

## Advanced Basic Theoretical Reinforcement Learning Result

By advanced, I mean that one would not find these results in a graduate level RL class, or even popular RL textbooks. The results are mostly found in RL research papers.

By basic, I mean that these results seem general enough that they can be used as building block for more complicated result.

Theory usually progresses one lemma at a time, so I thought it is a good idea to collate these results into one document for easy reference.

For each result, if a cleanly written proof is available, then I link to it. Otherwise, I provide the proof myself.

The theoretical result in each section focuses on one theme, such as model-based RL.

## 1 Relating performances of a policy in 2 dynamics model to differences between the 2 models

### 1.1 trajectory-level performance vs. per-state state-action values

Given 2 dynamics model  $M$  and  $\widehat{M}$ , how can we related the trajectory-level performance of a policy  $\pi$  in the 2 models to the differences in per-state state-action values when states are sampled from either dynamics model? Formal result is in subsection 3.1 [1].

### 1.2 trajectory-level performance vs. per-state per-action model outputs

Given 2 dynamics model  $M$  and  $\widehat{M}$ , how can we related the trajectory-level performance of a policy  $\pi$  in the 2 models to the difference in per-state per-action differences in output of the 2 models? Formal result is in subsection 3.2 [1].

## 2 Relating per-state action distribution and future state visitation distribution

Given two policies  $\pi_1$  and  $\pi_2$ , how can we related their per-state difference in action distribution to the differences in their future discounted state visitation distribution? Formal result is in subsection 3.3 [2].

### 3 Formal Results

#### 3.1 Formal Result for subsection 1.1

$M(\cdot|s, a)$  denotes the distribution on the next state given the current state and action.

$S_t^{\pi, M}$  denotes the random variable of the states at step  $t$  when we execute policy  $\pi$  on the dynamic model starting from  $S_0$ .

$P_X$  denotes the density function for the random variable  $X$ .

$\rho^{\pi, M}$  denotes the discounted distribution of the states visited by  $\pi$  on  $M$ . That is,  $\rho^{\pi, M} = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t P_{S_t^{\pi, M}}$ .

$$V^{\pi, M} = E_{s_0 \sim P_{S_0}} [V^{\pi, M}(s_0)].$$

The rest of the notation should be standard in the RL world. Please refer to section 2 in the paper.

$$\text{Let } G^{\pi, \widehat{M}, M}(s, a) = E_{\hat{s}' \sim \widehat{M}(\cdot|s, a)} [V^{\pi, \widehat{M}}(\hat{s}')] - E_{s' \sim M(\cdot|s, a)} [V^{\pi, \widehat{M}}(s')]$$

For any policy  $\pi$  and dynamical model  $M, \widehat{M}$ , we have:

$$V^{\pi, \widehat{M}} - V^{\pi, M} = \frac{\gamma}{1 - \gamma} E_{\substack{S \sim \rho^{\pi, M} \\ A \sim \pi(\cdot|S)}} [G^{\pi, \widehat{M}, M}(S, A)]$$

Proof is available in subsection 4.1

#### 3.2 Formal result for subsection 1.2

The notation is the same as in subsection 3.1. Suppose that  $V^{\pi, \widehat{M}}$  is  $L$ -Lipschitz, that is:

$$\forall s, s' \in S, |V^{\pi, \widehat{M}}(s) - V^{\pi, \widehat{M}}(s')| \leq L \|s - s'\|$$

For any policy  $\pi$  and deterministic dynamical model  $M, \widehat{M}$ , we have:

$$|V^{\pi, \widehat{M}} - V^{\pi, M}| \leq \frac{\gamma}{1 - \gamma} L E_{\substack{S \sim \rho^{\pi, M} \\ A \sim \pi(\cdot|S)}} [\|\widehat{M}(S, A) - M(S, A)\|]$$

Proof is available in subsection 4.2

#### 3.3 Formal result for section 2

$$\text{Let } d_{\pi} = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t d_{\pi}^t.$$

Given any two policy  $\pi_1$  and  $\pi_2$  such that  $E_{s \sim d_{\pi_1}} [D_{TV}(\pi_1(\cdot|s), \pi_2(\cdot|s))] \leq \alpha$  then we have:

$$\|d_{\pi_1} - d_{\pi_2}\|_1 \leq \frac{2\alpha}{1 - \gamma}$$

The proof is clearly presented in the appendix of [2].

## 4 Proofs

### 4.1 Proof for subsection 3.1

Let  $W_j(s)$  be the cumulative reward when we use dynamical model  $M$  for  $j$  steps and then use  $\widehat{M}$ , that is:

$$W_j(s) = \underset{\substack{\forall t \geq 0, A_t \sim \pi(\cdot | S_t) \\ \forall j \geq t \geq 0, S_{t+1} \sim M(\cdot | S_t, A_t) \\ \forall t > j, S_{t+1} \sim \widehat{M}(\cdot | S_t, A_t)}}{E} \left[ \sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) \middle| S_0 = s \right]$$

By definition,  $W_{\infty}(s) = V^{\pi, M}(s)$  and  $W_0(s) = V^{\pi, \widehat{M}}(s)$ , then:

$$V^{\pi, M}(s) - V^{\pi, \widehat{M}}(s) = \sum_{j=0}^{\infty} (W_{j+1}(s) - W_j(s))$$

We now decompose  $W_j(s)$  into sum of rewards obtained when sampling under  $M$  and  $\widehat{M}$ :

$$\begin{aligned} W_j(s) &= \underset{\substack{\forall t \geq 0, A_t \sim \pi(\cdot | S_t) \\ \forall j \geq t \geq 0, S_{t+1} \sim M(\cdot | S_t, A_t) \\ \forall t > j, S_{t+1} \sim \widehat{M}(\cdot | S_t, A_t)}}{E} \left[ \sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) \middle| S_0 = s \right] \\ &= \underset{\substack{\forall t \geq 0, A_t \sim \pi(\cdot | S_t) \\ \forall j \geq t \geq 0, S_{t+1} \sim M(\cdot | S_t, A_t) \\ \forall t > j, S_{t+1} \sim \widehat{M}(\cdot | S_t, A_t)}}{E} \left[ \sum_{t=0}^j \gamma^t R(S_t, A_t) \middle| S_0 = s \right] + \underset{\substack{\forall t \geq 0, A_t \sim \pi(\cdot | S_t) \\ \forall j \geq t \geq 0, S_{t+1} \sim M(\cdot | S_t, A_t) \\ \forall t > j, S_{t+1} \sim \widehat{M}(\cdot | S_t, A_t)}}{E} \left[ \sum_{t=j+1}^{\infty} \gamma^t R(S_t, A_t) \middle| S_0 = s \right] \\ &= \underset{\substack{\forall t \geq 0, A_t \sim \pi(\cdot | S_t) \\ \forall j \geq t \geq 0, S_{t+1} \sim M(\cdot | S_t, A_t) \\ \forall t > j, S_{t+1} \sim \widehat{M}(\cdot | S_t, A_t)}}{E} \left[ \sum_{t=0}^j \gamma^t R(S_t, A_t) \middle| S_0 = s \right] + \underset{\substack{S_j \sim \rho_{S_j}^{\pi, M}(\cdot | S_0 = s) \\ A_j \sim \pi(\cdot | S_j)}}{E} \left[ \underset{\widehat{S}_{j+1} \sim \widehat{M}(\cdot | S_j, A_j)}{E} \left[ \gamma^{j+1} V^{\pi, \widehat{M}}(\widehat{S}_{j+1}) \right] \right] \end{aligned}$$

Proof of the last equality is below:

$$\begin{aligned}
& \begin{matrix} E \\ \forall t \geq 0, A_t \sim \pi(\cdot | S_t) \\ \forall j \geq t \geq 0, S_{t+1} \sim M(\cdot | S_t, A_t) \\ \forall t > j, S_{t+1} \sim \widehat{M}(\cdot | S_t, A_t) \end{matrix} \left[ \sum_{t=j+1}^{\infty} \gamma^t R(S_t, A_t) \middle| S_0 = s \right] \\
&= \sum_{a_0} \pi(a_0 | s) \sum_{s_1} M(s_1 | s, a_0) \dots \sum_{a_{j-1}} \pi(a_{j-1} | s_{j-1}) \sum_{s_j} M(s_j | s_{j-1}, a_{j-1}) \begin{matrix} E \\ t=j, A_t \sim \pi(\cdot | s_j) \\ \forall t > j, A_t \sim \pi(\cdot | S_t) \\ S_{t+1} \sim \widehat{M}(\cdot | S_t, A_t) \end{matrix} \left[ \sum_{t=j+1}^{\infty} \gamma^t R(S_t, A_t) \middle| S_j = s_j \right] \\
&= \sum_{a_0} \sum_{s_1} \dots \sum_{a_{j-1}} \sum_{s_j} \pi(a_0 | s) M(s_1 | s, a_0) \dots \pi(a_{j-1} | s_{j-1}) M(s_j | s_{j-1}, a_{j-1}) \begin{matrix} E \\ t=j, A_t \sim \pi(\cdot | s_j) \\ \forall t > j, A_t \sim \pi(\cdot | S_t) \\ S_{t+1} \sim \widehat{M}(\cdot | S_t, A_t) \end{matrix} \left[ \sum_{t=j+1}^{\infty} \gamma^t R(S_t, A_t) \middle| S_j = s_j \right] \\
&= \sum_{a_0, s_1, \dots, a_{j-1}} \sum_{s_j} P(a_0, s_1, \dots, a_{j-1}, s_j | S_0 = s) \begin{matrix} E \\ t=j, A_t \sim \pi(\cdot | s_j) \\ \forall t > j, A_t \sim \pi(\cdot | S_t) \\ S_{t+1} \sim \widehat{M}(\cdot | S_t, A_t) \end{matrix} \left[ \sum_{t=j+1}^{\infty} \gamma^t R(S_t, A_t) \middle| S_j = s_j \right] \\
&= \sum_{s_j} \sum_{a_0, s_1, \dots, a_{j-1}} P(a_0, s_1, \dots, a_{j-1}, s_j | S_0 = s) \begin{matrix} E \\ t=j, A_t \sim \pi(\cdot | s_j) \\ \forall t > j, A_t \sim \pi(\cdot | S_t) \\ S_{t+1} \sim \widehat{M}(\cdot | S_t, A_t) \end{matrix} \left[ \sum_{t=j+1}^{\infty} \gamma^t R(S_t, A_t) \middle| S_j = s_j \right] \\
&= \sum_{s_j} P(s_j | S_0 = s) \begin{matrix} E \\ t=j, A_t \sim \pi(\cdot | s_j) \\ \forall t > j, A_t \sim \pi(\cdot | S_t) \\ S_{t+1} \sim \widehat{M}(\cdot | S_t, A_t) \end{matrix} \left[ \sum_{t=j+1}^{\infty} \gamma^t R(S_t, A_t) \middle| S_j = s_j \right] \\
&= \sum_{s_j} \rho_{S_j}^{\pi, M}(s_j | S_0 = s) \begin{matrix} E \\ t=j, A_t \sim \pi(\cdot | s_j) \\ \forall t > j, A_t \sim \pi(\cdot | S_t) \\ S_{t+1} \sim \widehat{M}(\cdot | S_t, A_t) \end{matrix} \left[ \sum_{t=j+1}^{\infty} \gamma^t R(S_t, A_t) \middle| S_j = s_j \right] \\
&= \begin{matrix} E \\ S_j \sim \rho_{S_j}^{\pi, M}(\cdot | S_0 = s) \end{matrix} \left[ \begin{matrix} E \\ t=j, A_t \sim \pi(\cdot | s_j) \\ \forall t > j, A_t \sim \pi(\cdot | S_t) \\ S_{t+1} \sim \widehat{M}(\cdot | S_t, A_t) \end{matrix} \left[ \sum_{t=j+1}^{\infty} \gamma^t R(S_t, A_t) \middle| S_j = s_j \right] \right] \\
&= \begin{matrix} E \\ S_j \sim \rho_{S_j}^{\pi, M}(\cdot | S_0 = s) \\ A_j \sim \pi(\cdot | S_j) \end{matrix} \left[ \begin{matrix} E \\ A_t \sim \pi(\cdot | S_t) \\ S_{t+1} \sim \widehat{M}(\cdot | S_t, A_t) \end{matrix} \left[ \sum_{t=j+1}^{\infty} \gamma^t R(S_t, A_t) \middle| S_{j+1} = \widehat{M}(S_j, A_j) \right] \right] \\
&= \begin{matrix} E \\ S_j \sim \rho_{S_j}^{\pi, M}(\cdot | S_0 = s) \\ A_j \sim \pi(\cdot | S_j) \end{matrix} \left[ \begin{matrix} E \\ \widehat{S}_{j+1} \sim \widehat{M}(\cdot | S_j, A_j) \\ A_t \sim \pi(\cdot | S_t) \\ S_{t+1} \sim \widehat{M}(\cdot | S_t, A_t) \end{matrix} \left[ \sum_{t=j+1}^{\infty} \gamma^t R(S_t, A_t) \middle| S_{j+1} = \widehat{S}_{j+1} \right] \right] \\
&= \begin{matrix} E \\ S_j \sim \rho_{S_j}^{\pi, M}(\cdot | S_0 = s) \\ A_j \sim \pi(\cdot | S_j) \end{matrix} \left[ \begin{matrix} E \\ \widehat{S}_{j+1} \sim \widehat{M}(\cdot | S_j, A_j) \\ A_t \sim \pi(\cdot | S_t) \\ S_{t+1} \sim \widehat{M}(\cdot | S_t, A_t) \end{matrix} \left[ \gamma^{j+1} \sum_{t=j+1}^{\infty} \gamma^{t-j-1} R(S_t, A_t) \middle| S_{j+1} = \widehat{S}_{j+1} \right] \right] \\
&= \begin{matrix} E \\ S_j \sim \rho_{S_j}^{\pi, M}(\cdot | S_0 = s) \\ A_j \sim \pi(\cdot | S_j) \end{matrix} \left[ \begin{matrix} E \\ \widehat{S}_{j+1} \sim \widehat{M}(\cdot | S_j, A_j) \\ A_t \sim \pi(\cdot | S_t) \\ S_{t+1} \sim \widehat{M}(\cdot | S_t, A_t) \end{matrix} \left[ \gamma^{j+1} \sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) \middle| S_0 = \widehat{S}_{j+1} \right] \right] \\
&= \begin{matrix} E \\ S_j \sim \rho_{S_j}^{\pi, M}(\cdot | S_0 = s) \\ A_j \sim \pi(\cdot | S_j) \end{matrix} \left[ \begin{matrix} E \\ \widehat{S}_{j+1} \sim \widehat{M}(\cdot | S_j, A_j) \end{matrix} \left[ \gamma^{j+1} V^{\pi, \widehat{M}}(\widehat{S}_{j+1}) \right] \right]
\end{aligned}$$

By a similar argument:

$$W_{j+1}(s) = \begin{matrix} E \\ \forall t \geq 0, A_t \sim \pi(\cdot | S_t) \\ \forall j \geq t \geq 0, S_{t+1} \sim M(\cdot | S_t, A_t) \\ \forall t > j, S_{t+1} \sim \widehat{M}(\cdot | S_t, A_t) \end{matrix} \left[ \sum_{t=0}^j \gamma^t R(S_t, A_t) \middle| S_0 = s \right] + \begin{matrix} E \\ S_j \sim \rho_{S_j}^{\pi, M}(\cdot | S_0 = s) \\ A_j \sim \pi(\cdot | S_j) \end{matrix} \left[ S_{j+1} \sim M(\cdot | S_j, A_j) \left[ \gamma^{j+1} V^{\pi, \widehat{M}}(S_{j+1}) \right] \right]$$

Thus:

$$\begin{aligned} W_{j+1}(s) - W_j(s) &= \gamma^{j+1} \begin{matrix} E \\ S_j \sim \rho_{S_j}^{\pi, M}(\cdot | S_0 = s) \\ A_j \sim \pi(\cdot | S_j) \end{matrix} \left[ S_{j+1} \sim M(\cdot | S_j, A_j) \left[ V^{\pi, \widehat{M}}(S_{j+1}) \right] - \widehat{S}_{j+1} \sim \widehat{M}(\cdot | S_j, A_j) \left[ V^{\pi, \widehat{M}}(\widehat{S}_{j+1}) \right] \right] \\ \Rightarrow \sum_{j=0}^{\infty} [W_{j+1}(s) - W_j(s)] &= \sum_{j=0}^{\infty} \left[ \gamma^{j+1} \begin{matrix} E \\ S_j \sim \rho_{S_j}^{\pi, M}(\cdot | S_0 = s) \\ A_j \sim \pi(\cdot | S_j) \end{matrix} \left[ S_{j+1} \sim M(\cdot | S_j, A_j) \left[ V^{\pi, \widehat{M}}(S_{j+1}) \right] - \widehat{S}_{j+1} \sim \widehat{M}(\cdot | S_j, A_j) \left[ V^{\pi, \widehat{M}}(\widehat{S}_{j+1}) \right] \right] \right] \\ &= \sum_{j=0}^{\infty} \left[ \gamma^{j+1} \sum_{s'} \rho_{S_j}^{\pi, M}(s' | S_0 = s) \begin{matrix} E \\ A_j \sim \pi(\cdot | s') \end{matrix} \left[ S_{j+1} \sim M(\cdot | s', A_j) \left[ V^{\pi, \widehat{M}}(S_{j+1}) \right] - \widehat{S}_{j+1} \sim \widehat{M}(\cdot | s', A_j) \left[ V^{\pi, \widehat{M}}(\widehat{S}_{j+1}) \right] \right] \right] \\ &= \frac{\gamma}{1-\gamma} \sum_{j=0}^{\infty} \left[ \gamma^j (1-\gamma) \sum_{s'} \rho_{S_j}^{\pi, M}(s' | S_0 = s) \begin{matrix} E \\ A_j \sim \pi(\cdot | s') \end{matrix} \left[ S_{j+1} \sim M(\cdot | s', A_j) \left[ V^{\pi, \widehat{M}}(S_{j+1}) \right] - \widehat{S}_{j+1} \sim \widehat{M}(\cdot | s', A_j) \left[ V^{\pi, \widehat{M}}(\widehat{S}_{j+1}) \right] \right] \right] \\ &= \frac{\gamma}{1-\gamma} \sum_{s'} \left[ \sum_{j=0}^{\infty} \gamma^j (1-\gamma) \rho_{S_j}^{\pi, M}(s' | S_0 = s) \begin{matrix} E \\ A_j \sim \pi(\cdot | s') \end{matrix} \left[ S_{j+1} \sim M(\cdot | s', A_j) \left[ V^{\pi, \widehat{M}}(S_{j+1}) \right] - \widehat{S}_{j+1} \sim \widehat{M}(\cdot | s', A_j) \left[ V^{\pi, \widehat{M}}(\widehat{S}_{j+1}) \right] \right] \right] \\ &= \frac{\gamma}{1-\gamma} \sum_{s'} \left[ \rho^{\pi, M}(s' | S_0 = s) \begin{matrix} E \\ A_j \sim \pi(\cdot | s') \end{matrix} \left[ S_{j+1} \sim M(\cdot | s', A_j) \left[ V^{\pi, \widehat{M}}(S_{j+1}) \right] - \widehat{S}_{j+1} \sim \widehat{M}(\cdot | s', A_j) \left[ V^{\pi, \widehat{M}}(\widehat{S}_{j+1}) \right] \right] \right] \\ &= \frac{\gamma}{1-\gamma} \begin{matrix} E \\ S_j \sim \rho_{S_j}^{\pi, M}(\cdot | S_0 = s) \\ A_j \sim \pi(\cdot | S_j) \end{matrix} \left[ S_{j+1} \sim M(\cdot | S_j, A_j) \left[ V^{\pi, \widehat{M}}(S_{j+1}) \right] - \widehat{S}_{j+1} \sim \widehat{M}(\cdot | S_j, A_j) \left[ V^{\pi, \widehat{M}}(\widehat{S}_{j+1}) \right] \right] \\ &= \frac{\gamma}{1-\gamma} \begin{matrix} E \\ S' \sim \rho^{\pi, M}(\cdot | S_0 = s) \\ A \sim \pi(\cdot | S) \end{matrix} \left[ S' \sim M(\cdot | S, A) \left[ V^{\pi, \widehat{M}}(S') \right] - \widehat{S}' \sim \widehat{M}(\cdot | S, A) \left[ V^{\pi, \widehat{M}}(\widehat{S}') \right] \right] \end{aligned}$$

Thus:

$$\begin{aligned}
V^{\pi, \widehat{M}} - V^{\pi, M} &= \mathbb{E}_{S_0 \sim \rho_{S_0}} \left[ V^{\pi, \widehat{M}}(S_0) - V^{\pi, M}(S_0) \right] \\
&= \mathbb{E}_{S_0 \sim \rho_{S_0}} \left[ \frac{\gamma}{1 - \gamma} \mathbb{E}_{\substack{S \sim \rho^{\pi, M}(\cdot | S_0 = s) \\ A \sim \pi(\cdot | S)}} \left[ \mathbb{E}_{S' \sim M(\cdot | S, A)} \left[ V^{\pi, \widehat{M}}(S') \right] - \mathbb{E}_{\widehat{S}' \sim \widehat{M}(\cdot | S, A)} \left[ V^{\pi, \widehat{M}}(\widehat{S}') \right] \right] \right] \\
&= \frac{\gamma}{1 - \gamma} \mathbb{E}_{S_0 \sim \rho_{S_0}} \left[ \mathbb{E}_{\substack{S \sim \rho^{\pi, M}(\cdot | S_0 = s) \\ A \sim \pi(\cdot | S)}} \left[ \mathbb{E}_{S' \sim M(\cdot | S, A)} \left[ V^{\pi, \widehat{M}}(S') \right] - \mathbb{E}_{\widehat{S}' \sim \widehat{M}(\cdot | S, A)} \left[ V^{\pi, \widehat{M}}(\widehat{S}') \right] \right] \right] \\
&= \frac{\gamma}{1 - \gamma} \mathbb{E}_{\substack{S \sim \rho^{\pi, M} \\ A \sim \pi(\cdot | S)}} \left[ \mathbb{E}_{S' \sim M(\cdot | S, A)} \left[ V^{\pi, \widehat{M}}(S') \right] - \mathbb{E}_{\widehat{S}' \sim \widehat{M}(\cdot | S, A)} \left[ V^{\pi, \widehat{M}}(\widehat{S}') \right] \right] \\
&= \frac{\gamma}{1 - \gamma} \mathbb{E}_{\substack{S \sim \rho^{\pi, M} \\ A \sim \pi(\cdot | S)}} \left[ G^{\pi, \widehat{M}, M}(S, A) \right]
\end{aligned}$$

## 4.2 Proof for subsection 3.2

From subsection 3.1 :

$$\begin{aligned}
V^{\pi, \widehat{M}} - V^{\pi, M} &= \frac{\gamma}{1 - \gamma} \mathbb{E}_{\substack{S \sim \rho^{\pi, M} \\ A \sim \pi(\cdot | S)}} \left[ \mathbb{E}_{\widehat{s}' \sim \widehat{M}(\cdot | s, a)} \left[ V^{\pi, \widehat{M}}(\widehat{s}') \right] - \mathbb{E}_{s' \sim M(\cdot | s, a)} \left[ V^{\pi, \widehat{M}}(s') \right] \right] \\
&= \frac{\gamma}{1 - \gamma} \mathbb{E}_{\substack{S \sim \rho^{\pi, M} \\ A \sim \pi(\cdot | S)}} \left[ V^{\pi, \widehat{M}}(\widehat{M}(s, a)) - V^{\pi, \widehat{M}}(M(s, a)) \right] \\
\Rightarrow |V^{\pi, \widehat{M}} - V^{\pi, M}| &\leq \frac{\gamma}{1 - \gamma} \mathbb{E}_{\substack{S \sim \rho^{\pi, M} \\ A \sim \pi(\cdot | S)}} \left[ |V^{\pi, \widehat{M}}(\widehat{M}(s, a)) - V^{\pi, \widehat{M}}(M(s, a))| \right] \\
&\leq \frac{\gamma}{1 - \gamma} L \mathbb{E}_{\substack{S \sim \rho^{\pi, M} \\ A \sim \pi(\cdot | S)}} \left[ \left\| \widehat{M}(s, a) - M(s, a) \right\| \right]
\end{aligned}$$

## References

- [1] H. Xu, Y. Li, Y. Tian, T. Darrell, and T. Ma, “Algorithmic framework for model-based reinforcement learning with theoretical guarantees,” *CoRR*, vol. abs/1807.03858, 2018. [Online]. Available: <http://arxiv.org/abs/1807.03858>
- [2] W. Sun, G. J. Gordon, B. Boots, and J. A. Bagnell, “Dual policy iteration,” *CoRR*, vol. abs/1805.10755, 2018. [Online]. Available: <http://arxiv.org/abs/1805.10755>