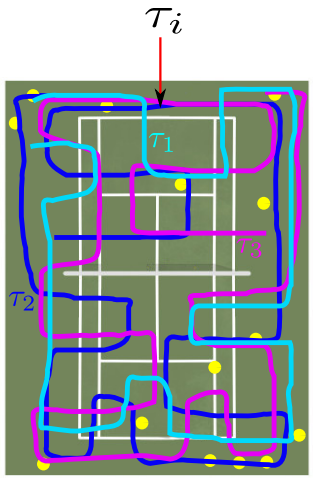
Direct Policy Search vs Reinforcement Learning

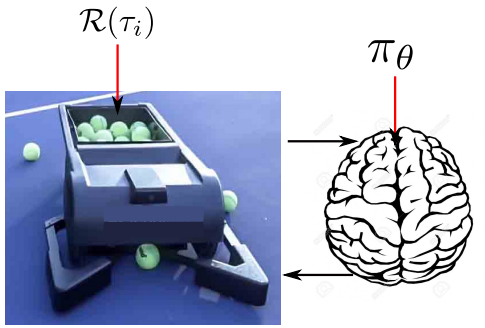
Introduction

Olivier Sigaud

Sorbonne Université
http://people.isir.upmc.fr/sigaud

Outline





- Follow-up of the lecture about policy search methods
- Distinction between two approaches: policy gradient, direct policy search
- Lessons about the latter, comparisons between both, and combinations of both

A topic that matters to me

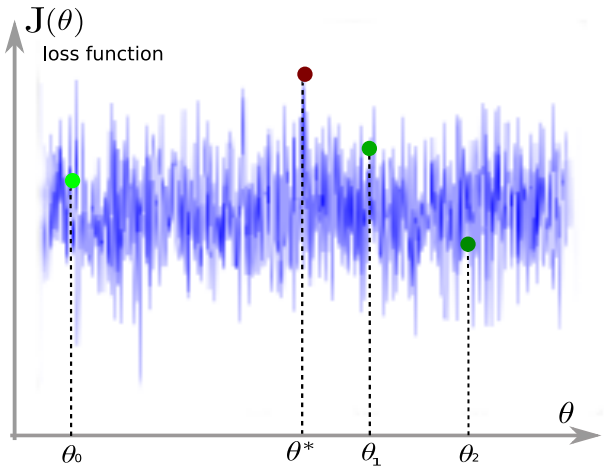

Stulp, F. and Sigaud, O. (2012) Path integral policy improvement with covariance matrix adaptation. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1–8, Edinburgh, Scotland


Stulp, F. and Sigaud, O. (2013) Robot skill learning: From reinforcement learning to evolution strategies. *Paladyn Journal of Behavioral Robotics*, 4(1):49–61


Sigaud, O. and Stulp, F. (2019) Policy search in continuous action domains: an overview. *Neural Networks*, 113:28–40


Sigaud, O. (2022) Combining evolution and deep reinforcement learning for policy search: a survey. *arXiv preprint arXiv:2203.14009*

(Truly) Random Search



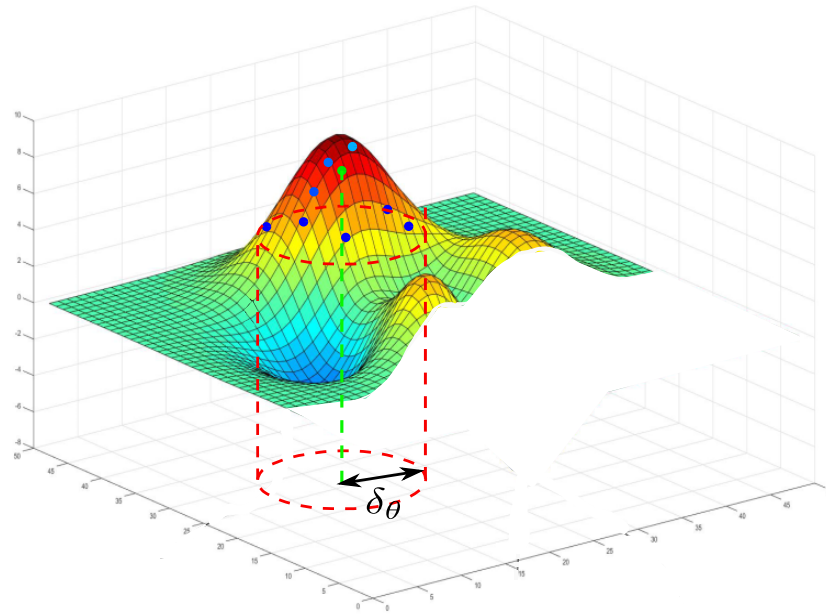
- Select $\boldsymbol{\theta}_i$ randomly
- Evaluate $J(\boldsymbol{\theta}_i)$
- If $J(\boldsymbol{\theta}_i)$ is the best so far, keep $\boldsymbol{\theta}_i$
- Loop until $J(\boldsymbol{\theta}_i) > target$ (maximize reward)

- Of course, this is not efficient if the space of $\boldsymbol{\theta}$ is large
- General "blind" algorithm, no assumption on $J(\boldsymbol{\theta})$
- We can do better if $J(\boldsymbol{\theta})$ shows some local regularity

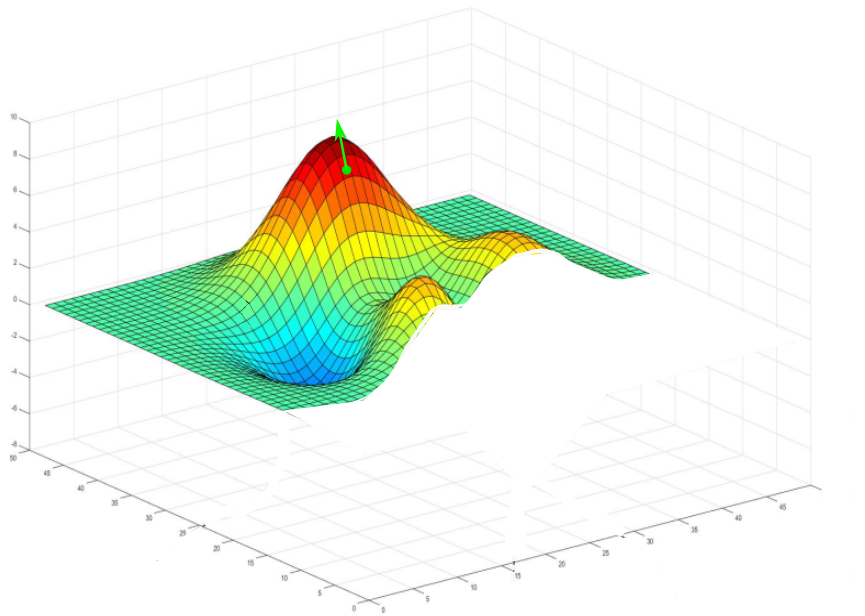
ISIR
INSTITUT DES SYSTÈMES INTELLIGENTS ET DE ROBOTIQUE

Two approaches to function optimization

- Assumption: The function is locally smooth, unknown good solutions are close to known good solutions



Variation - selection

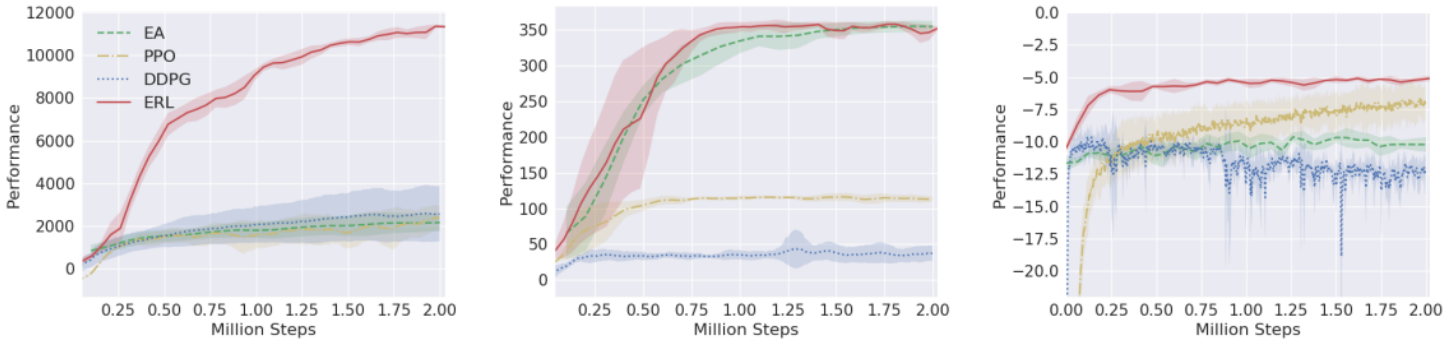


Gradient ascent

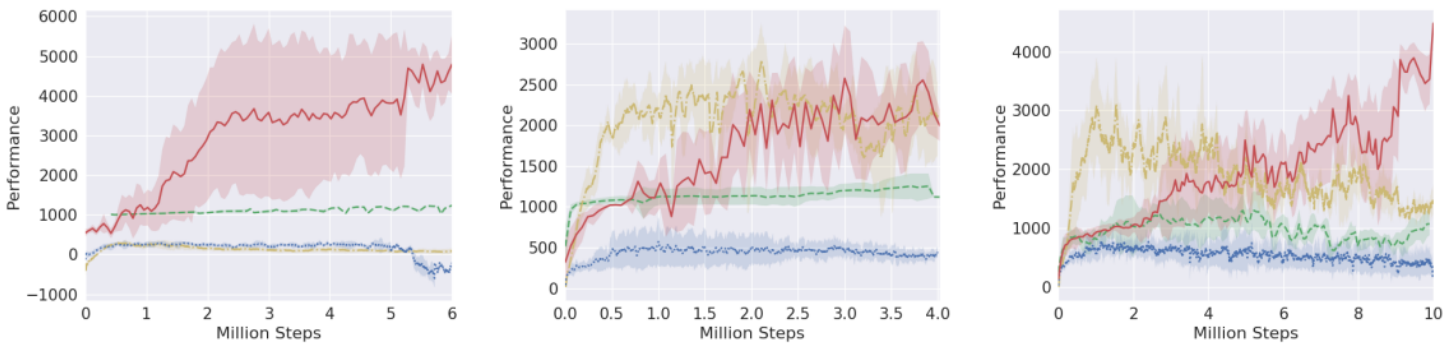
- DPS: Variation - selection: Performing variations and evaluating them
- PG: Gradient ascent: Following the gradient from analytical knowledge

Sigaud, O. & Stulp, F. (2019) Policy search in continuous action domains: an overview. Neural Networks

Why is the comparison relevant?



(a) HalfCheetah (b) Swimmer (c) Reacher

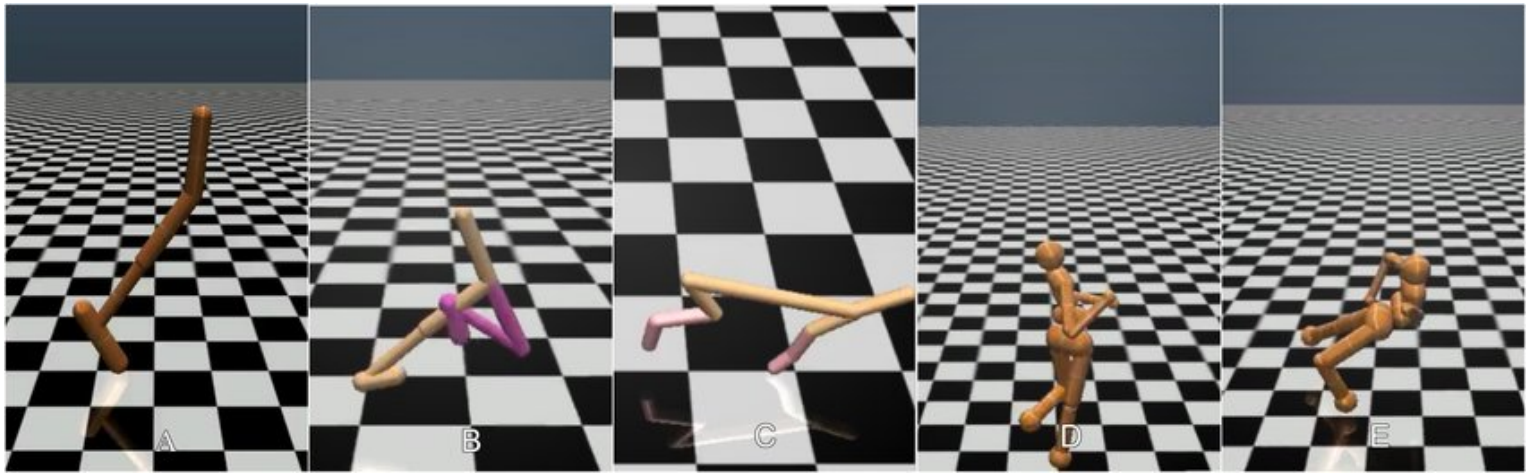


(d) Ant (e) Hopper (f) Walker2D

- A quick look at several papers shows that it is by no way obvious that PPO outperforms ESs

Shauharda Khadka and Kagan Tumer. Evolution-guided policy gradient in reinforcement learning. In *Neural Information Processing Systems*, 2018

Growing recognition in ML conferences



- Deep neuroevolution is competitive with deep RL in challenging benchmarks
- In wall clock time, but not in samples
- Opportunity for a deeper understanding of the inner mechanisms

Tim Salimans, Jonathan Ho, Xi Chen, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017

Mania, H., Guy, A., and Recht, B. (2018) Simple random search of static linear policies is competitive for reinforcement learning. *Advances in Neural Information Processing Systems*, 31

Topic of the lecture

- Deep RL methods seem to be far more sample efficient: Why is this so?
- Potential explanations:
  - The gradient gives the direction of steepest ascent: it improves faster (no!)
  - Gradient ascent does not need sampling. It uses analytical knowledge of the function under optimization to improve it (no!)
  - RL searches a better space: it uses information at each state action pair, versus the whole episode for direct policy search methods (env. dependent)
  - RL methods can reuse more samples than variation-selection methods yes, if off-policy!
- Approach: investigate each potential reason to see whether it holds in practice
- Then move to combinations

Outline of next lessons

- Comparing direct policy search and RL:
  1. Quick presentation of direct policy search methods
  2. More details on policy gradient updates
  3. Comparing improvement steps
  4. Comparing optimization spaces, role of replay
- Combining direct policy search and RL:
  1. To better optimize policies
  2. To better optimize actions
  3. To ensure more diversity
  4. To optimize something else

Any question?

Send mail to: Olivier.Sigaud@isir.upmc.fr

Khadka, S. and Tumer, K. (2018).

Evolution-guided policy gradient in reinforcement learning.
In *Neural Information Processing Systems*.

Mania, H., Guy, A., and Recht, B. (2018).

Simple random search of static linear policies is competitive for reinforcement learning.
Advances in Neural Information Processing Systems, 31.

Salimans, T., Ho, J., Chen, X., and Sutskever, I. (2017).

Evolution strategies as a scalable alternative to reinforcement learning.
arXiv preprint arXiv:1703.03864.

Sigaud, O. (2022).

Combining evolution and deep reinforcement learning for policy search: a survey.
ACM Transactions in Evolutionary Learning and Optimization.

Sigaud, O. and Stulp, F. (2019).

Policy search in continuous action domains: an overview.
Neural Networks, 113:28–40.

Stulp, F. and Sigaud, O. (2012).

Path integral policy improvement with covariance matrix adaptation.
In *Proceedings of the 29th International Conference on Machine Learning*, pages 1–8, Edinburgh, Scotland.

Stulp, F. and Sigaud, O. (2013).

Robot skill learning: From reinforcement learning to evolution strategies.
Paladyn Journal of Behavioral Robotics, 4(1):49–61.