POLITECNICO DI MILANO
SCHOOL OF INDUSTRIAL AND INFORMATION ENGINEERING

Reinforcement Learning in Configurable Environments: an Information Theoretic approach

Supervisor: Prof. Marcello Restelli
Co-supervisor: Dott. Alberto Maria Metelli

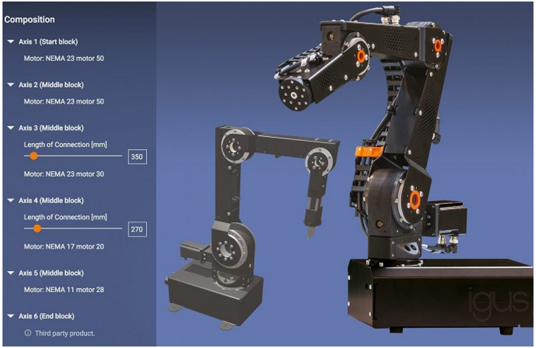
Emanuele Ghelfi, 875550

20 Dec, 2018

Contents

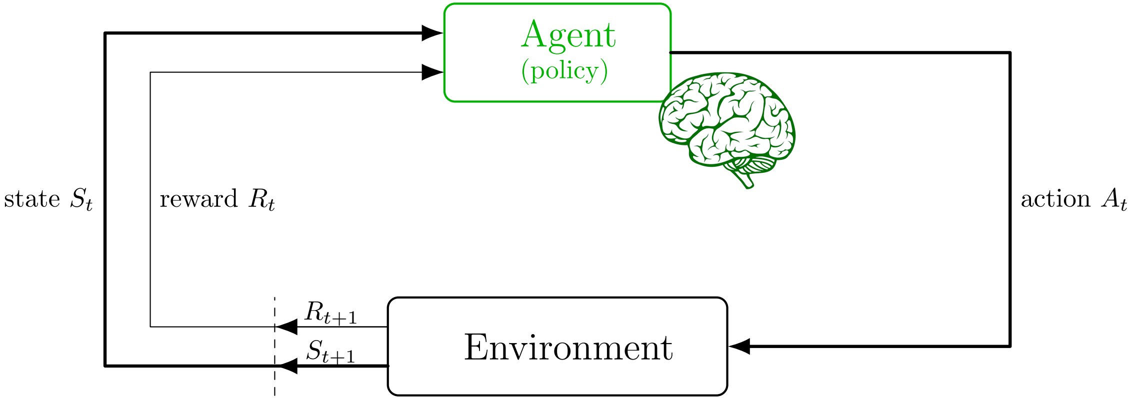
Motivations - Configurable Environments

- Configure environmental parameters in a principled way.



Reinforcement Learning

Reinforcement Learning (Sutton and Barto, 1998) considers sequential decision making problems.



Policy Search

Definition (Policy)

A policy is a function $\pi_{\boldsymbol{\theta}} : \mathcal{S} \to \Delta(\mathcal{A})$ that maps states to probability distributions over actions.

Definition (Performance)

The performance of a policy is defined as:

$$J^{\pi} = J(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\substack{a_t \sim \pi(\cdot|s_t) \\ s_{t+1} \sim P(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{H-1} \gamma^t R(s_t, a_t, s_{t+1}) \right] . \tag{1}$$

Definition (MDP solution)

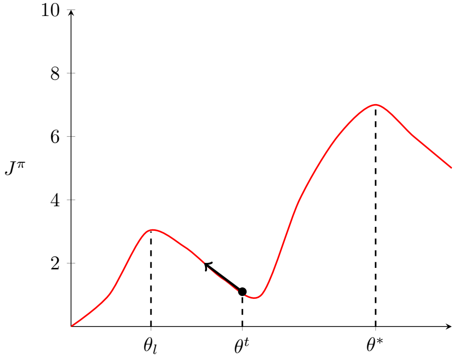
An MDP is solved when we find the best performing policy:

$$\boldsymbol{\theta}^* \in \arg\max_{\boldsymbol{\theta} \in \Theta} J(\boldsymbol{\theta}) . \tag{2}$$

Gradient vs Trust Region

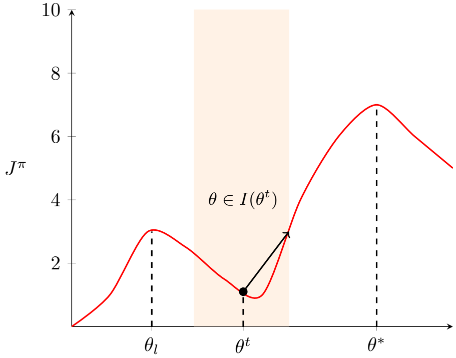
Gradient methods optimize performance acting with gradient updates:

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \lambda \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^t).$$



Trust region methods perform a constrained optimization:

$$\max_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \qquad \text{s.t. } \boldsymbol{\theta} \in I(\boldsymbol{\theta}^t).$$

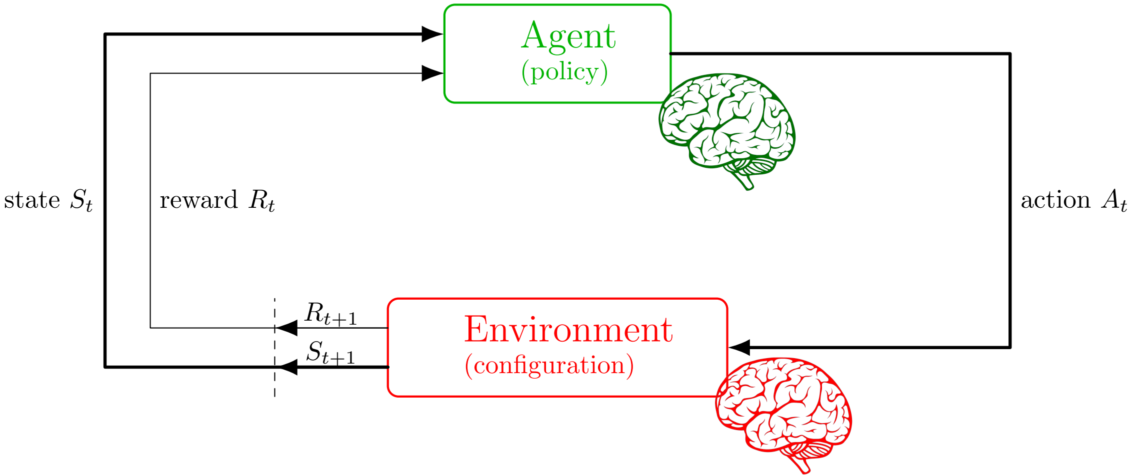


Trust Region methods

- Relative Entropy Policy Search (REPS) (Peters et al., 2010).
- Trust Region Policy Optimization (TRPO) (Schulman et al., 2015).
- Proximal Policy Optimization (PPO) (Schulman et al., 2017).
- Policy Optimization via Importance Sampling (POIS) (Metelli et al., 2018b).

Configurable MDP

A Configurable Markov Decision Process (Metelli et al., 2018a) (**CMDP**) is an MDP extension.



Configurable MDP

Definition (CMDP performance)

The performance of a model-policy pair is:

$$J^{P,\pi} = J(\boldsymbol{\omega}, \boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\substack{a_t \sim \pi(\cdot|s_t) \\ s_{t+1} \sim P(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{H-1} \gamma^t R(s_t, a_t, s_{t+1}) \right]. \tag{3}$$

Definition (CMDP Solution)

The CMDP solution is the model-policy pair maximizing the performance:

$$\boldsymbol{\omega}^*, \boldsymbol{\theta}^* \in \mathop{\arg\max}_{\boldsymbol{\omega} \in \Omega, \boldsymbol{\theta} \in \Theta} J(\boldsymbol{\omega}, \boldsymbol{\theta}). \tag{4}$$

State of the Art

Safe Policy Model Iteration (Metelli et al., 2018a):

- Safe Approach (Pirotta et al., 2013) for solving CMDPs:

$$\underbrace{J^{P',\pi'} - J^{P,\pi}}_{\text{Performance improvement}} \geq B(P',\pi') = \underbrace{\frac{\mathbb{A}_{P,\pi,\mu}^{P',\pi} + \mathbb{A}_{P,\pi,\mu}^{P,\pi'}}{1-\gamma}}_{\text{Advantage term}} - \underbrace{\frac{\gamma \Delta q^{P,\pi} D}{2(1-\gamma)^2}}_{\text{Dissimilarity Penalization}} . \quad (5)$$

Limitations:

- **Finite** state-actions space.
- **Full knowledge** of the environment dynamics.
- High sample complexity.

Relative Entropy Model Policy Search

We present **REMPS**, a novel algorithm for **CMDPs**:

- **Information Theoretic** approach.
- **Optimization** and **Projection**.
- **Approximated** models.
- **Continous** state and action spaces.

We optimize the **Average Reward**:

$$J^{P,\pi} = \liminf_{H \to +\infty} \mathop{\mathbb{E}}_{\substack{a_t \sim \pi(\cdot|s_t) \\ s_{t+1} \sim P(\cdot|s_t, a_t)}} \left[\frac{1}{H} \sum_{t=0}^{H-1} R(s_t, a_t, s_{t+1}) \right].$$

REMPS - Optimization

- We define the following constrained optimization problem:

Primal

$$\max_{d} \mathbb{E}_d[R(s, a, s')]$$

subject to:

$$D_{KL}(d || d^{P,\pi}) \leq \epsilon$$
$$\mathbb{E}_d[1] = 1 \,.$$

Dual

$$\min_{\eta} \eta \log \left(\mathbb{E}_{d^{P,\pi}} \left[e^{\left(\epsilon + \frac{R(s,a,s')}{\eta} \right)} \right] \right)$$

subject to:

$$\eta \geq 0.$$

- $d^{P,\pi}$: **sampling distribution**.
- ϵ: **Trust Region**.
- d: stationary distribution over state, action and next-state.

REMPS - Optimization

Primal

$$\boxed{\max_{d} \mathbb{E}_d[R(s, a, s')] \qquad \text{Objective Function}}$$

subject to:

$$D_{KL}(d||d^{P,\pi}) \leq \epsilon$$
$$\mathbb{E}_d[1] = 1\,.$$

- $d^{P,\pi}$: sampling distribution.
- ϵ: **Trust Region**.
- d: stationary distribution.

REMPS - Optimization

Primal

$$\max_{d} \mathbb{E}_d[R(s,a,s')]$$

subject to:

$$\boxed{D_{KL}(d||d^{P,\pi}) \leq \epsilon \qquad \text{Trust Region}}$$

$$\mathbb{E}_d[1] = 1\,.$$

- $d^{P,\pi}$: sampling distribution.
- ϵ: **Trust Region**.
- d: stationary distribution.

REMPS - Optimization

Primal

$$\max_d \mathbb{E}_d[R(s,a,s')]$$

subject to:

$$D_{KL}(d||d^{P,\pi}) \leq \epsilon$$

$$\boxed{\mathbb{E}_d[1] = 1\,. \qquad d \text{ is well formed}}$$

- $d^{P,\pi}$: sampling distribution.
- ϵ: **Trust Region**.
- d: stationary distribution.

REMPS - Optimization Solution

Theorem (REMPS solution)

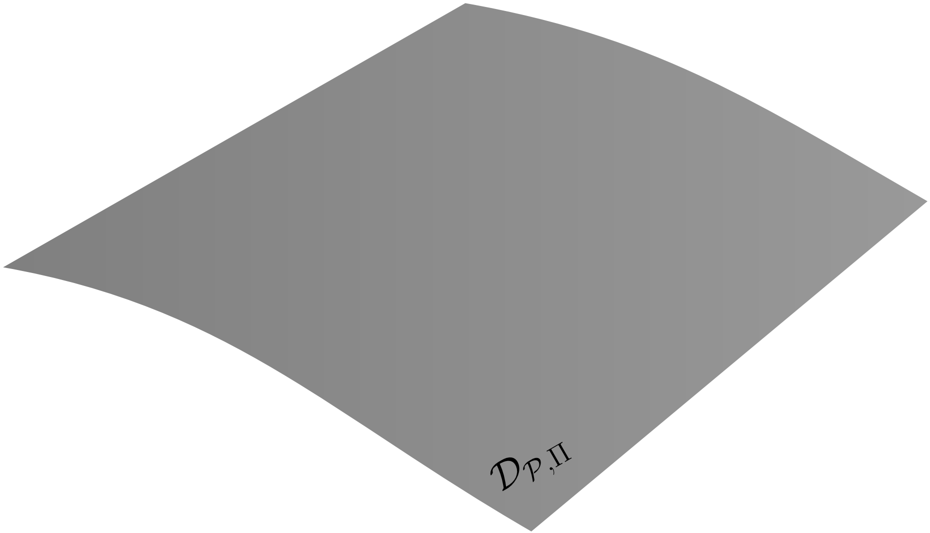
The solution of the REMPS problem is:

$$d(s,a,s') = \frac{d^{P,\pi}(s,a,s') \exp\left(\frac{R(s,a,s')}{\eta}\right)}{\int_{\mathcal{S}} \int_{\mathcal{A}} \int_{\mathcal{S}} d^{P,\pi}(s,a,s') \exp\left(\frac{R(s,a,s')}{\eta}\right) \mathrm{d}s \mathrm{d}a \mathrm{d}s'} \tag{6}$$

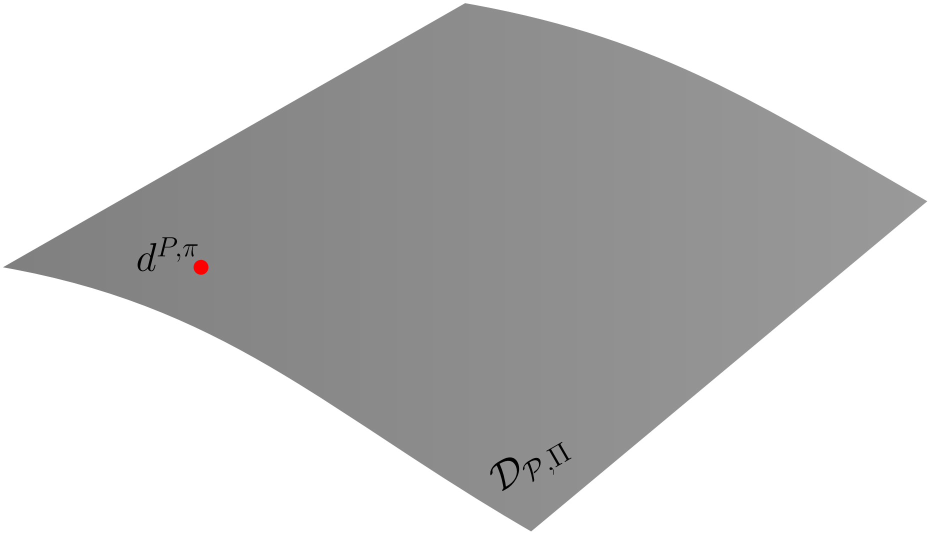
where η is the minimizer of the dual problem.

- Probability of (s,a,s') weighted exponentially with the reward.

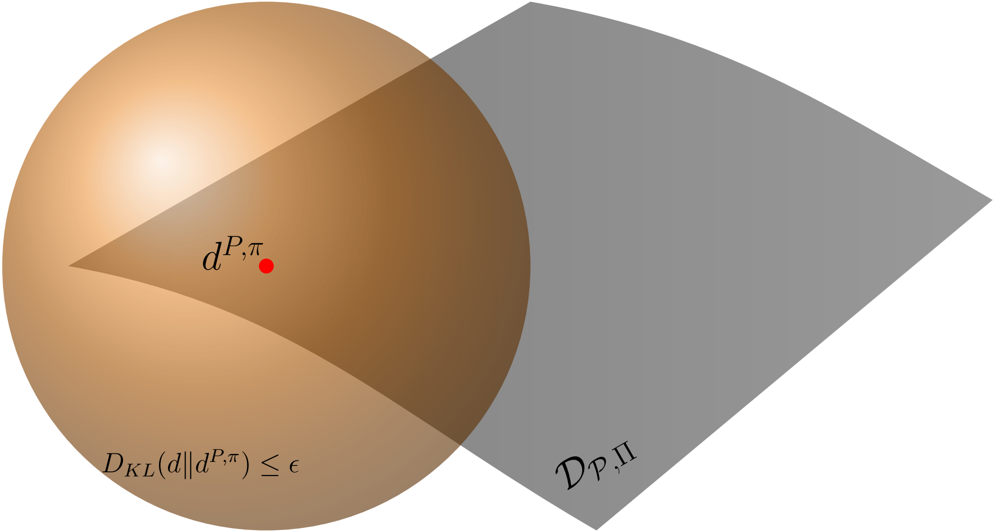
REMPS - Projection



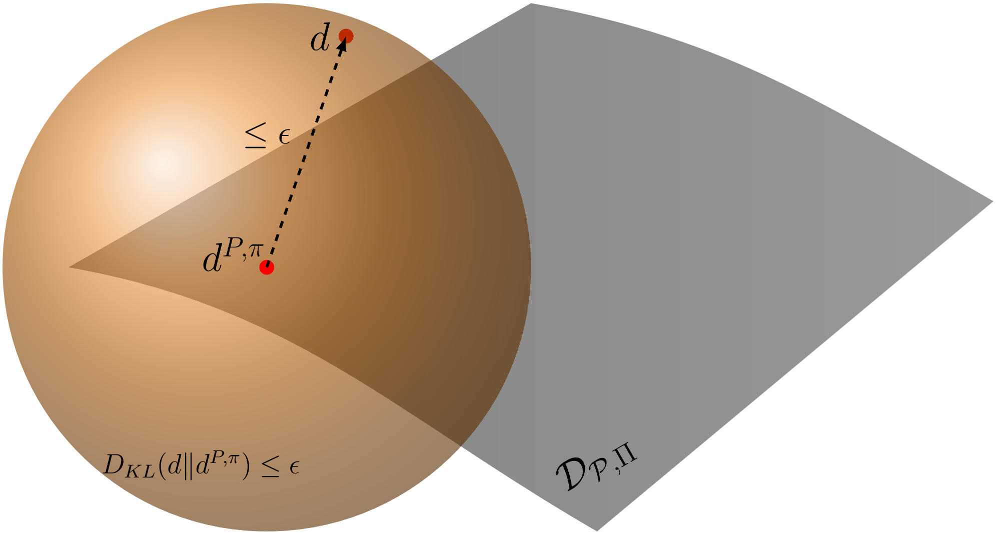
REMPS - Projection



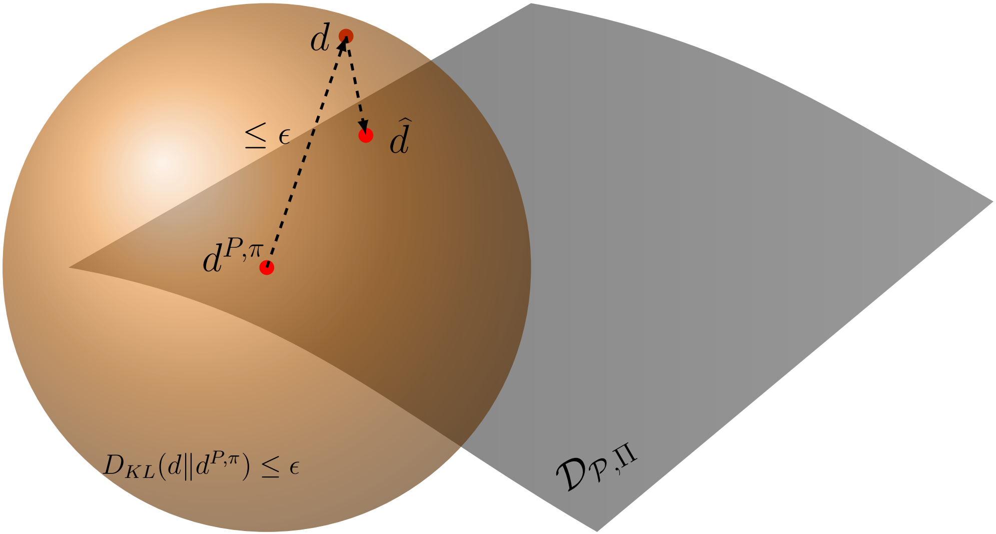
REMPS - Projection



REMPS - Projection



REMPS - Projection



REMPS - Projection

d-Projection

- Discrete state-action spaces.
- $\arg\min_{\boldsymbol{\theta}\in\Theta,\boldsymbol{\omega}\in\Omega} D_{KL}\left(d \| d^{\boldsymbol{\omega},\boldsymbol{\theta}}\right)$.

State-Kernel Projection

- Discrete action spaces.
- $\arg\min_{\boldsymbol{\theta}\in\Theta,\boldsymbol{\omega}\in\Omega} \mathbb{D}_{KL}(P^{\pi} \| P^{\pi_{\boldsymbol{\theta}}}_{\boldsymbol{\omega}})$.

Independent Projection

- Continuous state action spaces.
- $\arg\min_{\boldsymbol{\theta}\in\Theta} \mathbb{D}_{KL}(\pi' \| \pi_{\boldsymbol{\theta}})$.
- $\arg\min_{\boldsymbol{\omega}\in\Omega} \mathbb{D}_{KL}(P' \| P_{\boldsymbol{\omega}})$.

Algorithm

Algorithm 1 Relative Entropy Model Policy Search

for t = 0,1,... until convergence **do**
  Collect samples from $\pi_{\boldsymbol{\theta}_t}, P_{\boldsymbol{\omega}_t}$
  Obtain η^*, the minimizer of the dual problem. ▷ **Optimization**
  Project *d* according to the projection strategy. ▷ **Projection**
    a. *d*-Projection;
    b. State-Kernel Projection;
    c. Independent Projection.
  Update Policy.
  Update Model.
end for
return Policy-Model Pair $(\pi_{\boldsymbol{\theta}_t}, P_{\boldsymbol{\omega}_t})$

Finite–Sample Analysis

How much can differ the ideal performance from the approximated one?

- d solution of Optimization with ∞ samples.
- $\widetilde{d}$ solution of Optimization with N samples.
- $\widetilde{d}'$ solution of Optimization and Projection with N samples.

$$J_d - J_{\widetilde{d}'} = \underbrace{J_d - J_{\widetilde{d}}}_{\text{OPT}} + \underbrace{J_{\widetilde{d}} - J_{\widetilde{d}'}}_{\text{PROJ}}. \tag{7}$$

Finite–Sample Analysis

How much can differ the ideal performance from the approximated one?

$$J_d - J_{\widetilde{d'}} \leq r_{\max}\psi(N)\sqrt{\frac{8v\log\frac{2eN}{v} + 8\log\frac{8}{\delta}}{N}} \tag{8}$$

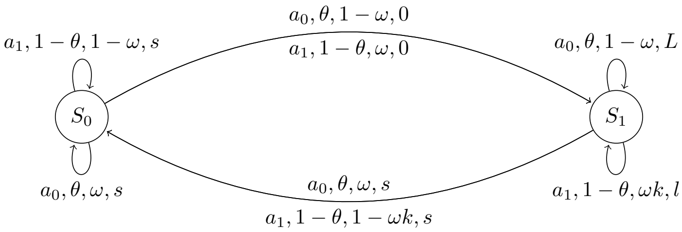
$$+ r_{\max}\phi\sqrt[4]{\frac{8v\log\frac{2eN}{v} + 8\log\frac{8}{\delta}}{N}} \tag{9}$$

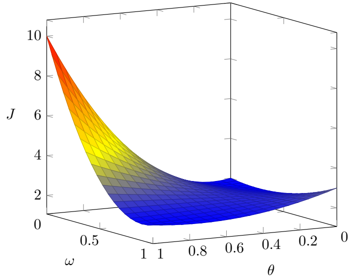
$$+ \sqrt{2}r_{\max}\sup_{d\in\mathcal{D}_{d^{P,\pi}}}\inf_{\overline{d}\in\mathcal{D}_{\mathcal{P},\Pi}}\sqrt{D_{KL}(d\|\overline{d})}. \tag{10}$$

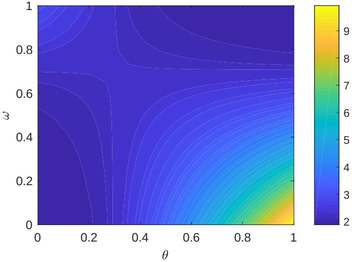
Experimental Results - Chain

Motivations:

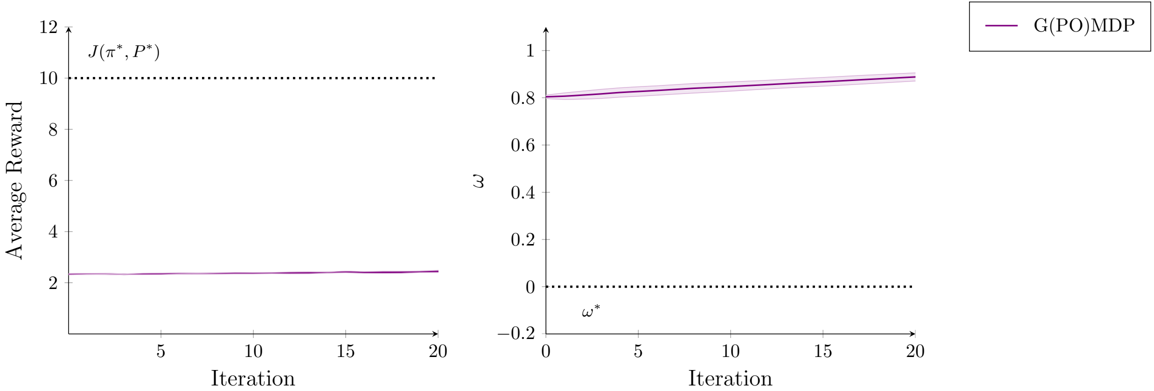
- Visualize the behaviour of **REMPS**;
- Overcoming local minima;
- Configure transition function.



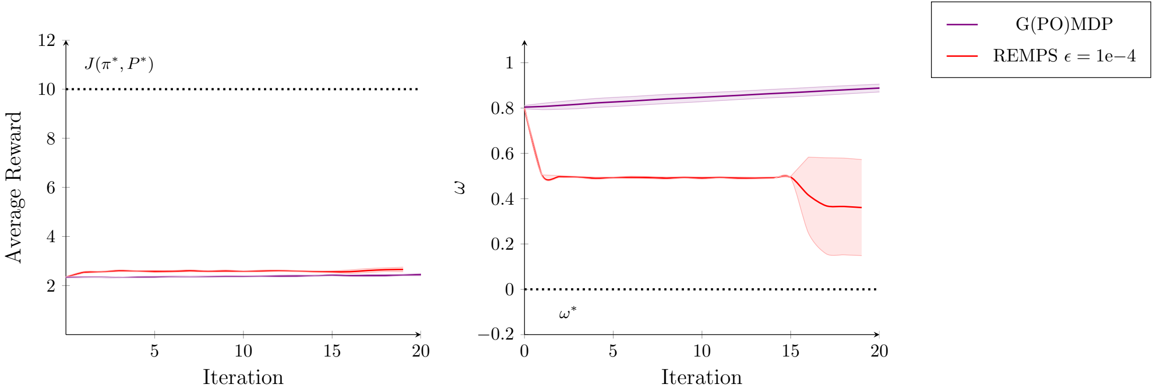




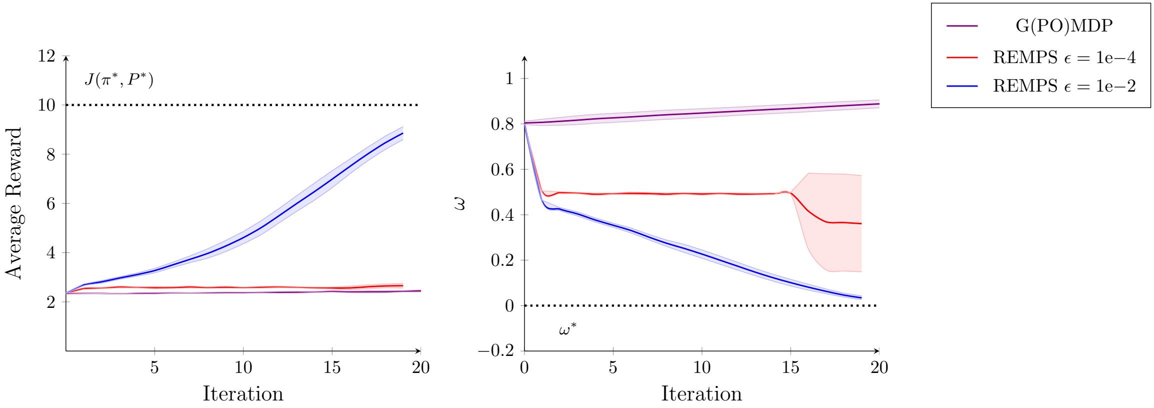
Experimental Results - Chain



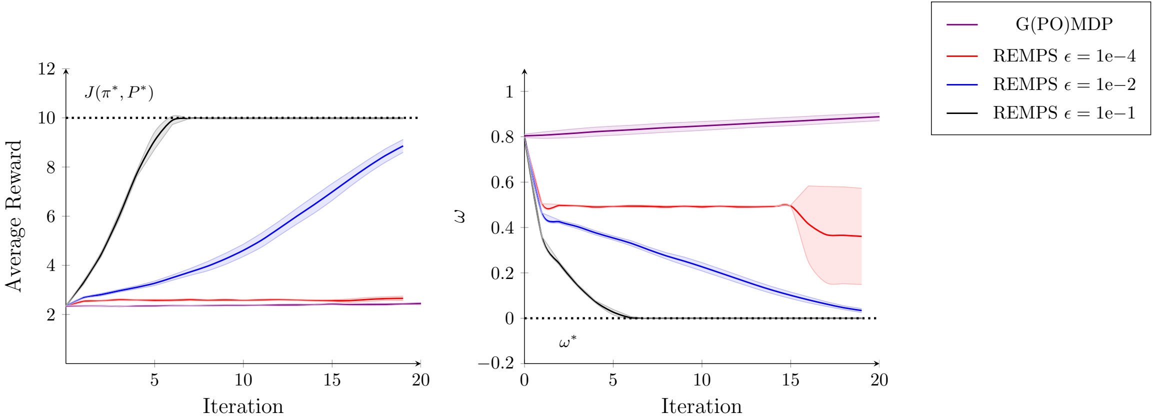
Experimental Results - Chain



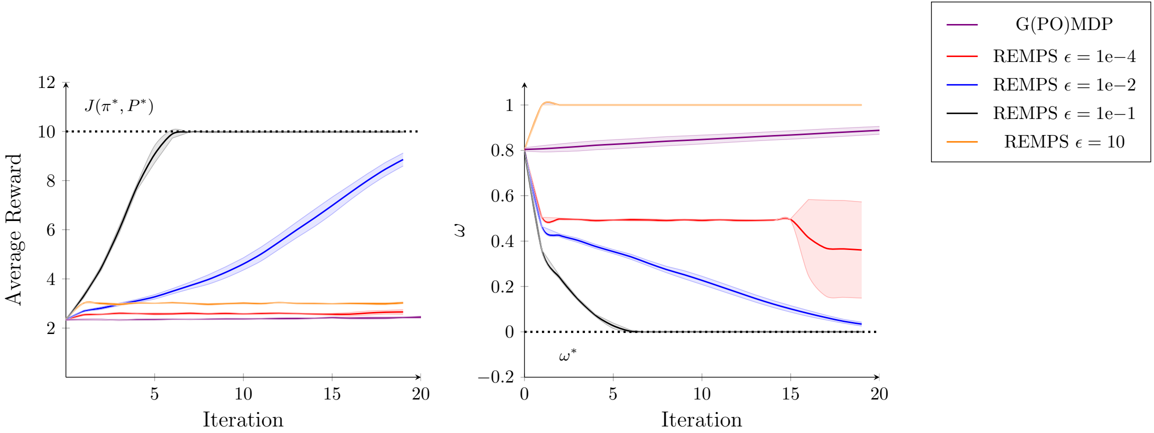
Experimental Results - Chain



Experimental Results - Chain



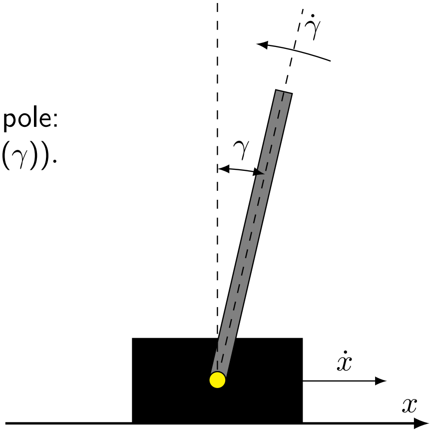
Experimental Results - Chain



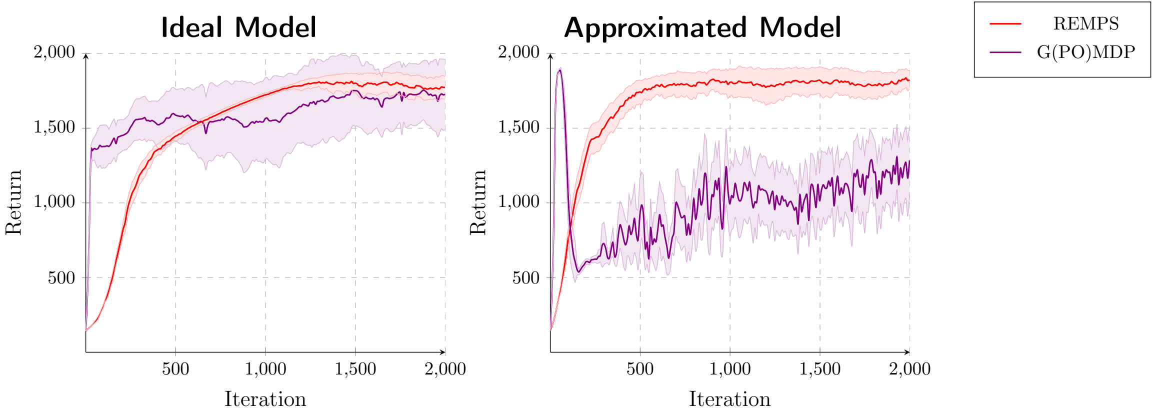
Experimental Results - Cartpole

- Standard RL benchmark;
- Configure cart **acceleration**;
- Approximated model.

Minimize acceleration balancing the pole:
$R(s, a, s') = 10 - \frac{\omega^2}{20} - 20 \cdot (1 - \cos(\gamma))$.



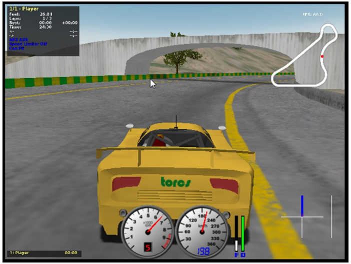
Experimental Results - Cartpole



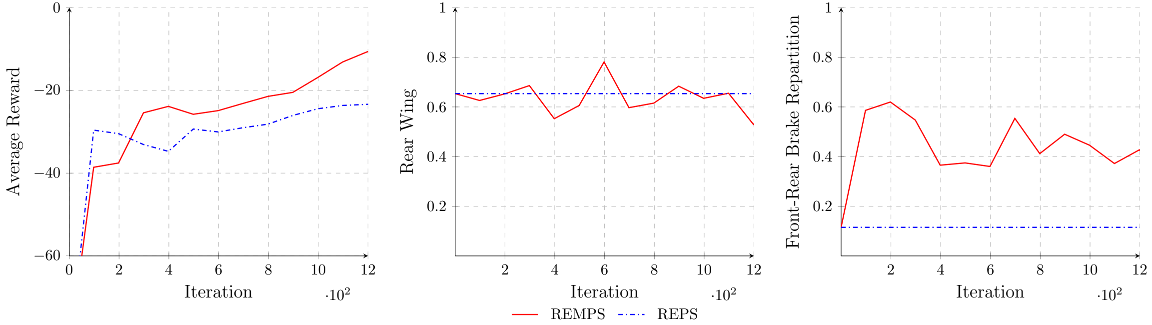
Experimental Results - TORCS

TORCS: Car Racing simulation (Bernhard Wymann, 2013; Loiacono et al., 2010).

- Autonomous driving and **Configuration**.
- **Continuous Control**.
- Approximated model.
- We configure the **rear wing** and the **brake system**.



Experimental Results - TORCS



Conclusions

- Contributions:
  - **REMPS** able to solve the model-policy learning problem.
  - **Finite-sample** analysis.
  - **Experimental** evaluation.
- Future research directions:
  - Adaptive KL-constraint.
  - Other divergences.
  - Finite-Time Analysis.
- Plan to submit at **ICML** 2019.

Conclusions

Thank you for your attention

Questions?

Bibliography I

[Bernhard Wymann 2013] Bernhard Wymann, Christophe Guionneau Christos Dimitrakakis Rémi Coulom Andrew S.: *TORCS, The Open Racing Car Simulator*. http://www.torcs.org. 2013

[Loiacono et al. 2010] Loiacono, D. ; Prete, A. ; Lanzi, P. L. ; Cardamone, L.: Learning to overtake in TORCS using simple reinforcement learning. In: *IEEE Congress on Evolutionary Computation*, July 2010, pp. 1–8. – pp. 1–8

[Metelli et al. 2018a] Metelli, Alberto M. ; Mutti, Mirco ; Restelli, Marcello: Configurable Markov Decision Processes. In: Dy, Jennifer (Ed.) ; Krause, Andreas (Ed.): *Proceedings of the 35th International Conference on Machine Learning* vol. 80. Stockholmsmässan, Stockholm Sweden : PMLR, 10–15 Jul 2018, pp. 3488–3497. pp. 3488–3497

[Metelli et al. 2018b] Metelli, Alberto M. ; Papini, Matteo ; Faccio, Francesco ; Restelli, Marcello: Policy Optimization via Importance Sampling. In: *arXiv:1809.06098 [cs, stat]* (2018), September. (2018), September

Bibliography II

[Peters et al. 2010] Peters, Jan ; Mulling, Katharina ; Altun, Yasemin: *Relative Entropy Policy Search*. 2010

[Pirotta et al. 2013] Pirotta, Matteo ; Restelli, Marcello ; Pecorino, Alessio ; Calandriello, Daniele: Safe Policy Iteration. In: Dasgupta, Sanjoy (Ed.) ; McAllester, David (Ed.): *Proceedings of the 30th International Conference on Machine Learning* vol. 28. Atlanta, Georgia, USA : PMLR, 17–19 Jun 2013, pp. 307–315. – pp. 307–315

[Schulman et al. 2015] Schulman, John ; Levine, Sergey ; Moritz, Philipp ; Jordan, Michael I. ; Abbeel, Pieter: Trust Region Policy Optimization. In: *arXiv:1502.05477 [cs]* (2015), February. (2015), February

[Schulman et al. 2017] Schulman, John ; Wolski, Filip ; Dhariwal, Prafulla ; Radford, Alec ; Klimov, Oleg: Proximal Policy Optimization Algorithms. In: *arXiv:1707.06347 [cs]* (2017), July. (2017), July

[Sutton and Barto 1998] Sutton, Richard S. ; Barto, Andrew G.: *Introduction to Reinforcement Learning*. 1st. Cambridge, MA, USA : MIT Press, 1998 Cambridge, MA, USA : MIT Press, 1998

d-projection

Projection of the stationary state distribution:

$$\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\omega}} = \arg \min_{\boldsymbol{\theta} \in \Theta, \boldsymbol{\omega} \in \Omega} D_{KL} \left(d(s, a, s') \| d^{\boldsymbol{\omega}, \boldsymbol{\theta}}(s, a, s') \right)$$

$$s.t. \; d^{\boldsymbol{\omega}, \boldsymbol{\theta}}(s) = \int_{\mathcal{S}} \int_{\boldsymbol{A}} d^{\boldsymbol{\omega}, \boldsymbol{\theta}}(s') \pi_{\boldsymbol{\theta}}(a|s') P_{\boldsymbol{\omega}}(s'|s, a) \mathrm{d}a \mathrm{d}s'.$$

Projection of the State Kernel

Projection of the state kernel:

$$
\begin{aligned}
\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\omega}} &= \arg \min_{\boldsymbol{\theta} \in \Theta, \boldsymbol{\omega} \in \Omega} \int_{\mathcal{S}} d(s) D_{KL}(P^{\pi}(\cdot|s) \| P_{\boldsymbol{\omega}}^{\pi_{\boldsymbol{\theta}}}(\cdot|s)) \mathrm{d}s \\
&= \arg \max_{\boldsymbol{\theta} \in \Theta, \boldsymbol{\omega} \in \Omega} \int_{\mathcal{S}} d(s) \int_{\mathcal{S}} P^{\pi}(s'|s) \log P_{\boldsymbol{\omega}}^{\pi_{\boldsymbol{\theta}}}(s'|s) \mathrm{d}s' \mathrm{d}s \\
&= \arg \max_{\boldsymbol{\theta} \in \Theta, \boldsymbol{\omega} \in \Omega} \int_{\mathcal{S}} d(s) \int_{\mathcal{S}} \int_{\mathcal{A}} \pi'(a|s) P'(s'|s,a) \log P_{\boldsymbol{\omega}}^{\pi_{\boldsymbol{\theta}}}(s'|s) \mathrm{d}s' \mathrm{d}s \\
&= \arg \max_{\boldsymbol{\theta} \in \Theta, \boldsymbol{\omega} \in \Omega} \int_{\mathcal{S}} \int_{\mathcal{A}} \int_{\mathcal{S}} d(s,a,s') \log \int_{\mathcal{A}} P_{\boldsymbol{\omega}}(s'|s,a') \pi_{\boldsymbol{\theta}}(a'|s) \mathrm{d}a' \mathrm{d}s \mathrm{d}a \mathrm{d}s'
\end{aligned}
$$

Independent Projection

Independent Projection of policy and model:

$$
\begin{aligned}
\widehat{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}\in\Theta} \int_{\mathcal{S}} d(s) D_{KL}(\pi'(\cdot|s) \| \pi_{\boldsymbol{\theta}}(\cdot|s)) \mathrm{d}s = && (11)\\
&= \arg\min_{\boldsymbol{\theta}\in\Theta} \int_{\mathcal{S}} d(s) \int_{\mathcal{A}} \pi'(a|s) \log \frac{\pi(a|s)}{\pi_{\boldsymbol{\theta}}(a|s)} \mathrm{d}a\mathrm{d}s = && (12)\\
&= \arg\min_{\boldsymbol{\theta}\in\Theta} \int_{\mathcal{S}} \int_{\mathcal{A}} \int_{\mathcal{S}} d(s,a,s') \log \frac{\pi'(a|s)}{\pi_{\boldsymbol{\theta}}(a|s)} \mathrm{d}s\mathrm{d}a\mathrm{d}s' = && (13)\\
&= \arg\max_{\boldsymbol{\theta}\in\Theta} \int_{\mathcal{S}} \int_{\mathcal{A}} \int_{\mathcal{S}} d(s,a,s') \log \pi_{\boldsymbol{\theta}}(a|s) \mathrm{d}s\mathrm{d}a\mathrm{d}s', && (14)
\end{aligned}
$$

Projection

Theorem (Joint bound)

Let us denote with $d^{P,\pi}$ the stationary distribution induced by the model P and policy π and $d^{P',\pi'}$ the stationary distribution induced by the model P' and policy π'. Let us assume that the reward is uniformly bounded, that is for $s, s' \in \mathcal{S}$, $a \in \mathcal{A}$ it holds that $|R(s,a,s')| < R_{\max}$. The norm of the difference of performance can be upper bounded as:

$$|J^{P,\pi} - J^{P',\pi'}| \leq R_{\max}\sqrt{2D_{KL}(d^{P,\pi} \| d^{P',\pi'})}. \tag{15}$$

Projection

Theorem (Disjoint bound)

Let us denote with $d^{P,\pi}$ the stationary distribution induced by the model P and policy π, $d^{P',\pi'}$ the stationary distribution induced by the model P' and policy π'. Let us assume that the reward is uniformly bounded, that is for $s, s' \in \mathcal{S}$, $a \in \mathcal{A}$ it holds that $|R(s,a,s')| < R_{\max}$. If (P', π') admits group invertible state kernel $P'^{\pi'}$ the norm of the difference of performance can be upper bounded as:

$$|J^{P,\pi} - J^{P',\pi'}| \leq R_{\max}\, c_1 \mathbb{E}_{s,a\sim d^{P,\pi}} \left[\sqrt{2D_{KL}(\pi' \| \pi)} + \sqrt{2D_{KL}(P' \| P)} \right], \tag{16}$$

where $c_1 = 1 + ||A'^{\#}||_{\infty}$ and $A'^{\#}$ is the group inverse of the state kernel $P'^{\pi'}$.

G(PO)MDP - Model extension

$$J^{P,\pi} = \int p_{\boldsymbol{\theta},\boldsymbol{\omega}}(\tau) G(\tau) \mathrm{d}\tau$$

$$\nabla_{\boldsymbol{\omega}} J^{P,\pi} = \int p_{\boldsymbol{\theta},\boldsymbol{\omega}}(\tau) \nabla_{\boldsymbol{\omega}} \log p(\tau) G(\tau) \mathrm{d}\tau \tag{17}$$

$$= \int p_{\boldsymbol{\theta},\boldsymbol{\omega}}(\tau) \left(\sum_{k=0}^{H} \log P_{\boldsymbol{\omega}}(s_{k+1}|s_k, a_k) \right) G(\tau) \tag{18}$$

$$\widehat{\nabla_{\boldsymbol{\omega}} J^{P\pi}}_{G(PO)MDP} = \langle \sum_{l=0}^{H} \left(\sum_{k=l}^{H} \nabla_{\boldsymbol{\omega}} \log P_{\boldsymbol{\omega}}(s_{k+1}|s_k, a_k) \right) \left(\gamma^l R(s_l, a_l, s_{l+1}) \right) \rangle_N \tag{19}$$