RL seminar #3: Policy gradient

Maksim Kretov

MIPT, Deep learning lab & iPavlov.ai

11 Nov 2017

Outline

Class information
- Assignments

Policy gradient
- Reducing variance
- Other

Value function methods
- Double Q-learning
- Generalized advantage estimation
- Other

Next steps

Table of Contents

Assignments

Coding (Programming assignment #2)

Deadline: 24 Nov 2017 (Thursday)

Quiz

Deadline: 10 Nov 2017

Questions

Few questions.

Table of Contents

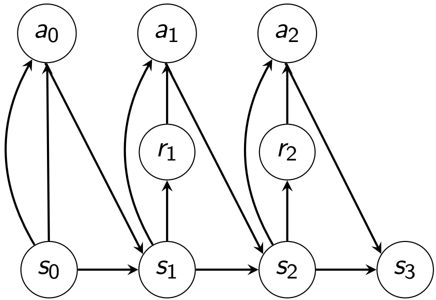
Policy Gradient: reducing variance

Question: How causality is used for variance reduction?

Let's define trajectory τ:

$\tau = \{s_0, a_0, s_1, a_1, r_1(s_0, a_0), ..s_{T-1}, a_{T-1}, r_{T-1}(s_{T-2}, a_{T-2}), s_T\}$

$d\tau = ds_0 da_0 ds_1 da_1 ... ds_{T-1} da_{T-1} ds_T$



Policy Gradient: reducing variance

Derivation of REINFORCE rule

Consider formula for expected reward:

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}[R(\tau)] = \int p(\tau|\theta) R(\tau) d\tau$$

$$p(\tau|\theta) = p(s_0) \prod_{t=0}^{T-1} \pi_\theta(a_t|s_t) p(s_{t+1}|s_t, a_t)$$

$$\nabla_\theta J(\theta) = \int R(\tau) \nabla_\theta p(\tau|\theta) d\tau = \mathbb{E}_\tau[R(\tau) \nabla_\theta \log p(\tau|\theta)]$$

$$\nabla_\theta \log p(\tau|\theta) = \sum_t \nabla_\theta \log \pi_\theta(a_t|s_t) \quad \text{and} \quad R(\tau) = \sum_{r_t \in \tau} r_t$$

Policy Gradient: reducing variance

REINFORCE rule

We derived the following equation:

$$\nabla_\theta J(\theta) = \mathbb{E}_\tau \Big[\sum_{t'=0}^{T-1} r_{t'} \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) \Big]$$

Let's consider one term in the sum over t':

$$\mathbb{E}_\tau \Big[r_{t'} \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) \Big] =$$

$$= \mathbb{E}_\tau \Big[r_{t'} \textstyle\sum_{t=0}^{t'-1} \nabla_\theta \log \pi_\theta(a_t|s_t) + r_{t'} \sum_{t=t'}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) \Big]$$

And now pay attention to the last term.

Policy Gradient: reducing variance

REINFORCE rule

Let's show that last term is zero, considering integrals over a_t:

$$\mathbb{E}_\tau\left[r_{t'}\sum_{t=t'}^{T-1}\nabla_\theta\log\pi_\theta(a_t|s_t)\right]=\sum_{t=t'}^{T-1}\mathbb{E}_\tau[r_{t'}\nabla_\theta\log\pi_\theta(a_t|s_t)]$$

Integration over s_T can be performed, because the only term depending on s_T is $p(s_T|a_{T-1},s_{T-1})$, this gives just a term 1. We can then integrate sequentially starting from a_{T-1}:

$$\mathbb{E}_{\tau\backslash a_t}[r_{t'}\int\pi_\theta(a_t|s_t)\nabla_\theta\log\pi_\theta(a_t|s_t)da_t]=$$

$$\sum_{t=t'}^{T-1}\mathbb{E}_{\tau\backslash a_t}[r_{t'}\nabla_\theta\int\pi_\theta(a_t|s_t)da_t]=0$$

Policy Gradient: reducing variance

REINFORCE rule

For the first term this is not true:

$$\mathbb{E}_\tau \left[r_{t'} \sum_{t=0}^{t'-1} \nabla_\theta \log \pi_\theta(a_t|s_t) \right] \neq 0$$

What is the difference?

We cannot integrate over $a_{t'-1}$, because there is term $r_{t'}$. We also cannot start 'from the beginning', integrating over a_0 at first, because there is a term $p(s_1|a_0, s_0)$ there (mind the $\mathbb{E}$ operation).

Let's just leave it.

Policy Gradient: reducing variance

REINFORCE rule

Almost there:

$$\nabla_\theta J(\theta) = \mathbb{E}_\tau\Big[\sum_{t'=0}^{T-1} r_{t'} \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t)\Big] =$$

$$= \mathbb{E}_\tau\Big[\sum_{t'=0}^{T-1} r_{t'} \sum_{t=0}^{t'-1} \nabla_\theta \log \pi_\theta(a_t|s_t)\Big] =$$

Rearrange terms in two sums and finally:

$$= \mathbb{E}_\tau\Big[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) \sum_{t'=t+1}^{T-1} r_{t'}\Big] = \mathbb{E}_\tau\Big[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) R_t\Big]$$

Baselines

Question: Action-dependent baselines

Two main points:

1. Bias-variance trade-off
2. Not baseline, but a control variate

Bias-variance trade-off

Let y be random variable (our estimator) and true value of unknown variable is y_t:

$$MSE = \mathbb{E}[(y - y_t)^2] = \mathbb{E}[((y - \mathbb{E}y) + (\mathbb{E}y - y_t))^2] =$$

$$= \mathbb{E}[(y - \mathbb{E}y)^2] + (\mathbb{E}y - y_t)^2 = Variance + Bias^2$$

Baselines

Bias-variance trade-off (continued)

Focusing only on unbiased estimators, we may find the one with higher error (optimization becomes constrained):

$$\min_{y} MSE(y) \leq \min_{y:\mathbb{E}y=y_t} MSE(y)$$

Example

Let y_o be solution to constrained min. task for unbiased estimator and construct biased estimator y_b: $y_b = (1-\alpha)y_o$. Then:

$$MSE(y_b) = (1-\alpha)^2 MSE(y_o) + \alpha^2 y$$

If we select small enough α, then error will be lower for biased y_b.

Baselines

Control variates

$$\mathbb{E}[f(x)] = \mathbb{E}[f(x) - \mu g(x)] + \mu \mathbb{E}[g(x)]$$
$$\mathbb{E}[f(x)] \to \frac{1}{k} \sum_i (f(x_i) - \mu g(x_i)) + \mu \mathbb{E}[g(x)]$$

Such estimator of $f(x)$ is still unbiased so we can select such function $g(x)$ that further minimize variance provided that we can calculate $\mathbb{E}[g(x)] \Rightarrow$ action-dependent baselines.

Actor-Critic

Actor-Critic: approximate returns with approximate baseline fitted for sample returns.
Policy Gradient: sample returns and baseline (constant or fitted).

Question: AC in comparison with usual PG

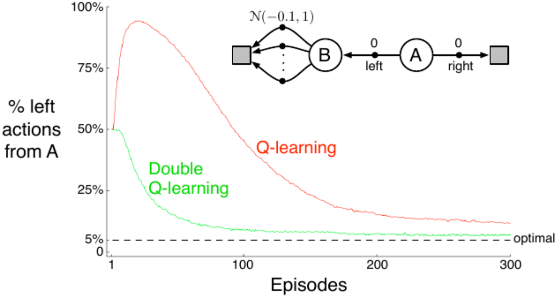
- '+': AC controls bias-variance (lower variance with critic).
- '+': AC provides additional control over training process (monitor values).
- '+': Don't need to sample till the end of episode to make training update.
- '-': More training parameters.
- '-': Need to think of architecture of critic.

Table of Contents

Double Q-learning

Question: Double Q-learning

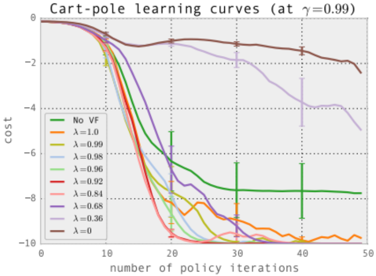
$$Q_1(s_t, a_t) \leftarrow Q_1(s_t, a_t) +$$
$$+\alpha(r_{t+1} + \gamma Q_2(s_{t+1}, argmax_a Q_1(S_{t+1}, a)) - Q_1(a_t, s_t)$$

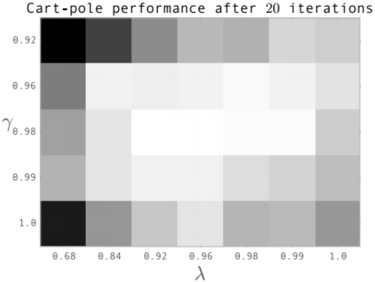


[0] Image source: https://goo.gl/SeJbSy (RL introduction book from R.Sutton and A.Barto)

Generalized advantage estimation

Question: GAE





[0] Image source: J. Schulman's PhD thesis `joschu.net/docs/thesis.pdf`

Other

Question: On-policy and off-policy methods.

On-policy algorithms

Require data to be collected under current policy.

Off-policy algorithms

Use data collected from any policy.

Table of Contents

Next steps

Plan for the week

- Quiz N3: 16 Nov (Thursday).
- Home assignment N2 issued (deadline 25 Nov).

Reading

Lectures 9-11 of CS294.
Please, post your questions about lectures in google doc:
`https://goo.gl/qN6jmJ`