Dynamic Mirror Descent based Model Predictive Control for Accelerating Robot Learning

Utkarsh A. Mishra[1,*], Soumya R. Samineni[1,*], Prakhar Goel[2], Chandravaran Kunjeti[3], Himanshu Lodha[1], Aman Singh[1], Aditya Sagi[1], Shalabh Bhatnagar[1] and Shishir Kolathaya[1]

APPENDIX

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 [1]). 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 [1] and [2], however with consideration and assumptions related to the convexity of the losses.

Proof: [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,

$$C\left(\mathbf{x_t},\mathbf{u_t}\right)=\sum_{h=0}^{H-1}\gamma^h c(x_{t,h},u_{t,h})+\gamma^H c_H(x_{t,H}) \tag{1}$$

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 [1]. 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 [1].

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

where, $|(p^h(x,u)-p^h_\phi(x,u))|\leq h\epsilon_f$ is inherited from Lemma B.2 in [1], the uncertainty in dynamics approximation. ■

Proof: [Proof of Theorem 1] From Lemma-1, we know that,

$$J\left(\tilde{\eta}_t\right)\leq J_r\left(\tilde{\eta}_t\right)+R_{f,H} \tag{2}$$

and subtracting $J_r\left(\eta^\star_t\right)$ from both sides of Eq (4) results in

$$J\left(\tilde{\eta}_t\right)-J_r\left(\eta^\star_t\right)\leq J_r\left(\tilde{\eta}_t\right)-J_r\left(\eta^\star_t\right)+R_{f,H} \tag{3}$$

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,

$$\sum_{t=0}^{T}\left(J\left(\tilde{\eta}_t\right)-J_r\left(\eta^\star_t\right)\right)\leq\sum_{t=0}^{T}\left(J_r\left(\tilde{\eta}_t\right)-J_r\left(\eta^\star_t\right)\right)+\sum_{t=0}^{T}R_{f,H} \tag{4}$$

$$Re_T\left(\boldsymbol{\eta}_T\right)\leq\sum_{t=0}^{T}\left(J_r\left(\tilde{\eta}_t\right)-J_r\left(\eta^\star_t\right)\right)+T\;R_{f,H} \tag{5}$$

[1] Department of Computer Science and Automation, Indian Institute of Science Bangalore
[2] Electronics and Communication Engineering Department, Manipal Institute of Technology India
[3] Electronics and Communication Engineering Department, National Institute of Technology Karnataka, Surathkal India
[*] These authors have contributed equally.

Based on [3], the DMD update rule directly results in

$$\sum_{t=0}^{T}\left(J\left(\tilde{\eta}_t\right)-J_r\left(\eta_t^{\star}\right)\right)\leq\frac{D_{\max}}{\alpha_{T+1}}+\frac{4M}{\alpha_T}W_{\Phi_t}\left(\boldsymbol{\eta}_T\right)+\frac{G_{\ell}^2}{2\sigma}\sum_{t=1}^{T}\alpha_t \tag{6}$$

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

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

which completes the proof. ■

REFERENCES

[1] M. Janner, J. Fu, M. Zhang, and S. Levine, "When to trust your model: Model-based policy optimization," in Advances in Neural Information Processing Systems (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[2] A. S. Morgan, D. Nandha, G. Chalvatzaki, C. D'Eramo, A. M. Dollar, and J. Peters, "Model predictive actor-critic: Accelerating robot skill acquisition with deep reinforcement learning," arXiv preprint arXiv:2103.13842, 2021.

[3] E. Hall and R. Willett, "Dynamical models and tracking regret in online convex programming," in International Conference on Machine Learning, pp. 579–587, PMLR, 2013.