Configurable Markov Decision Processes

Alberto Maria Metelli, Mirco Mutti and Marcello Restelli

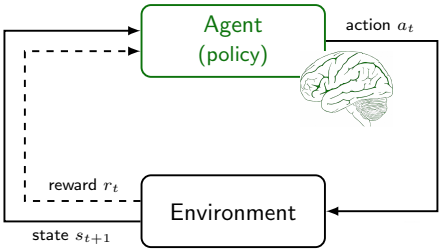
35th International Conference on Machine Learning
12th July 2018


FLEXBOX
EDOX
Swiss Watches
9
P ZERO


UP TO 70 % OFF

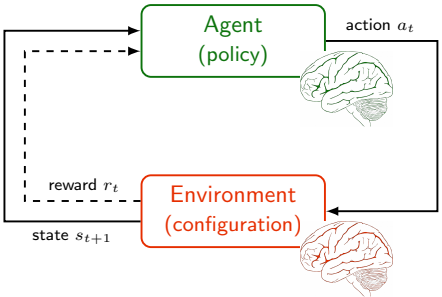
Configurable Environments

Markov Decision Process



Configurable Environments

Configurable Markov Decision Process



Configurable Markov Decision Process (Conf-MDP)

Definition

A Configurable Markov Decision Process (Conf-MDP) is a tuple:

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

- $(\mathcal{S}, \mathcal{A}, R, \gamma, \mu)$ is an MDP without the transition model
- $\mathcal{P}$ is the *model space*
- Π is the *policy space*

How to configure the model?

- The new problem:

$$P^*, \pi^* = \arg\max_{P \in \mathcal{P}, \pi \in \Pi} J_\mu^{P,\pi}$$

- In principle any Reinforcement Learning (RL) algorithm can be used
- Changing the environment must be done with care!
- We choose a *safe learning* (Kakade and Langford, 2002) method inspired by Safe Policy Iteration (SPI) (Pirotta et al., 2013)

Safe Policy Iteration (SPI)

- Idea: choose the new policy π' by optimizing a lower bound on the performance improvement

$$\underbrace{J_\mu^{\pi'} - J_\mu^{\pi}}_{\text{performance improvement}} \geq \underbrace{\frac{\mathbb{A}_{\pi,\mu}^{\pi'}}{1-\gamma}}_{\text{advantage}} - \underbrace{\frac{\|\pi' - \pi\|_\infty^2}{(1-\gamma)^3}}_{\text{dissimilarity penalization}}$$

- Rationale: update towards a policy with large advantage avoiding moving too far
- This rationale is at the basis of many successful algorithms (e.g., Peters et al., 2010; Schulman et al., 2015, 2017)

Safe Policy Model Iteration (SPMI)

Performance Improvement Bound

- Idea: choose the new policy π' and the new environment configuration P' by optimizing a lower bound on the performance improvement

$$\underbrace{J_\mu^{P',\pi'} - J_\mu^{P,\pi}}_{\text{performance improvement}} \geq B(P',\pi',P,\pi) = \underbrace{\frac{\mathbb{A}_{P,\pi,\mu}^{P',\pi}}{1-\gamma}}_{\text{model advantage}} + \underbrace{\frac{\mathbb{A}_{P,\pi,\mu}^{P,\pi'}}{1-\gamma}}_{\text{policy advantage}} - \underbrace{\frac{D_{P,\pi}^{P',\pi'}}{(1-\gamma)^3}}_{\text{dissimilarity penalization}}$$

- $D_{P,\pi}^{P',\pi'}$ is a dissimilarity term between (P,π) and (P',π').
- Rationale: update towards a policy and a model with large advantage without moving too far.

Safe Policy Model Iteration (SPMI)

Policy and Model Update

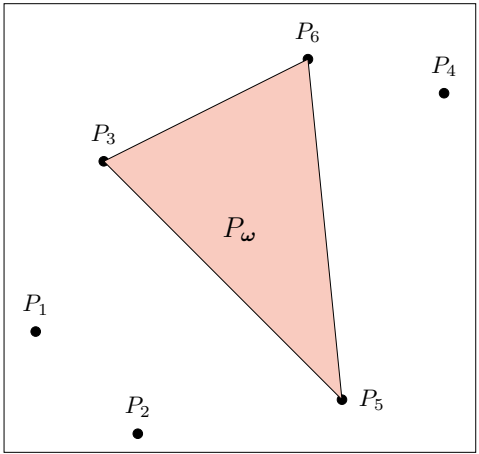
- How to select the next pair (P', π')?

$$\pi' = \alpha\overline{\pi} + (1-\alpha)\pi$$
$$P' = \beta\overline{P} + (1-\beta)P$$
$$\alpha^*, \beta^* = \arg\max_{\alpha,\beta\in[0,1]} B(P', \pi', P, \pi)$$

- guaranteed monotonic performance improvement
- *joint* and *adaptive* model and policy optimization

- How to represent the model space $\mathcal{P}$?
  - *unconstrained*: any model is valid
  - *parametric*: $P_{\boldsymbol{\omega}}$ parametric
    - e.g., $P_{\boldsymbol{\omega}}$ convex combination of set of vertex models

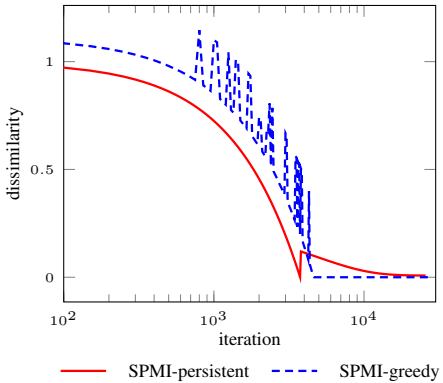
$$\mathcal{P} = \mathrm{co}(\boldsymbol{P}) \qquad \boldsymbol{P} = \{P_1, P_2, \ldots, P_M\}$$



Safe Policy Model Iteration (SPMI)

Target model selection

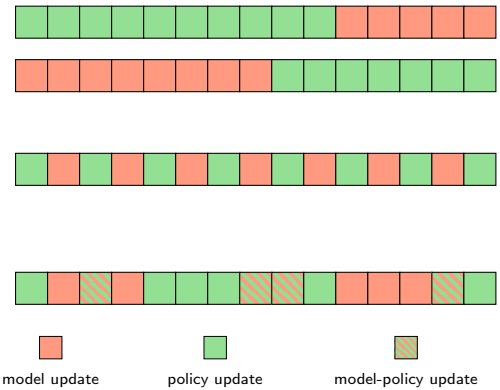
- How to select the target model $\overline{P}$?
  - *greedy*: select the model maximizing the advantage
  - *persistent*: keep the old target unless the greedy has larger bound value



Experimental Evaluation

Setting

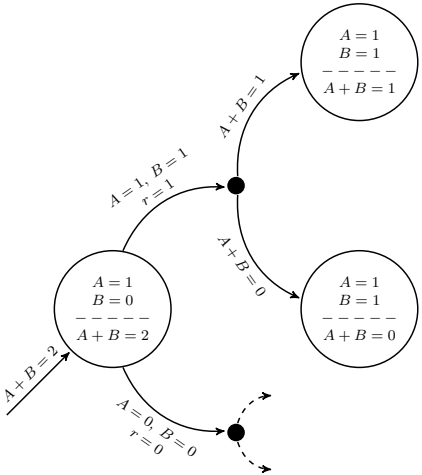
- Two illustrative domains (Student-Teacher domain, Racetrack simulator)
- Compared algorithms
  - *Sequential*: learn the model and the policy in sequence
    - SMI+SPI
    - SPI+SMI
  - *Alternated*: alternate one policy update and one model update
    - SPMI-alt
  - *Adaptive*: update policy or model or both based on the bound
    - SPMI
    - SPMI-sup



Experimental Evaluation

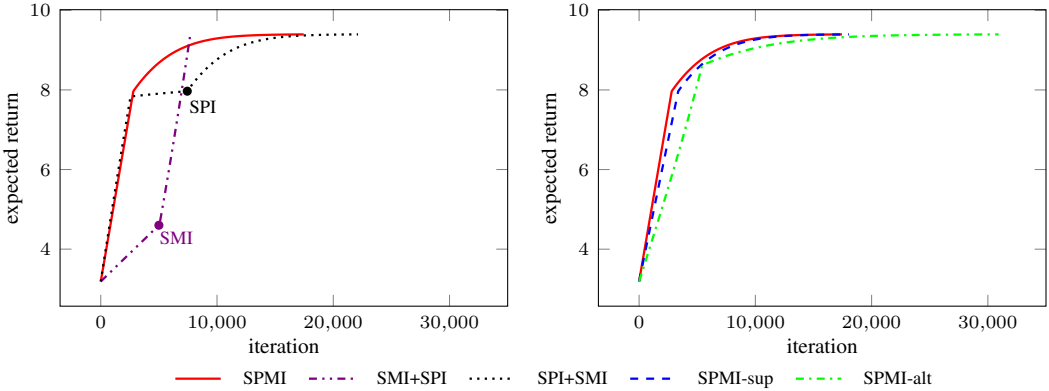
Student-Teacher domain

- Model of *concept learning* inspired by (Rafferty et al., 2011)
- An *online teaching platform* (environment) provides the student with a statement (e.g., A+B=2)
- The *student* (agent) performs an assignment of the literals (e.g., A=1, B=1)
- The student is rewarded when it performs a *consistent* assignment
- The student can change a limited number of literals at a time



Experimental Evaluation

Student-Teacher domain



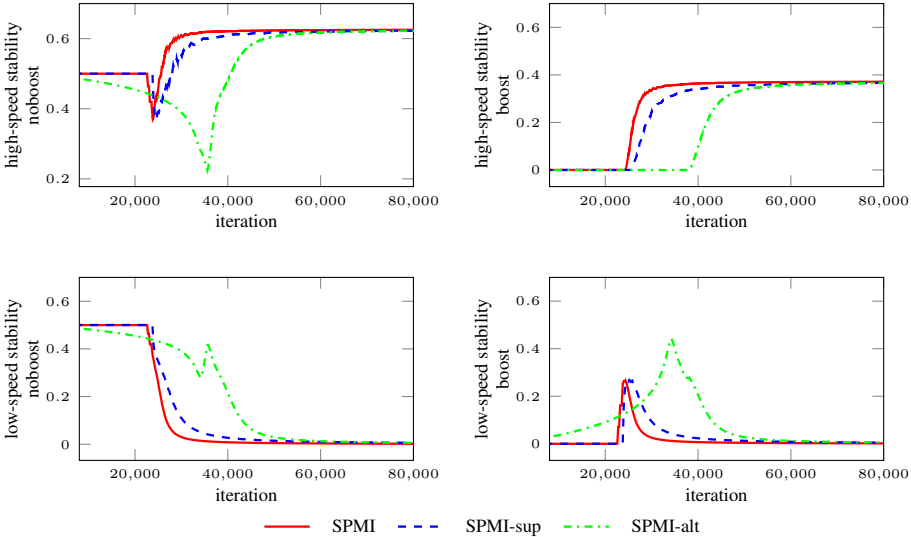
Experimental Evaluation

Racetrack simulator

- Simplified model of a car racing problem
- The *driver* (agent) wants to minimize the lap time
- The *track engineer* configures the car (environment) acting on:
  - Vehicle stability (high-speed vs low speed stability)
  - Engine power (boost vs no boost)

Experimental Evaluation

Racetrack simulator - 4 vertex models



Discussion

- Could we model *environment configuration* with the existing frameworks? *Not exactly...*
- **Unique agent**
  - Difficult to constrain the configuration space $\mathcal{P}$
  - The configuration action happens at the beginning of the episode
- **Multi-agent system**
  - Configuring the environment concerns with which MDP is more convenient to solve
  - Not really an interaction between learning agents

Conclusions

- Main contributions
  - New framework: *Conf-MDP*
  - Class of *safe learning algorithms*
  - Experimental evaluation
- Configuring the environment can improve the agent performance
- Future Works
  - Extension to parametric model $P_{\boldsymbol{\omega}}$ and policy $\pi_{\boldsymbol{\theta}}$
  - Identification of the agent policy space Π and the configuration space $\mathcal{P}$

Thank you for your attention!

albertomaria.metelli@polimi.it

Poster #88 @ Hall B

Tonight 06:15 - 09:00 PM

References

Bowerman, B. L. (1974). *Nonstationary Markov decision processes and related topics in nonstationary Markov chains*. PhD thesis, Iowa State University.

Ciosek, K. A. and Whiteson, S. (2017). Offer: Off-environment reinforcement learning. In *AAAI*, pages 1819–1825.

Florensa, C., Held, D., Wulfmeier, M., Zhang, M., and Abbeel, P. (2017). Reverse curriculum generation for reinforcement learning. In *Conference on Robot Learning*, pages 482–495.

Garcia, A. and Smith, R. L. (2000). Solving nonstationary infinite horizon dynamic optimization problems. *Journal of Mathematical Analysis and Applications*, 244(2):304–317.

Givan, R., Leach, S., and Dean, T. (1997). Bounded parameter markov decision processes. In Steel, S. and Alami, R., editors, *Recent Advances in AI Planning*, pages 234–246. Springer Berlin Heidelberg.

Kakade, S. and Langford, J. (2002). Approximately optimal approximate reinforcement learning. In *Proceedings of the 19th International Conference on Machine Learning*, volume 2, pages 267–274.

Peters, J., Mülling, K., and Altun, Y. (2010). Relative entropy policy search. In *AAAI*, pages 1607–1612. Atlanta.

Pirotta, M., Restelli, M., Pecorino, A., and Calandriello, D. (2013). Safe policy iteration. In *Proceedings of the 30th International Conference on International Conference on Machine Learning*, volume 28 of *ICML'13*, pages 307–315.

References (cont.)

Rafferty, A. N., Brunskill, E., Griffiths, T. L., and Shafto, P. (2011). Faster teaching by pomdp planning. In *AIED*, pages 280–287. Springer.

Satia, J. K. and Lave Jr, R. E. (1973). Markovian decision processes with uncertain transition probabilities. *Operations Research*, 21(3):728–740.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. (2015). Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, ICML'15, pages 1889–1897.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, pages 1057–1063.

White III, C. C. and Eldeib, H. K. (1994). Markov decision processes with imprecise transition probabilities. *Operations Research*, 42(4):739–749.

Related Works

- Environment not-known or changes naturally
  - MDPs with imprecise probabilities (Satia and Lave Jr, 1973; White III and Eldeib, 1994)
  - Bounded parameters MDPs (Givan et al., 1997)
  - Non-stationary MDPs (Bowerman, 1974; Garcia and Smith, 2000)
- Environment changes to favor learning
  - Simulator configuration (Florensa et al., 2017)
  - Change initial state (Ciosek and Whiteson, 2017)

Value functions - 1

- Advantage functions

$$A^{P,\pi}(s,a,s') = U^{P,\pi}(s,a,s') - Q^{P,\pi}(s,a)$$
$$A^{P,\pi}(s,a) = Q^{P,\pi}(s,a) - V^{P,\pi}(s)$$
$$\tilde{A}^{P,\pi}(s,a,s') = U^{P,\pi}(s,a,s') - V^{P,\pi}(s)$$

- Relative advantage functions

$$A^{P',\pi}_{P,\pi}(s,a) = \int P'(s'|s,a) A^{P,\pi}(s,a,s') \mathrm{d}s'$$
$$A^{P,\pi'}_{P,\pi}(s) = \int \pi'(a|s) A^{P,\pi}(s,a) \mathrm{d}a$$
$$A^{P',\pi'}_{P,\pi}(s) = \int \int P'(s'|s,a)\pi'(a|s) \tilde{A}^{P,\pi}(s,a,s') \mathrm{d}a \mathrm{d}s'$$

Value functions - 2

Lemma

The following equality relates the relative advantage functions:

$$A_{P,\pi}^{P',\pi'}(s) = A_{P,\pi}^{P,\pi'}(s) + \int \pi'(a|s) A_{P,\pi}^{P',\pi}(s,a) \mathrm{d}a$$

- Expected relative advantage functions

$$\mathbb{A}_{P,\pi,\mu}^{P',\pi} = \int \int d_{\mu}^{P,\pi}(s)\pi(a|s) A_{P,\pi}^{P',\pi}(s,a) \mathrm{d}a \mathrm{d}s$$

$$\mathbb{A}_{P,\pi,\mu}^{P,\pi'} = \int d_{\mu}^{P,\pi}(s) A_{P,\pi}^{P,\pi'}(s) \mathrm{d}s$$

$$\mathbb{A}_{P,\pi,\mu}^{P',\pi'} = \int d_{\mu}^{P,\pi}(s) A_{P,\pi}^{P',\pi'}(s) \mathrm{d}s$$

Bound on the γ-discounted distributions

Corollary

Let (P, π) and (P', π') be two model-policy pairs, the ℓ^1-norm of the difference between the γ- discounted state distributions can be upper bounded as:

$$\left\| d_\mu^{P',\pi'} - d_\mu^{P,\pi} \right\|_1 \leq \mathbb{E}_{\substack{s \sim d_\mu^{P,\pi} \\ a \sim \pi(\cdot|s)}} \left[\left\| P'(\cdot|s,a) - P(\cdot|s,a) \right\|_1 + \left\| \pi'(\cdot|s) - \pi(\cdot|s) \right\|_1 \right]$$

Performance Improvement

Theorem

The performance improvement of model- policy pair (P', π') over (P, π) is given by:

$$J_\mu^{P',\pi'} - J_\mu^{P,\pi} = \frac{1}{1-\gamma} \int d_\mu^{P',\pi'}(s) A_{P,\pi}^{P',\pi'}(s) \mathrm{d}s$$

Dissimilarity Term

$$D_{P,\pi}^{P',\pi'} = D_{\mathbb{E}}^{\pi',\pi}\left(D_{\infty}^{\pi',\pi} + D_{\infty}^{P',P}\right) + D_{\mathbb{E}}^{P',P}\left(D_{\infty}^{\pi',\pi} + \gamma D_{\infty}^{P',P}\right)$$

$$D_{\mathbb{E}}^{\pi',\pi} = \mathbb{E}_{s\sim d_{\mu}^{P,\pi}} \left\|\pi'(\cdot|s) - \pi(\cdot|s)\right\|_1$$

$$D_{\infty}^{\pi',\pi} = \sup_{s\in\mathcal{S}} \left\|\pi'(\cdot|s) - \pi(\cdot|s)\right\|_1$$

$$D_{\mathbb{E}}^{P',P} = \mathbb{E}_{\substack{s\sim d_{\mu}^{P,\pi} \\ a\sim\pi(\cdot|s)}} \left\|P'(\cdot|s,a) - P(\cdot|s,a)\right\|_1$$

$$D_{\infty}^{P',P} = \sup_{s\in\mathcal{S},a\in\mathcal{A}} \left\|P'(\cdot|s,a) - P(\cdot|s,a)\right\|_1$$

Theoretical Analysis

- $P_{\boldsymbol{\omega}}$ convex combination of set of vertex models $\boldsymbol{P} = \{P_1, P_2, \ldots, P_M\}$

Theorem

For any transition model $P_{\boldsymbol{\omega}} \in \mathrm{co}(\boldsymbol{P})$ *it holds that* $\mathbb{A}^{P_{\boldsymbol{\omega}}}_{P_{\boldsymbol{\omega}^*},\mu} \leq 0$. *Moreover, for all* $P_{\boldsymbol{\omega}} \in \mathrm{co}\big(\{P_i \in \boldsymbol{P} : \omega_i^* > 0\}\big)$, *it holds that* $\mathbb{A}^{P_{\boldsymbol{\omega}}}_{P_{\boldsymbol{\omega}^*},\mu} = 0$.

- Sufficient condition for optimality: all the expected relative advantages must be non-positive

P-gradient Theorem

- Extend *policy gradient* (Sutton et al., 2000) for model learning:

$$\nabla_{\boldsymbol{\omega}} J_{\mu}^{P_{\boldsymbol{\omega}}} = \int d_{\mu}^{P_{\boldsymbol{\omega}}}(s)\pi(a|s)\nabla_{\boldsymbol{\omega}} P_{\boldsymbol{\omega}}(s'|s,a)U^{P_{\boldsymbol{\omega}}}(s,a,s')\mathrm{d}s'\mathrm{d}a\mathrm{d}s.$$

$$U^{P_{\boldsymbol{\omega}}}(s,a,s') = \mathbb{E}\left[\sum_{t=0}^{+\infty}\gamma^t r_{t+1}\middle| s_0 = s, a_0 = a, s_1 = s'\right]$$

$$\underbrace{J_\mu^{P',\pi'} - J_\mu^{P,\pi}}_{\text{performance improvement}} \geq B(P',\pi',P,\pi) = \underbrace{\frac{\mathbb{A}_{P,\pi,\mu}^{P',\pi}}{1-\gamma}}_{\text{model advantage}} + \underbrace{\frac{\mathbb{A}_{P,\pi,\mu}^{P,\pi'}}{1-\gamma}}_{\text{policy advantage}} - \underbrace{\frac{D_{\text{sup}}}{(1-\gamma)^3}}_{\text{dissimilarity penalization}}$$

$$D_{\text{sup}} = \left\|\pi' - \pi\right\|_\infty^2 + \gamma \left\|P' - P\right\|_\infty^2 + (1+\gamma) \left\|\pi' - \pi\right\|_\infty \left\|P' - P\right\|_\infty$$

- Number of iterations to convergence

Problem	SPMI	SPMI-sup	SPMI-alt	SPI+SMI	SPI+SMI
2-1-1-2	16234	18054	30923	22130	**7705**
2-1-2-2	**2839**	3194	5678	**2839**	12973
2-2-1-2	20345	18287	>50000	39722	**10904**
2-2-2-2	**12025**	14315	>50000	>50000	15257
2-3-1-2	14187	13391	**11772**	>50000	12183
3-1-1-2	15410	17929	22707	31122	**14257**
3-1-2-2	**3313**	3313	8434	**3313**	22846
3-1-3-2	**2945**	3435	5891	**2945**	18090

Experimental Evaluation

Racetrack simulator - 2 vertex models