TRPO and ACKTR

Olivier Sigaud

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

TRPO

Outline

- More PG with baselines: 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
- Then PPO (a quick overview of two versions)

Surrogate return objective

- The standard policy gradient algorithm for stochastic policies is:

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

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

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

  where $\pi_{\boldsymbol{\theta}old}$ is the policy at the previous iteration
- Because $\nabla_{\boldsymbol{\theta}} \log f(\boldsymbol{\theta})|_{\boldsymbol{\theta}old} = \frac{\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})|_{\boldsymbol{\theta}old}}{f(\boldsymbol{\theta}old)} = \nabla_{\boldsymbol{\theta}}(\frac{f(\boldsymbol{\theta})}{f(\boldsymbol{\theta}old)})|_{\boldsymbol{\theta}old}$

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

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

The policy gradient is on-policy

- The policy gradient calculation assumes that the training trajectories are obtained from the policy we are optimizing:
- Reminder: we want to find $\text{argmax}_{\boldsymbol{\theta}} \sum_{\tau} P(\tau, \boldsymbol{\theta}) \psi(\tau)$
- We use

$$P(\tau^{(i)}, \boldsymbol{\theta}_{samp}) = \prod_{t=1}^{H} p(s_{t+1}^{(i)} | s_t^{(i)}, a_t^{(i)}) . \pi_{\boldsymbol{\theta}_{samp}}(a_t^{(i)} | s_t^{(i)})$$

- Here, by definition, $\pi_{\boldsymbol{\theta}_{samp}}(a_t^{(i)} | s_t^{(i)})$ is the policy which generated the trajectories
- Then we take the gradient and get the policy gradient formula
- If we want to optimize another policy $\pi_{\boldsymbol{\theta}_{other}}(a_t^{(i)} | s_t^{(i)})$, the derivation is wrong

Importance sampling

- How can we estimate an expectation of a function over a distribution $\boldsymbol{\theta}_1$ if we know it from another distribution $\boldsymbol{\theta}_2$?

$$\begin{aligned}\mathbb{E}_{x\sim\boldsymbol{\theta}_1}[f(x)] &= P(x|\boldsymbol{\theta}_1)f(x) \\ &= \frac{P(x|\boldsymbol{\theta}_1)}{P(x|\boldsymbol{\theta}_2)}P(x|\boldsymbol{\theta}_2)f(x) \\ &= \frac{P(x|\boldsymbol{\theta}_1)}{P(x|\boldsymbol{\theta}_2)}\mathbb{E}_{x\sim\boldsymbol{\theta}_2}[f(x)]\end{aligned}$$

- $\frac{P(x|\boldsymbol{\theta}_1)}{P(x|\boldsymbol{\theta}_2)}$ is the importance sampling term
- In policy gradient methods, the distributions of interest are $\pi_{\boldsymbol{\theta}_{samp}}$ and $\pi_{\boldsymbol{\theta}_{other}}$.

Importance sampling: application to TRPO

- We sampled data from $\pi_{\boldsymbol{\theta}_{samp}}$
- We want to optimize another policy $\pi_{\boldsymbol{\theta}_{other}}$,
- We can rewrite

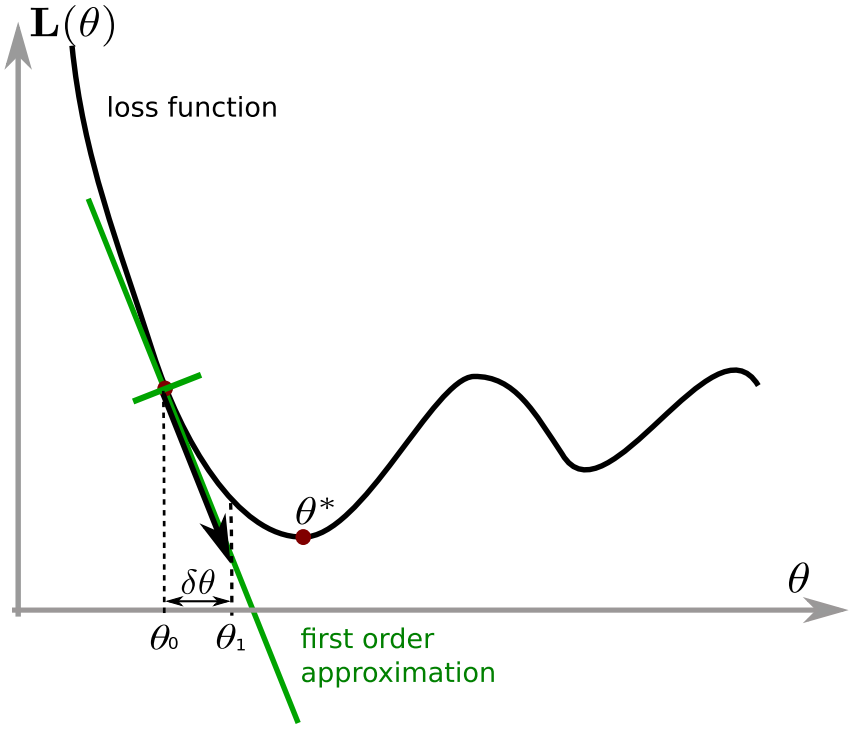
$$P(\tau^{(i)}, \boldsymbol{\theta}_{other}) = \prod_{t=1}^{H} p(s_{t+1}^{(i)}|s_t^{(i)}, a_t^{(i)}).\pi_{\boldsymbol{\theta}_{other}}(a_t^{(i)}|s_t^{(i)}).\frac{\pi_{\boldsymbol{\theta}_{samp}}(a_t^{(i)}|s_t^{(i)})}{\pi_{\boldsymbol{\theta}_{samp}}(a_t^{(i)}|s_t^{(i)})}$$

- Or

$$P(\tau^{(i)}, \boldsymbol{\theta}_{other}) = \prod_{t=1}^{H} p(s_{t+1}^{(i)}|s_t^{(i)}, a_t^{(i)}).\frac{\pi_{\boldsymbol{\theta}_{other}}(a_t^{(i)}|s_t^{(i)})}{\pi_{\boldsymbol{\theta}_{samp}}(a_t^{(i)}|s_t^{(i)})}.\pi_{\boldsymbol{\theta}_{samp}}(a_t^{(i)}|s_t^{(i)})$$

- The term $\frac{\pi_{\boldsymbol{\theta}_{other}}(a_t^{(i)}|s_t^{(i)})}{\pi_{\boldsymbol{\theta}_{samp}}(a_t^{(i)}|s_t^{(i)})}$ is the importance sampling term
- In TRPO, $\pi_{\boldsymbol{\theta}_{sample}} = \pi_{\boldsymbol{\theta}_{old}}$, $\pi_{\boldsymbol{\theta}_{other}} = \pi_{\boldsymbol{\theta}}$
- We apply the same derivation as for the policy gradient...
- We get $Loss^{IS}(\boldsymbol{\theta}) = \mathbb{E}_t[\frac{\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)}{\pi_{\boldsymbol{\theta}old}(\mathbf{a}_t|\mathbf{s}_t)}\hat{A}_{\boldsymbol{\phi}}^{\pi_{\boldsymbol{\theta}}}]$

Trust region



- The gradient of a function is only accurate close to the point where it is calculated
- $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ is only accurate close to the current policy $\pi_{\boldsymbol{\theta}}$
- Thus, when updating, $\pi_{\boldsymbol{\theta}}$ must not move too far away from a "trust region" around $\pi_{\boldsymbol{\theta}_{old}}$

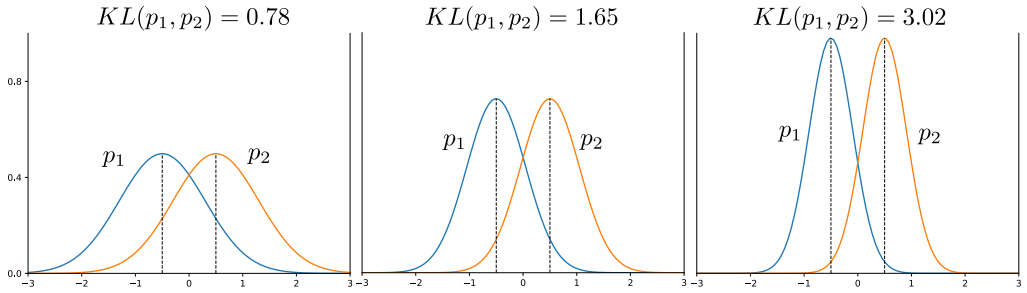
Kakade, S. & Langford, J. (2002) Approximately optimal approximate reinforcement learning. In *ICML*, volume 2, pages 267–274

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
- $$\max_{\boldsymbol{\theta}} \mathbb{E}_t\left[\frac{\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t)}{\pi_{\boldsymbol{\theta}old}(\mathbf{a}_t|\mathbf{s}_t)} A_{\boldsymbol{\phi}}^{\pi_{\boldsymbol{\theta}old}}(\mathbf{s}_t, \mathbf{a}_t)\right]$$
  $$\text{subject to } \mathbb{E}_t[KL(\pi_{\boldsymbol{\theta}old}(.|\mathbf{s})||\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t))] \leq \epsilon$$
- 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*

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

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*

Advantage estimation

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

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

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

- Equivalent to a mean KL divergence constraint between $V^{\pi_{\boldsymbol{\theta}}}_{\boldsymbol{\phi}}$ and $V^{\pi_{\boldsymbol{\theta}}}_{\boldsymbol{\phi}_{old}}$
- Very similar to target critic in DQN, DDPG... Can be implemented in the same way

Properties

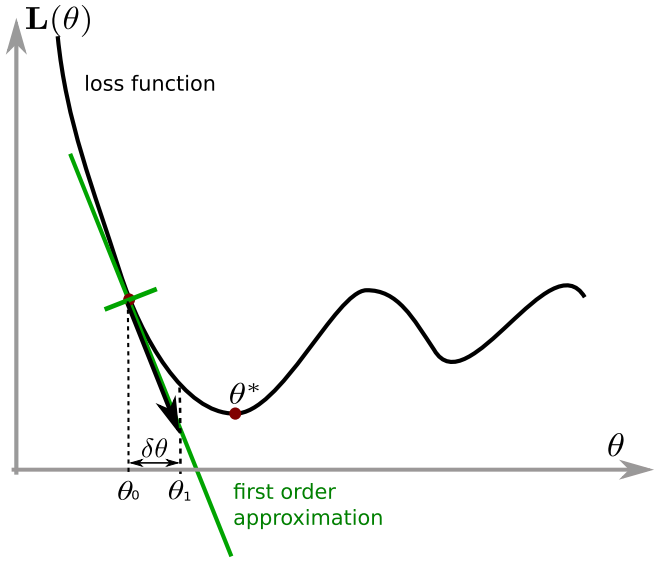
- Moves slowly away from current policy
- Key: use of line search to deal with the gradient step size
- More stable than DDPG, performs well in practice, but less sample efficient
- Conjugate gradient approach not provided in standard tensor gradient librairies, thus not much used
- Greater impact of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

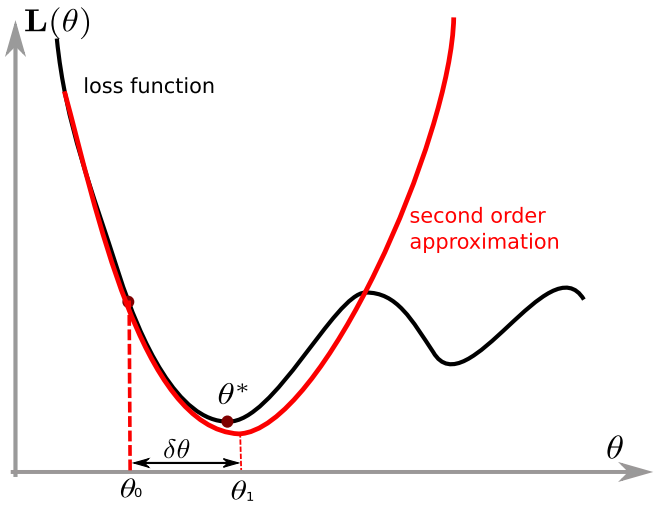

Jan Peters and Stefan Schaal. Natural actor-critic. *Neurocomputing*, 71 (7-9):1180–1190, 2008


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

ACKTR

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?

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

Kakade, S. and Langford, J. (2002).

Approximately optimal approximate reinforcement learning.
In *ICML*, volume 2, pages 267–274.

Kakade, S. M. (2001).

A natural policy gradient.
In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, pages 1531–1538. MIT Press.

Peters, J., Mülling, K., and Altun, Y. (2010).

Relative entropy policy search.
In *AAAI*, pages 1607–1612. Atlanta.

Peters, J. and Schaal, S. (2008).

Natural actor-critic.
Neurocomputing, 71(7-9):1180–1190.

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.

Schulman, J., Levine, S., Abbeel, P., Jordan, M. I., and Moritz, P. (2015).

Trust region policy optimization.
In Bach, F. R. and Blei, D. M., editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1889–1897. JMLR.org.

Wu, Y., Mansimov, E., Grosse, R. B., Liao, S., and Ba, J. (2017).

Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation.
In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5279–5288.