From Policy Gradient to Actor-Critic methods

TRPO and ACKTR

Olivier Sigaud

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

Outline

- Two algorithms are presented: TRPO and ACKTR
- Two aspects distinguish TRPO:
  - Surrogate return objective
  - Natural policy gradient
- A small difference with ACKTR:
  - Using Kronecker Factored Approximated Curvature to estimate the natural gradient

Surrogate return objective

- The standard policy gradient algorithm for stochastic policies is:

$$\nabla_\theta J(\theta) = \mathbb{E}_t[\nabla_\theta \text{log} \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\hat{A}_\phi]$$

- This gradient is obtained from differentiating $Loss^{PG}(\theta) = \mathbb{E}_t[\text{log}\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\hat{A}_\phi]$
- But we obtain the same gradient from differentiating

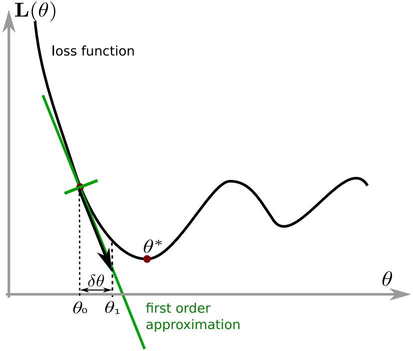
$$Loss^{IS}(\theta) = \mathbb{E}_t[\frac{\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)}{\pi_{\theta_{old}}(\mathbf{a}_t|\mathbf{s}_t)}\hat{A}_\phi]$$

where $\pi_{\theta_{old}}$ is the policy at the previous iteration

- Because $\nabla_\theta \text{log} f(\theta)|_{\theta_{old}} = \frac{\nabla_\theta f(\theta)|_{\theta_{old}}}{f(\theta_{old})} = \nabla_\theta(\frac{f(\theta)}{f(\theta_{old})})|_{\theta_{old}}$

- Another view based on importance sampling
- See John Schulmann's Deep RL bootcamp lecture #5
https://www.youtube.com/watch?v=SQtOI9jsrJ0 (8')

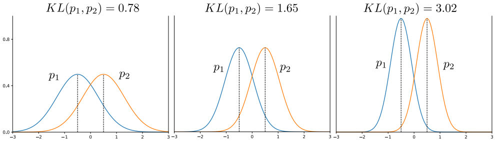
Trust region



- The gradient of a function is only accurate close to that function
- The gradient of the surrogate objective is only accurate close to the current policy π_θ
- Thus, when updated, the new policy must not move too far away from a "trust region" around the current policy

Kakade, S. & Langford, J. (2002) Approximately optimal approximate reinforcement learning. In *ICML*, volume 2, pages 267–274

Natural Policy Gradient



$KL(p_1, p_2) = 0.78$ $KL(p_1, p_2) = 1.65$ $KL(p_1, p_2) = 3.02$

- One way to constrain two stochastic policies to stay close is constraining their KL divergence
- The KL divergence is smaller when the variance is larger
- Under fixed KL constraint, it is easier to move the mean further away when the variance is large
- Thus the mean policy converges first, then the variance is reduced
- Ensures a large enough amount of exploration noise
- Other properties presented in the Pierrot et al. (2018) paper

Sham M. Kakade. A natural policy gradient. In *Advances in neural information processing systems*, pp. 1531–1538, 2002

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*

Trust Region Policy Optimization

- Theory: monotonous improvement towards the optimal policy (Assumptions do not hold in practice)
- To ensure small steps, TRPO uses a natural gradient update instead of standard gradient
- Minimize Kullback-Leibler divergence to previous policy
- $$\max_{\theta} \mathbb{E}_t[\frac{\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)}{\pi_{\theta_{old}}(\mathbf{a}_t|\mathbf{s}_t)} A_{\pi_{\theta_{old}}}(\mathbf{s}_t, \mathbf{a}_t)]$$
  $$\text{subject to } \mathbb{E}_t[KL(\pi_{\theta_{old}}(.|\mathbf{s})||\pi_\theta(\mathbf{a}_t|\mathbf{s}_t))] \leq \delta$$
- In TRPO, optimization performed using a conjugate gradient method to avoid approximating the Fisher Information matrix

Schulman, J., Levine, S., Moritz, P., Jordan, M. I., & Abbeel, P. (2015) Trust Region Policy Optimization. *CoRR, abs/1502.05477*

Advantage estimation

- To get $\hat{A}_\phi$, an empirical estimate of $V^\pi(s)$ is needed
- TRPO uses a MC estimate approach through regression, but constrains it (as for the policy):

$$\min_\phi \sum_{n=0}^{N} ||V_\phi(s_n) - V(s_n)||^2$$

$$\text{subject to } \frac{1}{N}\sum_{n=0}^{N} \frac{||V_\phi(s_n) - V_{\phi_{old}}(s_n)||^2}{2\sigma^2} \leq \epsilon$$

- Equivalent to a mean KL divergence constraint between V_ϕ and $V_{\phi_{old}}$

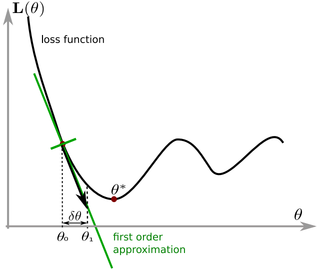
Properties

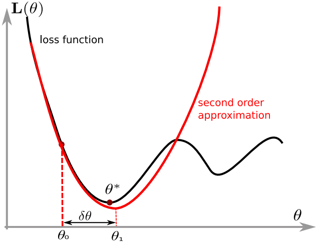
- Moves slowly away from current policy
- Key: use of line search to deal with the gradient step size
- More stable than DDPG, performs well in practice, but less sample efficient
- Conjugate gradient approach not provided in standard tensor gradient librairies, thus not much used
- Greater impact of PPO
- Related work: NAC, REPS

Jan Peters and Stefan Schaal. Natural actor-critic. *Neurocomputing*, 71 (7-9):1180–1190, 2008

Jan Peters, Katharina Mülling, and Yasemin Altun. Relative entropy policy search. In *AAAI*, pp. 1607–1612. Atlanta, 2010

First order versus second order derivative





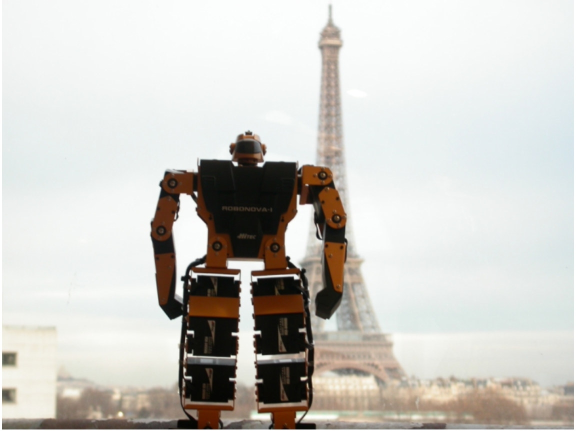
- In first order methods, need to define a step size
- Second order methods provide a more accurate approximation
- They also provide a true minimum, when the Hessian matrix is symmetric positive-definite matrix (SPD)
- In both cases, the derivative is very local
- The trust region constraint applies too

ACKTR

- K-FAC: Kronecker Factored Approximated Curvature: efficient estimate of natural gradient
- Using block diagonal estimations of the Hessian matrix, to do better than first order
- ACKTR: TRPO with K-FAC natural gradient calculation
- But closer to actor-critic updates (see PPO)
- The per-update cost of ACKTR is only 10% to 25% higher than SGD
- Improves sample efficiency
- Not much excitement: less robust gradient approximation?
- Next lesson: PPO

Yuhuai Wu, Elman Mansimov, Shun Liao, Roger Grosse, and Jimmy Ba (2017) Scalable trust-region method for deep reinforcement learning using Kronecker-factored approximation. *arXiv preprint arXiv:1708.05144*

Any question?



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

Sham Kakade and John Langford.
Approximately optimal approximate reinforcement learning.
In *ICML*, volume 2, pp. 267–274, 2002.

Sham M. Kakade.
A natural policy gradient.
In *Advances in neural information processing systems*, pp. 1531–1538, 2002.

Jan Peters and Stefan Schaal.
Natural actor-critic.
Neurocomputing, 71(7-9):1180–1190, 2008.

Jan Peters, Katharina Mülling, and Yasemin Altun.
Relative entropy policy search.
In *AAAI*, pp. 1607–1612. Atlanta, 2010.

Thomas Pierrot, Nicolas Perrin, and Olivier Sigaud.
First-order and second-order variants of the gradient descent: a unified framework.
arXiv preprint arXiv:1810.08102, 2018.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel.
Trust region policy optimization.
CoRR, abs/1502.05477, 2015.

Yuhuai Wu, Elman Mansimov, Shun Liao, Roger Grosse, and Jimmy Ba.
Scalable trust-region method for deep reinforcement learning using Kronecker-factored approximation.
arXiv preprint arXiv:1708.05144, 2017.