a) = \sum_{k=0}^\infty \gamma^k \beta_k^\pi(s',a',s,a)$. To derive the policy gradient theorem for non-linear RL, let us first look at the gradient of the occupancy measure w.r.t. the policy parameters. We can obtain the following equation,

$$\nabla_\theta \mu^\pi(s,a) = \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} \mu^\pi(s')\beta^\pi(s',a',s,a)\nabla_\theta \pi(a'|s'), \tag{6}$$

Proof of Eqn. 6. Recall that

$$\mu^\pi(s,a) = \sum_{t=0}^\infty \gamma^t \sum_{(s_0,a_0,\cdots,s_{t-1},a_{t-1})\in(\mathcal{S}\times\mathcal{A})^t} q(s_0) \prod_{i=0}^{t-1} P(s_{i+1}|s_i,a_i) \prod_{j=0}^{t} \pi(a_j|s_j) \tag{7}$$

where $s_t = s$ and $a_t = a$. Taking derivative w.r.t. θ on both sides, we have

$$\nabla_\theta \mu^\pi(s,a) = \sum_{t=0}^{\infty} \gamma^t \sum_{(s_0,a_0,\cdots,s_{t-1},a_{t-1})\in(\mathcal{S}\times\mathcal{A})^t} q(s_0) \prod_{i=0}^{t-1} P(s_{i+1}|s_i,a_i) \nabla_\theta \prod_{j=0}^{t} \pi(a_j|s_j) \tag{8}$$

(using product rule) (9)

$$= \sum_{t=0}^{\infty} \gamma^t \sum_{(s_0,a_0,\cdots,s_{t-1},a_{t-1})\in(\mathcal{S}\times\mathcal{A})^t} q(s_0) \prod_{i=0}^{t-1} P(s_{i+1}|s_i,a_i) \sum_{k=0}^{t} \prod_{\substack{j=0\\ j\neq k}}^{t} \pi(a_j|s_j) \nabla_\theta \pi(a_k|s_k) \tag{10}$$

$$= \sum_{t=0}^{\infty} \gamma^t \sum_{k=0}^{t} \sum_{(s_0,a_0,\cdots,s_{t-1},a_{t-1})\in(\mathcal{S}\times\mathcal{A})^t} q(s_0) \prod_{i=0}^{t-1} P(s_{i+1}|s_i,a_i) \prod_{\substack{j=0\\ j\neq k}}^{t} \pi(a_j|s_j) \nabla_\theta \pi(a_k|s_k) \tag{11}$$

$$= \sum_{t=0}^{\infty} \gamma^t \sum_{k=0}^{t} \sum_{s_k\in\mathcal{S}} \underbrace{\sum_{(s_0,a_0,\cdots,s_{k-1},a_{k-1})\in(\mathcal{S}\times\mathcal{A})^k} q(s_0) \prod_{i=0}^{k-1} P(s_{i+1}|s_i,a_i)\pi(a_i|s_i)}_{\alpha_k^\pi(s_k)} \tag{12}$$

$$\sum_{a_k\in\mathcal{A}} \nabla_\theta \pi(a_k|s_k) \underbrace{\sum_{(s_{k+1},a_{k+1},\cdots,s_{t-1},a_{t-1})\in(\mathcal{S}\times\mathcal{A})^k} \prod_{i=k+1}^{t} P(s_i|s_{i-1},a_{i-1})\pi(a_i|s_i)}_{\beta_{t-k}^\pi(s_k,a_k,s_t,a_t)} \tag{13}$$

$$= \sum_{t=0}^{\infty} \sum_{k=0}^{t} \sum_{(s_k,a_k)\in\mathcal{S}\times\mathcal{A}} \gamma^k \alpha_k^\pi(s_k) \gamma^{t-k} \beta_{t-k}^\pi(s_k,a_k,s_t,a_t) \nabla_\theta \pi(a_k|s_k) \tag{14}$$

Note that $\alpha_k^\pi(s)$ is essentially the probability of transiting to state s in exactly k steps from the initial state $s_0 \sim q$ under policy π, so $\mu^\pi(s) = \sum_{k=0}^{\infty} \gamma^k \alpha_k^\pi(s)$. Therefore, we can write

$$\nabla_\theta \mu^\pi(s,a) = \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} \sum_{m=0}^{\infty} \gamma^m \alpha_m^\pi(s') \sum_{n=0}^{\infty} \gamma^n \beta_n^\pi(s',a',s,a) \nabla_\theta \pi(a'|s') \tag{15}$$

$$= \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} \mu^\pi(s') \beta^\pi(s',a',s,a) \nabla_\theta \pi(a'|s') \tag{16}$$

which concludes the proof. □

A.3 Occupation Measure Bootstraping

Lemma. *For all policy π and kernel P, the iterative sequence given by*

$$\mu_{n+1} := q + \gamma P^\pi \mu_n, \quad \forall n \in \mathbb{N},$$

converges linearly to d^π.

Proof. We first prove, $\mu^\pi \in \mathbb{R}^{\mathcal{S}}$ can be written as

$$\mu^\pi = q^T (I - \gamma P^\pi)^{-1} = q^T \sum_{n=0}^{\infty} (P^\pi)^n$$

$$\implies \gamma \mu^\pi P^\pi = \Big(q^T \sum_{n=0}^{\infty} (\gamma P^\pi)^n \Big) \gamma P^\pi = \mu^\pi - q.$$

We conclude that we have

$$\mu^\pi = q + \gamma \mu^\pi P^\pi.$$

Now, we have

$$\begin{aligned}\|\mu^\pi - \mu^\pi_{n+1}\|_1 &= \|q + \gamma\mu^\pi P^\pi - q - \gamma\mu_n P^\pi\|_1, \qquad \text{(from definition)}\\ &= \gamma\|(\mu^\pi - \mu_n)P^\pi\|_1\\ &\leq \gamma\sum_{s'}\sum_{s}|\mu^\pi(s) - \mu_n(s)|P(s'|s)\\ &= \gamma\sum_{s}|\mu^\pi(s) - \mu_n(s)|\\ &= \gamma\|\mu^\pi - \mu^\pi_n\|_1.\end{aligned}$$

This proves the claim. Note that convergence in not in L_∞ norm but L_1 norm instead. □

Under appropriate step size sequence $\{\eta_t\}$, the update rule $\mu(s_t) = \mu(s_t) + \eta_t \Big[\, q(s_t) + \gamma\mu(s_{t+1}) - \mu(s_t) \,\Big]$ convergence to d^π_P when state sequence $\{s_t\}$ is generated under policy π, kernel P and initial state distribution μ Borkar (2008).