**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 **??** [**?**].

## 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 **??** [**?**].

# 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? [**?**].

# 3 Formal Results

## 3.1 Formal Result for ??

$M(.|s,a)$ 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, $\rho^{\pi,M} = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t P_{S_t^{\pi,M}}$.

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

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

Let $G^{\pi,\widehat{M},M}(s,a) = \underset{\hat{s}' \sim \widehat{M}(.|s,a)}{E}\left[V^{\pi,\widehat{M}}(\hat{s}')\right] - \underset{s' \sim M(.|s,a)}{E}\left[V^{\pi,\widehat{M}}(s')\right]$

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

$$V^{\pi,\widehat{M}} - V^{\pi,M} = \frac{\gamma}{1-\gamma} \underset{\substack{S \sim \rho^{\pi,M} \\ A \sim \pi(.|S)}}{E}\left[G^{\pi,\widehat{M},M}(S,A)\right]$$

Proof is available in **??**

## 3.2 Formal result for ??

The notation is the same as in **??**. 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:

$$|V^{\pi,\widehat{M}} - V^{\pi,M}| \leq \frac{\gamma}{1-\gamma} L \underset{\substack{S \sim \rho^{\pi,M} \\ A \sim \pi(.|S)}}{E}\left[\left\|\widehat{M}(S,A) - M(S,A)\right\|\right]$$

Proof is available in **??**

## 3.3 Formal result for ??

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 $E_{s\sim d_{\pi_1}}\left[D_{TV}\left(\pi_1(\cdot|s), \pi_2(\cdot|s)\right)\right] \leq \alpha$ 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 [**?**].

# 4 Proofs

## 4.1 Proof for ??

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

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

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 $\widehat{M}$:

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

Proof of the last equality is below:

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

$$=\sum_{a_0}\pi(a_0|s)\sum_{s_1}M(s_1|s,a_0)\ldots\sum_{a_{j-1}}\pi(a_{j-1}|s_{j-1})\sum_{s_j}M(s_j|s_{j-1},a_{j-1})\mathop{E}_{\substack{t=j, A_t\sim\pi(.|s_j)\\ \forall t>j, A_t\sim\pi(.|S_t)\\ S_{t+1}\sim\widehat{M}(.|S_t,A_t)}}\left[\sum_{t=j+1}^{\infty}\gamma^t R(S_t,A_t)\Big|S_j=s_j\right]$$

$$=\sum_{a_0}\sum_{s_1}\ldots\sum_{a_{j-1}}\sum_{s_j}\pi(a_0|s)M(s_1|s,a_0)\ldots\pi(a_{j-1}|s_{j-1})M(s_j|s_{j-1},a_{j-1})\mathop{E}_{\substack{t=j, A_t\sim\pi(.|s_j)\\ \forall t>j, A_t\sim\pi(.|S_t)\\ S_{t+1}\sim\widehat{M}(.|S_t,A_t)}}\left[\sum_{t=j+1}^{\infty}\gamma^t R(S_t,A_t)\Big|S_j=s_j\right]$$

$$=\sum_{a_0,s_1,\ldots,a_{j-1}}\sum_{s_j}P(a_0,s_1,\ldots,a_{j-1},s_j|S_0=s)\mathop{E}_{\substack{t=j, A_t\sim\pi(.|s_j)\\ \forall t>j, A_t\sim\pi(.|S_t)\\ S_{t+1}\sim\widehat{M}(.|S_t,A_t)}}\left[\sum_{t=j+1}^{\infty}\gamma^t R(S_t,A_t)\Big|S_j=s_j\right]$$

$$=\sum_{s_j}\sum_{a_0,s_1,\ldots,a_{j-1}}P(a_0,s_1,\ldots,a_{j-1},s_j|S_0=s)\mathop{E}_{\substack{t=j, A_t\sim\pi(.|s_j)\\ \forall t>j, A_t\sim\pi(.|S_t)\\ S_{t+1}\sim\widehat{M}(.|S_t,A_t)}}\left[\sum_{t=j+1}^{\infty}\gamma^t R(S_t,A_t)\Big|S_j=s_j\right]$$

$$=\sum_{s_j}P(s_j|S_0=s)\mathop{E}_{\substack{t=j, A_t\sim\pi(.|s_j)\\ \forall t>j, A_t\sim\pi(.|S_t)\\ S_{t+1}\sim\widehat{M}(.|S_t,A_t)}}\left[\sum_{t=j+1}^{\infty}\gamma^t R(S_t,A_t)\Big|S_j=s_j\right]$$

$$=\sum_{s_j}\rho_{S_j}^{\pi,M}(s_j|S_0=s)\mathop{E}_{\substack{t=j, A_t\sim\pi(.|s_j)\\ \forall t>j, A_t\sim\pi(.|S_t)\\ S_{t+1}\sim\widehat{M}(.|S_t,A_t)}}\left[\sum_{t=j+1}^{\infty}\gamma^t R(S_t,A_t)\Big|S_j=s_j\right]$$

$$=\mathop{E}_{S_j\sim\rho_{S_j}^{\pi,M}(.|S_0=s)}\left[\mathop{E}_{\substack{t=j, A_t\sim\pi(.|S_j)\\ \forall t>j, A_t\sim\pi(.|S_t)\\ S_{t+1}\sim\widehat{M}(.|S_t,A_t)}}\left[\sum_{t=j+1}^{\infty}\gamma^t R(S_t,A_t)\Big|S_j=s_j\right]\right]$$

$$=\mathop{E}_{\substack{S_j\sim\rho_{S_j}^{\pi,M}(.|S_0=s)\\ A_j\sim\pi(.|S_j)}}\left[\mathop{E}_{\substack{A_t\sim\pi(.|S_t)\\ S_{t+1}\sim\widehat{M}(.|S_t,A_t)}}\left[\sum_{t=j+1}^{\infty}\gamma^t R(S_t,A_t)\Big|S_{j+1}=\widehat{M}(S_j,A_j)\right]\right]$$

$$=\mathop{E}_{\substack{S_j\sim\rho_{S_j}^{\pi,M}(.|S_0=s)\\ A_j\sim\pi(.|S_j)}}\left[\mathop{E}_{\widehat{S}_{j+1}\sim\widehat{M}(.|S_j,A_j)}\left[\mathop{E}_{\substack{A_t\sim\pi(.|S_t)\\ S_{t+1}\sim\widehat{M}(.|S_t,A_t)}}\left[\sum_{t=j+1}^{\infty}\gamma^t R(S_t,A_t)\Big|S_{j+1}=\widehat{S}_{j+1}\right]\right]\right]$$

$$=\mathop{E}_{\substack{S_j\sim\rho_{S_j}^{\pi,M}(.|S_0=s)\\ A_j\sim\pi(.|S_j)}}\left[\mathop{E}_{\widehat{S}_{j+1}\sim\widehat{M}(.|S_j,A_j)}\left[\mathop{E}_{\substack{A_t\sim\pi(.|S_t)\\ S_{t+1}\sim\widehat{M}(.|S_t,A_t)}}\left[\gamma^{j+1}\sum_{t=j+1}^{\infty}\gamma^{t-j-1}R(S_t,A_t)\Big|S_{j+1}=\widehat{S}_{j+1}\right]\right]\right]$$

$$=\mathop{E}_{\substack{S_j\sim\rho_{S_j}^{\pi,M}(.|S_0=s)\\ A_j\sim\pi(.|S_j)}}\left[\mathop{E}_{\widehat{S}_{j+1}\sim\widehat{M}(.|S_j,A_j)}\left[\mathop{E}_{\substack{A_t\sim\pi(.|S_t)\\ S_{t+1}\sim\widehat{M}(.|S_t,A_t)}}\left[\gamma^{j+1}\sum_{t=0}^{\infty}\gamma^{t}R(S_t,A_t)\Big|S_0=\widehat{S}_{j+1}\right]\right]\right]$$

$$=\mathop{E}_{\substack{S_j\sim\rho_{S_j}^{\pi,M}(.|S_0=s)\\ A_j\sim\pi(.|S_j)}}\left[\mathop{E}_{\widehat{S}_{j+1}\sim\widehat{M}(.|S_j,A_j)}\left[\gamma^{j+1}V^{\pi,\widehat{M}}(\widehat{S}_{j+1})\right]\right]$$

By a similar argument:

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

Thus:

$$W_{j+1}(s) - W_j(s) = \gamma^{j+1}\underset{\substack{S_j\sim\rho^{\pi,M}_{S_j}(.|S_0=s)\\ A_j\sim\pi(.|S_j)}}{E}\left[\underset{S_{j+1}\sim M(.|S_j,A_j)}{E}\left[V^{\pi,\widehat{M}}(S_{j+1})\right] - \underset{\widehat{S}_{j+1}\sim\widehat{M}(.|S_j,A_j)}{E}\left[V^{\pi,\widehat{M}}(\widehat{S}_{j+1})\right]\right]$$

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

Thus:

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

## 4.2 Proof for ??

From **??** :

$$
\begin{aligned}
V^{\pi,\widehat{M}} - V^{\pi,M} &= \frac{\gamma}{1-\gamma}\underset{\substack{S\sim\rho^{\pi,M}\\ A\sim\pi(.|S)}}{E}\left[\underset{\hat{s}'\sim\widehat{M}(.|s,a)}{E}\left[V^{\pi,\widehat{M}}(\hat{s}')\right] - \underset{s'\sim M(.|s,a)}{E}\left[V^{\pi,\widehat{M}}(s')\right]\right] \\
&= \frac{\gamma}{1-\gamma}\underset{\substack{S\sim\rho^{\pi,M}\\ A\sim\pi(.|S)}}{E}\left[V^{\pi,\widehat{M}}(\widehat{M}(s,a)) - V^{\pi,\widehat{M}}(M(s,a))\right]
\end{aligned}
$$

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