Preference-based Reinforcement Learning

Róbert Busa-Fekete[1,2], Weiwei Cheng[1], Eyke Hüllermeier[1], Balázs Szörényi[2,3], Paul Weng[3]

[1]Computational Intelligence Group, Philipps-Universität Marburg

[2]Research Group on Artificial Intelligence of the Hungarian Academy of Sciences and University of Szeged (RGAI)

[3]INRIA Lille - Nord Europe, SequeL project, 40 avenue Halley, 59650 Villeneuve d'Ascq, France

[4]Laboratory of Computer Science, Université Pierre et Marie Curie

EWRL 2013/ Dagstuhl

August 6, 2013

Preference-based Reinforcement Learning

- Motivating example: Medical treatment design

Direct policy search (DPS)

- Evolutionary direct policy search approach

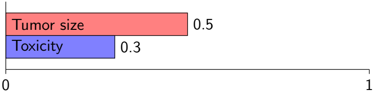
Preference-based Evolutionary Direct policy search (PB-EDPS)

- Preference-based Racing (PBR)

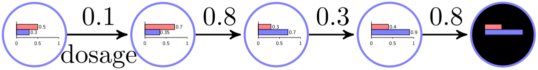
- Many problems where it is hard to define a reasonable reward function
  - task of driving [Abbeel and Ng, 2004]
  - medical treatment design [Zhao et al., 2009]
- Aggregation of rewards: one may not always be willing/able to combine rewards
  - Multi-objective reinforcement learning
- Episodic setup: $\mathbf{h}$ following policy π, $\mathbf{h}'$ following policy π'
- Given $\mathbf{h}$ and $\mathbf{h}'$, it might be easier to decide which one is preferred (at least in some problems)
- The piece of information we want to learn from is preferences over simulations!

Motivating example: medical treatment design [Zhao et al., 2009]

- Virtual patient with cancer
- State captures some essential factors in cancer treatment



- Episodic setup: an episode corresponds to a treatment of a patient over six months
- The action is the dosage level itself



- Transitions:
  - The tumor is constantly growing (without treatment or if the dosage is too low)
  - The higher dose selected, the higher toxicity evolves, and the more tumor growth is inhibited.
  - The higher the toxicity and the tumor size, the higher the probability of the patient's death.

Motivating example [Zhao et al., 2009]

- Terminal state: end of sixth month or patient dies
- The reward is defined based on the wellness of patient
  - tumor size: $\nearrow$ -5, $\rightarrow$ 5, $\searrow$ 15
  - toxicity level: $\nearrow$ -5, $\rightarrow$ 0, $\searrow$ 5
  - The reward assigned to **death** is -60
- Based on the wellness of patients, it is straightforward to define a **preference relation** over treatments
  - Given two trajectories $\mathbf{h}_1$ and $\mathbf{h}_2$ generated by following two different treatments
  - 

  - 

  - Otherwise Pareto dominance
    - $\mathbf{h}_1 \succ \mathbf{h}_2$ if the tumor size AND the toxicity level both are smaller under $\mathbf{h}_1$

Point of departure: preferences

- There is no reward function (hard to define a reasonable one) and the goal is not to find a reward function!!!
- The piece of information we want to learn from is preferences over trajectories!
- Partial order $\prec$ over trajectories $\mathbf{h} \in \mathcal{H}^{(T)}$
  - From a tutor or an expert
  - Extracted from trajectories

Decision model

- Decision model: "lifting" the preference relation $\prec$ on $\mathcal{H}^{(T)}$ to a preference relation $\ll$ on the space of policies
- Intermezzo:
  each policy π generate a probability distribution over the set of trajectories (for a fixed MDP) which is denoted by $\mathbf{P}_\pi$
  - policy $\equiv$ random variable whose realizations are trajectories
- $s(\pi, \pi') = \mathbb{E}_{\mathbf{h} \sim \mathbf{P}_\pi, \mathbf{h}' \sim \mathbf{P}_{\pi'}} [\mathbb{I}\{\mathbf{h} \prec \mathbf{h}'\}]$
  - Probability of that π' beats π
- Ordinal decision model

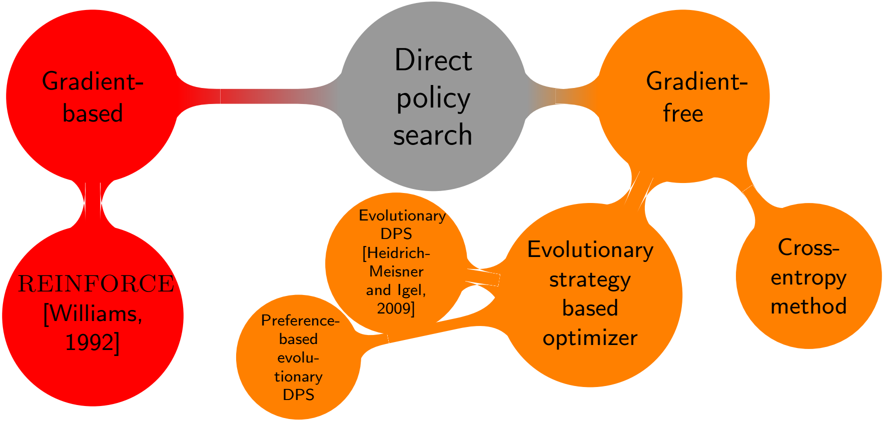
$$\pi \ll \pi' \text{ if and only if } s(\pi', \pi) < s(\pi, \pi')$$

- Alternative decision model?

Preference-based Evolutionary direct policy search

Direct policy search (DPS)

1. Parametric policy space: $\Pi = \{\pi_\Theta | \Theta \in \mathbb{R}^d\}$, for example the space of linear policies: $\pi_{\mathbf{w}}(\mathbf{s}) = \mathbf{w}^T\mathbf{s}$ if $\mathcal{S} \subseteq \mathbb{R}^d$
2. The policy search can be viewed as an optimization task: Π is the search space, some policy evaluation is the target function



Evolutionary direct policy search approach

- [Heidrich-Meisner and Igel, 2009]
- Covariance Matrix Adaptation Evolution Strategy (CMA-ES)[Hansen and Kern, 2004]
  - It maintains a distribution over the solution space (in this case over the space of policies)
- Expected total reward is optimized that can be estimated based on finite set of trajectories $\{\mathbf{h}_1, \ldots, \mathbf{h}_n\} \sim \mathbf{P}_\pi$ as

$$\widehat{\rho}_\pi^{(n)} = \frac{1}{n} \sum_{i=1}^{n} V(\mathbf{h}_i)$$

where $V(.)$ is the cummulative reward

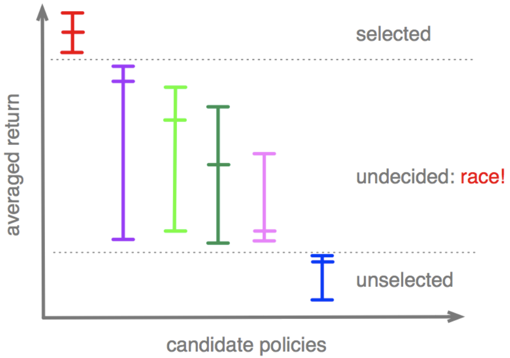
Evolutionary direct policy search [Heidrich-Meisner and Igel, 2009]

- Repeat these three steps until convergence
  1. Generate a population of candidate solutions (in this case, a set of policies with different parameters).
     - $\pi_{\Theta_1}, \ldots, \pi_{\Theta_\lambda}$ where $\Theta_1, \ldots, \Theta_\lambda \sim \mathcal{N}(\mathbf{m}, \Sigma)$
  2. Evaluate the candidate solutions (estimate the performance of the policies based on simulations $\{\mathbf{h}_1, \ldots, \mathbf{h}_n\} \sim \mathbf{P}_{\pi_{\Theta_i}}$).

$$\widehat{\rho}^{(n)}_{\pi_{\Theta_i}} = \frac{1}{n} \sum_{i=1}^{n} V(\mathbf{h}_i)$$

     and select the best μ individuals
  3. Update $\mathbf{m}$ and Σ by using the parameters of best μ individuals/policies

Racing algorithm



(a) [Heidrich-Meisner and Igel, 2009]

- In the bandit literature, these algorithms are called PAC bandits

Basic idea

1. Direct motivation: the Evolution strategy optimizers need only ranking, but they do not need the function values themselves
2. GOAL: devise a racing algorithm that utilizes only pairwise comparison of random samples (in this case trajectories) and is able to select the best policies with respect to the decision model ($\ll$)
3. This naturally gives rise to a preference-based policy search method

Recall the decision model

- $s(\pi, \pi') = \mathbb{E}_{\mathbf{h} \sim \mathbf{P}_\pi, \mathbf{h}' \sim \mathbf{P}_{\pi'}} [\mathbb{I}\{\mathbf{h} \prec \mathbf{h}'\}]$
- Ordinal decision model

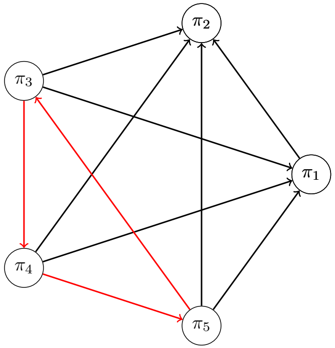
$$\pi \ll \pi' \text{ if and only if } s(\pi', \pi) < s(\pi, \pi')$$

- There can be preferential cycles
  - $\pi \ll \pi'$ AND $\pi' \ll \pi''$ AND $\pi'' \ll \pi$
  - "select the best options" is not a well-defined task
- Practical solution: surrogate ranking model
  - Given $\pi_1, \ldots, \pi_K$

$$\pi_i \ll_C \pi_j \Leftrightarrow d_i < d_j$$

    where $d_i = |\{k : \pi_k \ll \pi_i, k \neq i\}|$
  - It is a complete preorder since it has a numeric representation (d_i)
  - Unfortunately, the preference relation $\ll_C$ depends on the set of policies considered

An example for the surrogate ranking model



- edge $\Leftrightarrow \pi_i \ll \pi_j$
- $\ll_C$
  - $d_2 = 4$
  - $d_1 = 3$
  - $d_3 = d_4 = d_5 = 1$

Concentration property of $\bar{s}(.,.)$

- $s(\pi, \pi') = \mathbb{E}_{\mathbf{h} \sim \mathbf{P}_\pi, \mathbf{h}' \sim \mathbf{P}_{\pi'}} [\mathbb{I}\{\mathbf{h} \prec \mathbf{h}'\}]$
- $\pi \ll \pi'$ if and only if $s(\pi', \pi) < s(\pi, \pi')$
- $\pi_i \ll_C \pi_j \Leftrightarrow d_i < d_j$ where $d_i = |\{k : \pi_k \ll \pi_i, k \neq i\}|$
- $s(\pi, \pi')$ can be estimated based on finite sets of trajectories $\{\mathbf{h}_1, \ldots, \mathbf{h}_n\} \sim \mathbf{P}_\pi$ and $\{\mathbf{h}'_1, \ldots, \mathbf{h}'_n\} \sim \mathbf{P}_{\pi'}$ as

$$\bar{s}(\pi, \pi') = \frac{1}{nn'} \sum_{i=1}^{n} \sum_{j=1}^{n'} \mathbb{I}\{\mathbf{h}_i \prec \mathbf{h}'_j\}$$

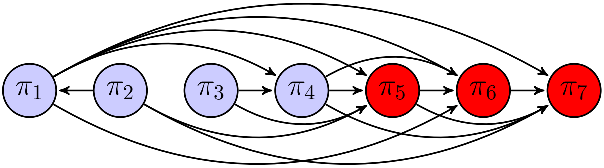
- Hoeffding-bound for U-statistics, two-sample case
- Hoeffding, 1963, §5b: For any $\epsilon > 0$

$$\mathbf{P}\left(\left|\bar{s}(\pi, \pi') - s(\pi, \pi')\right| \geq \epsilon\right) \leq 2 \exp\left(-2 \min(n, n')\epsilon^2\right)$$

- Empirical Bernstein-bound?

Preference-based racing

- We have an *efficient estimator* for $s(\pi_i, \pi_j)$
- We can calculate confidence interval for $\bar{s}(\pi_i, \pi_j)$
- $K = 7, K' = 3$
  edge $\Leftrightarrow$ $\bar{s}(\pi_i, \pi_j)$ is significantly bigger than $\bar{s}(\pi_j, \pi_i)$



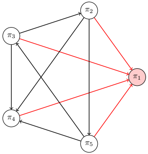
- Expected sample complexity: Even-Dar et al. [2002]
  $(\Delta_{i,j} = |1/2 - s(\pi_i, \pi_j)|)$

Preference-based evolutionary direct policy search

- Repeat these three steps until convergence
  1. Generate a population of candidate solutions (in this case, a set of policies with different parameters).
     - $\pi_{\Theta_1}, \ldots, \pi_{\Theta_\lambda}$ where $\Theta_1, \ldots, \Theta_\lambda \sim \mathcal{N}(\mathbf{m}, \Sigma)$
  2. Select the best μ individuals by using Preference-based Racing algorithm
  3. Update $\mathbf{m}$ and Σ by using the parameters of best μ individuals/policies

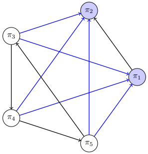
The relation of $\ll$ and $\ll_C$ (only locally valid)

π_i is a Condorcet winner among a set of policies $\pi_1, \dots, \pi_K$ if $\pi_\ell \ll \pi_i$ for all $\ell \neq i$

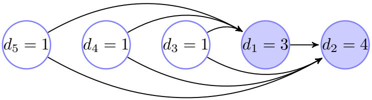


- If the Condorcet winner exists, it is the largest element of $\ll_C$

Smith set is the smallest non-empty set $\mathcal{D} \subset \{\pi_1, \dots, \pi_K\}$ satisfying $\pi_k \ll \pi_i$ for all $\pi_i \in \mathcal{D}$ and $\pi_j \in \{\pi_1, \dots, \pi_K\} \setminus \mathcal{D}$



- **Proposition** Let $\Pi = \{\pi_1, \dots, \pi_K\}$ be a set of random variables for which there exists a Smith set $\mathcal{D}$ of size $K_\mathcal{D}$. Then for any $\pi_i \in \mathcal{D}$ and $\pi_j \in \Pi \setminus \mathcal{D}$, $\pi_j \ll_C \pi_i$.



Issues to be discussed

- The existence of global optima
- If there exists a global Condorcet winner, under what assumptions we can find it (w.h.p) by using Evolution strategy along with Preference-based racing algorithm?
- Hoeffding-bound is loose: the use of Clopper-Pearson-type confidence bound for trinomial random variables

Bibliography I

P. Abbeel and A.Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the 21th International conference on Machine Learning*, pages ??–??, 2004.

E. Even-Dar, S. Mannor, and Y. Mansour. PAC bounds for multi-armed bandit and markov decision processes. In *Proceedings of the 15th Annual Conference on Computational Learning Theory*, pages 255–270, 2002.

N. Hansen and S. Kern. Evaluating the CMA evolution strategy on multimodal test functions. In *Parallel Problem Solving from Nature-PPSN VIII*, pages 282–291, 2004.

V. Heidrich-Meisner and C. Igel. Hoeffding and Bernstein races for selecting policies in evolutionary direct policy search. In *Proceedings of the 26th International Conference on Machine Learning*, pages 401–408, 2009.

J. Hemelrijk. Note on Wilcoxon's two-sample test when ties are present. *The Annals of Mathematical Statistics*, 23(1):133–135, 1952.

Bibliography II

R.J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, 1992.

Y. Zhao, M.R. Kosorok, and D. Zeng. Reinforcement learning design for cancer clinical trials. *Statistics in Medicine*, 28(26):3294–3315, 2009.

Preference-based racing for $\ll_C$

- One can estimate $s(\pi, \pi')$ based on finite set of trajectories
- $\{\mathbf{h}_1, \dots \mathbf{h}_n\} \sim \mathbf{P}_\pi$ and $\left\{\mathbf{h}'_1, \dots \mathbf{h}'_{n'}\right\} \sim \mathbf{P}_{\pi'}$

$$\bar{s}(\pi, \pi') = \frac{1}{nn'} \sum_{i=1}^{n} \sum_{j=1}^{n'} \mathbb{I}\{\mathbf{h}_i \prec \mathbf{h}'_j\}$$

- Incomparable trajectories: solution by [Hemelrijk, 1952]

$$\mathbb{I}^{\prec}\{x, x'\} = \begin{cases} 1 & \text{if} \quad x \prec x' \\ 0 & \text{if} \quad x' \prec x \\ 1/2 & \text{otherwise} \end{cases}$$

- Probabilistic interpretation: if two samples are incomparable, then we select one of them being preferred with probability 1/2
- $s(\pi, \pi') = 1 - s(\pi', \pi)$

Preference-based racing: optimization view

- Preference-based case: $\mathrm{PBR}(\pi_1, \ldots, \pi_K, K', n_{\max}, \delta)$

$$\operatorname*{argmax}_{I \subseteq \{1,\ldots,K\}:\, |I|=K'} \sum_{i \in I} \sum_{j \neq i} \mathbb{I}\{\pi_j \ll_C \pi_i\}$$

  with probability at least $1 - \delta$

- Since $s(\pi_i, \pi_j) = 1 - s(\pi_j, \pi_i)$

$$\operatorname*{argmax}_{I \subseteq \{1,\ldots,K\}:\, |I|=K'} \sum_{i \in I} \sum_{j \neq i} \mathbb{I}\{s(\pi_j, \pi_i) > 1/2\} \tag{1}$$

- We have an *efficient estimator* of $s(\pi_i, \pi_j)$

Algorithm 1 $\mathrm{PBR}(\pi_1, \ldots, \pi_K, K', n_{\max}, \delta)$

$A = \{(i,j) \mid 1 \le i,j \le K\}$, $n = 0$
while $(n \le n_{\max}) \wedge (|A| > 0)$ **do**
  for all i appearing in A **do**
    $\mathbf{h}_i^{(n)} \sim \mathcal{M}$ and π_i ▷ Generate trajectories
  end for
  for all $(i,j) \in A$ **do**
    Update $\bar{s}_{i,j} = \frac{1}{n^2} \sum_{\ell=1}^{n} \sum_{\ell'=1}^{n} \mathbb{I}\{\mathbf{h}_i^{(\ell)} \prec \mathbf{h}_j^{(\ell')}\}$
    $c_{i,j} = \sqrt{\frac{1}{2n} \log \frac{2K^2 n_{\max}}{\delta}}$, $u_{i,j} = \widehat{s}_{i,j} + c_{i,j}$, $\ell_{i,j} = \widehat{s}_{i,j} - c_{i,j}$
  end for
  for $i = 1 \to K$ **do**
    $z_i = |\{j : u_{i,j} < 1/2, j \ne i\}|$, $o_i = |\{j : \ell_{i,j} > 1/2, j \ne i\}|$
  end for
  $C = \left\{i : K - K' < \left|\{j : K - z_j < o_i\}\right|\right\}$ ▷ select
  $D = \left\{i : K' < \left|\{j : K - o_j < z_i\}\right|\right\}$ ▷ discard
  for $(i,j) \in A$ **do**
    if $(i, j \in C \cup D) \vee (1/2 \notin [\ell_{i,j}, u_{i,j}])$ **then**
      $A = A \setminus (i,j)$ ▷ Do not update $\hat{s}_{i,j}$ any more
    end if
  end for
  $n = n + 1$
end while
return the top-K' options for which the most $\bar{s}_{i,j}$ above $1/2$

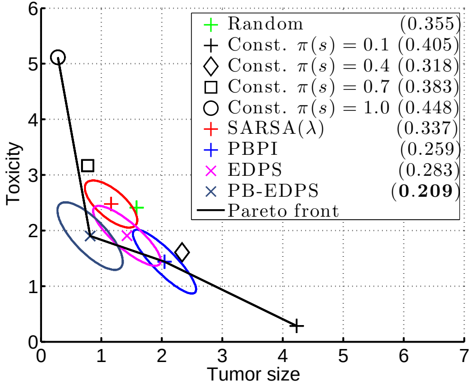
Cancer treatment

1. State space consists of toxicity level and tumor size (X, Y)
2. Linear policy space
3. Each policy search method were trained 100 times and each policy were evaluated on 300 virtual patients
4. 6-months treatment
5. Transitions: $X_{t+1} = X_t + \Delta X_t$ and $Y_{t+1} = Y_t + \Delta Y_t$

$$\Delta Y_t = [a_1 \cdot \max(X_t, X_0) - b_1 \cdot (D_t - d_1)] \times \mathbb{I}\{Y_t > 0\}$$
$$\Delta X_t = a_2 \cdot \max(Y_t, Y_0) - b_2 \cdot (D_t - d_2)$$

6. Probability of death: $1 - \exp(-\exp(c_0 + c_1 X_t + c_2 Y_t))$

Cancer treatment



(b) 100 repetitions of training process