LiRA: Light-Robust Adversary for Model-based Reinforcement Learning

Taisuke Kobayashi[1]

Abstract— **This study addresses a new adversarial learning framework to make reinforcement learning robust moderately and not conservative too much. To this end, the adversarial learning is first rederived with variational inference. In addition, *light robustness*, which allows for maximizing robustness within an acceptable performance degradation, is utilized as a constraint. As a result, the proposed framework, so-called LiRA, can automatically adjust adversary level, balancing robustness and conservativeness. The behaviors of LiRA are confirmed in numerical simulations.**

I. Introduction

Reinforcement learning (RL) attracts increasing attention as methodology for robots to learn stochastic control policies that enable them to accomplish a given task by trial and error [1]. In particular, model-based RL has high expectations for real-world robot applications due to its excellent sample efficiency [2], [3], and various robot applications have been reported recently. Although the framework proposed in this study can be applied to model-free RL, for the sake of simplicity, this paper limits its focus to model-based RL.

In addition to the sample efficiency (and, of course, control performance), recent RL studies have often aim to maximize robustness to prepare for unexpected events/behaviors in a faced enviornment. A naive approach is domain randomization on simulations [4]–[6], which encompasses a variety of optimal policies by learning from data experienced in simulations driven by various simulation parameters. Although this approach has had much success in recent years, it should be inefficient unless data are collected in parallel using plenty of computational resources.

As a more efficient approach, an adversarial learning can be considered, in which an adversary is introduced that actively interferes with the robot's task accomplishment [7]–[9]. Such active disturbances can produce complex and challenging experiences more efficiently than random ones, but this approach inevitably inherits the mode collapse issue in adversarial learning [10], [11], making learning the optimal policy unstable. In addition, in real-world robots, excessive disturbances might damage themselves and/or objects in the faced environment.

Furthermore, one of the adverse effects of maximizing robustness (caused in both of the above) is to make the policy too conservative [12]–[14]. In other words, since robots have to take into account contingencies that may not occur in practice, they tend to choose safer and more secure actions. If the maximum disturbance intensity could appropriately be designed in advance, such conservativeness would be controlled at the minimum. However, as the relationship between the policy performance and the disturbance intensity is nonlinear and varies from situation to situation, this solution is infeasible.

In this context, this study proposes a new adversarial learning framework, so-called LiRA (see Fig. 1). LiRA aims to improve the robustness moderately while mitigating learning collapse and policy conservativeness. To this end, adversarial learning is first rederived according to variational inference [15]. Then, as a new definition of robustness, *light robustness* [16] is integrated with it. This relaxes full robustness by imposing an inequality constraint that limits the degradation from ideal performance due to disturbances within a specified threshold (this corresponds to the conservativeness). Through the Lagrangian method [17], this inequality constraint is transformed into a loss function for numerical optimization, allowing for automatic tuning of the adversary level.

The behaviors of the proposed LiRA are confirmed by numerical simulations using Mujoco. LiRA is not too conservative in control performance only with the nominal noise, and the degradation of control performance in response to disturbance intensity is suppressed. In addition, according to the specified conservativeness, the proposed LiRA attempts to increase the robustness by increasing the adversary level for the condition with weaker disturbance sensitivity, and vice versa. Such an auto-tuning capability also yields self-paced curriculum learning [18], [19], where disturbances are suppressed during the under-performance phase and reinforced as learning progresses. These results indicate that LiRA can achieve a good balance between robustness and conservativeness.

II. Preliminaries

A. Model-based reinforcement learning

In RL [1], an agent aims to gain the maximum rewards from an environment in the future. For mathematical formulation of the relationship between the agent and environment, Markov decision process (MDP) is basically assumed, i.e. $(\mathcal{S}, \mathcal{A}, p_e, r)$. Here, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $p_e : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is the state transition probability (a.k.a. dynamics), and $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the reward function to evaluate each transition.

*This work was supported by JST, PRESTO Grant Number JPMJPR20C3, Japan.

[1] T. Kobayashi is with the National Institute of Informatics (NII) and with The Graduate University for Advanced Studies (SOKENDAI), 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan kobayashi@nii.ac.jp

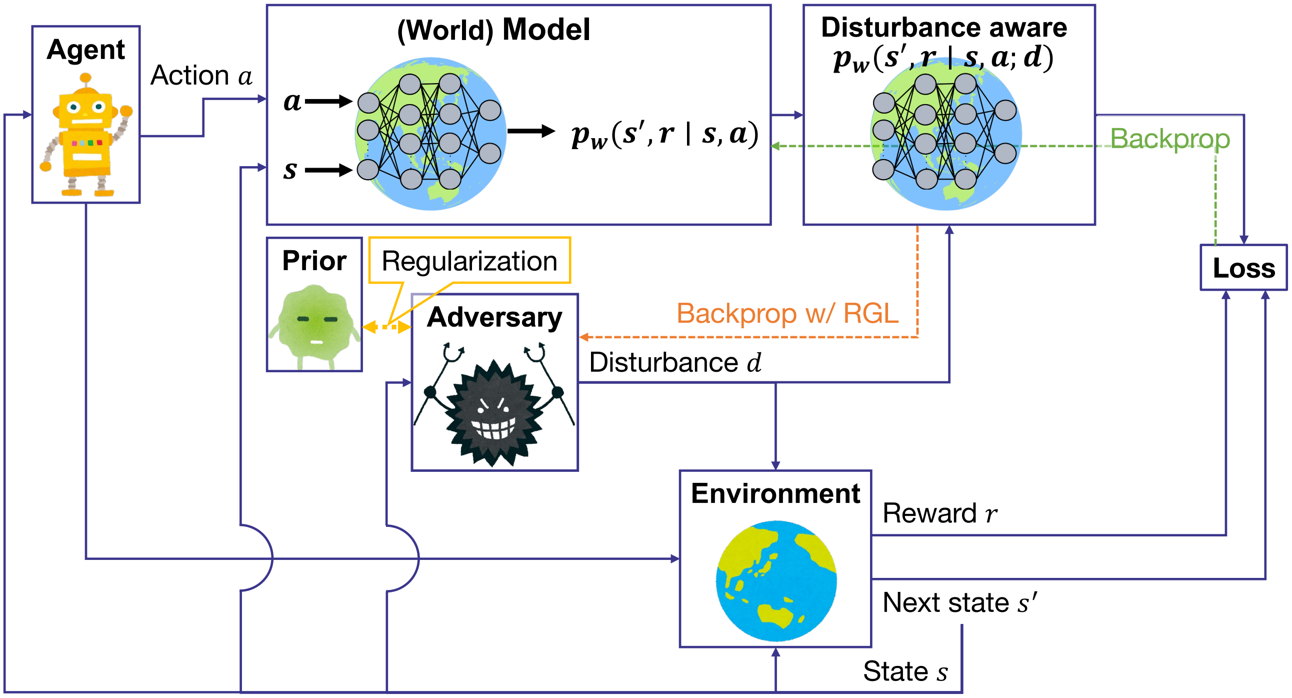


Fig. 1: Proposed framework: LiRA

Under MDP, this study solves the following optimization problem for acquiring the optimal (stochastic) policy $\pi^* : \mathcal{S} \to \mathcal{A}$, which should be able to accomplish the task defined by r, at the discrete time step $t \in \mathbb{N}$.

$$\pi^*(a_t \mid s_t) = \arg \max_{\pi(a_t|s_t)} \sum_{k=0}^{H} r_{t+k} \tag{1}$$
$$\text{s.t.} \begin{cases} r_{t+k} & = r(s_{t+k}, a_{t+k}, s'_{t+k}) \\ s'_{t+k} & \sim p_e(s_{t+k+1} \mid s_{t+k}, a_{t+k}) \\ a_{t+k} & \sim \pi(a_{t+k} \mid s_{t+k}) \end{cases}$$

where $H \in \mathbb{N}$ denotes the horizon indicating how far into the future to be considered.

If the agent knows p_e and r accurately, this problem can numerically be solved using MPC (in this paper, AccelMPPI [20] is employed). Therefore, model-based RL algorithms explicitly approximate them as a model (p_w in this paper) using, for example, deep neural networks with parameters θ [2], [21].

$$\theta^* = \arg \min_{\theta} \mathbb{E}_{p_e, r, \pi}[-\ln p_w(s', r \mid s, a; \theta)] \tag{2}$$
$$\text{s.t.} \begin{cases} r & = r(s, a, s') \\ s' & \sim p_e(s' \mid s, a) \\ a & \sim \pi(a \mid s) \end{cases}$$

Note that, in this paper, r is alternatively represented as a probability conditioned on s and a except s' for simplicity. The obtained p_w replaces p_r and r in eq. (1).

B. Adversarial learning

To make π^* obtained through the above optimization problem robust, worst-case scenarios for the objective function (i.e. the sum of predicted reward, called the return) are basically considered [22], [23]. However, this incurs an extra computational cost at inference time compared to the case simply using expected value of the prediction. In this study, therefore, robust control is achieved by making p_w robust during learning, referring to the result reported in the literature [24] that control performance becomes implicitly robust if p_w is optimized in consideration of events that are rare in reality.

To make p_w fully robust, it is effective to intentionally allow the agent to experience the rare events. That is, this purpose can be achieved by adversarial learning with the following min-max problem, instead of eq. (2):

$$\theta^*, \phi^* = \arg \min_{\theta} \max_{\phi} \mathbb{E}_{\tilde{p}_e, \tilde{r}, \pi, \varpi}[-\ln p_w(s', r \mid s, a; \theta)] \tag{3}$$
$$\text{s.t.} \begin{cases} r & = \tilde{r}(s, a, s'; d) \\ s' & \sim \tilde{p}_e(s' \mid s, a; d) \\ a & \sim \pi(a \mid s) \\ d & \sim \varpi(d \mid s; \phi) \end{cases}$$

where $\varpi : \mathcal{S} \to \mathcal{D}$ denotes the learnable adversary (or, disturbance generator) with parameters ϕ. Note that the disturbance d should have upper and lower bounds and be distributed around zero to avoid the collapse and bias of the original environment. Therefore, is is assumed that $d \in [-d^{\max}, d^{\max}]^{|\mathcal{D}|}$ with $d^{\max}$ the maximum disturbance intensity, which is specified in advance as a hyperparameter.

As a remark, ϖ can be conditioned on a in addition to s, but in that case, the effects of d are too strong because d

can easily interfere with a by determining later than a. As this study is in favor of moderate robustness, a was omitted from the condition for ϖ. In addition, the way d acts on the environment is a discussion for robotic applications, so it was also omitted in this paper in favor of theory.

III. LiRA: Light-robust adversary

A. Adversarial learning with variational inference

First, adversarial learning defined in eq. (3) is redefined based on variational inference as a basis for incorporating the light robustness introduced in the next section. Specifically, we can focus on the fact that the disturbance d is usually unobservable and regarded to be the latent variable. That is, by introducing the disturbance-aware model and the prior, the following evidence lower bound is derived according to Jensen's inequality.

$$\begin{aligned}&\ln p_w(s', r \mid s, a; \theta)\\ =&\ln \mathbb{E}_{\varpi(d)}[p_w(s', r \mid s, a; d, \theta)]\\ \geq&\mathbb{E}_{\varpi(d \mid s)}[\ln p_w(s', r \mid s, a; d, \theta)] - \mathrm{KL}(\varpi(d \mid s; \phi) || \varpi(d))\\ \geq& \inf_{\varpi(d \mid s; \phi)} [\ln p_w(s', r \mid s, a; d, \theta)] - \mathrm{KL}(\varpi(d \mid s; \phi) || \varpi(d))\end{aligned} \tag{4}$$

where $\mathrm{KL}(\cdot || \cdot)$ denotes Kullback-Leibler divergence between two probabilities. When maximizing this lower bound, the adversary tries to minimize $\ln p_w(s', r \mid s, a; d, \theta)$ to take into account the rare events, while regularizing $\varpi(d\ mids; \phi) \to \varpi(d)$. That is, the min-max problem in eq. (3) seems to be extended by adding the regularization.

The degree of regularization is commonly adjusted by introducing the gain $\beta \geq 0$ [25] ($\beta = 0$ means no regularization as like eq. (3)). Although this regularization is expected to reduce the generation of excessive disturbances, how it (β, more specifically) affects to the balance between robustness and conservativeness is unclear. For developing an auto-tuning mechanism of β (or other alternative gain) to achieve the desired balance, some kind of criteria that can be intuitively specified by the user are needed.

In addition, we need to focus on the fact that the newly introduced disturbance-aware model, $p_w(s', r \mid s, a; d, \theta)$, cannot be used when d is unknown, so it is used only for adversarial learning. Only if an additional disturbance estimator could be introduced, it could be used even during inference to enable the domain adaptation. In that light, as the disturbance-marginalized model, $p_w(s', r \mid s, a; \theta)$, is not included in the above lower bound, it should be additionally optimized by explicitly considering some kind of conditions.

B. Integration with light robustness

To address these remaining issues, the *light robustness* [16] is integrated with the maximization problem of the above lower bound. The light robustness establishes a tolerance for performance degradation due to disturbances and assigns that constraint to the optimization problem. This allows for a more intuitive setting since the "relative" tolerance can be defined, in comparison to, for example, the degree of regularization to the prior and the absolute predictive performance of the model.

Specifically, for any state (and action), the following constraint is applied.

$$\begin{aligned}&- \ln p_w(s', r \mid s, a; \theta) \leq - \ln p_w(s', r \mid s, a; d, \theta) + \rho\\ &\underbrace{\ln p_w(s', r \mid s, a; d, \theta) - \ln p_w(s', r \mid s, a; \theta) - \rho + \Delta(s; \eta)}_{\delta(s', r, s, a, d)}\\ &= 0\end{aligned} \tag{5}$$

where, $\rho \geq 0$ denotes the tolerance of performance degradation and $\Delta \geq 0$ denotes the slack variable, which represents the different between the left and right sides and is approximated by parameters η. The first line is the inequality according to the original light robustness, and the second line is a clarification for this study. Note that although it may be temporarily violated depending on the initialization of the models, $\ln p_w(s', r \mid s, a; d, \theta) \leq \ln p_w(s', r \mid s, a; \theta)$ holds in general because the likelihood is higher when conditioned with more necessary information.

This constraint can be converted into the corresponding regularization term via Lagrangian with an auto-tuned gain λ, as shown in the literature [17]. In summary, the proposed LiRA solves the following optimization problem to suppress conservativeness while making the model moderately robust.

$$\begin{aligned}\theta^*, \phi^* = \arg \min_\theta \max_\phi \mathbb{E}_{\tilde{p}_e, \tilde{r}, \pi, \varpi}[&-\lambda \ln p_w(s', r \mid s, a; \theta)\\ &- (1 - \lambda) \ln p_w(s', r \mid s, a; d, \theta)\\ &- \beta \mathrm{KL}(\varpi(d \mid s; \phi) || \varpi(d))]\end{aligned} \tag{6}$$

$$\text{s.t.} \begin{cases} r &= \tilde{r}(s, a, s'; d)\\ s' &\sim \tilde{p}_e(s' \mid s, a; d)\\ a &\sim \pi(a \mid s)\\ d &\sim \varpi(d \mid s; \phi)\end{cases}$$

where, λ (and η for approximating Δ) can be optimized according to the literature [17]. Note that λ is a Lagrange multiplier, so $\lambda \in \mathbb{R}$ holds, but if $\lambda < 0$ and $\lambda > 1$, θ will be learned in a direction that degrates the predictive performance of the models. Since this is not in line with the original purpose of model learning, λ is restricted within $[0, 1]$ in this study.

If the disturbance is too strong, $\delta > 0$ is likely to occur, leading to $\lambda \to 1$. As a result, the adversarial learning for the disturbance-aware model is suppressed, and the adversary is dominantly regularized to its prior, weakening the disturbance. On the other hand, if the disturbance is too weak, $\delta < 0$ and $\lambda \to 0$ are expected, and the adversarial learning is activated to strengthen the disturbance. In this way, LiRA automatically adjusts the disturbance intensity to be moderately robust and not exceed the specified tolerance (a.k.a. conservativeness).

IV. Numerical verification

A. Task

Numerical simulations are conducted to verify the behaviors of the robot when learned with the proposed LiRA.

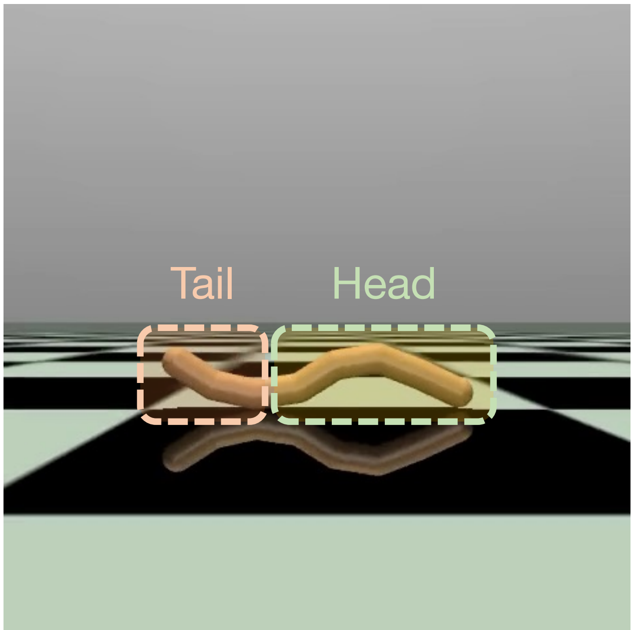


Fig. 2: Worm-type robot

Mujoco is used as the simulator, and a task is to maximize the forward speed of a worm-type robot (see Fig. 2). This robot is 1 m long with 6–8 joints at equal intervals: 4 controllable joints on the head side and 2–4 joints for disturbance on the tail side.

The fewer the number of joints for disturbance, the smaller the effect of the disturbance and thus the less the need to limit the disturbance intensity; in other words, the more the number of joints for disturbance, the more the task performance degradation and the more the need to limit the disturbance intensity. That is, if the tolerance of the performance degradation is fixed for all the conditions, the disturbance intensities should converge to different values.

B. Results

First, the following three conditions are compared in terms of the control performance after learning the above task under three types of disturbance.

- *Nominal* ($\lambda = 1$, $\beta = \infty$):
  The model is learned only with the prior.
- *Full* ($\lambda = 0$, $\beta = 0$):
  The model is learned in a fully adversarial manner.
- *Proposal* (λ is optimized, $\beta = 10^{-3}$):
  The model is learned with LiRA.

Note that the prior in this task is given as Gaussian (a.k.a. nominal noise). During learning/inference, the robot is controlled with AccelMPPI [20], but exploration noise is added to action only during learning. The three types of disturbance are as follows: the first is the nominal noise and is the weakest; the second is a composite of several Brownian noises; and the third is scaled more than the second.

TABLE I: Statistics of λ auto-tuned by LiRA

Tail2	Tail3	Tail4
0.474±0.279	0.630±0.282	0.694±0.274

The test results are shown in Fig. 3. It can be found that *Nominal* without adversarial learning achieved high performance with small disturbance, but it was vulnerable to large ones. *Full*, which always learns adversarially to the maximum extent possible, maintained its performance independent of the disturbance intensity, but its basic performance was low and conservative. Compared to *Nominal*, *Proposal* (a.k.a. LiRA) was able to suppress the performance degradation caused by the disturbance intensity, was less conservative than *Full*, and succeeded in achieving a well-balanced and stable learning.

The auto-tuning process of λ by LiRA is depicted in Fig. 4 with its statistics summarized in Table I. Note again that the smaller λ is, the stronger the disturbance intensity. The fewer the number of disturbance joints and the smaller the original disturbance influence, the smaller λ waws, giving priority to adversarial learning. Conversely, as the number of joints for external disturbances increased, λ became larger, suggesting that the disturbance intensity was suppressed to prevent the performance degradation.

In addition, λ was once large from the beginning to the middle of learning progress, and after a while, it became smaller and converged to each value suitable for the task. This result suggests the emergence of some kind of self-paced curriculum that once simplifies the task to an easily predictable situation and then makes the task more difficult by gradually increasing the disturbance. Indeed, it can be said that the derivation of LiRA is partially similar to the self-paced learning methods [18], [19], and therefore, LiRA might lead to such an additional value.

V. Conclusion

This study proposed a new adversarial learning framework, so-called LiRA. LiRA enabled RL agents to improve the robustness of policy moderately while mitigating learning collapse and policy conservativeness. To this end, adversarial learning was reformulateed using the variational inference and the light robustness. As a result, LiRA achieved a good balance between robustness and conservativeness in comparison to the cases with/without the previous adversary, which fully prevents model learning. In the near future, LiRA will be utilized to learn moderately robust robot policies in the real world.

References

[1] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[2] K. Chua, R. Calandra, R. McAllister, and S. Levine, "Deep reinforcement learning in a handful of trials using probabilistic dynamics models," in *Advances in Neural Information Processing Systems*, 2018, pp. 4754–4765.

[3] G. Williams, P. Drews, B. Goldfain, J. M. Rehg, and E. A. Theodorou, "Information-theoretic model predictive control: Theory and applications to autonomous driving," *IEEE Transactions on Robotics*, vol. 34, no. 6, pp. 1603–1622, 2018.

[4] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2017, pp. 23–30.

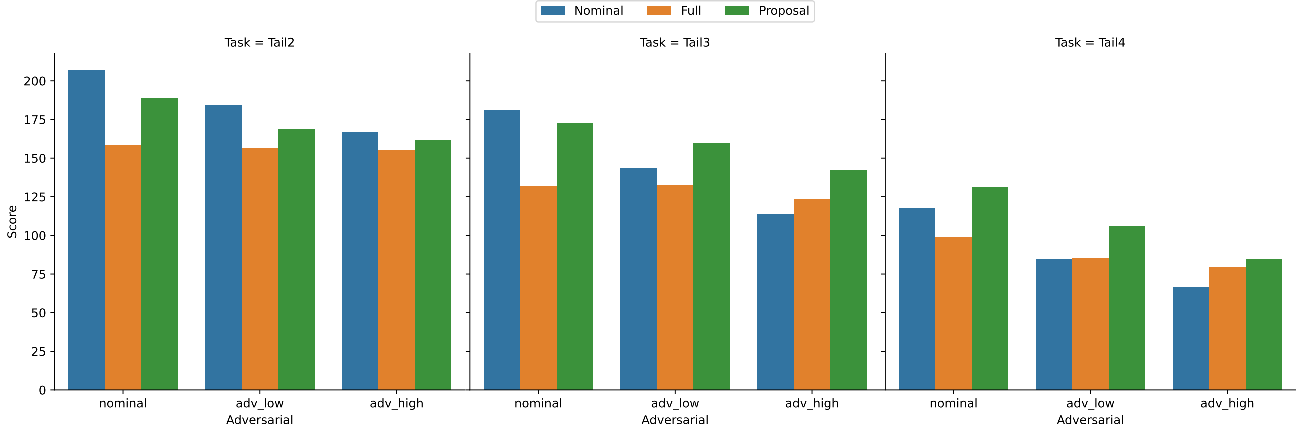
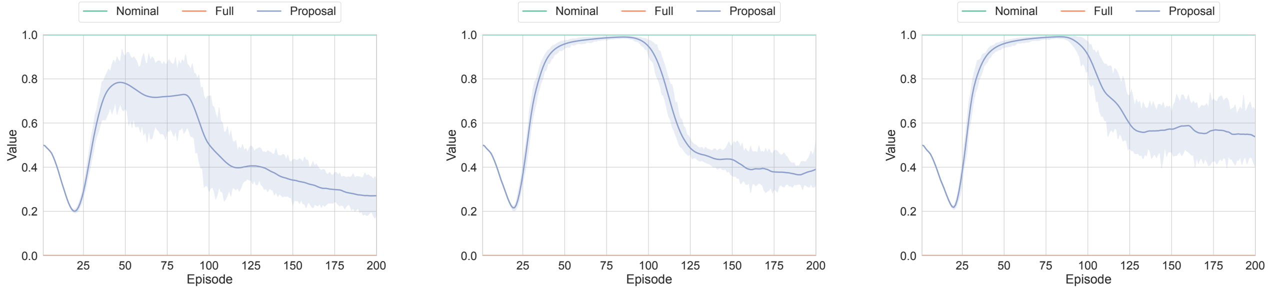


Fig. 3: Control performance on three types of disturbance



(a) Tail2

(b) Tail3

(c) Tail4

Fig. 4: Learning curves of λ auto-tuned by LiRA

[5] F. Ramos, R. C. Possas, and D. Fox, "Bayessim: adaptive domain randomization via probabilistic inference for robotics simulators," in *Robotics: Science and Systems*, 2019.

[6] F. Muratore, T. Gruner, F. Wiese, B. Belousov, M. Gienger, and J. Peters, "Neural posterior domain randomization," in *Conference on Robot Learning*. PMLR, 2022, pp. 1532–1542.

[7] L. Pinto, J. Davidson, R. Sukthankar, and A. Gupta, "Robust adversarial reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2817–2826.

[8] A. Gleave, M. Dennis, C. Wild, N. Kant, S. Levine, and S. Russell, "Adversarial policies: Attacking deep reinforcement learning," in *International Conference on Learning Representations*, 2020.

[9] P. Zhai, J. Luo, Z. Dong, L. Zhang, S. Wang, and D. Yang, "Robust adversarial reinforcement learning with dissipation inequation constraint," in *AAAI Conference on Artificial Intelligence*, vol. 36, no. 5, 2022, pp. 5431–5439.

[10] A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton, "Veegan: Reducing mode collapse in gans using implicit variational learning," *Advances in neural information processing systems*, vol. 30, 2017.

[11] K. Liu, W. Tang, F. Zhou, and G. Qiu, "Spectral regularization for combating mode collapse in gans," in *IEEE/CVF international conference on computer vision*, 2019, pp. 6382–6390.

[12] M. Petrik and R. H. Russel, "Beyond confidence regions: Tight bayesian ambiguity sets for robust mdps," *Advances in neural information processing systems*, vol. 32, 2019.

[13] M. Lechner, A. Amini, D. Rus, and T. A. Henzinger, "Revisiting the adversarial robustness-accuracy tradeoff in robot learning," *IEEE Robotics and Automation Letters*, vol. 8, no. 3, pp. 1595–1602, 2023.

[14] J. Huang, H. J. Choi, and N. Figueroa, "Trade-off between robustness and rewards adversarial training for deep reinforcement learning under large perturbations," *IEEE Robotics and Automation Letters*, vol. 8, no. 12, pp. 8018–8025, 2023.

[15] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations*, 2014.

[16] M. Fischetti and M. Monaci, "Light robustness," *Robust and online large-scale optimization: Models and techniques for transportation systems*, pp. 61–84, 2009.

[17] T. Kobayashi, "Soft actor-critic algorithm with truly-satisfied inequality constraint," *arXiv preprint arXiv:2303.04356*, 2023.

[18] M. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," *Advances in neural information processing systems*, vol. 23, 2010.

[19] P. Klink, C. D'Eramo, J. R. Peters, and J. Pajarinen, "Self-paced deep reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9216–9227, 2020.

[20] T. Kobayashi and K. Fukumoto, "Real-time sampling-based model predictive control based on reverse kullback-leibler divergence and its adaptive acceleration," *arXiv preprint arXiv:2212.04298*, 2022.

[21] T. Kobayashi and R. Watanuki, "Sparse representation learning with modified q-vae towards minimal realization of world model," *Advanced Robotics*, vol. 37, no. 13, pp. 807–827, 2023.

[22] J. Köhler, R. Soloperto, M. A. Müller, and F. Allgöwer, "A computationally efficient robust model predictive control framework for uncertain nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 66, no. 2, pp. 794–801, 2020.

[23] M. Zanon and S. Gros, "Safe reinforcement learning using robust mpc," *IEEE Transactions on Automatic Control*, vol. 66, no. 8, pp. 3638–3652, 2020.

[24] T. Aotani and T. Kobayashi, "Cooperative transport by manipulators with uncertainty-aware model-based reinforcement learning," in *IEEE/SICE International Symposium on System Integration*. IEEE, 2024, pp. 959–964.

[25] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework." in *International Conference on Learning Representations*, 2017.