Direct Policy Search vs Reinforcement Learning

Does policy gradient perform better steps?

Olivier Sigaud

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

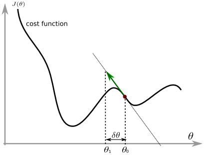
Reminder

- Deep RL methods seem to be far more sample efficient: Why is this so?
- Potential explanations:
  - 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!)
  - … more in the next lesson …
- Approach: investigate each potential reason to see whether it holds in practice

INSTITUT DES SYSTÈMES INTELLIGENTS ET DE ROBOTIQUE

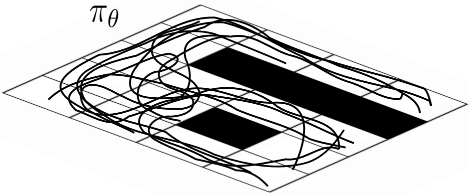
Strengths of the policy gradient approach

- In principle, the gradient ascent approach is superior for two reasons:
  - Standard gradient computation requires no sampling
  - The gradient provides the best direction of improvement



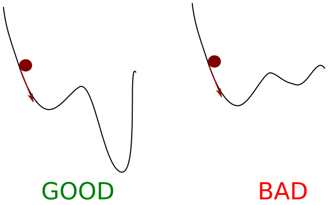
  - The optimum of $\mathbf{J}(\boldsymbol{\theta}+\delta\boldsymbol{\theta})$ over $\delta\boldsymbol{\theta}$ is reached when $\dfrac{\partial \mathbf{J}(\boldsymbol{\theta}+\delta\boldsymbol{\theta})}{\partial \delta\boldsymbol{\theta}} = 0$
  - First order approx: $\mathbf{J}(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}\mathbf{J}(\boldsymbol{\theta})^T \delta\boldsymbol{\theta} + \nu \delta\boldsymbol{\theta}^T \delta\boldsymbol{\theta}$ + higher order terms
  - $\dfrac{\partial \mathbf{J}(\boldsymbol{\theta}+\delta\boldsymbol{\theta})}{\partial \delta\boldsymbol{\theta}} \sim \nabla_{\boldsymbol{\theta}}\mathbf{J}(\boldsymbol{\theta})^T \delta\boldsymbol{\theta} + 2\nu\delta\boldsymbol{\theta} \rightarrow \delta\boldsymbol{\theta}^* = -\alpha \nabla_{\boldsymbol{\theta}}\mathbf{J}(\boldsymbol{\theta})$
- But in practice, we get an approximated gradient from sampled trajectories, so this is not true

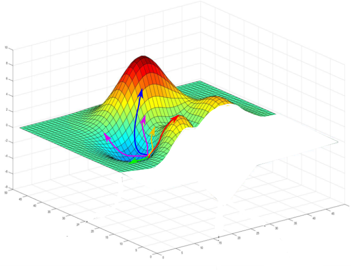
No sampling in policy gradient?



- Computing a gradient is an analytical derivation (no sampling)
- But in policy search, the cost function $J(\boldsymbol{\theta})$ is known through sampling
- $$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{m} \sum_{i=1}^{m} \sum_{t=1}^{H} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_t^{(i)} | s_t^{(i)}) R(\tau^{(i)}) \quad (1)$$
- But to apply (1), we need m trajectories

Advanced and adaptive gradient ascent





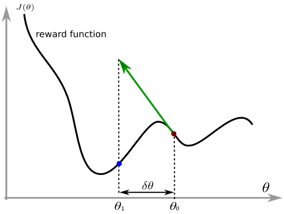
- Deep RL libraries and algorithms do more than plain gradient ascent
- Do advanced gradient ascent techniques improve sample efficiency?
- Adaptive gradient ascent methods: Adam, RMSProp, Momentum, Nesterov...
- Advanced gradient ascent methods: natural gradient, Gauss-Newton, ...
- General message: no free lunch, some technique improves in some context
- In practice, Adam often wins, but the best optimizer is problem-dependent

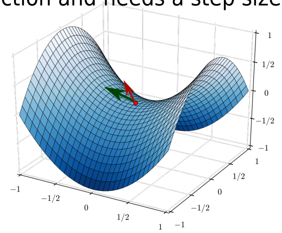
Ruder, S. (2016) An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*

Pierrot, T., Perrin, N., & Sigaud, O. (2018) First-order and second-order variants of the gradient descent: a unified framework. *arXiv preprint arXiv:1810.08102*

Limits of the PG approach

A gradient provides a local direction and needs a step size



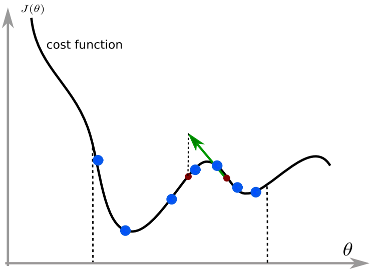


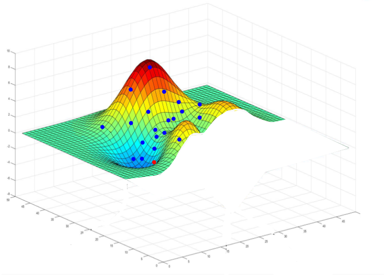
If the step size is too large, may result in loss

If the direction is slightly wrong, may result in loss

- Gradient descent techniques are blind, improvement depends on landscape
- The gradient step only ensures very local improvement (need for a trust region)
- The policy gradient is inaccurate: variance from sampling, sum over local gradients, ...
- Wrong direction of improvement, inaccurate step size tuning

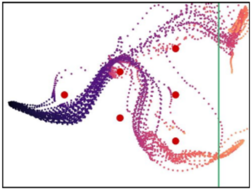
Advantages of direct policy search





- The generated solutions are evaluated before selection!
- They do not need a trust region
- Thus they may perform large jumps in policy space (e.g. in flat landscapes)
- They investigate more potential solutions at each step

Lower variance in direct policy search?



- From Salimans: in RL, variance accumulates along each individual action
- So, the longer the trajectories, the more variance (a lower γ helps)
- In direct policy search, variance does not grow with the length of trajectories, because a trajectory is a sample
- So direct policy search is more advantageous in problems with longer trajectories

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

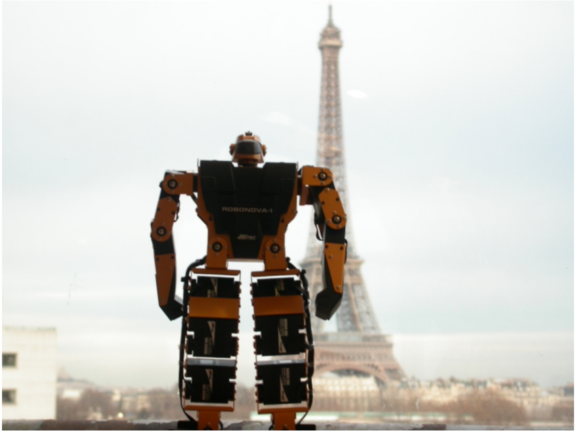
Robustness

- According to Lehman, ES optimize an expectation of reward over the population, thus they are more robust
- Direct policy search methods do not depend on Markov property, they are more general (POMDPs, etc.)
- They do not need the policy function to be differentiable: they are derivative-free
- But derivative-free approaches need $\mathcal{O}(d)$ more iterations than derivative-based approaches (Nesterov)
- The gradient step leverages:
  - Knowledge of the policy structure
  - Working on separate state-action pairs, not just trajectories (next lesson)

Lehman, J., Chen, J., Clune, J., and Stanley, K. O. ES is more than just a traditional finite-difference approximator. *arXiv preprint arXiv:1712.06568*, 2017

Nesterov, Y. and Spokoiny, V. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017

Any question?



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

Lehman, J., Chen, J., Clune, J., and Stanley, K. O. (2017).
ES is more than just a traditional finite-difference approximator.
arXiv preprint arXiv:1712.06568.

Nesterov, Y. and Spokoiny, V. (2017).
Random gradient-free minimization of convex functions.
Foundations of Computational Mathematics, 17(2):527–566.

Pierrot, T., Perrin, N., and Sigaud, O. (2018).
First-order and second-order variants of the gradient descent: a unified framework.
arXiv preprint arXiv:1810.08102.

Ruder, S. (2016).
An overview of gradient descent optimization algorithms.
arXiv preprint arXiv:1609.04747.

Salimans, T., Ho, J., Chen, X., and Sutskever, I. (2017).
Evolution strategies as a scalable alternative to reinforcement learning.
arXiv preprint arXiv:1703.03864.