

A APPENDIX: PROOFS OF LEMMA 1 AND THEOREM 1

Let us consider \mathcal{M} and \mathcal{M}_r as the approximated and real MDP with dynamics model f_ϕ and f respectively. Let the total variation distance between them be bounded by ϵ_f (see (Janner et al., 2019)). This dynamics model predicts both the next state distribution and rewards. The corresponding MPC objective is represented as J and J_r respectively. Here, J denotes that the costs are calculated from the approximated reward function setting whereas J_r is obtained from rollouts in the true MDP. Now, we will derive the bounds on the performance improvement in a similar way as demonstrated in (Janner et al., 2019) and (Morgan et al., 2021), however with consideration and assumptions related to the convexity of the losses.

Proof 1 (Proof of Lemma 1) For any stochastic dynamics model f and reward function r , considering the cost of a trajectory in an MDP with policy π_η and value function V_ζ is given by,

$$C(\mathbf{x}_t, \mathbf{u}_t) = \sum_{h=0}^{H-1} \gamma^h c(x_{t,h}, u_{t,h}) + \gamma^H c_H(x_{t,H}) \quad (16)$$

where, γ is the discount factor, $c(x_{t,h}, u_{t,h}) = -r(x_{t,h}, u_{t,h})$ and c_H is the terminal cost calculated as $-V_\zeta(x_{t,H})$. Let c_{max} be the bound on this cost.

Now, to realize the maximum improvement in the approximated MDP while using the policy parameters $(\tilde{\eta}_t)$, obtained from the shift model, we use a formulation motivated by the bound formulated in Lemma B.3 in (Janner et al., 2019). We consider p_ϕ as the discounted state-action visitation corresponding to f_ϕ (similarly p for f) and superscript h to resemble the notations of (Janner et al., 2019).

$$\begin{aligned} & J(x_t, \tilde{\eta}_t) - J_r(x_t, \tilde{\eta}_t) \\ &= \mathbb{E}_{\mathbf{u}_t \sim \pi_{\tilde{\eta}_t}, \mathbf{x}_t \sim f_\phi} \left[\sum_{h=0}^H \gamma^h c(x_{t,h}, u_{t,h}) + \gamma^H c_H(x_{t,H}) \right] \\ & \quad - \mathbb{E}_{\mathbf{u}_t \sim \pi_{\tilde{\eta}_t}, \mathbf{x}_t \sim f} \left[\sum_{h=0}^H \gamma^h c(x_{t,h}, u_{t,h}) + \gamma^H c_H(x_{t,H}) \right] \\ &= \sum_{\mathbf{x}_t, \mathbf{u}_t} (p_\phi(x, u) - p(x, u)) c(x, u) \\ &\leq \sum_{\mathbf{x}_t, \mathbf{u}_t} \sum_{h=0}^{H-1} \gamma^h (p_\phi^h(x_{t,h}, u_{t,h}) - p^h(x_{t,h}, u_{t,h})) c(x_{t,h}, u_{t,h}) \\ & \quad + \gamma^H (p_\phi^H(x_{t,H}, u_{t,H}) - p^H(x_{t,H}, u_{t,H})) V_\zeta(x_{t,H}) \\ &\leq 2 c_{max} \sum_{h=0}^{H-1} \gamma^h h \epsilon_f + \gamma^H 2 V_{max} H \epsilon_f \\ &= 2 c_{max} \frac{(H-1)\gamma^{H+1} - H\gamma^H + \gamma}{(1-\gamma)^2} \epsilon_f + \gamma^H 2 V_{max} H \epsilon_f \end{aligned}$$

where, $|(p^h(x, u) - p_\phi^h(x, u))| \leq h \epsilon_f$ is inherited from Lemma B.2 in (Janner et al., 2019), the uncertainty in dynamics approximation.

Proof 2 (Proof of Theorem 1) From Lemma-1, we know that,

$$J(\tilde{\eta}_t) \leq J_r(\tilde{\eta}_t) + R_{f,H} \quad (17)$$

and subtracting $J_r(\eta_t^*)$ from both sides of Eq (4) results in

$$J(\tilde{\eta}_t) - J_r(\eta_t^*) \leq J_r(\tilde{\eta}_t) - J_r(\eta_t^*) + R_{f,H} \quad (18)$$

where LHS corresponds to the instantaneous regret incurred by rollouts on approximate MDP (with J) using shifted parameters $(\tilde{\eta}_t)$ and on true MDP (with J_r) using the DMD-optimized parameters (η_t) .

Now, to get the cumulative regret for T decision steps, both sides of Eq (5) should be summed over T and can be shown as,

$$\sum_{t=0}^T (J(\tilde{\eta}_t) - J_r(\eta_t^*)) \leq \sum_{t=0}^T (J_r(\tilde{\eta}_t) - J_r(\eta_t^*)) + \sum_{t=0}^T R_{f,H} \quad (19)$$

$$Re_T(\boldsymbol{\eta}_T) \leq \sum_{t=0}^T (J_r(\tilde{\eta}_t) - J_r(\eta_t^*)) + T R_{f,H} \quad (20)$$

Based on (Hall & Willett, 2013), the DMD update rule directly results in

$$\sum_{t=0}^T (J(\tilde{\eta}_t) - J_r(\eta_t^*)) \leq \frac{D_{\max}}{\alpha_{T+1}} + \frac{4M}{\alpha_T} W_{\Phi_t}(\boldsymbol{\eta}_T) + \frac{G_\ell^2}{2\sigma} \sum_{t=1}^T \alpha_t \quad (21)$$

Substituting Eq (8) in Eq (7), we finally get the bound on the maximum regret as

$$Re_T(\boldsymbol{\eta}_T) \leq \frac{D_{\max}}{\alpha_{T+1}} + \frac{4M}{\alpha_T} W_{\Phi_t}(\boldsymbol{\eta}_T) + \frac{G_\ell^2}{2\sigma} \sum_{t=1}^T \alpha_t + T R_{f,H},$$

which completes the proof.