Revisit Policy Optimization in Matrix Form

Sitao Luan[1,2], **Xiao-Wen Chang**[1], **Doina Precup**[1,2,3]
{sitao.luan@mail, chang@cs, dprecup@cs}.mcgill.ca
[1]McGill University; [2]Mila; [3]DeepMind

Abstract

In tabular case, when the reward and environment dynamics are known, policy evaluation can be written as $\boldsymbol{V}_{\boldsymbol{\pi}} = (I - \gamma P_{\boldsymbol{\pi}})^{-1}\boldsymbol{r}_{\boldsymbol{\pi}}$, where $P_{\boldsymbol{\pi}}$ is the state transition matrix given policy $\boldsymbol{\pi}$ and $\boldsymbol{r}_{\boldsymbol{\pi}}$ is the reward signal given $\boldsymbol{\pi}$. What annoys us is that $P_{\boldsymbol{\pi}}$ and $\boldsymbol{r}_{\boldsymbol{\pi}}$ are both mixed with $\boldsymbol{\pi}$, which means every time when we update $\boldsymbol{\pi}$, they will change together. In this paper, we leverage the notation from [4] to disentangle $\boldsymbol{\pi}$ and environment dynamics which makes optimization over policy more straightforward. We show that policy gradient theorem [3] and TRPO [2] can be written in a more general notation framework and such framework has good potential to be extended to model-based reinforcement learning.

1 Introduction

1.1 Bellman Equation in Matrix Form

Markov decision process (MDP) is a framework to model the learning process that the agent learns from the interaction with the environment [3]. The interaction happens in discrete time steps, $t = 0, 1, 2, 3, \cdots$. At step t, given a state $S_t = s_t \in \mathcal{S}$, the agent picks an action $a_t \in \mathcal{A}(s_t)$ according to a policy $\boldsymbol{\pi}(\cdot|s_t)$, which is a rule of choosing actions given a state. Then, at time $t+1$, the environmental dynamics $p : \mathcal{S} \times \mathcal{R} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ takes the agent to a new state $S_{t+1} = s_{t+1} \in \mathcal{S}$ and provide a numerical reward $R_{t+1} = r_{t+1}(s_t, a_t, s_{t+1}) \in \mathbb{R}$. Such a sequence of interactions gives us a trajectory $\tau = \{S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \cdots\}$. Our objective is to find an optimal policy to maximize the expected long-term discounted cumulative rewards $\boldsymbol{V}_{\boldsymbol{\pi}}(s) = E_{\boldsymbol{\pi}}[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}|S_t = s]$ for each state s or $\boldsymbol{Q}_{\boldsymbol{\pi}}(s, a) = E_{\boldsymbol{\pi}}[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}|S_t = s, a_t = a]$ for each state-action pair (s, a), where γ is the discount factor. The Bellman equation for $\boldsymbol{V}_{\boldsymbol{\pi}}(s)$ can be written as follows:

$$\begin{aligned}
\boldsymbol{V}_{\boldsymbol{\pi}}(s) &= \sum_a \boldsymbol{\pi}(a|s)\sum_{s',r} p(s', r|s, a)[r + \gamma \boldsymbol{V}_{\boldsymbol{\pi}}(s')] \\
&= \sum_a \boldsymbol{\pi}(a|s)\sum_{s',r} r \cdot p(s', r|s, a) + \sum_a \boldsymbol{\pi}(a|s)\sum_{s',r} p(s', r|s, a)\gamma \boldsymbol{V}_{\boldsymbol{\pi}}(s') \\
&= \sum_a \boldsymbol{\pi}(a|s) r(s, a) + \gamma \sum_{s'}\{\sum_a \boldsymbol{\pi}(a|s) \cdot p(s'|s, a)\}\boldsymbol{V}_{\boldsymbol{\pi}}(s') \\
&= \boldsymbol{r}_{\boldsymbol{\pi}}(s) + \gamma \sum_{s'} P_{\boldsymbol{\pi}}(s'|s)\boldsymbol{V}_{\boldsymbol{\pi}}(s') \quad \text{for all } s \in \mathcal{S}
\end{aligned} \tag{1}$$

Where $\boldsymbol{r}_{\boldsymbol{\pi}}(s)$ is the expected immediate reward at state s under $\boldsymbol{\pi}$ and $P_{\boldsymbol{\pi}}(s'|s)$ is the transition probability of moving from s to s' under $\boldsymbol{\pi}$. Notice that the transition probability is a combination of

Optimization Foundations for Reinforcement Learning Workshop at NeurIPS 2019, Vancouver, Canada.

policy and the environment dynamics. We can write $\boldsymbol{V}_{\boldsymbol{\pi}}, \boldsymbol{r}_{\boldsymbol{\pi}}, P_{\boldsymbol{\pi}}$ in matrix form

$$\boldsymbol{V}_{\boldsymbol{\pi}} = \begin{bmatrix} \boldsymbol{V}_{\boldsymbol{\pi}}(s_1) \\ \boldsymbol{V}_{\boldsymbol{\pi}}(s_2) \\ \vdots \\ \boldsymbol{V}_{\boldsymbol{\pi}}(s_n) \end{bmatrix}, \ \boldsymbol{r}_{\boldsymbol{\pi}} = \begin{bmatrix} \boldsymbol{r}_{\boldsymbol{\pi}}(s_1) \\ \boldsymbol{r}_{\boldsymbol{\pi}}(s_2) \\ \vdots \\ \boldsymbol{r}_{\boldsymbol{\pi}}(s_n) \end{bmatrix} = \begin{bmatrix} \sum_a \boldsymbol{\pi}(a|s_1) r(s_1, a) \\ \sum_a \boldsymbol{\pi}(a|s_2) r(s_2, a) \\ \vdots \\ \sum_a \boldsymbol{\pi}(a|s_n) r(s_n, a) \end{bmatrix}, \ P_{\boldsymbol{\pi}} = \begin{bmatrix} P_{\boldsymbol{\pi}}(s_1|s_1) & P_{\boldsymbol{\pi}}(s_2|s_1) & \cdots & P_{\boldsymbol{\pi}}(s_n|s_1) \\ P_{\boldsymbol{\pi}}(s_1|s_2) & P_{\boldsymbol{\pi}}(s_2|s_2) & \cdots & P_{\boldsymbol{\pi}}(s_n|s_2) \\ \vdots & \vdots & \ddots & \vdots \\ P_{\boldsymbol{\pi}}(s_1|s_n) & P_{\boldsymbol{\pi}}(s_2|s_n) & \cdots & P_{\boldsymbol{\pi}}(s_n|s_n) \end{bmatrix} \tag{2}$$

Then the Bellman equation can be rewritten in matrix form

$$\boldsymbol{V}_{\boldsymbol{\pi}} = \boldsymbol{r}_{\boldsymbol{\pi}} + \gamma P_{\boldsymbol{\pi}} \boldsymbol{V}_{\boldsymbol{\pi}}$$

Assume that $P_{\boldsymbol{\pi}}$ and $\boldsymbol{r}_{\boldsymbol{\pi}}$ are given and $I - \gamma P_{\boldsymbol{\pi}}$ is nonsingular, then

$$\boldsymbol{V}_{\boldsymbol{\pi}} = (I - \gamma P_{\boldsymbol{\pi}})^{-1} \boldsymbol{r}_{\boldsymbol{\pi}} \tag{3}$$

For later uses, we define the vector $\boldsymbol{Q}_{\pi} = [\boldsymbol{Q}(s_1, \cdot)^T, \boldsymbol{Q}(s_2, \cdot)^T, \ldots, \boldsymbol{Q}(s_n, \cdot)^T]^T$.

1.2 Problem

One problem with (3) is that $P_{\boldsymbol{\pi}}$ and $\boldsymbol{r}_{\boldsymbol{\pi}}$ are both dependent on $\boldsymbol{\pi}$, which means every time when $\boldsymbol{\pi}$ changes, we need to reconstruct them. In addition, although $\boldsymbol{V}_{\boldsymbol{\pi}}$ is a function of $\boldsymbol{\pi}$, we cannot write the function in terms of $\boldsymbol{\pi}$ explicitly. What we desire is to have

$$\boldsymbol{V}_{\boldsymbol{\pi}} = f(\boldsymbol{\pi}|P, r) \quad \text{or} \quad \boldsymbol{V}_{\boldsymbol{\pi}_\theta} = f(\boldsymbol{\pi}_\theta|P, r) \tag{4}$$

where $\boldsymbol{\pi}_\theta$ is a parameterized policy and P, r only depend on environment dynamics, instead of $\boldsymbol{V}_{\boldsymbol{\pi}} = f(\boldsymbol{\pi}|P_{\boldsymbol{\pi}}, \boldsymbol{r}_{\boldsymbol{\pi}})$ or $\boldsymbol{V}_{\boldsymbol{\pi}_\theta} = f(\boldsymbol{\pi}_\theta|P_{\boldsymbol{\pi}_\theta}, r_{\boldsymbol{\pi}_\theta})$. Then, we can directly write our objective as a function of $\boldsymbol{\pi}$, which can make optimization more straightforward and efficient, just as follows,

$$\operatorname*{argmax}_{\boldsymbol{\pi}} \boldsymbol{V}_{\boldsymbol{\pi}} = f(\boldsymbol{\pi}|P, r) \quad \text{or} \quad \operatorname*{argmax}_{\theta} \boldsymbol{V}_{\boldsymbol{\pi}_\theta} = f(\boldsymbol{\pi}_\theta|P, r) \tag{5}$$

and solve it via gradient descent or other techniques. The notations from [4] is helpful to construct (4)(5). We introduce the formulation in the following sections and try to rewrite TRPO [2] with it.

2 Notation and Preliminary

2.1 Notation and Properties

For simplicity, we assume both the number of states in $\mathcal{S}$ and the number of actions in $\mathcal{A}$ are finite. We define the notations as follows [4]:

- $|\mathcal{S}|$ and $|\mathcal{A}|$ denote the number of states in $\mathcal{S}$ and the number of actions in $\mathcal{A}$, respectively.
- $P \in \mathbb{R}^{|S||A| \times |S|}$ is a transition matrix whose entries are $P_{(sa,s')} = p(s'|s, a)$, where $p(s'|s, a) \geq 0, \ \sum_{s'} p(s'|s, a) = 1$, for all s and a, *i.e.*,

$$P\mathbf{1}_{|S|} = \mathbf{1}_{|S||A|}, \quad \mathbf{1}_{|S|} = [1, 1, \ldots, 1]^T \in \mathbb{R}^{|S| \times 1} \tag{6}$$

- $\boldsymbol{\pi} \in \mathbb{R}^{|S||A| \times 1}$ is a stationary policy, whose entries are $\boldsymbol{\pi}_{(sa)} = \boldsymbol{\pi}(a|s)$, where $\sum_a \boldsymbol{\pi}_{(sa)} = 1$. It is convenient to rewrite the policy as a matrix $\Pi \in \mathbb{R}^{|S| \times |S||A|}$, where $\Pi_{(s,s'a)} = \boldsymbol{\pi}_{(sa)}$ if $s' = s$, otherwise 0, *i.e.*

$$\Pi = \operatorname{diag}(\boldsymbol{\pi}(\cdot|s_1)^T, \cdots, \boldsymbol{\pi}(\cdot|s_{|S|})^T), \ \boldsymbol{\pi}(\cdot|s_i)^T \in \mathbb{R}^{1 \times |A|}. \tag{7}$$

  $\Xi \in \mathbb{R}^{|S| \times |S||A|}$ is an auxiliary (marginalization) matrix defined by

$$\Xi = \operatorname{diag}(\mathbf{1}_{|A|}^T, \cdots, \mathbf{1}_{|A|}^T) \tag{8}$$

  It is easy to verify that

$$\Pi \mathbf{1}_{|S||A|} = \mathbf{1}_{|S|}, \quad \Pi \Xi^T = I_{|S|} \tag{9}$$

  We can easily reconstruct the state-to-state transition matrix by $P_{\boldsymbol{\pi}} = \Pi P \in \mathbb{R}^{|S| \times |S|}$, and state-action-to-state-action transition matrix by $P'_{\boldsymbol{\pi}} = P\Pi \in \mathbb{R}^{|S||A| \times |S||A|}$. From (6) and (9),

$$P_{\boldsymbol{\pi}} \mathbf{1}_{|S|} = \Pi P \mathbf{1}_{|S|} = \Pi \mathbf{1}_{|S||A|} = \mathbf{1}_{|S|}$$
$$P'_{\boldsymbol{\pi}} \mathbf{1}_{|S||A|} = P\Pi \mathbf{1}_{|S||A|} = P \mathbf{1}_{|S|} = \mathbf{1}_{|S||A|}$$

- $\boldsymbol{\rho}_0 \in \mathbb{R}^{|S|\times 1}$ is the initial state distribution. $\boldsymbol{\mu}_{\boldsymbol{\pi}} \in \mathbb{R}^{|S|\times 1}$ is the steady-state distribution of π given environment P which satisfies $\boldsymbol{\mu}_{\boldsymbol{\pi}}^T \Pi P = \boldsymbol{\mu}_{\boldsymbol{\pi}}^T$. $\boldsymbol{\rho}_{\boldsymbol{\pi}}^T = \boldsymbol{\rho}_0^T \sum_{i=0}^{\infty} \gamma^i (\Pi P)^i = \boldsymbol{\rho}_0^T (I - \gamma \Pi P)^{-1}$ is the discounted visitation frequency vector under $\boldsymbol{\pi}$ and
$$\boldsymbol{\rho}_{\boldsymbol{\pi}}^T \mathbf{1} = \boldsymbol{\rho}_0^T \sum_{i=0}^{\infty} \gamma^i (\Pi P)^i \mathbf{1} = \boldsymbol{\rho}_0^T (\sum_{i=0}^{\infty} \gamma^i) \mathbf{1} = 1/(1-\gamma)$$
- $\boldsymbol{r} \in \mathbb{R}^{|S||A|\times 1}$ is the average reward vector whose entries are $\boldsymbol{r}_{(sa)} = r(s,a) = E[r|s,a] = \sum_{s',r} r \cdot p(s', r|s, a)$, which specify the average reward obtained when taking action a in state s. From the definition of $\boldsymbol{r}_{\boldsymbol{\pi}}$ in (2) and Π in (7), it is easy to see that
$$\boldsymbol{r}_{\boldsymbol{\pi}} = \Pi \boldsymbol{r} \tag{10}$$

Now we can rewrite the Bellman equation. From (2), we have
$$\boldsymbol{V}_{\boldsymbol{\pi}} = (I - \gamma P_{\boldsymbol{\pi}})^{-1} \boldsymbol{r}_{\boldsymbol{\pi}} = (I - \gamma \Pi P)^{-1} \Pi \boldsymbol{r} = \sum_{i=0}^{\infty} \gamma^i (\Pi P)^i \Pi \boldsymbol{r} = \Pi \boldsymbol{r} + \gamma \Pi P \boldsymbol{V}_{\boldsymbol{\pi}} \tag{11}$$
Similarly, we have $\boldsymbol{Q}_{\boldsymbol{\pi}} = \sum_{i=0}^{\infty} \gamma^i (P\Pi)^i \boldsymbol{r} = \boldsymbol{r} + \gamma P \Pi \boldsymbol{Q}_{\boldsymbol{\pi}}$. One can easily check that the relations between $\boldsymbol{V}_{\boldsymbol{\pi}}$ and $\boldsymbol{Q}_{\boldsymbol{\pi}}$ are as follows,
$$\boldsymbol{V}_{\boldsymbol{\pi}} = \Pi \boldsymbol{Q}_{\boldsymbol{\pi}},\ \boldsymbol{Q}_{\boldsymbol{\pi}} = \sum_{i=0}^{\infty} \gamma^i (P\Pi)^i \boldsymbol{r} = \boldsymbol{r} + \gamma P [\sum_{i=0}^{\infty} \gamma^i (P\Pi)^i \Pi \boldsymbol{r}] = \boldsymbol{r} + \gamma P \boldsymbol{V}_{\boldsymbol{\pi}} \tag{12}$$

2.2 A Recap of TRPO

We do a simple recap of TRPO [2] in this subsection. We will show how these results can easily be obtained when written in matrix form and how to extend them.

For policy-based reinforcement learning algorithms, if we use policy gradient with a fixed learning rate to update the policy, it always happens that the learning rate is sometimes too large that we will get a worse policy. This oscillation makes the training unstable and the convergence slow. So we want to find a way that our policy is guaranteed to improve after each policy update even the improvement might be quite small at some time. The objective $\eta(\boldsymbol{\pi})$ is defined as follows,
$$\eta(\boldsymbol{\pi}) = \mathbb{E}_{s_0, a_0, \ldots} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t) \right],\ \text{where } s_0 \sim \boldsymbol{\rho}_0(s_0), a_t \sim \boldsymbol{\pi}(a_t|s_t), s_{t+1} \sim p(s_{t+1}|s_t, a_t) \tag{13}$$
With the definition of the advantage function $\boldsymbol{A}_{\boldsymbol{\pi}}(s,a)$:
$$\boldsymbol{A}_{\boldsymbol{\pi}}(s,a) = \boldsymbol{Q}_{\boldsymbol{\pi}}(s,a) - \boldsymbol{V}_{\boldsymbol{\pi}}(s) \tag{14}$$
$\eta(\tilde{\boldsymbol{\pi}})$ can be split into two parts,
$$\eta(\tilde{\boldsymbol{\pi}}) = \eta(\boldsymbol{\pi}) + \mathbb{E}_{s_0, a_0, \cdots \sim \tilde{\boldsymbol{\pi}}} \left[\sum_{t=0}^{\infty} \gamma^t \boldsymbol{A}_{\boldsymbol{\pi}}(s_t, a_t) \right] = \eta(\boldsymbol{\pi}) + \sum_s \boldsymbol{\rho}_{\tilde{\boldsymbol{\pi}}}(s) \sum_a \tilde{\boldsymbol{\pi}}(a|s) \boldsymbol{A}_{\boldsymbol{\pi}}(s,a) \tag{15}$$
where $\boldsymbol{\rho}_{\tilde{\boldsymbol{\pi}}}(s) = \sum_{i=0}^{\infty} \gamma^i p(s_i = s|\tilde{\boldsymbol{\pi}}, P, \boldsymbol{\rho}_0^T)$ is the discounted visitation frequency of s under $\tilde{\boldsymbol{\pi}}$. A local approximation to $\eta(\tilde{\boldsymbol{\pi}})$ is constructed as follows:
$$L_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}}) = \eta(\boldsymbol{\pi}) + \sum_s \boldsymbol{\rho}_{\boldsymbol{\pi}}(s) \sum_a \tilde{\boldsymbol{\pi}}(a|s) \boldsymbol{A}_{\boldsymbol{\pi}}(s,a) \tag{16}$$
It satisfies two important properties when $\boldsymbol{\pi}_\theta$ is parameterized by θ:
$$L_{\boldsymbol{\pi}_{\theta_0}}(\boldsymbol{\pi}_{\theta_0}) = \eta(\boldsymbol{\pi}_{\theta_0}),\ \ \nabla_\theta L_{\boldsymbol{\pi}_{\theta_0}}(\boldsymbol{\pi}_\theta)\Big|_{\theta=\theta_0} = \nabla_\theta \eta(\boldsymbol{\pi}_\theta)\Big|_{\theta=\theta_0} \tag{17}$$
Given the total variation divergence for two discrete distributions $D_{TV}(p\|q) = \frac{1}{2}\sum_i \left|p_i - q_i\right|$, we define $D_{\mathrm{TV}}^{\max}(\boldsymbol{\pi}, \tilde{\boldsymbol{\pi}}) = \max_s D_{\mathrm{TV}}(\boldsymbol{\pi}(\cdot|s)\|\tilde{\boldsymbol{\pi}}(\cdot|s))$. It can be proved that
$$\eta(\tilde{\boldsymbol{\pi}}) \geq L_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}}) - \frac{4\epsilon\gamma}{(1-\gamma)^2}\alpha^2,\ \text{where } \alpha = D_{\mathrm{TV}}^{\max}(\boldsymbol{\pi}, \tilde{\boldsymbol{\pi}}),\ \epsilon = \max_{s,a} |\boldsymbol{A}_{\boldsymbol{\pi}}(s,a)| \tag{18}$$
With $D_{\mathrm{TV}}(p\|q)^2 \leq D_{\mathrm{KL}}(p\|q)$, we define $D_{\mathrm{KL}}^{\max}(\boldsymbol{\pi}, \tilde{\boldsymbol{\pi}}) = \max_s D_{\mathrm{KL}}(\boldsymbol{\pi}(\cdot|s)\|\tilde{\boldsymbol{\pi}}(\cdot|s))$. Then we have
$$\eta(\tilde{\boldsymbol{\pi}}) \geq L_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}}) - C D_{\mathrm{KL}}^{\max}(\boldsymbol{\pi}, \tilde{\boldsymbol{\pi}}),\ \text{where } C = \frac{4\epsilon\gamma}{(1-\gamma)^2} \tag{19}$$
Let $M_i(\boldsymbol{\pi}) = L_{\boldsymbol{\pi}_i}(\boldsymbol{\pi}) - C D_{\mathrm{KL}}^{\max}(\boldsymbol{\pi}_i, \boldsymbol{\pi})$, we have
$$\eta(\boldsymbol{\pi}_{i+1}) \geq M_i(\boldsymbol{\pi}_{i+1}), \eta(\boldsymbol{\pi}_i) = M_i(\boldsymbol{\pi}_i) \Rightarrow \eta(\boldsymbol{\pi}_{i+1}) - \eta(\boldsymbol{\pi}_i) \geq M_i(\boldsymbol{\pi}_{i+1}) - M_i(\boldsymbol{\pi}_i)$$
So if we define $\boldsymbol{\pi}_{i+1}$ as $\boldsymbol{\pi}_{i+1} = \underset{\boldsymbol{\pi}}{\operatorname{argmax}}\ M_i(\boldsymbol{\pi})$, then
$$\eta(\boldsymbol{\pi}_0) \leq \eta(\boldsymbol{\pi}_1) \leq \eta(\boldsymbol{\pi}_2) \leq \cdots \tag{20}$$
Therefore, $M_i(\boldsymbol{\pi})$ becomes a surrogate function that we want to maximize.

2.3 Relations Between Several Distance Measures

The total variation distance between two probability measures μ and ν on a sigma-algebra $\mathcal{F}$ of subsets of the sample space Ω is defined as $\delta(\mu,\nu) = \sup_{A\in\mathcal{F}} \left|\mu(A) - \nu(A)\right|$. The total variation distance is related to the Kullback–Leibler divergence by Pinsker's inequality:

$$\delta(\mu,\nu) \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\mu,\nu)}$$

And from [1], we have

$$\delta(\mu,\nu) = \frac{1}{2}\|\mu-\nu\|_1 = \frac{1}{2}\sum_{\omega\in\Omega} |\mu(\omega) - \nu(\omega)|$$

Thus,

$$\|\mu-\nu\|_1^2 \leq 2 D_{\mathrm{KL}}(\mu,\nu)$$

3 TRPO in Matrix Form

In this section, we first write TRPO in matrix form and we will see there are more ways to find the local approximation as (16). We derive several interesting properties of these approximations. The norm $\|\cdot\|$ we use in this section is 1-norm. Note that Π (matrix) is just a rewriting of $\boldsymbol{\pi}$ (vector). They represent the same policy. This relation is the same for parameterized Π_θ and $\boldsymbol{\pi}_\theta$.

With notations introduced in Section 2.1, η can be written as

$$\eta(\boldsymbol{\pi}) = \boldsymbol{\rho}_0^\top \boldsymbol{V}_{\boldsymbol{\pi}} = \boldsymbol{\rho}_0^\top \sum_{i=0}^{\infty} \gamma^i (\Pi P)^i \Pi \boldsymbol{r}$$

From (14), we see that the advantage function $\boldsymbol{A}_{\boldsymbol{\pi}}$, the vector form of $\boldsymbol{A}_{\boldsymbol{\pi}}(s,a)$, can be written as

$$\boldsymbol{A}_{\boldsymbol{\pi}} = \boldsymbol{Q}_{\boldsymbol{\pi}} - \Xi^T \boldsymbol{V}_{\boldsymbol{\pi}}$$

From (9),(11) and (12), it is easy to verify that $\Pi \boldsymbol{A}_{\boldsymbol{\pi}} = 0$. Then

$$(I - \gamma P\Pi)^{-1} \boldsymbol{A}_{\boldsymbol{\pi}} = \sum_{i=0}^{\infty} \gamma^i (P\Pi)^i \boldsymbol{A}_{\boldsymbol{\pi}} = \boldsymbol{A}_{\boldsymbol{\pi}} + \sum_{i=1}^{\infty} \gamma^i (P\Pi)^i \boldsymbol{A}_{\boldsymbol{\pi}} = \boldsymbol{A}_{\boldsymbol{\pi}} \tag{21}$$

Since $\eta(\tilde{\boldsymbol{\pi}}) = \boldsymbol{\rho}_0^T \boldsymbol{V}_{\tilde{\boldsymbol{\pi}}}$ amd $\eta(\boldsymbol{\pi}) = \boldsymbol{\rho}_0^T \boldsymbol{V}_{\boldsymbol{\pi}}$, (15) can easily be shown as follows,

$$\begin{aligned}
\mathbb{E}_{s_0,a_0,\cdots\sim\tilde{\boldsymbol{\pi}}}\left[\textstyle\sum_{t=0}^{\infty} \gamma^t \boldsymbol{A}_{\boldsymbol{\pi}}(s_t,a_t)\right] &= \boldsymbol{\rho}_0^T \tilde{\Pi}(I-\gamma P\tilde{\Pi})^{-1} \boldsymbol{A}_{\boldsymbol{\pi}} = \boldsymbol{\rho}_0^T \tilde{\Pi}(I-\gamma P\tilde{\Pi})^{-1}[\boldsymbol{Q}_{\boldsymbol{\pi}} - \Xi^T \boldsymbol{V}_{\boldsymbol{\pi}}] \\
&= \boldsymbol{\rho}_0^T \tilde{\Pi}(I-\gamma P\tilde{\Pi})^{-1}[\boldsymbol{r} + \gamma P \boldsymbol{V}_{\boldsymbol{\pi}} - \Xi^T \boldsymbol{V}_{\boldsymbol{\pi}}] \\
&= \boldsymbol{\rho}_0^T \tilde{\Pi} \sum_{i=0}^{\infty} \gamma^i (P\tilde{\Pi})^i \boldsymbol{r} + \boldsymbol{\rho}_0^T \tilde{\Pi} \textstyle\sum_{i=0}^{\infty} \gamma^i (P\tilde{\Pi})^i \gamma P \boldsymbol{V}_{\boldsymbol{\pi}} - \boldsymbol{\rho}_0^T \tilde{\Pi} \sum_{i=0}^{\infty} \gamma^i (P\Pi)^i \Xi^T \boldsymbol{V}_{\boldsymbol{\pi}} \\
&= \boldsymbol{\rho}_0^T \textstyle\sum_{i=0}^{\infty} \gamma^i (\tilde{\Pi}P)^i \tilde{\Pi} \boldsymbol{r} + \boldsymbol{\rho}_0^T \sum_{i=0}^{\infty} \gamma^{i+1} (\tilde{\Pi}P)^{i+1} \boldsymbol{V}_{\boldsymbol{\pi}} - \boldsymbol{\rho}_0^T \sum_{i=0}^{\infty} \gamma^i (\Pi P)^i \boldsymbol{V}_{\boldsymbol{\pi}} \\
&= \boldsymbol{\rho}_0^T \boldsymbol{V}_{\tilde{\boldsymbol{\pi}}} - \boldsymbol{\rho}_0^T \boldsymbol{V}_{\boldsymbol{\pi}} = \eta(\tilde{\boldsymbol{\pi}}) - \eta(\boldsymbol{\pi})
\end{aligned}$$

Thus (4) can be written in the following form

$$\eta(\tilde{\boldsymbol{\pi}}) = \eta(\boldsymbol{\pi}) + \boldsymbol{\rho}_0^T \tilde{\Pi}(I-\gamma P\tilde{\Pi})^{-1} A_{\boldsymbol{\pi}} = \eta(\boldsymbol{\pi}) + \boldsymbol{\rho}_0^T (I-\gamma\tilde{\Pi}P)^{-1} \tilde{\Pi} A_{\boldsymbol{\pi}} \equiv \eta(\boldsymbol{\pi}) + f_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}})$$

It is easy to see that $f_{\boldsymbol{\pi}}(\boldsymbol{\pi}) = 0$. Suppose $\tilde{\Pi} = \Pi + d\Pi$ ($d\Pi \to 0$ and $d\Pi \mathbf{1}_{|S||A|\times 1} = 0$) and if we want $\eta(\tilde{\boldsymbol{\pi}}) \geq \eta(\boldsymbol{\pi})$, we should have $f_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}}) = f_{\boldsymbol{\pi}}(\boldsymbol{\pi} + d\boldsymbol{\pi}) \geq 0$. Note that

$$\begin{aligned}
f_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}}) = f_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}}) - f_{\boldsymbol{\pi}}(\boldsymbol{\pi}) &= \left. df_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}})\right|_{\tilde{\boldsymbol{\pi}}=\boldsymbol{\pi}} \\
&= \mathrm{trace}\left(\boldsymbol{\rho}_0^T (I-\gamma\tilde{\Pi}P)^{-1} \gamma (d\Pi) P (I-\gamma\tilde{\Pi}P)^{-1} \tilde{\Pi} \boldsymbol{A}_{\boldsymbol{\pi}} + \boldsymbol{\rho}_0^T (I-\gamma\tilde{\Pi}P)^{-1} (d\Pi) \boldsymbol{A}_{\boldsymbol{\pi}}\right)\Big|_{\tilde{\boldsymbol{\pi}}=\boldsymbol{\pi}} \qquad (22) \\
&= \mathrm{trace}\left(\boldsymbol{\rho}_0^T (I-\gamma\Pi P)^{-1} (d\Pi) \boldsymbol{A}_{\boldsymbol{\pi}}\right) = \mathrm{trace}\left(\boldsymbol{A}_{\boldsymbol{\pi}} \boldsymbol{\rho}_0^T (I-\gamma\Pi P)^{-1} d\Pi\right)
\end{aligned}$$

Therefore,

$$\left(\nabla_{\tilde{\Pi}} f_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}})\Big|_{\tilde{\boldsymbol{\pi}}=\boldsymbol{\pi}}\right)^T = \boldsymbol{A}_{\boldsymbol{\pi}} \boldsymbol{\rho}_0^T (I-\gamma\Pi P)^{-1} \tag{23}$$

There are 6 ways to set the approximation function as (16):

$$L^1_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}}) = \eta(\boldsymbol{\pi}) + \boldsymbol{\rho}_0^T\tilde{\Pi}(I-\gamma P\tilde{\Pi})^{-1}\boldsymbol{A}_{\boldsymbol{\pi}} = \eta(\boldsymbol{\pi}) + \boldsymbol{\rho}_0^T(I-\gamma\tilde{\Pi}P)^{-1}\tilde{\Pi}\boldsymbol{A}_{\boldsymbol{\pi}} \text{ (policy gradient)}$$
$$L^2_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}}) = \eta(\boldsymbol{\pi}) + \boldsymbol{\rho}_0^T\Pi(I-\gamma P\tilde{\Pi})^{-1}\boldsymbol{A}_{\boldsymbol{\pi}}$$
$$L^3_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}}) = \eta(\boldsymbol{\pi}) + \boldsymbol{\rho}_0^T\tilde{\Pi}(I-\gamma P\Pi)^{-1}\boldsymbol{A}_{\boldsymbol{\pi}}$$
$$L^4_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}}) = \eta(\boldsymbol{\pi}) + \boldsymbol{\rho}_0^T(I-\gamma\Pi P)^{-1}\tilde{\Pi}\boldsymbol{A}_{\boldsymbol{\pi}} \text{ (TRPO)}$$
$$L^5_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}}) = \eta(\boldsymbol{\pi}) + \boldsymbol{\rho}_0^T(I-\gamma\tilde{\Pi}P)^{-1}\Pi\boldsymbol{A}_{\boldsymbol{\pi}} = \eta(\boldsymbol{\pi}) \text{ (trivial)}$$
$$L^6_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}}) = \eta(\boldsymbol{\pi}) + \boldsymbol{\rho}_0^T\Pi(I-\gamma P\Pi)^{-1}\boldsymbol{A}_{\boldsymbol{\pi}} = \eta(\boldsymbol{\pi}) + \boldsymbol{\rho}_0^T(I-\gamma\Pi P)^{-1}\Pi\boldsymbol{A}_{\boldsymbol{\pi}} = \eta(\boldsymbol{\pi}) \text{ (trivial)}$$

We will discuss $L^2_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}})$, $L^3_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}})$ and $L^4_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}})$ in the following subsections.

3.1 Approximation Function $L^2_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}})$

For $L^2_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}}) = \eta(\boldsymbol{\pi}) + \boldsymbol{\rho}_0^T\Pi(I-\gamma P\tilde{\Pi})^{-1}\boldsymbol{A}_{\boldsymbol{\pi}}$, it is easy to see that $L^2_{\boldsymbol{\pi}}(\boldsymbol{\pi}) = 0$. For parametrized $\boldsymbol{\pi}_\theta$, to test (17), we have

$$\begin{aligned} dL^2_{\boldsymbol{\pi}_{\theta_0}}(\boldsymbol{\pi}_\theta)\Big|_{\theta=\theta_0} &= d\boldsymbol{\rho}_0^T\Pi_{\theta_0}(I-\gamma P\Pi_\theta)^{-1}\boldsymbol{A}_{\boldsymbol{\pi}_{\theta_0}}\Big|_{\theta=\theta_0} \\ &= \operatorname{trace}\Big(\boldsymbol{\rho}_0^T\Pi_{\theta_0}(I-\gamma P\Pi_\theta)^{-1}\gamma P(d\Pi_\theta)(I-\gamma P\Pi_\theta)^{-1}\boldsymbol{A}_{\boldsymbol{\pi}_{\theta_0}}\Big)\Big|_{\theta=\theta_0} \\ &= \operatorname{trace}\Big(\boldsymbol{A}_{\boldsymbol{\pi}_{\theta_0}}\boldsymbol{\rho}_0^T\gamma\Pi_{\theta_0}P(I-\gamma\Pi_{\theta_0}P)^{-1}(d\Pi_\theta)\Big)\Big|_{\theta=\theta_0} \end{aligned} \tag{24}$$

It is easy to see

$$\left(\nabla_\theta L^2_{\boldsymbol{\pi}_{\theta_0}}(\boldsymbol{\pi}_\theta)\Big|_{\theta=\theta_0}\right)^T = \boldsymbol{A}_{\boldsymbol{\pi}_{\theta_0}}\boldsymbol{\rho}_0^T\gamma\Pi_{\theta_0}P(I-\gamma\Pi_{\theta_0}P)^{-1}\frac{d\Pi_\theta}{d\theta}\Big|_{\theta=\theta_0} \neq \left(\nabla_\theta\eta(\boldsymbol{\pi}_\theta)\Big|_{\theta=\theta_0}\right)^T \tag{25}$$

which means $L^2_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}})$ does not match $\eta(\boldsymbol{\pi})$ to the first order. To calculate the difference between $L^2_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}})$ and $\eta(\boldsymbol{\pi})$, we have

$$\eta(\tilde{\boldsymbol{\pi}}) - L^2_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}}) = \boldsymbol{\rho}_0^T(\tilde{\Pi}-\Pi)(I-\gamma P\tilde{\Pi})^{-1}\boldsymbol{A}_{\boldsymbol{\pi}} = \boldsymbol{\rho}_0^T(d\Pi)(I-\gamma P\tilde{\Pi})^{-1}\boldsymbol{A}_{\boldsymbol{\pi}}$$

Then we have

$$\begin{aligned} \left\|\eta(\tilde{\boldsymbol{\pi}}) - L^2_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}})\right\| &= \left\|\boldsymbol{\rho}_0^T(d\Pi)(I-\gamma P\tilde{\Pi})^{-1}\boldsymbol{A}_{\boldsymbol{\pi}}\right\| \leq \left\|\boldsymbol{\rho}_0^T\right\|\|d\Pi\|\left\|\textstyle\sum_{i=0}^{\infty}\gamma^i(P\tilde{\Pi})^i\right\|\|\boldsymbol{A}_{\boldsymbol{\pi}}\| \\ &\leq \frac{\sqrt{2D^{\max}_{\mathrm{KL}}(\boldsymbol{\pi},\tilde{\boldsymbol{\pi}})}}{1-\gamma}\|\boldsymbol{A}_{\boldsymbol{\pi}}\| \end{aligned}$$

3.2 Approximation Function $L^3_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}})$

For $L^3_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}}) = \eta(\boldsymbol{\pi}) + \boldsymbol{\rho}_0^T\tilde{\Pi}(I-\gamma P\Pi)^{-1}\boldsymbol{A}_{\boldsymbol{\pi}}$, it is easy to see that $L^3_{\boldsymbol{\pi}}(\boldsymbol{\pi}) = 0$. Moreover, to test (16), we have

$$\begin{aligned} dL^3_{\boldsymbol{\pi}_{\theta_0}}(\boldsymbol{\pi}_\theta)\Big|_{\theta=\theta_0} &= d\boldsymbol{\rho}_0^T\Pi_\theta(I-\gamma P\Pi_{\theta_0})^{-1}\boldsymbol{A}_{\boldsymbol{\pi}_{\theta_0}}\Big|_{\theta=\theta_0} \\ &= \operatorname{trace}\Big(\boldsymbol{\rho}_0^Td\Pi_\theta\boldsymbol{A}_{\boldsymbol{\pi}_{\theta_0}}\Big)\Big|_{\theta=\theta_0} = \operatorname{trace}\Big(\boldsymbol{A}_{\boldsymbol{\pi}_{\theta_0}}\boldsymbol{\rho}_0^Td\Pi_\theta\Big)\Big|_{\theta=\theta_0} \end{aligned} \tag{26}$$

And

$$\left(\nabla_\theta L^3_{\boldsymbol{\pi}_{\theta_0}}(\boldsymbol{\pi}_\theta)\Big|_{\theta=\theta_0}\right)^T = \boldsymbol{A}_{\boldsymbol{\pi}_{\theta_0}}\boldsymbol{\rho}_0^T\frac{d\Pi_\theta}{d\theta}\Big|_{\theta=\theta_0} \neq \nabla_\theta\eta(\boldsymbol{\pi}_\theta)\Big|_{\theta=\theta_0} \tag{27}$$

which means $L^3_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}})$ does not match $\eta(\boldsymbol{\pi})$ to the first order. But it is easy to check,

$$\nabla_\theta L^2_{\boldsymbol{\pi}_{\theta_0}}(\boldsymbol{\pi}_\theta)\Big|_{\theta=\theta_0} + \nabla_\theta L^3_{\boldsymbol{\pi}_{\theta_0}}(\boldsymbol{\pi}_\theta)\Big|_{\theta=\theta_0} = \nabla_\theta\eta(\boldsymbol{\pi}_\theta)\Big|_{\theta=\theta_0} = \nabla_\theta L^4_{\boldsymbol{\pi}_{\theta_0}}(\boldsymbol{\pi}_\theta)\Big|_{\theta=\theta_0} \tag{28}$$

which means the gradient of $L^2_{\boldsymbol{\pi}_{\theta_0}}(\boldsymbol{\pi}_\theta)$ and $L^3_{\boldsymbol{\pi}_{\theta_0}}(\boldsymbol{\pi}_\theta)$ is a gradient decomposition of $\eta(\boldsymbol{\pi}_\theta)$ at θ_0. To calculate the difference between $L^3_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}})$ and $\eta(\boldsymbol{\pi})$, we have

$$\begin{aligned} \eta(\tilde{\boldsymbol{\pi}}) - L^3_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}}) &= \boldsymbol{\rho}_0^T\tilde{\Pi}\left((I-\gamma P\tilde{\Pi})^{-1} - (I-\gamma P\Pi)^{-1}\right)\boldsymbol{A}_{\boldsymbol{\pi}} = \boldsymbol{\rho}_0^T\tilde{\Pi}\left((I-\gamma P\Pi)^{-1}(\gamma Pd\Pi)(I-\gamma P\Pi)^{-1}\right)\boldsymbol{A}_{\boldsymbol{\pi}} \\ &= \boldsymbol{\rho}_0^T\tilde{\Pi}\left((I-\gamma P\Pi)^{-1}(\gamma Pd\Pi)\right)\boldsymbol{A}_{\boldsymbol{\pi}} \end{aligned}$$

Then we have

$$\left\|\eta(\tilde{\boldsymbol{\pi}}) - L^3_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}})\right\| = \left\|\boldsymbol{\rho}_0^T \tilde{\Pi}\left((I-\gamma P\Pi)^{-1}(\gamma P d\Pi)\right)\boldsymbol{A}_{\boldsymbol{\pi}}\right\| \le \left\|\boldsymbol{\rho}_0^T\right\|\|d\Pi\|\left\|\sum_{i=0}^{\infty}\gamma^{i+1}(P\Pi)^i P\right\|\left\|\tilde{\Pi}\right\|\|\boldsymbol{A}_{\boldsymbol{\pi}}\|$$
$$= \frac{\gamma\|d\Pi\|\|\boldsymbol{A}_{\boldsymbol{\pi}}\|}{1-\gamma} \le \frac{\gamma\sqrt{2D_{\mathrm{KL}}^{\max}(\boldsymbol{\pi},\tilde{\boldsymbol{\pi}})}\|\boldsymbol{A}_{\boldsymbol{\pi}}\|}{1-\gamma}$$

3.3 Approximation Function $L^4_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}})$

If we set

$$L_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}}) = \eta(\boldsymbol{\pi}) + \boldsymbol{\rho}_0^T(I-\gamma P\Pi)^{-1}\tilde{\Pi}A_{\boldsymbol{\pi}} = \eta(\boldsymbol{\pi}) + f_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}})$$

Then,

$$\begin{aligned}\eta(\tilde{\boldsymbol{\pi}}) - L^4_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}}) &= \boldsymbol{\rho}_0^T\left((I-\gamma P\tilde{\Pi})^{-1} - (I-\gamma P\Pi)^{-1}\right)\tilde{\Pi}A_{\boldsymbol{\pi}}\\ &= \boldsymbol{\rho}_0^T\left((I-\gamma P(\Pi+d\Pi))^{-1} - (I-\gamma P\Pi)^{-1}\right)(\Pi+d\Pi)A_{\boldsymbol{\pi}}\\ &= \boldsymbol{\rho}_0^T(I-\gamma P\Pi)^{-1}(\gamma P d\Pi)(I-\gamma P\Pi)^{-1}d\Pi A_{\boldsymbol{\pi}}\end{aligned}$$

Then we have

$$\begin{aligned}\left\|\eta(\tilde{\boldsymbol{\pi}}) - L^4_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}})\right\| &= \left\|\boldsymbol{\rho}_0^T(I-\gamma P\Pi)^{-1}(\gamma P d\Pi)(I-\gamma P\Pi)^{-1}d\Pi A_{\boldsymbol{\pi}}\right\|\\ &\le \left\|\boldsymbol{\rho}_0^T\right\|\left\|(I-\gamma P\Pi)^{-1}\right\|\left\|\gamma P\right\|\|d\Pi\|\left\|(I-\gamma P\Pi)^{-1}\right\|\|d\Pi\|\|A_{\boldsymbol{\pi}}\|\\ &= \frac{\gamma\|d\Pi\|^2\|A_{\boldsymbol{\pi}}\|}{(1-\gamma)^2} \le \frac{2\gamma D_{\mathrm{KL}}^{\max}(\boldsymbol{\pi},\tilde{\boldsymbol{\pi}})\|A_{\boldsymbol{\pi}}\|}{(1-\gamma)^2}\end{aligned}$$

which is the same as (19).

Another way to consider this inequality is

$$\begin{aligned}\left\|\eta(\tilde{\boldsymbol{\pi}}) - L^4_{\boldsymbol{\pi}}(\tilde{\boldsymbol{\pi}})\right\| &= \left\|\boldsymbol{\rho}_0^T\left((I-\gamma P\tilde{\Pi})^{-1} - (I-\gamma P\Pi)^{-1}\right)\tilde{\Pi}A_{\boldsymbol{\pi}}\right\|\\ &\le \left\|\boldsymbol{\rho}_0^T\left((I-\gamma P\tilde{\Pi})^{-1}\right)\tilde{\Pi}A_{\boldsymbol{\pi}}\right\| + \left\|\boldsymbol{\rho}_0^T\left((I-\gamma P\Pi)^{-1}\right)\tilde{\Pi}A_{\boldsymbol{\pi}}\right\|\\ &\le \left\|\boldsymbol{\rho}_0^T\right\|\left\|(I-\gamma P\tilde{\Pi})^{-1}\right\|\|d\Pi\|\|A_{\boldsymbol{\pi}}\| + \left\|\boldsymbol{\rho}_0^T\right\|\left\|(I-\gamma P\Pi)^{-1}\right\|\|d\Pi\|\|A_{\boldsymbol{\pi}}\|\\ &= \frac{2\|d\Pi\|\|A_{\boldsymbol{\pi}}\|}{1-\gamma} \le \frac{2\sqrt{2D_{\mathrm{KL}}^{\max}(\boldsymbol{\pi},\tilde{\boldsymbol{\pi}})}\|\boldsymbol{A}_{\boldsymbol{\pi}}\|}{1-\gamma}\end{aligned}$$

Consider these two bounds: if $\sqrt{D_{\mathrm{KL}}^{\max}(\boldsymbol{\pi},\tilde{\boldsymbol{\pi}})} > \frac{\sqrt{2}(1-\gamma)}{\gamma}$, the second bound is tighter; otherwise, the first bound is tighter.

4 Other Potential Applications

There are some other applications of this set of notations, *e.g.*

- If we know $\nabla_{\boldsymbol{\pi}_\theta}\eta(\boldsymbol{\pi}_\theta)$ and want to get the $\nabla_\theta\eta(\boldsymbol{\pi}_\theta)$, *i.e.* when we know the optimal direction to update policy but we do not know how to control the parameters to make the policy turn to this direction, we can do

  $$\min_{\Delta\theta} d(\Pi_{opt}, \Pi_{\theta+\Delta\theta} - \Pi_\theta),$$

  where $d(\cdot,\cdot)$ is a distance measure, Π_{opt} is the optimal direction to update policy.
- Instead of estimating value function $\boldsymbol{V}_{\boldsymbol{\pi}}$, we can estimate the environment dynamics P, which is independent of $\boldsymbol{\pi}$. Each transition information in the trajectory is valuable no matter the reward signal is detected or not. Then, we can use the estimated $\hat{P}$ to update π directly. The estimation (or construction) of $\hat{P}$ is not affected by the update of policy.

References

[1] D. A. Levin and Y. Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[2] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897, 2015.

[3] R. S. Sutton and A. G. Barto. Reinforcement learning: an introduction. 2018.

[4] T. Wang, M. Bowling, and D. Schuurmans. Dual representations for dynamic programming and reinforcement learning. In *2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*, pages 44–51. IEEE, 2007.