From Policy Gradient to Actor-Critic methods

TRPO and ACKTR

Olivier Sigaud

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

- Start from algorithms close to PG: TRPO and ACKTR
- Three aspects distinguish TRPO:
  - Surrogate return objective
  - Natural policy gradient
  - Conjugate gradient approach
- Differences in ACKTR:
  - Approximate second order gradient descent (Hessian)
  - Using Kronecker Factored Approximated Curvature
Surrogate return objective

- The standard policy gradient algorithm for stochastic policies is:
  \[ \nabla_{\theta} J(\theta) = \mathbb{E}_t[\nabla_{\theta} \log \pi_\theta(a_t|s_t) \hat{A}^{\pi_\theta}_{\phi}] \]

- This gradient is obtained from differentiating
  \[ \text{Loss}^{PG}(\theta) = \mathbb{E}_t[\log \pi_\theta(a_t|s_t) \hat{A}^{\pi_\theta}_{\phi}] \]

- But we obtain the same gradient from differentiating
  \[ \text{Loss}^{IS}(\theta) = \mathbb{E}_t[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta old}(a_t|s_t)} \hat{A}^{\pi_\theta}_{\phi}] \]

  where \( \pi_{\theta old} \) is the policy at the previous iteration

- Because
  \[ \nabla_{\theta} \log f(\theta)|_{\theta old} = \frac{\nabla_{\theta} f(\theta)|_{\theta old}}{f(\theta old)} = \nabla_{\theta} \left( \frac{f(\theta)}{f(\theta old)} \right)|_{\theta old} \]

- Another view based on importance sampling

- See John Schulmann’s Deep RL bootcamp lecture #5
  https://www.youtube.com/watch?v=SQt019jsrJ0 (8’
The gradient of a function is only accurate close to the point where it is calculated.

\[ \nabla_{\theta} J(\theta) \] is only accurate close to the current policy \( \pi_{\theta} \).

Thus, when updating, \( \pi_{\theta} \) must not move too far away from a “trust region” around \( \pi_{\theta_{old}} \).

Natural Policy Gradient

- 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.


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

\[
\max_{\theta} \mathbb{E}_t \left[ \frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta old}(a_t|s_t)} A_{\phi}^{\pi_{\theta old}}(s_t, a_t) \right]
\]

subject to \(\mathbb{E}_t [KL(\pi_{\theta old}(\cdot|s)||\pi_{\theta}(a_t|s_t))] \leq \delta\)

- In TRPO, optimization performed using a conjugate gradient method to avoid approximating the Fisher Information matrix

Advantage estimation

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

$$\min_{\phi} \sum_{n=0}^{N} \left\| V^\pi_\phi(s_n) - V^\pi_\theta(s_n) \right\|^2$$

subject to $\frac{1}{N} \sum_{n=0}^{N} \frac{\left\| V^\pi_\phi(s_n) - V^\pi_\theta(s_n) \right\|^2}{2\sigma^2} \leq \epsilon$

- Equivalent to a mean KL divergence constraint between $V^\pi_\phi$ and $V^\pi_\theta_{old}$.
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 libraries, thus not much used
- Greater impact of PPO
- Related work: NAC, REPS


From Policy Gradient to Actor-Critic methods

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 (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 the 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

Any question?

Send mail to: Olivier.Sigaud@upmc.fr
Sham Kakade and John Langford.
Approximately optimal approximate reinforcement learning.

Sham M. Kakade.
A natural policy gradient.

Jan Peters and Stefan Schaal.
Natural actor-critic.

Jan Peters, Katharina Mülling, and Yasemin Altun.
Relative entropy policy search.

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

Trust region policy optimization.

Scalable trust-region method for deep reinforcement learning using Kronecker-factored approximation.