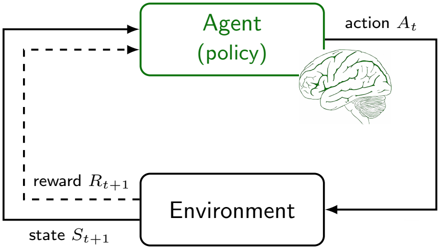
Reinforcement Learning in Configurable Continuous Environments

Alberto Maria Metelli, Emanuele Ghelfi and Marcello Restelli

36th International Conference on Machine Learning
13th June 2019

Non-Configurable Environments

Markov Decision Process
(MDP, Puterman, 2014)



$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, \gamma, \mu, p)$$

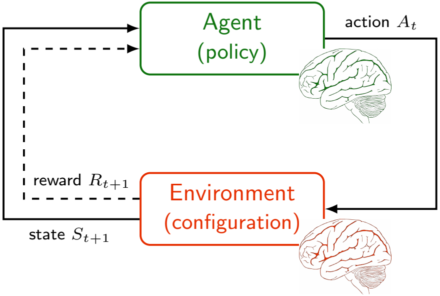
$$S_0 \sim \mu,\ A_t \sim \pi_{\boldsymbol{\theta}}(\cdot|S_t),\ S_{t+1} \sim p(\cdot|S_t, A_t)$$

- Learn the policy parameters $\boldsymbol{\theta}$ under the fixed environment p

$$\boldsymbol{\theta}^* = \operatorname*{arg\,max}_{\boldsymbol{\theta} \in \Theta} J(\boldsymbol{\theta}) = \mathbb{E}\left[\sum_{t=0}^{+\infty} \gamma^t R_{t+1}\right]$$


Sauber F1 Team
RENAULT
EDOX
FLEX-BOX
EDOX
Swiss Watches
9
Schweiz · Suisse · Svizzera
P ZERO

Configurable Markov Decision Process
(Conf-MDP, Metelli et al., 2018)



$$\mathcal{CM} = (\mathcal{S}, \mathcal{A}, r, \gamma, \mu, \mathcal{P}, \Pi)$$

$$S_0 \sim \mu,\ A_t \sim \pi_{\boldsymbol{\theta}}(\cdot|S_t),\ S_{t+1} \sim p_{\boldsymbol{\omega}}(\cdot|S_t, A_t)$$

- Learn the policy parameters $\boldsymbol{\theta}$ together with the environment configuration $\boldsymbol{\omega}$

$$\boldsymbol{\theta}^*, \boldsymbol{\omega}^* = \operatorname*{arg\,max}_{\boldsymbol{\theta}\in\Theta, \boldsymbol{\omega}\in\Omega} J(\boldsymbol{\theta}, \boldsymbol{\omega}) = \mathbb{E}\left[\sum_{t=0}^{+\infty} \gamma^t R_{t+1}\right]$$

State of the Art

- **Safe Policy Model Iteration** (SPMI, Metelli et al., 2018)
  - Optimize a lower bound of the performance improvement
- Limitations
  - **Finite** state-actions spaces
  - **Full knowledge** of the environment dynamics
- Similar approaches Keren et al. (2017) and Silva et al. (2018)

Relative Entropy Model Policy Search (REMPS)

Optimization

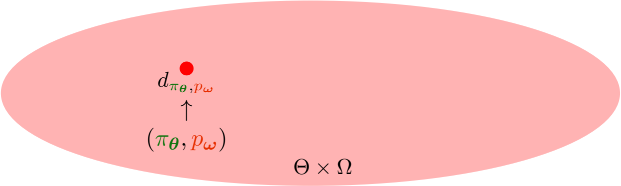
Find a new *stationary distribution* d' in a *trust region* centered in $d_{\pi_{\boldsymbol{\theta}}, p_{\boldsymbol{\omega}}}$

$$\max_{d'} J_{d'} = \mathop{\mathbb{E}}_{S,A,S' \sim d'} [r(S,A,S')]$$
$$\text{s.t. } D_{\mathrm{KL}}(d' \| d_{\pi_{\boldsymbol{\theta}}, p_{\boldsymbol{\omega}}}) \le \kappa,$$

Projection

Find a policy $\pi_{\boldsymbol{\theta}'}$ and configuration $p_{\boldsymbol{\omega}'}$ inducing a stationary distribution close to d'

$$\min_{\boldsymbol{\theta}' \in \Theta, \boldsymbol{\omega}' \in \Omega} D_{\mathrm{KL}} \left(d' \| d_{\pi_{\boldsymbol{\theta}'}, p_{\boldsymbol{\omega}'}} \right)$$



Optimization

Find a new *stationary distribution* d' in a *trust region* centered in $d_{\pi_{\boldsymbol{\theta}}, p_{\boldsymbol{\omega}}}$

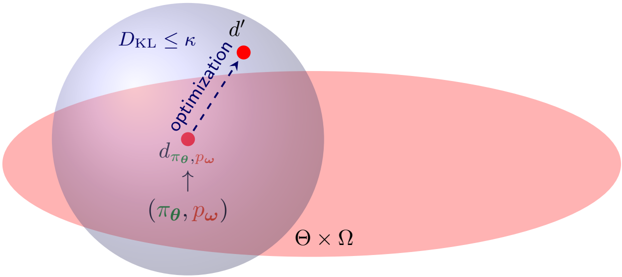
$$\max_{d'} J_{d'} = \mathop{\mathbb{E}}_{S,A,S' \sim d'} \left[r(S, A, S') \right]$$

$$\text{s.t. } D_{\mathrm{KL}}(d' \| d_{\pi_{\boldsymbol{\theta}}, p_{\boldsymbol{\omega}}}) \leq \kappa,$$

Projection

Find a policy $\pi_{\boldsymbol{\theta}'}$ and configuration $p_{\boldsymbol{\omega}'}$ inducing a stationary distribution close to d'

$$\min_{\boldsymbol{\theta}' \in \Theta, \boldsymbol{\omega}' \in \Omega} D_{\mathrm{KL}} \left(d' \| d_{\pi_{\boldsymbol{\theta}'}, p_{\boldsymbol{\omega}'}} \right)$$



Relative Entropy Model Policy Search (REMPS)

Optimization

Find a new *stationary distribution* d' in a *trust region* centered in $d_{\pi_{\boldsymbol{\theta}}, p_{\boldsymbol{\omega}}}$

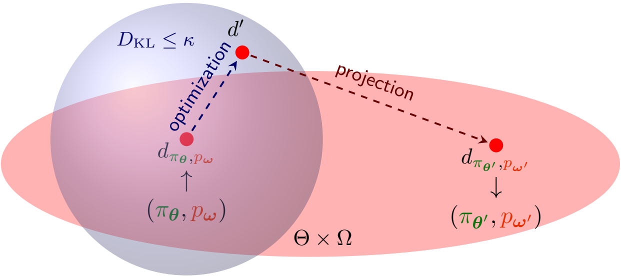
$$\max_{d'} J_{d'} = \mathop{\mathbb{E}}_{S,A,S' \sim d'} \left[r(S,A,S') \right]$$

$$\text{s.t. } D_{\mathrm{KL}}(d' \| d_{\pi_{\boldsymbol{\theta}}, p_{\boldsymbol{\omega}}}) \le \kappa,$$

Projection

Find a policy $\pi_{\boldsymbol{\theta}'}$ and configuration $p_{\boldsymbol{\omega}'}$ inducing a stationary distribution close to d'

$$\min_{\boldsymbol{\theta}' \in \Theta, \boldsymbol{\omega}' \in \Omega} D_{\mathrm{KL}} \left(d' \| d_{\pi_{\boldsymbol{\theta}'}, p_{\boldsymbol{\omega}'}} \right)$$



Relative Entropy Model Policy Search (REMPS)

Optimization

Find a new *stationary distribution* d' in a *trust region* centered in $d_{\pi_{\boldsymbol{\theta}}, p_{\boldsymbol{\omega}}}$

$$\max_{d'} J_{d'} = \mathop{\mathbb{E}}_{S,A,S' \sim d'} \left[r(S,A,S') \right]$$

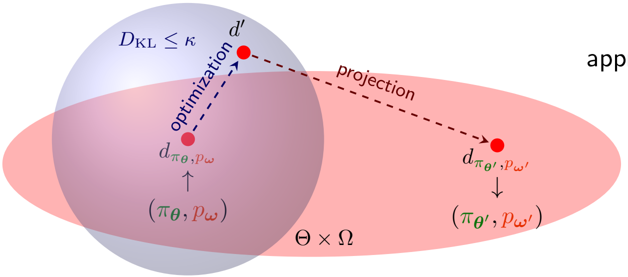
$$\text{s.t. } D_{\mathrm{KL}}(d' \| d_{\pi_{\boldsymbol{\theta}}, p_{\boldsymbol{\omega}}}) \leq \kappa,$$

Projection

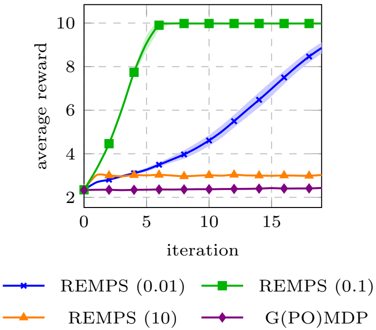
Find a policy $\pi_{\boldsymbol{\theta}'}$ and configuration $\boxed{p_{\boldsymbol{\omega}'}}$ inducing a stationary distribution close to d'

$$\min_{\boldsymbol{\theta}' \in \Theta, \boldsymbol{\omega}' \in \Omega} D_{\mathrm{KL}} \left(d' \| d_{\pi_{\boldsymbol{\theta}'}, p_{\boldsymbol{\omega}'}} \right)$$

Can also be an approximated model $\widehat{p}$

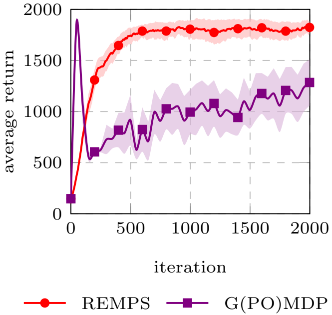


Chain Domain



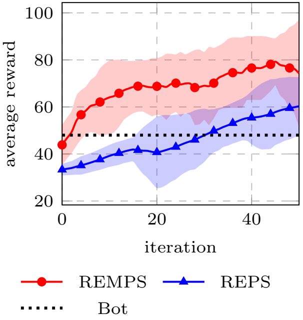
Cartpole

- Configure the *cart force*



TORCS

- Configure the *front-rear wing orientation* and *brake repartition*



Thank You for Your Attention!

- Poster **Pacific Ballroom #37**
- Code: `github.com/albertometelli/remps`
- Web page: `albertometelli.github.io/ICML2019-REMPS`
- Contact: `albertomaria.metelli@polimi.it`

References

Keren, S., Pineda, L., Gal, A., Karpas, E., and Zilberstein, S. (2017). Equi-reward utility maximizing design in stochastic environments. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4353–4360.

Metelli, A. M., Mutti, M., and Restelli, M. (2018). Configurable markov decision processes. In Dy, J. G. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 3488–3497. PMLR.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Silva, R., Melo, F. S., and Veloso, M. (2018). What if the world were different? gradient-based exploration for new optimal policies. *EPiC Series in Computing*, 55:229–242.