Control Regularization for Reduced Variance Reinforcement Learning

Richard Cheng[1], Abhinav Verma[2], Gabor Orosz[3], Swarat Chaudhuri[2], Yisong Yue[1], Joel W. Burdick[1]

[1]California Institute of Technology, [2]Rice University, [3]University of Michigan, Ann Arbor

Correspondence: rcheng@caltech.edu
Code at: *https://github.com/rcheng805/CORE-RL*

Overview and Motivation

- Reinforcement learning focuses on finding an agent's policy that maximizes long-term reward through trial and error
  - This trial-and-error approach has been successful for learning complex control tasks, but it is sample inefficient, unsafe, and has high variance.
- To be useful, reinforcement learning must *reliably* find good solutions with reasonable sample efficiency

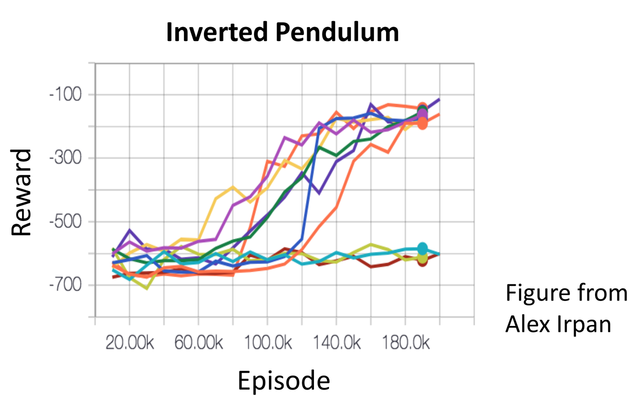
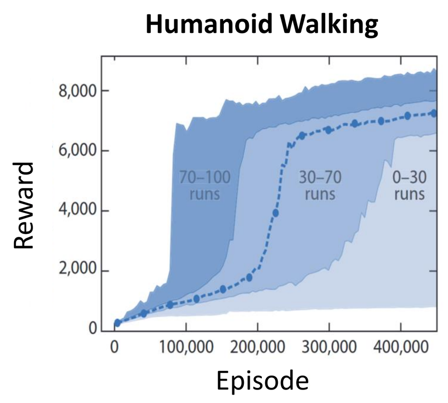


Figure from Alex Irpan



(Recht, 2018)

- This work introduces a regularization method that uses a control prior to significantly reduce variance in learning, improve sample efficiency, and improve safety

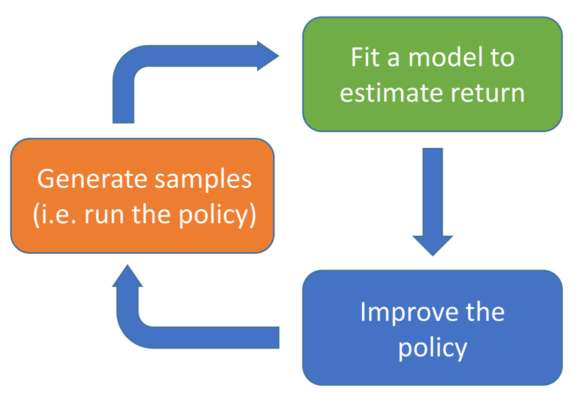
Background and Problem Formulation

Find Policy to Maximize Reward

$$\pi(a|s): S \times A \to [0,1]$$

$$\pi^* = \max_\pi J(\pi) = \max_\pi \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right]$$

$$\tau: (s_t, a_t, \ldots, s_{t+N}, a_{t+N})$$



Adapted from Sergey Levine

- Learn through sampled trajectories (no model required)

Policy Gradient: $\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[\nabla_\theta \log \pi_\theta(\tau) Q^\pi(\tau)]$

$$\approx \sum_{i=1}^{N}\sum_{t=1}^{T}[\nabla_\theta \log \pi_\theta(s_{i,t}, a_{i,t}) Q^\pi(s_{i,t}, a_{i,t})]$$

$$\theta_{k+1} = \theta_k + \alpha \nabla_\theta J(\theta_k)$$

What if we have a system model?

How to intelligently utilize prior knowledge?

- Model-free RL methods suffer from **high variance** in learning and sample inefficiency (Islam et al. 2017; Henderson et al. 2018; Recht 2019)

Control Regularization

Assume with have a *crude* control prior, $u_{prior}(s)$, synthesized from some system model:

$$s_{t+1} = f_{known}(s_t, a_t) + f_{unknown}(s_t, a_t)$$

Let us incorporate this control prior by blending it with the learned controller, $u_{\theta_k}(s)$:

$$u_k(s) = \frac{1}{1+\lambda} u_{\theta_k}(s) + \frac{\lambda}{1+\lambda} u_{prior}(s) \qquad (6)$$

where λ is a regularization parameter that weights the control prior against the RL control.

Lemma 1. *The policy $u_k(s)$ in Equation (6) is the solution to the following regularized optimization problem:*

$$\bar{u}_k(s) = \underset{u}{\operatorname{argmin}} \|u(s) - \bar{u}_{\theta_k}(s)\|_\Sigma + \lambda \|u(s) - u_{prior}(s)\|_\Sigma, \quad \forall s \in S \qquad (7)$$

which can be equivalently expressed as the constrained optimization problem,

$$\bar{u}_k(s) = \underset{u}{\operatorname{argmin}} \|u(s) - \bar{u}_{\theta_k}(s)\|_\Sigma \quad \text{s.t.} \quad \|u(s) - u_{prior}(s)\|_\Sigma \le \bar{\mu}(\lambda) \quad \forall s \in S \qquad (8)$$

where $\bar{\mu}$ constrains the policy search.

Bias-Variance Tradeoff and State-Space Interpretation

Control regularization reduces the variance arising from the policy gradient by a factor $\frac{1}{(1+\lambda)^2}$

High variance in policy gradients translates into high variance in policy learning

$$\pi_{\theta_{k+1}} = \pi_{\theta_k} + \alpha \frac{d\pi_{\theta_k}}{d\theta} \nabla_\theta J(\theta_k) + O(\Delta\theta^2)$$

$$\operatorname{var}_\theta[\pi_{\theta_{k+1}}] \approx \alpha^2 \frac{d\pi_{\theta_k}}{d\theta} \nabla_\theta J(\theta_k) \frac{d\pi_{\theta_k}^T}{d\theta}$$

for $\alpha \ll 1$

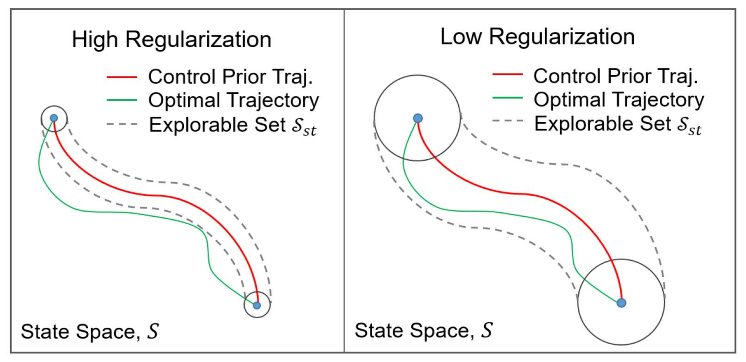
Control regularization may bias the learned policy, if regularization is high and the control prior is poor

$$D_{TV}(\pi_k, \pi_{opt}) \ge D_{sub} - \frac{1}{1+\lambda} D_{TV}(\pi_{\theta_k}, \pi_{prior})$$

$$D_{TV}(\pi_k, \pi_{opt}) \le \frac{\lambda}{1+\lambda} D_{sub} \quad \text{as} \quad k \to \infty$$

where $D_{TV}(\cdot,\cdot)$ represents the total variation distance between two policies, and $D_{sub} = D_{TV}(\pi_{opt}, \pi_{prior})$

State-Space Interpretation:



- The explorable region of state space is denoted by the set $\mathcal{S}_{st}$, which grows as λ decreases. Thus, higher regularization more heavily constrains exploration
- The difference between the control prior trajectory optimal trajectory (i.e. D_{sub}) *may* bias the final policy depending on the explorable region.

Control Prior Synthesis and Stability Properties

From a stability point of view, the control prior should maximize robustness to disturbances and model uncertainty. We treat the RL control, u_{θ_k}, as a performance maximizing "disturbance" to the control prior, u_{prior}.

- The regularized policy takes advantage of stability properties of the robust control prior, and the performance optimization properties of the RL controller.

Suppose we have system dynamics described by: $\dot{s} = f_c(s, a)$,

which is linearized with *bounded disturbance*, $d(s,a)$: $\dot{s} = As + B_2 a + d(s,a)$

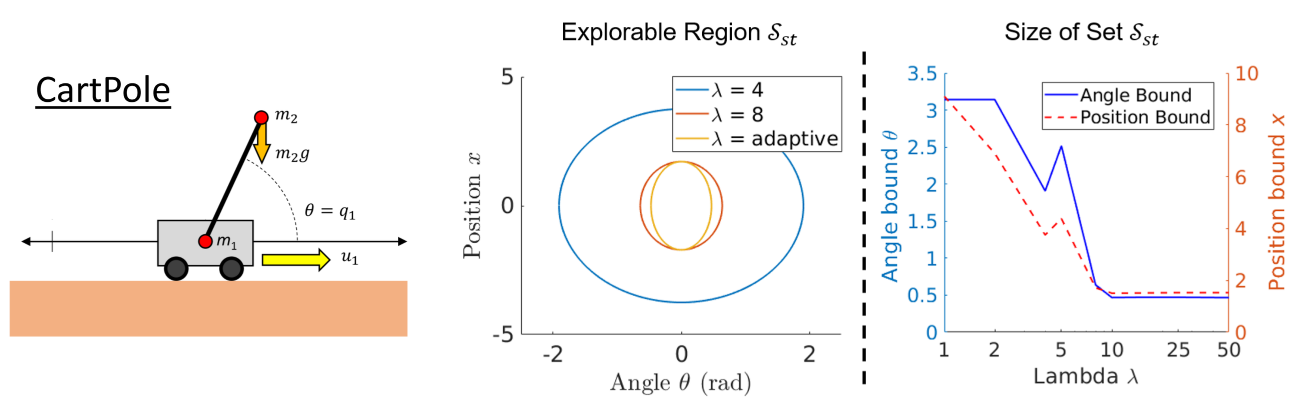
Theorem 2. Assume *a stabilizing $\mathcal{H}_\infty$ control prior within the set $\mathcal{C}$ for the dynamical system (14). Then asymptotic stability and forward invariance of the set $\mathcal{S}_{st} \subseteq \mathcal{C}$*

$$\mathcal{S}_{st}: \left\{ s \in \mathbb{R}^n : \|s\|_2 \le \frac{1}{\sigma_m(\zeta_k)}\left(2\|P\|_2 C_D + \frac{2}{1+\lambda}\|PB_2\|_2 C_\pi\right), s \in \mathcal{C} \right\}$$

is guaranteed under the regularized policy (5) for all $s \in \mathcal{C}$.

The set $\mathcal{S}_{st}$ contracts as we:

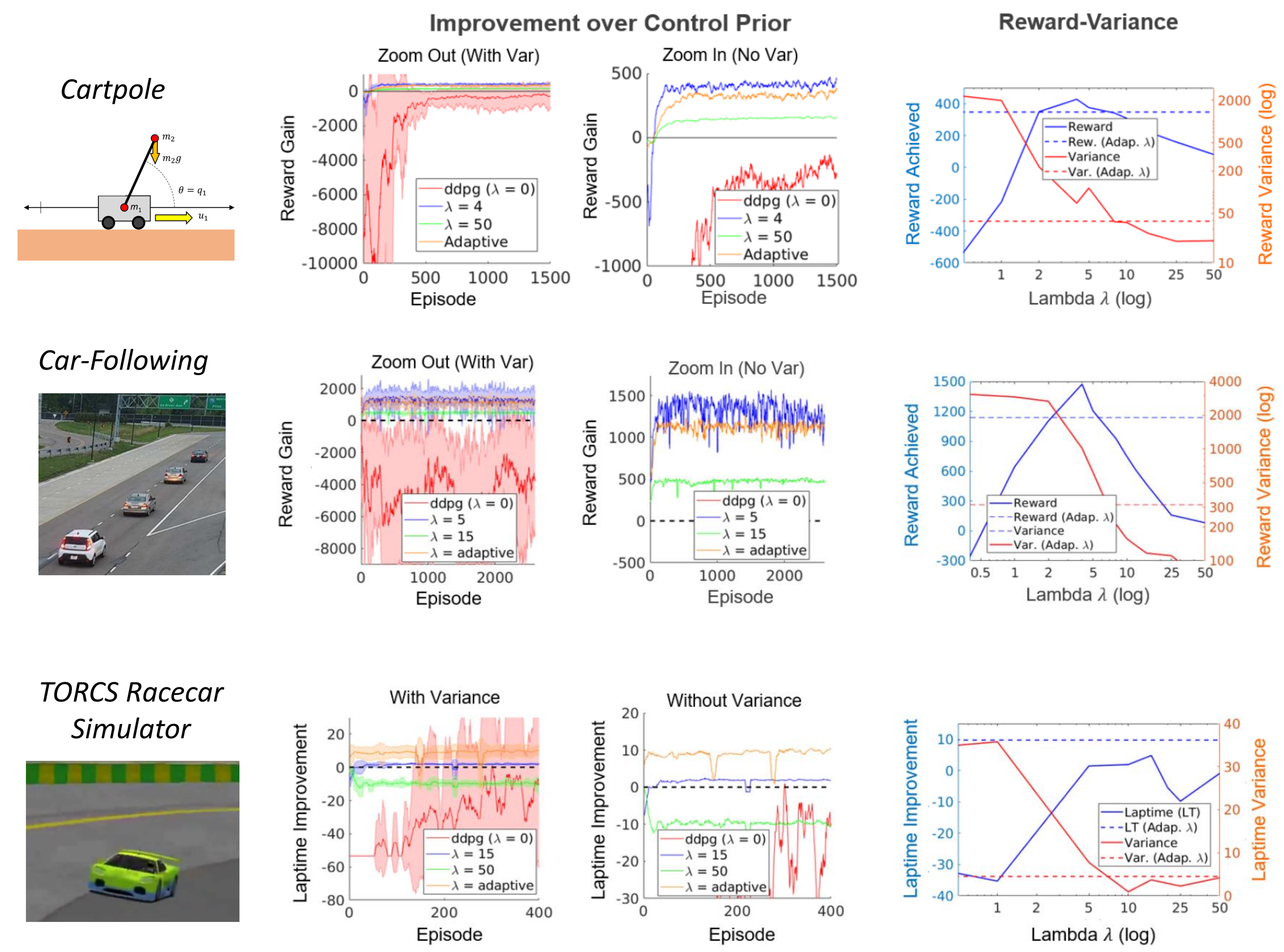
- Increase robustness of the control prior (increase $\sigma_m(\zeta_k)$)
- Decrease our dynamic uncertainty/nonlinearity C_D
- Increase weighting λ on the control prior



The point is ***not*** to say that $\mathcal{H}_\infty$ control provides the best control prior, but rather to show that regularization allows us to "capture" stability properties from a robust control prior.

Empirical Results

We validated our ideas on three problems, using two baseline RL algorithms – Deep Deterministic Policy Gradients (DDPG) and Proximal Policy Optimization (PPO). DDPG results are shown below (see paper for PPO results):



<u>**PROS:** With intermediate regularization, we observe</u>

- Significant improvement in reward (better than both the control prior and baseline algorithm)
- Faster learning (in car-following setting with fixed dataset size, regularization required to learn)
- Substantially reduced variance in learning
- Safety of the learned controller

<u>**CONS:** High regularization leads to</u>

- Significant bias of the reward towards the control prior
- Potentially lower reward in some runs (though unregularized learning is much more unreliable).

Adaptive Regularization Weighting

The regularization weight, λ, should be strong when the learned controller is highly uncertain, and should decrease as we become more confident in the learned controller

- A proxy for confidence in the learned control is error in the value function (i.e. TD-error).

$$|\delta^\pi(s_t)| = |r_{t+1} + \gamma Q^\pi(s_{t+1}, a_{t+1}) - Q^\pi(s_t, a_t)|,$$

This TD-error approximates how poorly the RL algorithm predicts the value of a given state. If it is high, we rely heavily on the control prior. We map this error to regularization weight:

$$\lambda(s_t) = \lambda_{\max}\left(1 - e^{-C|\delta(s_{t-1})|}\right)$$

Lower λ result when the value function predictions are accurate

Conclusion

Control regularization greatly reduces variance in learning, and can significantly improve performance and learning efficiency of RL

- It allows us to capture safety/stability properties from a robust control prior

Important issues that remain to be tackled are:

- Incorporating a changing control prior into the RL framework,
- Analyzing how poor of a control prior can be used while still benefiting learning,
- Improving the adaptive regularization strategy.

Acknowledgments

This work was funded in part by Raytheon under the Learning-to-Fly program and by DARPA under the Physics-Infused AI program

References (Partial List)

1. T.P. Lillicrap, J.J. Hunt, A Pritzel, N Heess, T Erez, Y Tassa, D Silver, D Wierstra. *Continuous Control with Deep Reinforcement Learning* (2016).
2. E Greensmith, P Bartlett, J Baxter. *Variance Reduction Techniques for Gradient Estimates in Reinforcement Learning* (2004).
3. R Sutton, D McAllester, S.P. Singh, Y Mansour. *Policy Gradient Methods for Reinforcement Learning with Function Approximations* (1999).
4. *B Recht. A Tour of Reinforcement Learning: The View from Continuous Control* (2019).
5. R Islam, P Henderson, M Gomrokchi, D Precup. *Reproducibility of Benchmarked Deep Reinforcement Learning of Tasks for Continuous Control* (2017).
6. T Johannink, S Bahl, A Nair, J Luo, A Kumar, M Loskyll, J.O. Aparicio, E Solowjow, S Levine. *Residual Reinforcement Learning for Robot Control* (2018).