Differentiable Trust Region Layers for Deep Reinforcement Learning

International Conference on Learning Representations 2021

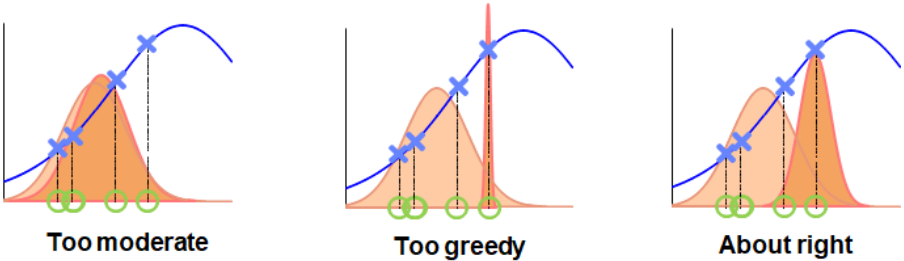
Fabian Otto[1,3] Philipp Becker[2] Ngo Anh Vien[1] Hanna Ziesche[1] Gerhard Neumann[2]

[1]Bosch Center for Artificial Intelligence, Tübingen, Germany
[2]Karlsruhe Institute of Technology, Karlsruhe, Germany
[3]Eberhard-Karls University of Tübingen, Tübingen, Germany

BOSCH

Motivation

Why Trust Regions?



- Popular approach in deep reinforcement learning, e. g. from TRPO [Schulman et al., 2015], PPO [Schulman et al., 2017], and MPO [Abdolmaleki et al., 2018]

Trust Region Optimization

We want to optimize the surrogate objective from TRPO [Schulman et al., 2015]

$$\max_{\theta} \mathbb{E}_{(s,a)\sim\pi_{\theta_{\text{old}}}} \left[\frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s)} A^{\pi_{\theta_{\text{old}}}}(s,a) \right]$$
$$\text{s.t.} \quad \mathbb{E}_{s\sim\pi_{\theta_{\text{old}}}} [d(\pi_{\theta_{\text{old}}}(\cdot|s)), \pi_\theta(\cdot|s))] \leq \epsilon$$

Trust Region Optimization

We want to optimize the surrogate objective from TRPO [Schulman et al., 2015]

$$\max_{\theta} \mathbb{E}_{(s,a)\sim\pi_{\theta_{\text{old}}}} \left[\frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s)} A^{\pi_{\theta_{\text{old}}}}(s,a) \right]$$
$$\text{s.t.} \quad d(\pi_{\theta_{\text{old}}}(\cdot|s)), \pi_\theta(\cdot|s)) \leq \epsilon \quad \forall s \in \mathcal{S}$$

External | Fabian Otto, Philipp Becker, Ngo Anh Vien, Hanna Ziesche, Gerhard Neumann | 2021-03-26

Trust Region Optimization

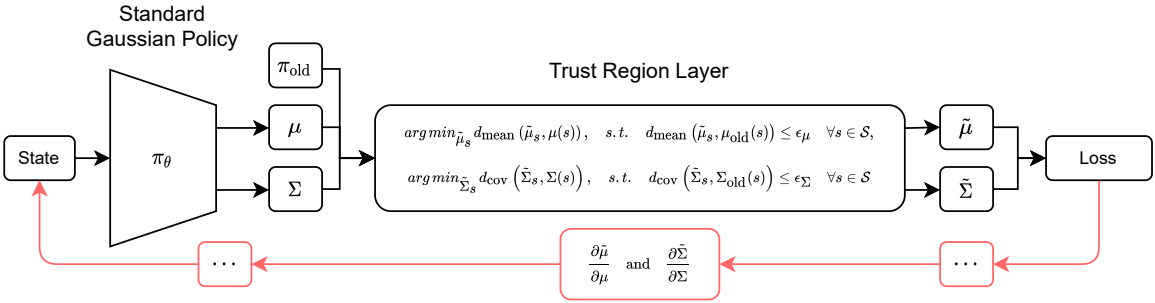
We want to optimize the surrogate objective from TRPO [Schulman et al., 2015]

$$\max_{\theta} \mathbb{E}_{(s,a)\sim\pi_{\theta_{\text{old}}}} \left[\frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s)} A^{\pi_{\theta_{\text{old}}}}(s,a) \right]$$
$$\text{s.t.} \quad d(\pi_{\theta_{\text{old}}}(\cdot|s)), \pi_\theta(\cdot|s)) \leq \epsilon \quad \forall s \in \mathcal{S}$$

For Gaussian policies we can split the trust region as shown by MPO [Abdolmaleki et al., 2018]

$$\max_{\theta} \mathbb{E}_{(s,a)\sim\pi_{\theta_{\text{old}}}} \left[\frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s)} A^{\pi_{\theta_{\text{old}}}}(s,a) \right]$$
$$\text{s.t.} \quad d_{\text{mean}}(\pi_{\theta_{\text{old}}}(\cdot|s)), \pi_\theta(\cdot|s)) \leq \epsilon_\mu \quad \forall s \in \mathcal{S}$$
$$d_{\text{cov}}(\pi_{\theta_{\text{old}}}(\cdot|s)), \pi_\theta(\cdot|s)) \leq \epsilon_\Sigma \quad \forall s \in \mathcal{S}$$

Differentiable Trust Region Layers

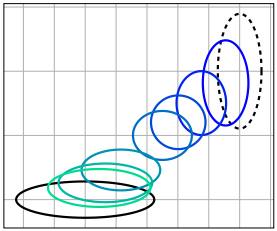


where $\pi_{\text{old}}(a|s) = \mathcal{N}(a|\mu_{\text{old}}(s), \Sigma_{\text{old}}(s))$ and $\pi_\theta(a|s) = \mathcal{N}(a|\mu(s), \Sigma(s))$.

- Allows to project π_θ into the trust region by finding parameters $\tilde{\mu}$ and $\tilde{\Sigma}$ that are closest to the original parameters μ and Σ while satisfying the trust region constraints

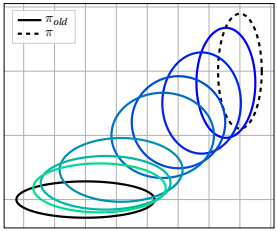
Similarity Measures

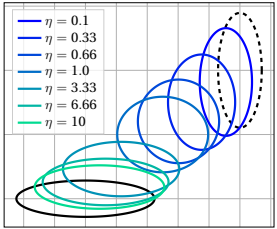
- Reverse KL Divergence
  - Mode seeking
  - Non-symmetric
  - No closed form solution
- Frobenius Norm
  - Symmetric
  - Offers closed form solution
- Wasserstein Distance
  - Symmetric
  - Offers closed form solution



KL

Frobenius





Wasserstein

Mean Projection

All three trust region layers make use of the same objective for the mean

$$\arg\min_{\tilde{\mu}} (\mu - \tilde{\mu})^{\mathrm{T}} \Sigma_{\mathrm{old}}^{-1} (\mu - \tilde{\mu}) \quad \text{s.t.} \quad (\mu_{\mathrm{old}} - \tilde{\mu})^{\mathrm{T}} \Sigma_{\mathrm{old}}^{-1} (\mu_{\mathrm{old}} - \tilde{\mu}) \leq \epsilon_{\mu}.$$

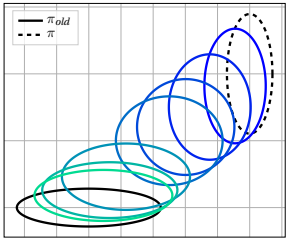
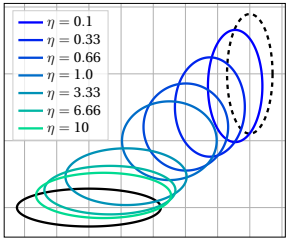
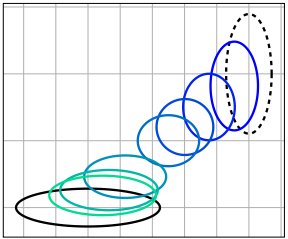
With the method of Lagrangian multipliers, we retrieve the projected mean $\tilde{\mu}$

$$\tilde{\mu} = \frac{\mu + \omega \mu_{\mathrm{old}}}{1 + \omega} \quad \text{with} \quad \omega = \sqrt{\frac{(\mu_{\mathrm{old}} - \mu)^{\mathrm{T}} \Sigma_{\mathrm{old}}^{-1} (\mu_{\mathrm{old}} - \mu)}{\epsilon_{\mu}}} - 1$$

$\Longrightarrow$ All computations are differentiable

Covariance Projection

$$\arg\min_{\tilde{\Sigma}_s} d_{\text{cov}}\left(\tilde{\Sigma}_s, \Sigma(s)\right), \quad \text{s.t.} \quad d_{\text{cov}}\left(\tilde{\Sigma}_s, \Sigma_{\text{old}}(s)\right) \leq \epsilon_\Sigma \quad \forall s \in \mathcal{S}$$

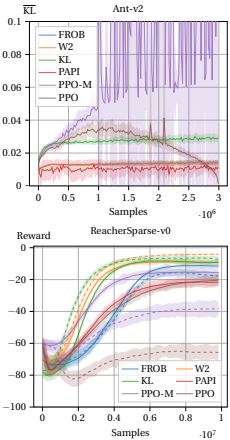
Frobenius	Wasserstein	KL
		
$\tilde{\Sigma} = \frac{\Sigma + \eta\Sigma_{\text{old}}}{1+\eta}$	$\tilde{\Sigma}^{1/2} = \frac{\Sigma^{1/2} + \eta\Sigma_{\text{old}}^{1/2}}{1+\eta}$	$\tilde{\Sigma}^{-1} = \frac{\Sigma^{-1} + \eta\Sigma_{\text{old}}^{-1}}{1+\eta}$

$\Longrightarrow$ All solutions are differentiable and Lagrangian multipliers η can be found efficiently

External | Fabian Otto, Philipp Becker, Ngo Anh Vien, Hanna Ziesche, Gerhard Neumann | 2021-03-26

Experiments

- Better or equal performance than PPO/PAPI on all Mujoco tasks
- KL projection layer is performing best
- Layers do not rely on implementations choices
- Policy changes are much more consistent based on chosen bound
- Adding entropy control improves performance further
- Layers are better suited for contextual covariances



External | Fabian Otto, Philipp Becker, Ngo Anh Vien, Hanna Ziesche, Gerhard Neumann | 2021-03-26

BOSCH

THANK YOU

Poster session:
Wed, May 5th, 2021, 9 a.m. - 11 a.m. (PDT)
Wed, May 5th, 2021, 6 p.m. - 8 p.m. (CEST)
Thu, May 6th, 2021, 12 a.m. - 2 a.m. (CST)

Correspondence: fabian.otto@bosch.com

BOSCH