# POLICY GRADIENT ALGORITHMS FROM PROBABILITY THEORY PERSPECTIVE

JULIAN WERGIELUK

ABSTRACT. This short note provides a concise description of the model architecture and learning algorithms of the agent developed in this project. We also report learning performance of the agent and provide a list of possible future model improvements.

## 1. DESCRIPTION OF THE LEARNING ALGORITHM

1.1. **Summary of the notation.** In this write-up I am going to cast the core ideas of policy gradient methods into a rigorous probabilistic framework. Much of the confusion and difficulty of understanding reinforcement learning comes from sloppy and incomplete mathematical notation.

Let's setup up the stage for the show and define a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Consider a Markov Decision Process (MDP) with measurable state and action spaces $(\mathbb{S}, \mathscr{S})$ and $(\mathbb{A}, \mathscr{A})$. A policy $\pi$ is a Markov kernel from $(\mathbb{S}, \mathscr{S})$ to $(\mathbb{A}, \mathscr{A})$, i.e. for each $s \in \mathbb{S}$, $\pi(. \mid s)$ is a probability distribution on $\mathscr{A}$. A trajectory

$$\tau = (S_0, A_0, R_1, S_1, A_1, R_2, \ldots)$$

encodes a sequence of states, actions and resulting numerical rewards pertaining to a policy $\pi$. The elements of $\tau$ are random variables, such that $S_0$ has the *start-state distribution* $\rho_0$ (which does not depend on the policy $\pi$), $\pi(.|s)$ is the conditional distribution of $A_i$ given $S_i = s$ for any $s \in \mathbb{S}$, and $\mathbb{P}(.|S_{i-1} = s, A_{i-1} = a)$ is the conditional distribution of $S_i$. The rewards $R_i$ are random variables taking values in $\mathbb{R}$. The return or discounted future reward $G_t$ at $i \in \mathbb{N}$ is defined as

$$G_i = \sum_{k=0}^{\infty} \gamma^k R_{i+k+1}, \tag{1}$$

where $\gamma$ is a hyperparameter from the interval $(0, 1)$, used to guarantee the convergence of the series (1).

Since $\tau$ is a sequence of random variables, and therefore itself a random variable taking values in the trajectory space

$$\mathbb{T} = \mathbb{S} \times \mathbb{A} \times (\mathbb{S} \times \mathbb{A} \times \mathbb{R})^{\infty},$$

we can sample from $\tau$ to obtain sequences of the form

$$(s_0, a_0, r_1, s_1, a_1, r_2, \ldots) \in \mathbb{T}.$$

These samples of $\tau$ are called *episodes* or *rollouts*.

The central object of interest is the mechanics of the world encoded in the transition probability measure $\mathbb{P}(. \mid s, a)$. For $i > 0$, it is the distribution of $S_i$ under the conditions $S_{i-1} = s$ and $A_{i-1} = a$ for fixed $s \in \mathbb{S}$ and $a \in \mathbb{A}$. This distribution does not depend on $i$, and, more importantly, does not depend on the policy $\pi$.

To simplify the setup, we assume that the reward $R_i$ at time $i$ is a deterministic function of the relevant states and actions. Depending on the problem at hand these are the three common choices encountered in the literature:

$$\begin{aligned} R_i &= \phi_1(S_i), \\ R_i &= \phi_2(A_{i-1}, S_i), \\ R_i &= \phi_3(S_{i-1}, A_{i-1}, S_i) \end{aligned}$$

$\phi_1$ is probably the most frequently used form.

In reinforcement learning, we parametrize the policy $\pi_\theta$ with an $\mathbb{R}^d$ vector $\theta$, $d \geq 1$. The goal is to find good values for $\theta$ which lead to high cumulative expected reward

$$J(\theta) = \mathbb{E}_\theta [G_0] = \mathbb{E}_\theta \left[ \sum_{k=0}^{\infty} \gamma^k R_{k+1} \right].$$

The subscript $\theta$ next to the expectation operator indicates that the underlying probability measure $\mathbb{P}_\theta$ combines both $\mathbb{P}$ and $\pi_\theta$.

1.2. **Value functions.** The *state-value function* is defined for each state $s \in \mathbb{S}$ as

$$V_\theta(s) = \mathbb{E}_\theta [G_0 | S_0 = s] = \mathbb{E}_\theta \left[ \sum_{k=0}^{\infty} \gamma^k R_{k+1} | S_0 = s \right]. \tag{2}$$

The *action-value function* is defined for each state-action pair $(s, a) \in \mathbb{S} \times \mathbb{A}$ as

$$Q_\theta(s, a) = \mathbb{E}_\theta [G_0 | S_0 = s, A_0 = a] = \mathbb{E}_\theta \left[ \sum_{k=0}^{\infty} \gamma^k R_{k+1} | S_0 = s, A_0 = a \right]. \tag{3}$$

The *advantage function* is the difference

$$A_\theta(s, a) = V_\theta(s) - Q_\theta(s, a). \tag{4}$$

1.3. **Episodic and perpetual tasks.** Reinforcement learning tasks can be partitioned into two categories, depending on whether the task can be completed. A task is called *episodic* if there is an end of the task, i.e. some notion of "being done". If there is no

such notion, the task is called *perpetual* [1]. Typical examples of episodic tasks are games. Most board games like chess and go, and computer games fall into this category.

It is important to note, that many episodic tasks can also run forever. For example, in chess, we can design a policy that "runs in circles" and does not pursue the goal of winning the game. Letting such a policy play against itself will result in an endless game.

What distinguishes an episodic task from a perpetual one, is the existence of a non-empty set of end states $\Delta$ in $\mathbb{S}$, aptly called the *coffin space*, for which we require that $\mathbb{P}(S_{i+1} = S_i | S_i \in \Delta) = 1$ and $\mathbb{P}(R_i = 0 | S_i \in \Delta) = 1$. This ensures that all the states in $\Delta$ are absorbing and that no additional reward can be accumulated after the task has been completed.

1.4. **Categorial and continuous action spaces.** The cardinality of the action space has a huge impact on the development of the theory of policy gradients as well as on the design and implementation of the policy optimization algorithms. Essentially, there are two cases we need to deal with. First, if the action space has finite number of elements, we call it *categorial*. Finite number of possible actions implies that the policy $\pi(.|s)$ evaluated at the state $s$, can be identified with a vector with $|\mathbb{A}|$ number of elements. Also, it is convenient to define the likelihood function $f$ of the policy $\pi$ as

$$f(a|s) = \pi(\{a\}|s),\ a \in \mathbb{A}.$$

$f(.|s)$ is the probability function of the probability measure $\pi(.|s)$.

The second case encountered in the literature is the *continuous* action space $\mathbb{A}$ which is a "nice" subset of $\mathbb{R}^{n_{\mathbb{A}}}$. Typical example of a continuous action space is a product of finite or infinite intervals in $\mathbb{R}^{n_{\mathbb{A}}}$. Let us assume that, for each state $s$, the measure $\pi(.|s)$ has a density $f(.|s)$, i.e. for an action set $\eta \in \mathscr{A}$ we have

$$\pi(\eta|s) = \int_{\eta} f(a|s)da.$$

As in the categorial case, we call $f(.|s)$ the likelihood function of the measure $\pi(.|s)$.

1.5. **Technical assumptions.** In the following list, we gather the technical assumptions.

**Assumption 1.1.** (1) *The reward function $\phi$ is measurable and bounded.*

[1] In the literature, people use the adjective "continuous". But we already have "continuous spaces", "continuous time", "continuous function", etc. The word "perpetual" is much better at conveying the property of running indefinitely long. See e.g. https://www.youtube.com/watch?v=t5TK1OxEj9A.

Let us state that our body of thought compiles without errors.

**Theorem 1.2.** *Suppose the conditions stated in Assumption 1.1 hold. Then we have:*

(1) *The value functions $V_\theta$, $Q_\theta$, and $A_\theta$ are well-defined.*

1.6. **Cumulative reward optimization.** One of our goals is to find the values of the parameter $\theta$ yielding a policy that consistently produces high cumulative expected award $J(\theta)$.

The gradient of $J(\theta)$ is given by

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_\theta [G_0] = \mathbb{E}_\theta \left[ G_0 \sum_{i=0}^{\infty} \nabla_\theta \log \pi_\theta(A_i|S_i) \right].$$

Using the Markov property we get

$$\nabla_\theta J(\theta) = \mathbb{E}_\theta \left[ \sum_{i=0}^{\infty} G_i \nabla_\theta \log \pi_\theta(A_i|S_i) \right].$$

Policy gradient with the state-value baseline

$$\nabla_\theta J(\theta) = \mathbb{E}_\theta \left[ \sum_{i=0}^{\infty} (G_i - V_\theta(S_i)) \nabla_\theta \log \pi_\theta(A_i|S_i) \right].$$

Policy gradient with the advantage function baseline

$$\nabla_\theta J(\theta) = \mathbb{E}_\theta \left[ \sum_{i=0}^{\infty} A_\theta(S_i, A_i) \nabla_\theta \log \pi_\theta(A_i|S_i) \right].$$

## 2. Training analysis

## 3. Ideas for future work

## References

[1] Erhan Çinlar. *Probability and Stochastics*. 1st ed. Springer Verlag, 2011, p. 572. ISBN: 0387878580.

*Email address*: julian.wergieluk@risklab.com