Direct Policy Search vs Reinforcement Learning

Comparing optimization spaces and sample reuse

Olivier Sigaud

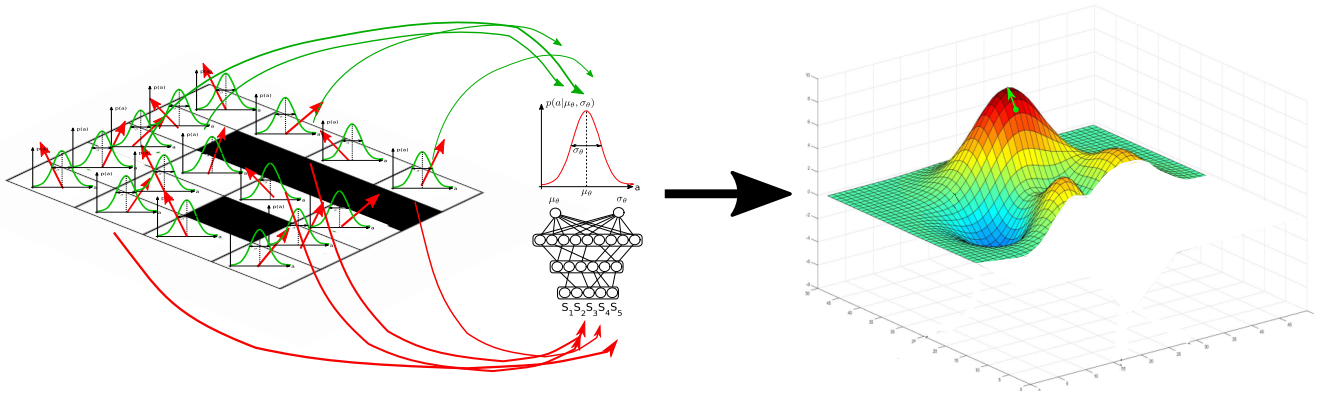
Sorbonne Université
http://people.isir.upmc.fr/sigaud

Reminder

- Deep RL methods seem to be far more sample efficient: Why is this so?
- Potential explanations:
  - ... see previous lesson...
  - 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 direct policy search methods yes, if off-policy!
- Approach: investigate each potential reason to see whether it holds in practice

Does deep RL search in a better space?

Where does search take place?



- Search comes from the exploration part
- In direct policy search, exploration acts directly in the $\boldsymbol{\theta}$ space
- The $s \times a$ space is generally much smaller than the $\boldsymbol{\theta}$ space
- The policy gradient defines a change in action probabilities in the $s \times a$ space
- Then this change is implemented (without search) in the $\boldsymbol{\theta}$ space
- Searching the $s \times a$ space might be faster than searching the $\boldsymbol{\theta}$ space
- This may be an advantage, in small enough $s \times a$ spaces

Policy perturbation

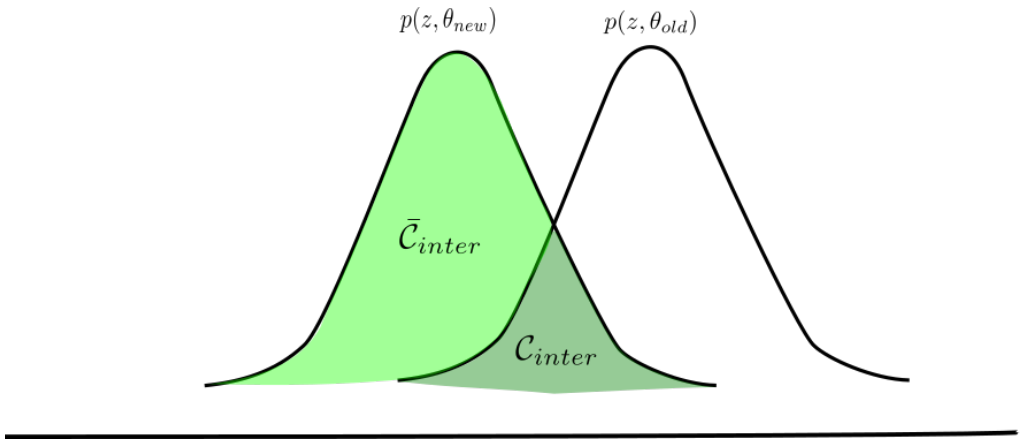
- Most deep RL algos explore in the $s \times a$ space: adding noise to actions
- But a few add noise to policy parameters
- Policy parameter noise is often better than action noise in deep RL
- Using policy parameter noise is closer to direct policy search (acts over $\boldsymbol{\theta}$)
- Better performance does not imply higher sample efficiency
- Need to better investigate the difference

Meire Fortunato, Mohammad Gheshlaghi Azar, Bilal Piot, Jacob Menick, Ian Osband, Alex Graves, Vlad Mnih, Remi Munos, Demis Hassabis, Olivier Pietquin, et al. Noisy networks for exploration. *arXiv preprint arXiv:1706.10295*, 2017

Matthias Plappert, Rein Houthooft, Prafulla Dhariwal, Szymon Sidor, Richard Y. Chen, Xi Chen, Tamim Asfour, Pieter Abbeel, and Marcin Andrychowicz. Parameter space noise for exploration. *arXiv preprint arXiv:1706.01905*, 2017

Can RL reuse samples better than direct policy search?

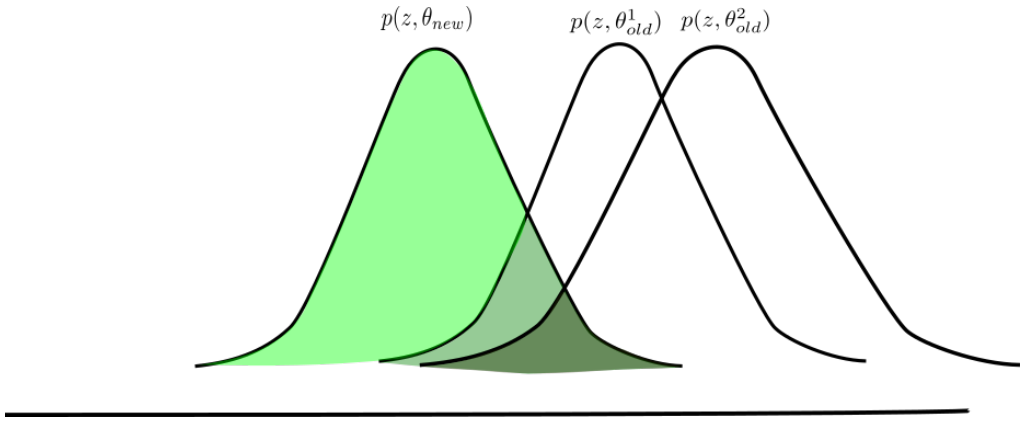
Sample reuse in ES: Importance Mixing



- A sample is a $< \boldsymbol{\theta}, J(\boldsymbol{\theta}) >$ pair
- No idea of recombining several "pieces of trajectories"
- To build a new generation, we can reuse samples from the previous one
- Importance Mixing does so: up to 90% gain in sample efficiency

Sun, Y., Wierstra, D., Schaul, T., & Schmidhuber, J. (2009) Efficient natural evolution strategies. In *Proceedings of the 11th Annual conference on Genetic and evolutionary computation* (pp. 539–546).: ACM.

Extended Importance Mixing



- Importance Mixing can be further improved over several generations
- The gain is marginal

Pourchot, A., Perrin, N., & Sigaud, O. (2018) Importance mixing: Improving sample reuse in evolutionary policy search methods. *arXiv preprint arXiv:1808.05832*

Reminder: being off-policy and using a replay buffer

- Being "off-policy" is a matter of degree
- Policies should not evolve too quickly (trust region, ...)
- Off-policy RL algorithms can use a replay buffer
- They keep data from olders (or expert) policies and train (again) from this older data
- Drastically improves sample efficiency
- But makes the algorithm more unstable
- Off-policy methods are more sample efficient than direct policy search methods, but they are more unstable

Summary

- Direct policy search:
  - robust, derivative-free, better over long horizon
  - poor sample reuse, low sample efficiency, need for evaluating many potentially poor policies
- Policy Gradient:
  - Uses analytical derivative of the policy function
  - Uses information from each step, not just trajectories
  - When off-policy, the replay buffer improves sample efficiency

Any question?

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

ISIR
INSTITUT
DES SYSTÈMES
INTELLIGENTS
ET DE ROBOTIQUE

Fortunato, M., Azar, M. G., Piot, B., Menick, J., Osband, I., Graves, A., Mnih, V., Munos, R., Hassabis, D., Pietquin, O., et al. (2017).
Noisy networks for exploration.
arXiv preprint arXiv:1706.10295.

Plappert, M., Houthooft, R., Dhariwal, P., Sidor, S., Chen, R. Y., Chen, X., Asfour, T., Abbeel, P., and Andrychowicz, M. (2017).
Parameter space noise for exploration.
arXiv preprint arXiv:1706.01905.

Pourchot, A., Perrin, N., and Sigaud, O. (2018).
Importance mixing: Improving sample reuse in evolutionary policy search methods.
arXiv preprint arXiv:1808.05832.

Sun, Y., Wierstra, D., Schaul, T., and Schmidhuber, J. (2009).
Efficient natural evolution strategies.
In *Proceedings of the 11th Annual conference on Genetic and evolutionary computation*, pages 539–546. ACM.