Control Regularization for Reduced Variance Reinforcement Learning

Richard Cheng, Abhinav Verma, Gabor Orosz, Swarat Chaudhuri
Yisong Yue, Joel W. Burdick

Caltech

UNIVERSITY OF MICHIGAN

RICE

Reinforcement Learning

Reinforcement learning (RL) studies how to use data from interactions with the environment to learn an optimal policy:

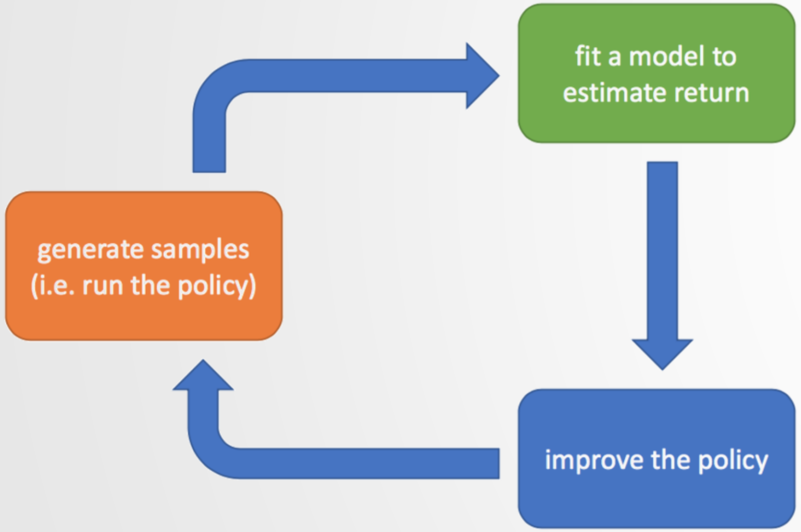


Figure from Sergey Levine

Policy: $\pi_\theta(a|s): S \times A \rightarrow [0,1]$

Reward Optimization: $$\max_\theta J(\theta) = \max_\theta \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_t^\infty \gamma^t \, r(s_t, a_t)\right]$$

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

Policy gradient-based optimization with no prior information:

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

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

Williams, 1992; Sutton et al. 1999
Baxter and Bartlett, 2000
Greensmith et al. 2004

Variance in Reinforcement Learning

RL methods suffer from high variance in learning
(Islam et al. 2017; Henderson et al. 2018)

Allows us to optimize policy with no prior information
(only sampled trajectories from interactions)

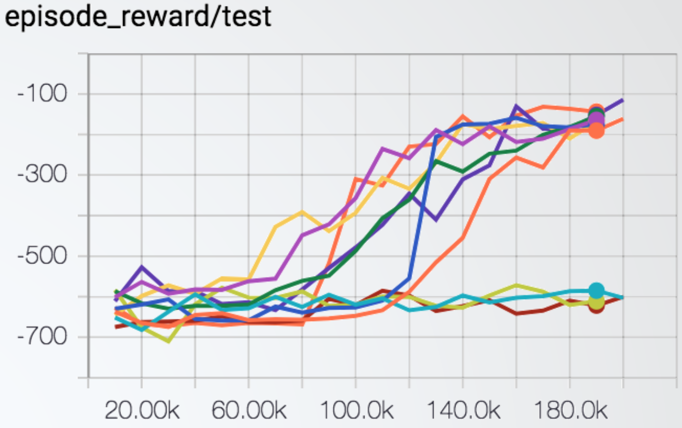


Figure from Alex Irpan

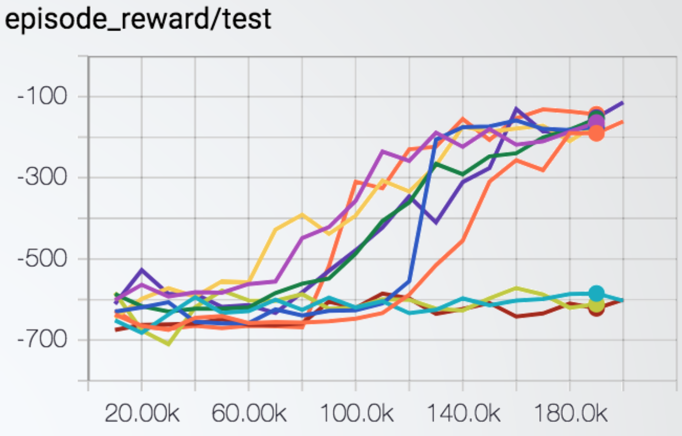
Inverted pendulum
10 random seeds

Greensmith et al. 2004, Zhao et al. 2012
Zhao et al. 2015; Thodoroff et al. 2018

Variance in Reinforcement Learning

RL methods suffer from high variance in learning
(Islam et al. 2017; Henderson et al. 2018)

Allows us to optimize policy with no prior information
(only sampled trajectories from interactions)



Inverted pendulum
10 random seeds

Figure from Alex Irpan

However, is this necessary or even desirable?

Cartpole

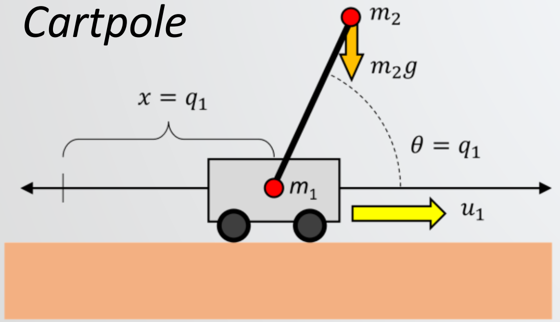


Figure from Kris Hauser

$$s_{t+1} \approx f(s_t) + g(s_t)a_t$$

LQR Controller

$$\boldsymbol{a = u_{prior}(s)}$$

Nominal controller is stable but based on:

- Error prone model
- Linearized dynamics

Greensmith et al. 2004, Zhao et al. 2012
Zhao et al. 2015; Thodoroff et al. 2018

Regularization with a Control Prior

Combine control prior, $u_{prior}(s)$, with learned controller, $u_{\theta_k}(s)$, sampled from $\pi_{\theta_k}(a|s)$

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

λ is a regularization parameter weighting the prior vs. the learned controller

π_{θ_k} learned in same manner with samples drawn from new distribution (e.g. $\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \pi} \left[\nabla_\theta \log \pi_\theta(\tau) Q^\pi(\tau) \right]$)

Under the assumption of Gaussian exploration noise (i.e. $\pi_\theta(a|s)$ has Gaussian distribution):

$$\overline{u}_k(s) = \arg\min_u \left\| u(s) - \overline{u}_{\theta_k} \right\|_\Sigma + \lambda ||u(s) - u_{prior}(s)||_\Sigma, \quad \forall s \in S$$

which can be equivalently expressed as the constrained optimization problem,

$$\overline{u}_k(s) = \arg\min_u \left\| u(s) - \overline{u}_{\theta_k} \right\|_\Sigma \quad \text{s.t.} \quad ||u(s) - u_{prior}(s)||_\Sigma \leq \tilde{\mu}(\lambda) \quad \forall s \in S,$$

Johannink et al. 2018; Silver et al. 2019

Interpretation of the Prior

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

Theorem 1. *Using the mixed policy above, variance from each policy gradient step is reduced by factor* $\frac{1}{(1+\lambda)^2}$.

However, this may introduce bias into the policy

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

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

where $D_{TV}(\cdot, \cdot)$ *represents the total variation distance between two policies.*

Interpretation of the Prior

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

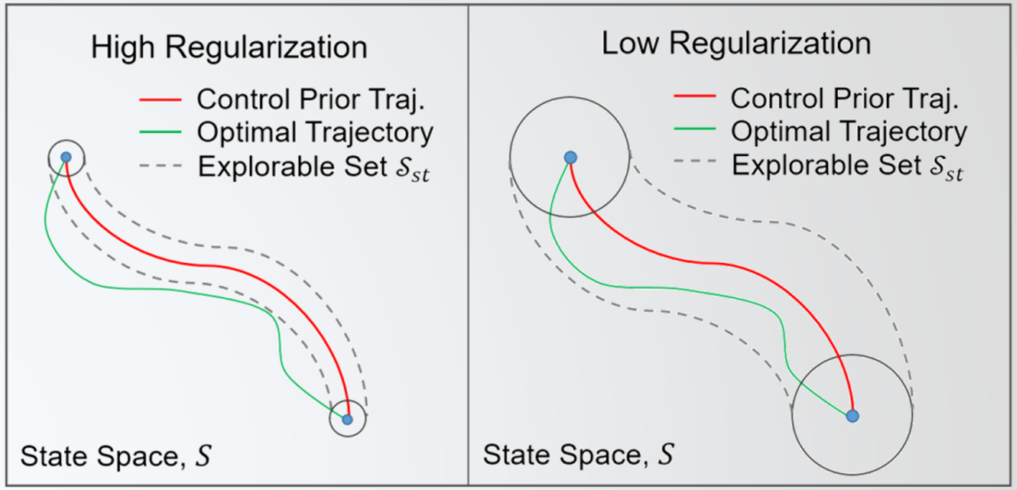
Theorem 1. *Using the mixed policy above, variance from each policy gradient step is reduced by factor* $\frac{1}{(1+\lambda)^2}$.

However, this may introduce bias into the policy

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

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

where $D_{TV}(\cdot, \cdot)$ *represents the total variation distance between two policies.*



(a) (b)

Strong regularization: The control prior heavily constrains exploration. Stabilize to the red trajectory, but miss green one.

Weak regularization: Greater room for exploration, but may not stabilize around red trajectory.

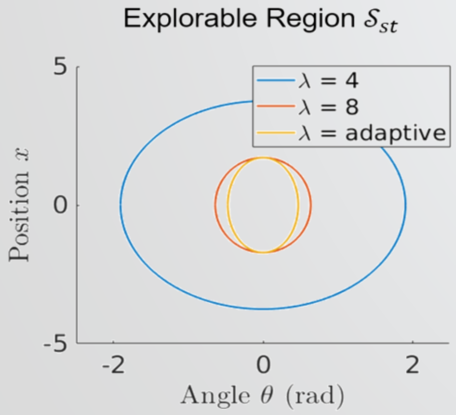
Stability Properties from the Prior

Regularization allows us to “capture” stability properties from a robust control prior

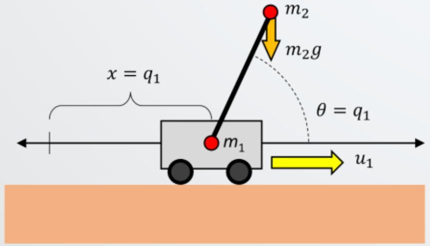
Theorem 2. Assume *a stabilizing $\mathcal{H}_\infty$ control prior within the set $\mathcal{C}$ for the dynamical system (14). Then asymptotic stability and forward invariance of the set $\mathcal{S}_{st} \subseteq \mathcal{C}$*

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

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



Cartpole



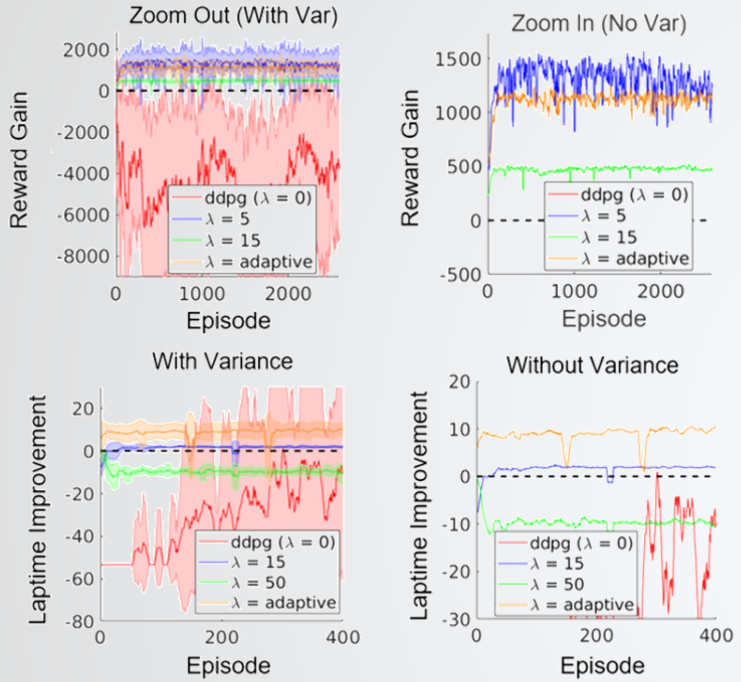
With a robust control prior, the regularized controller always remains near the equilibrium point, even during learning

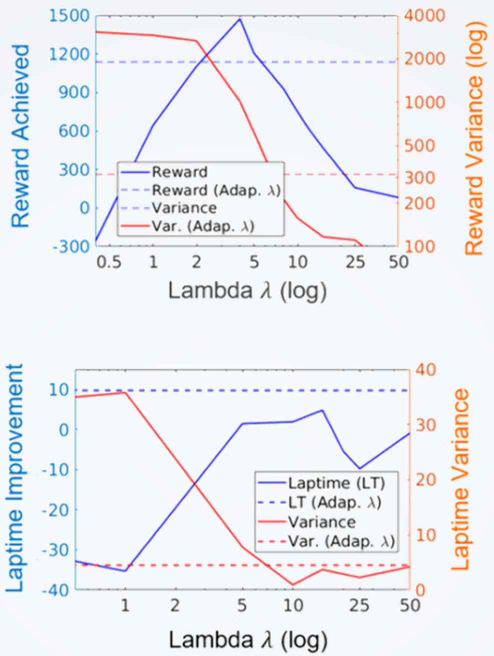
Results

Improvement over Control Prior

Reward-Variance

Car Following





TORCS RaceCar

Data gathered from chain of cars following each other. Goal is to optimize fuel-efficiency of the middle car.

Goal is to minimize laptime of simulated racecar

Control Regularization helps by providing:

- *Reduced variance*
- *Higher rewards*
- *Faster learning*
- *Potential safety guarantees*

However, high regularization also leads to potential bias

See Poster for similar results on CartPole domain

Code at: *https://github.com/rcheng805/CORE-RL*

Poster Number: 42