

Trust Region Policy Optimization (TRPO)

이동민

삼성전자 서울대 공동연구소
Jul 18, 2019

Lecturer

- 이동민
 - kid33629@gmail.com
- 허수빈
 - 62subin.h@gmail.com



CORE
Control + Optimization Research Lab

Outline

- Trust Region Policy Optimization (TRPO)
 - Learning process
 - Hyperparameter
 - Main loop
 - Train model
 - Train & TensorboardX
 - Learning curve & Test
- TRPO vs. PPO - Learning curve
- Supplement - Proximal Policy Optimization (PPO)
 - Train model

Trust Region Policy Optimization (TRPO)

- TRPO Algorithm

Algorithm 1 Trust Region Policy Optimization

- 1: Input: initial policy parameters θ_0 , initial value function parameters ϕ_0
- 2: Hyperparameters: KL-divergence limit δ , backtracking coefficient α , maximum number of backtracking steps K
- 3: **for** $k = 0, 1, 2, \dots$ **do**
- 4: Collect set of trajectories $\mathcal{D}_k = \{\tau_i\}$ by running policy $\pi_k = \pi(\theta_k)$ in the environment.
- 5: Compute rewards-to-go \hat{R}_t .
- 6: Compute advantage estimates, \hat{A}_t (using any method of advantage estimation) based on the current value function V_{ϕ_k} .
- 7: Estimate policy gradient as

$$\hat{g}_k = \frac{1}{|\mathcal{D}_k|} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^T \nabla_\theta \log \pi_\theta(a_t | s_t)|_{\theta_k} \hat{A}_t,$$

- 8: Use the conjugate gradient algorithm to compute

$$\hat{x}_k \approx \hat{H}_k^{-1} \hat{g}_k,$$

- where \hat{H}_k is the Hessian of the sample average KL-divergence.
- 9: Update the policy by backtracking line search with

$$\theta_{k+1} = \theta_k + \alpha^j \sqrt{\frac{2\delta}{\hat{x}_k^T \hat{H}_k \hat{x}_k}} \hat{x}_k,$$

- where $j \in \{0, 1, 2, \dots K\}$ is the smallest value which improves the sample loss and satisfies the sample KL-divergence constraint.
- 10: Fit value function by regression on mean-squared error:

$$\phi_{k+1} = \arg \min_{\phi} \frac{1}{|\mathcal{D}_k| T} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^T \left(V_\phi(s_t) - \hat{R}_t \right)^2,$$

- typically via some gradient descent algorithm.
- 11: **end for**
-

source : <https://spinningup.openai.com/en/latest/algorithms/trpo.html>

Trust Region Policy Optimization (TRPO)

- Learning process
 1. 상태에 따른 행동 선택
 2. 환경에서 선택한 행동으로 한 time step을 진행한 후, 다음 상태와 보상을 받음
 3. Sample (s, a, r) 을 trajectories set에 저장
 4. 일정 sample들이 모이면 trajectories set으로 Actor & Critic network 업데이트
 - Step 1 : Return 구하기
 - Step 2 : Critic network 업데이트
 - Step 3 : Surrogate loss를 통해 actor loss의 gradient 구하기
 - Step 4 : Conjugate gradient method를 통해 search direction 구하기
 - Step 5 : Step size와 maximal step 구하기
 - Step 6 : Backtracking line search를 통해 trust region안에서 actor network 업데이트

Trust Region Policy Optimization (TRPO)

- Actor network

```
4  class Actor(nn.Module):
5      def __init__(self, state_size, action_size, args):
6          super(Actor, self).__init__()
7          self.fc1 = nn.Linear(state_size, args.hidden_size)
8          self.fc2 = nn.Linear(args.hidden_size, args.hidden_size)
9          self.fc3 = nn.Linear(args.hidden_size, action_size)
10
11     def forward(self, x):
12         x = torch.tanh(self.fc1(x))
13         x = torch.tanh(self.fc2(x))
14
15         mu = self.fc3(x)
16         log_std = torch.zeros_like(mu)
17         std = torch.exp(log_std)
18
19         return mu, std
```

Trust Region Policy Optimization (TRPO)

- Actor network

```
4  class Actor(nn.Module):
5      def __init__(self, state_size, action_size, args):
6          super(Actor, self).__init__()
7          self.fc1 = nn.Linear(state_size, args.hidden_size)
8          self.fc2 = nn.Linear(args.hidden_size, args.hidden_size)
9          self.fc3 = nn.Linear(args.hidden_size, action_size)
10
11     def forward(self, x):
12         x = torch.tanh(self.fc1(x))
13         x = torch.tanh(self.fc2(x))
14
15         mu = self.fc3(x)
16         log_std = torch.zeros_like(mu)
17         std = torch.exp(log_std)
18
19         return mu, std
```

The number of parameters

Layer	Weight	Bias
1	$3 \times 64 = 192$	64
2	$64 \times 64 = 4096$	64
3	$64 \times 1 = 64$	1
Sum	4352	129
Total		4481

Trust Region Policy Optimization (TRPO)

- Critic network

```
21  class Critic(nn.Module):  
22      def __init__(self, state_size, args):  
23          super(Critic, self).__init__()  
24          self.fc1 = nn.Linear(state_size, args.hidden_size)  
25          self.fc2 = nn.Linear(args.hidden_size, args.hidden_size)  
26          self.fc3 = nn.Linear(args.hidden_size, 1)  
27  
28      def forward(self, x):  
29          x = torch.tanh(self.fc1(x))  
30          x = torch.tanh(self.fc2(x))  
31          value = self.fc3(x)  
32  
33      return value
```



Learning process

1. 상태에 따른 행동 선택

```
156 | mu, std = actor(torch.Tensor(state)) train.py  
157 | action = get_action(mu, std)
```

```
5   def get_action(mu, std): utils.py  
6       normal = Normal(mu, std)  
7       action = normal.sample()  
8  
9       return action.data.numpy()
```

- Normal(Gaussian) distribution example

```
mu = torch.Tensor([1, 0, -1])  
std = torch.Tensor([1., 1., 1.])  
  
from torch.distributions import Normal  
normal = Normal(mu, std)
```

```
x = normal.sample()  
print(x)  
...  
tensor([-0.2713,  0.3903, -0.1373])  
...
```



CORE
Control + Optimization Research Lab

Learning process

1. 상태에 따른 행동 선택

```
156     mu, std = actor(torch.Tensor(state)) train.py  
157     action = get_action(mu, std)
```

```
5     def get_action(mu, std): utils.py  
6         normal = Normal(mu, std)  
7         action = normal.sample()  
8  
9         return action.data.numpy()
```

- `Normal(mu, std)`
 - `Normal(Gaussian) distribution`에서 sampling을 할 경우, 분산을 일정하게 유지하면서 지속적인 exploration이 가능
 - `std`를 1로 고정함으로써 일정한 폭을 가지는 normal distribution에서 sampling

Learning process

2. 환경에서 선택한 행동으로 한 time step을 진행한 후, 다음 상태와 보상을 받음

```
159 | | | | next_state, reward, done, _ = env.step(action)
```

3. Sample (s, a, r) 을 trajectories set에 저장

```
139 | | | | trajectories = deque()  
161 | | | | mask = 0 if done else 1  
162 | | | |  
163 | | | | trajectories.append((state, action, reward, mask))
```

Learning process

4. 일정 sample들이 모이면 trajectories set으로 Actor & Critic network 업데이트

- Step 1 : Return 구하기

- time step 6까지 진행하고 episode가 끝났을 경우를 가정

$$G_1 = R_2 + \gamma R_3 + \gamma^2 R_4 + \gamma^3 R_5 + \gamma^4 R_6$$

$$G_2 = R_3 + \gamma R_4 + \gamma^2 R_5 + \gamma^3 R_6$$

$$G_3 = R_4 + \gamma R_5 + \gamma^2 R_6$$

$$G_4 = R_5 + \gamma R_6$$

$$G_5 = R_6$$

- 거꾸로 계산하며 계산해놓은 return값을 이용

$$G_5 = R_6$$

$$G_4 = R_5 + \gamma G_5$$

$$G_3 = R_4 + \gamma G_4$$

$$G_2 = R_3 + \gamma G_3$$

$$G_1 = R_2 + \gamma G_2$$

Learning process

4. 일정 sample들이 모이면 trajectories set으로 Actor & Critic network 업데이트

- Step 1 : Return 구하기

- 거꾸로 계산하며 계산해놓은 return값을 이용

$$G_5 = R_6$$

$$G_4 = R_5 + \gamma G_5$$

$$G_3 = R_4 + \gamma G_4$$

$$G_2 = R_3 + \gamma G_3$$

$$G_1 = R_2 + \gamma G_2$$

```
11     def get_returns(rewards, masks, gamma):  
12         returns = torch.zeros_like(rewards)  
13         running_returns = 0  
14  
15         for t in reversed(range(0, len(rewards))):  
16             running_returns = rewards[t] + masks[t] * gamma * running_returns  
17             returns[t] = running_returns  
18  
19         returns = (returns - returns.mean()) / returns.std()  
20  
21         return returns
```

utils.py



CORE
Control + Optimization Research Lab

Learning process

4. 일정 sample들이 모이면 trajectories set으로 Actor & Critic network 업데이트
 - Step 2 : Critic network 업데이트
 - Critic Loss

$$J_V(\phi) = \frac{(V_\phi(s) - R)^2}{\text{Prediction} \quad \text{Target}}$$

```
47      # -----
48      # step 2: update critic
49      criterion = torch.nn.MSELoss()
50
51      values = critic(torch.Tensor(states))
52      targets = returns.unsqueeze(1)
53
54      critic_loss = criterion(values, targets)
55      critic_optimizer.zero_grad()
56      critic_loss.backward()
57      critic_optimizer.step()
```

Learning process

- Learning process
 - 1. 상태에 따른 행동 선택
 - 2. 환경에서 선택한 행동으로 한 time step을 진행한 후, 다음 상태와 보상을 받음
 - 3. Sample (s, a, r) 을 trajectories set에 저장
 - 4. 일정 sample들이 모이면 trajectories set으로 Actor & Critic network 업데이트
 - Step 1 : Return 구하기 **Clear!**
 - Step 2 : Critic network 업데이트 **Clear!**
 - Step 3 : Surrogate loss를 통해 actor loss의 gradient 구하기
 - Step 4 : Conjugate gradient method를 통해 search direction 구하기
 - Step 5 : Step size와 maximal step 구하기
 - Step 6 : Backtracking line search를 통해 trust region안에서 actor network 업데이트

Learning process

- Optimization problem of theoretical TRPO → Surrogate objective function

$$\text{maximize}_{\theta} \quad \mathcal{L}_{\theta_{old}}(\theta) = \mathbb{E}_{s,a \sim \pi_{\theta_{old}}} \left[\frac{\pi_{\theta}(a|s)}{\pi_{\theta_{old}}(a|s)} A_{\pi_{\theta_{old}}}(s,a) \right]$$

$$s.t. \quad \bar{D}_{KL}^{\rho_{\theta_{old}}}(\theta_{old}, \theta) \leq \delta$$

- But how do we solve it? → Approximation!

$$\mathcal{L}_{\theta_{old}}(\theta) \approx g^T(\theta - \theta_{old}) \quad g \doteq \nabla_{\theta} \mathcal{L}_{\theta_{old}}(\theta) |_{\theta_{old}}$$

$$\bar{D}_{KL}^{\rho_{\theta_{old}}}(\theta_{old}, \theta) \approx \frac{1}{2}(\theta - \theta_{old})^T H(\theta - \theta_{old}) \quad H \doteq \nabla_{\theta}^2 \bar{D}_{KL}^{\rho_{\theta_{old}}}(\theta_{old}, \theta) |_{\theta_{old}}$$

Learning process

- Approximate optimization problem of practical TRPO

$$\text{maximize}_{\theta} \ g^T(\theta - \theta_{old})$$

$$s.t. \ \frac{1}{2}(\theta - \theta_{old})^T H(\theta - \theta_{old}) \leq \delta$$

- Approximate optimization problem of practical TRPO (argmax form)

$$\theta_{k+1} = \underset{\theta}{\text{argmax}} \ g^T(\theta - \theta_k)$$

$$s.t. \ \frac{1}{2}(\theta - \theta_k)^T H(\theta - \theta_k) \leq \delta$$

- Solution to approximate problem

$$\theta_{k+1} = \theta_k + \underbrace{\sqrt{\frac{2\delta}{g^T H^{-1} g}}}_{\text{Step size}} \underbrace{H^{-1} g}_{\text{Search direction}}$$

Learning process

- Solution to approximate problem

$$\theta_{k+1} = \theta_k + \underbrace{\sqrt{\frac{2\delta}{g^T H^{-1} g}} H^{-1} g}_{\text{Maximal step}}$$

- TRPO adds a modification to this update rule → Backtracking line search

$$\theta_{k+1} = \theta_k + \underbrace{\alpha^j \sqrt{\frac{2\delta}{g^T H^{-1} g}} H^{-1} g}_{\text{Backtracking coefficient}}$$

Learning process

❖ 정리

1) Get surrogate objective function

$$\mathbb{E}_{s,a \sim \pi_{\theta_{old}}} \left[\frac{\pi_{\theta}(a|s)}{\pi_{\theta_{old}}(a|s)} A_{\pi_{\theta_{old}}}(s,a) \right]$$

2) Find search direction, step size and maximal step through CGM and Hessian of KL

- Search direction : $H^{-1}g$
- Step size : $\sqrt{\frac{2\delta}{g^T H^{-1} g}}$
- Maximal step : $\sqrt{\frac{2\delta}{g^T H^{-1} g}} H^{-1}g$

3) Do line search on that direction inside trust region through Backtracking line search

$$\theta_{k+1} = \theta_k + \alpha^j \sqrt{\frac{2\delta}{g^T H^{-1} g}} H^{-1}g$$

Learning process

4. 일정 sample들이 모이면 trajectories set으로 Actor & Critic network 업데이트
 - Step 3 : Surrogate loss를 통해 actor loss의 gradient 구하기

```
59      # -----
60      # step 3: get gradient of actor loss through surrogate loss
61      mu, std = actor(torch.Tensor(states))
62      old_policy = get_log_prob(actions, mu, std)
```

train.py

```
23  def get_log_prob(actions, mu, std):      utils.py
24      normal = Normal(mu, std)
25      log_prob = normal.log_prob(actions)
26
27      return log_prob
```

- The probability density of the normal distribution

$$\pi_{\theta}(a|s) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(a - \mu_{\theta}(s))^2}{2\sigma^2}\right)$$

- Multiply the log on both sides

$$\log \pi_{\theta}(a|s) = -\log \sqrt{2\pi} - \log \sigma - \frac{(a - \mu_{\theta}(s))^2}{2\sigma^2}$$

Learning process

4. 일정 sample들이 모이면 trajectories set으로 Actor & Critic network 업데이트
 - Step 3 : Surrogate loss를 통해 actor loss의 gradient 구하기

```
59      # -----
60      # step 3: get gradient of actor loss through surrogate loss
61      mu, std = actor(torch.Tensor(states))
62      old_policy = get_log_prob(actions, mu, std)
63      actor_loss = surrogate_loss(actor, values, targets, states, old_policy.detach(), actions) train.py

29      def surrogate_loss(actor, values, targets, states, old_policy, actions): utils.py
30          mu, std = actor(torch.Tensor(states))
31          new_policy = get_log_prob(actions, mu, std)
32
33          advantages = targets - values
34
35          surrogate_loss = torch.exp(new_policy - old_policy) * advantages
36          surrogate_loss = surrogate_loss.mean()
37
38          return surrogate_loss
```

- surrogate_loss

$$\mathcal{L}_{\theta_{old}}(\theta) = \mathbb{E}_{s,a \sim \pi_{\theta_{old}}} \left[\frac{\pi_{\theta}(a|s)}{\pi_{\theta_{old}}(a|s)} A_{\pi_{\theta_{old}}}(s,a) \right]$$

Learning process

- 일정 sample들이 모이면 trajectories set으로 Actor & Critic network 업데이트
 - Step 3 : Surrogate loss를 통해 actor loss의 gradient 구하기

```
59      # -----
60      # step 3: get gradient of actor loss through surrogate loss
61      mu, std = actor(torch.Tensor(states))
62      old_policy = get_log_prob(actions, mu, std)
63      actor_loss = surrogate_loss(actor, values, targets, states, old_policy.detach(), actions)
64
65      actor_loss_grad = torch.autograd.grad(actor_loss, actor.parameters())
66      actor_loss_grad = flat_grad(actor_loss_grad)
```

train.py

- torch.autograd.grad(outputs, inputs) - actor.parameter : 4481(weight : 4352, bias : 129)

```
torch.autograd.grad(outputs, inputs, grad_outputs=None, retain_graph=None,
create_graph=False, only_inputs=True, allow_unused=False)
```

[SOURCE]

Computes and returns the sum of gradients of outputs w.r.t. the inputs.

- example

$$y = \sum_{i=1}^4 3 \times x_i$$

$$\nabla_x y = \left(\frac{\partial y}{\partial x_1}, \frac{\partial y}{\partial x_2}, \frac{\partial y}{\partial x_3}, \frac{\partial y}{\partial x_4} \right) = (3, 3, 3, 3)$$

Learning process

4. 일정 sample들이 모이면 trajectories set으로 Actor & Critic network 업데이트
 - Step 3 : Surrogate loss를 통해 actor loss의 gradient 구하기

```
59      # -----
60      # step 3: get gradient of actor loss through surrogate loss
61      mu, std = actor(torch.Tensor(states))
62      old_policy = get_log_prob(actions, mu, std)
63      actor_loss = surrogate_loss(actor, values, targets, states, old_policy.detach(), actions)
64
65      actor_loss_grad = torch.autograd.grad(actor_loss, actor.parameters())
66      actor_loss_grad = flat_grad(actor_loss_grad)
```

train.py

- `flat_grad(actor_loss_grad)`

utils.py

```
94  def flat_grad(grads):
95      grad_flatten = []
96      for grad in grads:
97          grad_flatten.append(grad.view(-1))
98      grad_flatten = torch.cat(grad_flatten)
99      return grad_flatten
```

Weight	Bias
3×64	64
64×64	64
64×1	1

Learning process

❖ 정리

- 1) Get surrogate objective function **Clear!**

$$\mathbb{E}_{s,a \sim \pi_{\theta_{old}}} \left[\frac{\pi_{\theta}(a|s)}{\pi_{\theta_{old}}(a|s)} A_{\pi_{\theta_{old}}}(s,a) \right]$$

- 2) Find search direction, step size and maximal step through CGM and Hessian of KL

- Search direction : $H^{-1}g$
- Step size : $\sqrt{\frac{2\delta}{g^T H^{-1} g}}$
- Maximal step : $\sqrt{\frac{2\delta}{g^T H^{-1} g}} H^{-1}g$

- 3) Do line search on that direction inside trust region through Backtracking line search

$$\theta_{k+1} = \theta_k + \alpha^j \sqrt{\frac{2\delta}{g^T H^{-1} g}} H^{-1}g$$



CORE
Control + Optimization Research Lab

Learning process

4. 일정 sample들이 모이면 trajectories set으로 Actor & Critic network 업데이트
 - Step 4 : Conjugate gradient method를 통해 search direction 구하기

```
68      # -----
69      # step 4: get search direction through conjugate gradient method
70      search_dir = conjugate_gradient(actor, states, actor_loss_grad.data, nsteps=10)
```

train.py

Learning process

❖ 정리

- 1) Get surrogate objective function **Clear!**

$$\mathbb{E}_{s,a \sim \pi_{\theta_{old}}} \left[\frac{\pi_{\theta}(a|s)}{\pi_{\theta_{old}}(a|s)} A_{\pi_{\theta_{old}}}(s,a) \right]$$

- 2) Find search direction, step size and maximal step through CGM and Hessian of KL

- Search direction : $H^{-1}g$ **Clear!**
- Step size : $\sqrt{\frac{2\delta}{g^T H^{-1} g}}$
- Maximal step : $\sqrt{\frac{2\delta}{g^T H^{-1} g}} H^{-1}g$

- 3) Do line search on that direction inside trust region through Backtracking line search

$$\theta_{k+1} = \theta_k + \alpha^j \sqrt{\frac{2\delta}{g^T H^{-1} g}} H^{-1}g$$



CORE
Control + Optimization Research Lab

Learning process

4. 일정 sample들이 모이면 trajectories set으로 Actor & Critic network 업데이트

- Step 5 : Step size와 maximal step 구하기

train.py

```
72 | # -----
73 | # step 5: get step size and maximal step
74 | gHg = (hessian_vector_product(actor, states, search_dir) * search_dir).sum(0, keepdim=True)
75 | step_size = torch.sqrt(2 * args.max_kl / gHg)[0]
76 | maximal_step = step_size * search_dir
```

- Step size : $\sqrt{\frac{2\delta}{g^T H^{-1} g}}$ ($Hx = g$), (max_kl δ : 0.01)
- Maximal step : $\sqrt{\frac{2\delta}{g^T H^{-1} g}} H^{-1} g$

Learning process

❖ 정리

- 1) Get surrogate objective function **Clear!**

$$\mathbb{E}_{s,a \sim \pi_{\theta_{old}}} \left[\frac{\pi_{\theta}(a|s)}{\pi_{\theta_{old}}(a|s)} A_{\pi_{\theta_{old}}}(s,a) \right]$$

- 2) Find search direction, step size and maximal step through CGM and Hessian of KL

- Search direction : $H^{-1}g$ **Clear!**

• Step size : $\sqrt{\frac{2\delta}{g^T H^{-1} g}}$ **Clear!**

• Maximal step : $\sqrt{\frac{2\delta}{g^T H^{-1} g}} H^{-1}g$ **Clear!**

- 3) Do line search on that direction inside trust region through Backtracking line search

$$\theta_{k+1} = \theta_k + \alpha^j \sqrt{\frac{2\delta}{g^T H^{-1} g}} H^{-1}g$$



Learning process

4. 일정 sample들이 모이면 trajectories set으로 Actor & Critic network 업데이트

- Step 6 : Backtracking line search를 통해 trust region안에서 actor network 업데이트

```
78      # -----
79      # step 6: perform backtracking line search and update actor in trust region
80      params = flat_params(actor)                                train.py

109     def flat_params(model):                                     utils.py
110         params = []
111         for param in model.parameters():
112             |   params.append(param.data.view(-1))
113         params_flatten = torch.cat(params)
114         return params_flatten
```

Learning process

4. 일정 sample들이 모이면 trajectories set으로 Actor & Critic network 업데이트

- Step 6 : Backtracking line search를 통해 trust region안에서 actor network 업데이트

```
78      # -----
79      # step 6: perform backtracking line search and update actor in trust region
80      params = flat_params(actor)
81
82      old_actor = Actor(state_size, action_size, args)
83      update_model(old_actor, params)
```

train.py

```
def update_model(model, new_params):
    index = 0
    for params in model.parameters():
        params_length = len(params.view(-1))
        new_param = new_params[index: index + params_length]
        new_param = new_param.view(params.size())
        params.data.copy_(new_param)
        index += params_length
```

utils.py

params_length 192
params_length 64
params_length 4096
params_length 64
params_length 64
params_length 1

Weight	Bias
$3 \times 64 = 192$	64
$64 \times 64 = 4096$	64
$64 \times 1 = 64$	1
4352	129

Learning process

4. 일정 sample들이 모이면 trajectories set으로 Actor & Critic network 업데이트
 - Step 6 : Backtracking line search를 통해 trust region안에서 actor network 업데이트

```
78      # -----
79      # step 6: perform backtracking line search and update actor in trust region
80      params = flat_params(actor)
81
82      old_actor = Actor(state_size, action_size, args)
83      update_model(old_actor, params)
84
85      backtracking_line_search(old_actor, actor, actor_loss, actor_loss_grad,
86                                old_policy, params, maximal_step, args.max_kl,
87                                values, targets, states, actions)
```

train.py

Learning process

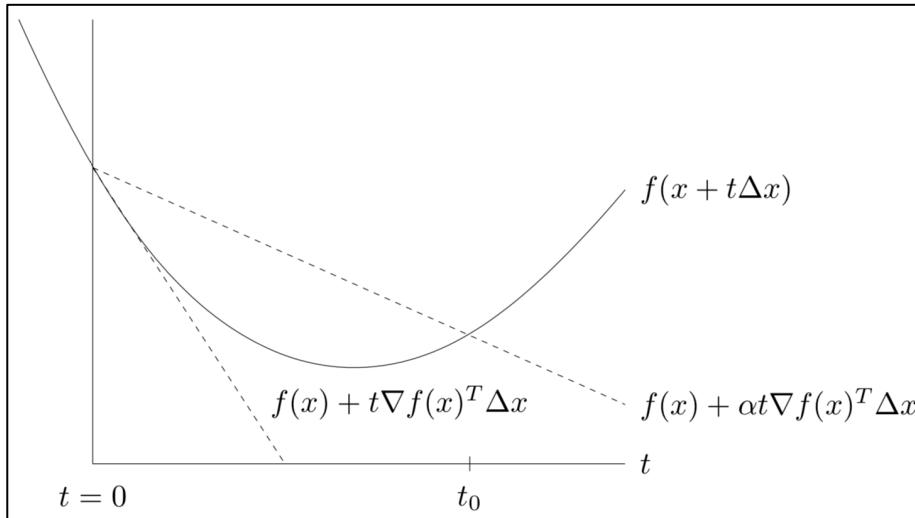
4. 일정 sample들이 모이면 trajectories set으로 Actor & Critic network 업데이트
 - Step 6 : Actor network를 업데이트하고, backtracking line search를 통해 trust region안에서 업데이트

Algorithm 9.2 Backtracking line search.

given a descent direction Δx for f at $x \in \text{dom } f$, $\alpha \in (0, 0.5)$, $\beta \in (0, 1)$.

$t := 1$.

while $f(x + t\Delta x) > f(x) + \alpha t \nabla f(x)^T \Delta x$, $t := \beta t$.



source : https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf

Hyperparameter

```
14 parser = argparse.ArgumentParser()
15 parser.add_argument('--env_name', type=str, default="Pendulum-v0")
16 parser.add_argument('--load_model', type=str, default=None)
17 parser.add_argument('--save_path', default='./save_model/', help=' ')
18 parser.add_argument('--render', action="store_true", default=False)
19 parser.add_argument('--gamma', type=float, default=0.99)
20 parser.add_argument('--hidden_size', type=int, default=64)
21 parser.add_argument('--critic_lr', type=float, default=1e-3)
22 parser.add_argument('--max_kl', type=float, default=1e-2)
23 parser.add_argument('--max_iter_num', type=int, default=500)
24 parser.add_argument('--total_sample_size', type=int, default=2048)
25 parser.add_argument('--log_interval', type=int, default=5)
26 parser.add_argument('--goal_score', type=int, default=-300)
27 parser.add_argument('--logdir', type=str, default='./logs',
28 | | | | | help='tensorboardx logs directory')
29 args = parser.parse_args()
```

Main loop

- Initialization
 - Seed - random number 고정
 - Actor network
 - TensorboardX
 - Recent rewards

```
119  def main():
120      env = gym.make(args.env_name)
121      env.seed(500)
122      torch.manual_seed(500)
123
124      state_size = env.observation_space.shape[0]
125      action_size = env.action_space.shape[0]
126      print('state size:', state_size)
127      print('action size:', action_size)
128
129      actor = Actor(state_size, action_size, args)
130      critic = Critic(state_size, args)
131      critic_optimizer = optim.Adam(critic.parameters(), lr=args.critic_lr)
132
133      writer = SummaryWriter(args.logdir)
134
135      recent_rewards = deque(maxlen=100)
136      episodes = 0
```

Main loop

- Episode 진행
 - Initialize trajectories set
 - 상태에 따른 행동 선택
 - 다음 상태와 보상을 받음
 - Trajectories set에 저장

```
138     for iter in range(args.max_iter_num):  
139         trajectories = deque()  
140         steps = 0  
141  
142         while steps < args.total_sample_size:  
143             done = False  
144             score = 0  
145             episodes += 1  
146  
147             state = env.reset()  
148             state = np.reshape(state, [1, state_size])  
149  
150             while not done:  
151                 if args.render:  
152                     env.render()  
153  
154                 steps += 1  
155  
156                 mu, std = actor(torch.Tensor(state))  
157                 action = get_action(mu, std)  
158  
159                 next_state, reward, done, _ = env.step(action)  
160  
161                 mask = 0 if done else 1  
162  
163                 trajectories.append((state, action, reward, mask))  
164  
165                 next_state = np.reshape(next_state, [1, state_size])  
166                 state = next_state  
167                 score += reward  
168  
169                 if done:  
170                     recent_rewards.append(score)
```

Main loop

- Train model
- Print & Visualize log
- Termination : 최근 100개의 episode의 평균 score가 -300보다 크다면
 - Save model
 - 학습 종료

```
172     actor.train(), critic.train()
173     train_model(actor, critic, critic_optimizer,
174                  | | |
174                  | | | trajectories, state_size, action_size)
175
176     writer.add_scalar('log/score', float(score), episodes)
177
178     if iter % args.log_interval == 0:
179         print('{0} iter | {0} episode | score_avg: {1:.2f}'.format(iter, episodes, np.mean(recent_rewards)))
180
181     if np.mean(recent_rewards) > args.goal_score:
182         if not os.path.isdir(args.save_path):
183             os.makedirs(args.save_path)
184
185         ckpt_path = args.save_path + 'model.pth'
186         torch.save(actor.state_dict(), ckpt_path)
187         print('Recent rewards exceed -300. So end')
188         break
```

Train model

- Trajectories → Numpy array
- Trajectories에 있는 2200개의 sample들을 각각 나눔
 - state - (2200, 3)
 - action - (2200, 1)
 - reward - (2200)
 - mask - (2200)

```
31  def train_model(actor, critic, critic_optimizer,
32  |   |   |   trajectories, state_size, action_size):
33  |   |   |       trajectories = np.array(trajectories)
34  |   |   |       states = np.vstack(trajectories[:, 0])
35  |   |   |       actions = list(trajectories[:, 1])
36  |   |   |       rewards = list(trajectories[:, 2])
37  |   |   |       masks = list(trajectories[:, 3])
38
39       actions = torch.Tensor(actions).squeeze(1)
40       rewards = torch.Tensor(rewards).squeeze(1)
41       masks = torch.Tensor(masks)
```

Train model

- **returns** - (2200)
- **values** - (2200, 1)
- **targets** - (2200, 1)

```
43     # -----
44     # step 1: get returns
45     returns = get_returns(rewards, masks, args.gamma)
46
47     # -----
48     # step 2: update critic
49     criterion = torch.nn.MSELoss()
50
51     values = critic(torch.Tensor(states))
52     targets = returns.unsqueeze(1)
53
54     critic_loss = criterion(values, targets)
55     critic_optimizer.zero_grad()
56     critic_loss.backward()
57     critic_optimizer.step()
```

Train model

- **old_policy** - (2200, 1)
- **actor_loss_grad** - (4481)
- **search_dir** - (4481)

```
59      # -----
60      # step 3: get gradient of actor loss through surrogate loss
61      mu, std = actor(torch.Tensor(states))
62      old_policy = get_log_prob(actions, mu, std)
63      actor_loss = surrogate_loss(actor, values, targets, states, old_policy.detach(), actions)
64
65      actor_loss_grad = torch.autograd.grad(actor_loss, actor.parameters())
66      actor_loss_grad = flat_grad(actor_loss_grad)
67
68      # -----
69      # step 4: get search direction through conjugate gradient method
70      search_dir = conjugate_gradient(actor, states, actor_loss_grad.data, nsteps=10)
```

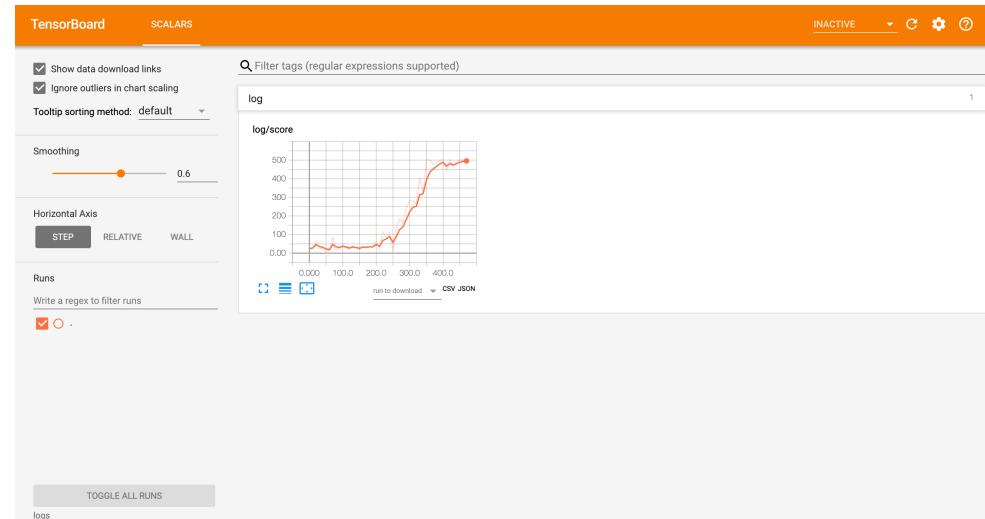
Train model

- **params** - (4481)

```
72      # -----
73      # step 5: get step size and maximal step
74      gHg = (hessian_vector_product(actor, states, search_dir) * search_dir).sum(0, keepdim=True)
75      step_size = torch.sqrt(2 * args.max_kl / gHg)[0]
76      maximal_step = step_size * search_dir
77
78      # -----
79      # step 6: perform backtracking line search and update actor in trust region
80      params = flat_params(actor)
81
82      old_actor = Actor(state_size, action_size, args)
83      update_model(old_actor, params)
84
85      backtracking_line_search(old_actor, actor, actor_loss, actor_loss_grad,
86                                old_policy, params, maximal_step, args.max_kl,
87                                values, targets, states, actions)
```

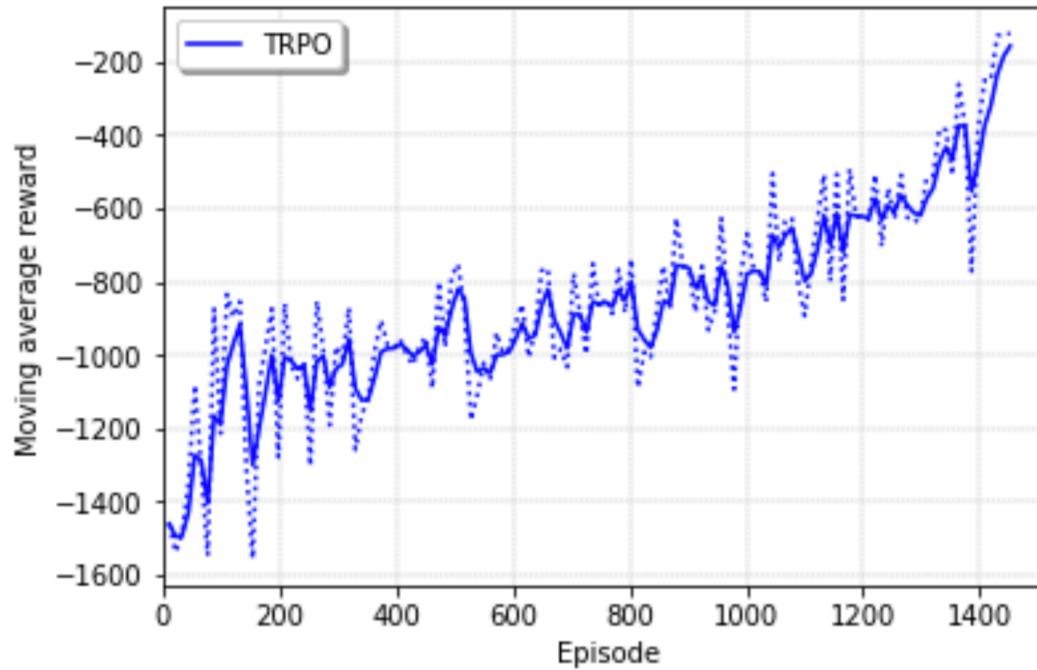
Train & TensorboardX

- Terminal A - train 실행
 - conda activate env_name
 - python train.py
- Terminal B - tensorboardX 실행
 - conda activate env_name
 - tensorboard --logdir logs
 - (웹에서) localhost:6006



Learning curve & Test

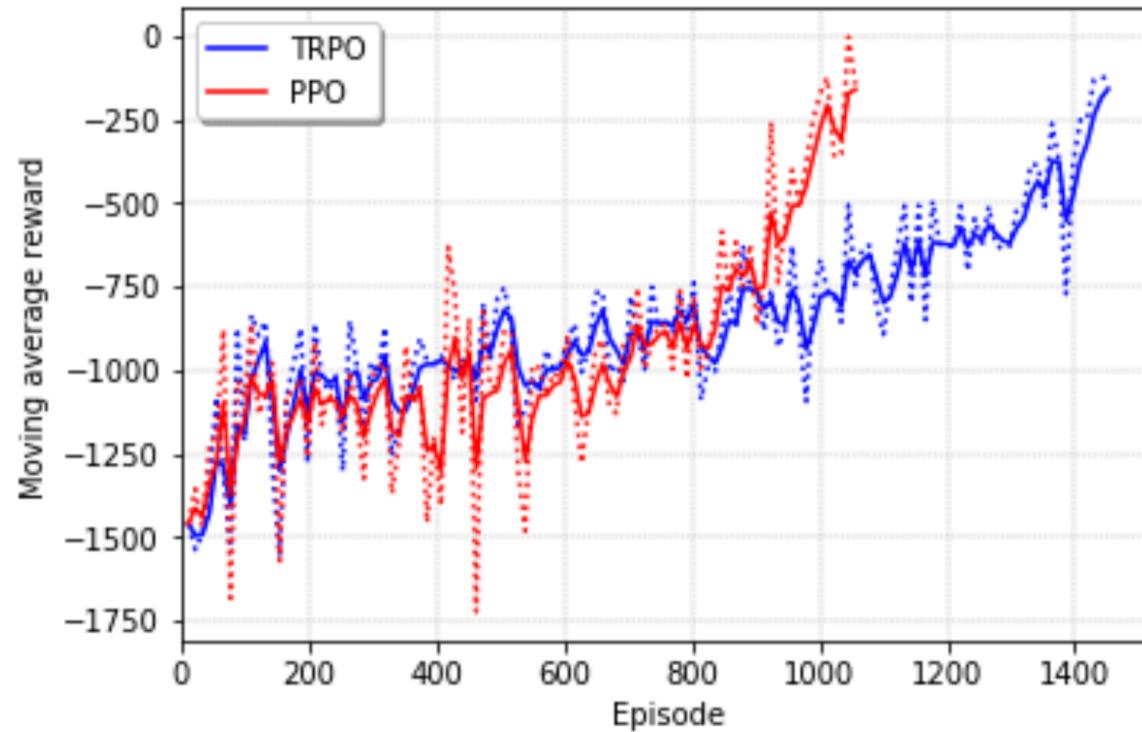
- Learning curve



- Test
 - `python test.py`

TRPO vs. PPO

- Learning curve



Thank you



CORE
Control + Optimization Research Lab

Supplement - PPO

- Train model - Mini batch 형태로 Actor & Critic network 업데이트
 - Mini batch로 나누기

```
52     criterion = torch.nn.MSELoss()
53
54     n = len(states)
55     arr = np.arange(n)
56
57     for _ in range(args.model_update_num):
58         np.random.shuffle(arr)
59
60         for i in range(n // args.batch_size):
61             mini_batch_index = arr[args.batch_size * i : args.batch_size * (i + 1)]
62             mini_batch_index = torch.LongTensor(mini_batch_index)
63
64             states_samples = torch.Tensor(states)[mini_batch_index]
65             actions_samples = torch.Tensor(actions)[mini_batch_index]
```

- $n = 2200$
- Batch size = 64

```
args.batch_size * i 0
args.batch_size * (i + 1) 64

args.batch_size * i 64
args.batch_size * (i + 1) 128

args.batch_size * i 128
args.batch_size * (i + 1) 192

.
.

args.batch_size * i 1984
args.batch_size * (i + 1) 2048

args.batch_size * i 2048
args.batch_size * (i + 1) 2112

args.batch_size * i 2112
args.batch_size * (i + 1) 2176
```

Supplement - PPO

- Train model - Mini batch 형태로 Actor & Critic network 업데이트
 - Critic Loss

$$J_V(\phi) = \frac{(V_\phi(s) - R)^2}{\text{Prediction} \quad \text{Target}}$$

```
67 | # get critic loss
68 | values_samples = critic(states_samples)
69 | targets_samples = returns.unsqueeze(1)[mini_batch_index]
70 |
71 | critic_loss = criterion(values_samples, targets_samples)
```

Supplement - PPO

- The probability ratio $r_t(\theta)$

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$$

- Surrogate objective function of TRPO

$$\mathcal{L}(\theta) = \mathbb{E}_t \left[\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)} \hat{A}_t \right] = [r_t(\theta) \hat{A}_t]$$

- Main objective function of PPO

$$\mathcal{L}^{CLIP}(\theta) = \mathbb{E}_t [\min(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t)]$$

Supplement - PPO

- Train model - Mini batch 형태로 Actor & Critic network 업데이트
 - Actor Loss (clip param : 0.2)

$$J_{\pi}(\theta) = \min(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t)$$

```
73     # get actor loss
74     actor_loss, ratio, advantages_samples = surrogate_loss(actor, values_samples, targets_samples,
75     states_samples, old_policy.detach(),
76     actions_samples, mini_batch_index)
77
78     clipped_ratio = torch.clamp(ratio,
79     1.0 - args.clip_param,
80     1.0 + args.clip_param)
81     clipped_actor_loss = clipped_ratio * advantages_samples
82
83     actor_loss = -torch.min(actor_loss, clipped_actor_loss).mean()
```

Supplement - PPO

- Train model - Mini batch 형태로 Actor & Critic network 업데이트
 - Update actor & critic

```
85          # update actor & critic
86          loss = actor_loss + 0.5 * critic_loss
87
88          critic_optimizer.zero_grad()
89          loss.backward(retain_graph=True)
90          critic_optimizer.step()
91
92          actor_optimizer.zero_grad()
93          loss.backward()
94          actor_optimizer.step()
```