

Abstract

Intro

Related work

{ method }

Exp

Citations

Journal # of building  
engineering  
Building simulation

5 days to open the app

{ How is our work  
better than others }

⇒ Train on March  
Test on Jan and April

⇒ Train on Dec  
Test on Nov and Dec. (slight overlap)

⇒ Exp 13, 14, 15 ⇒ Allows Bert to  
and sparse reward.

↓  
Vanilla policy gradient on sparse reward

Exp 14 ⇒ Transfer learning ~~using~~ at agent level.

~~Exp~~  
Exp 15 ⇒ Transfer learning at env.-model level.

Exp 14 ⇒ Not transfer learning at all here.  
Exp 15 ⇒ Model is for constant pricing  
scenario on boptest-hydro-  
-heat fact.

↳ Model is learnt using Dynamic  
pricing

↳ Gets from experiment 2 (policy and  
model based  
RL with  
Dynamic pricing  
scenario)

( $\Rightarrow$  We may want to add an experiment where pretrained agent for constant pricing scenario is used)

$\Rightarrow$  methodology

$\Rightarrow$  Same or model based RL with policy gradient / Reinforce

$\Rightarrow$  Transfer learning

actions are added

NN

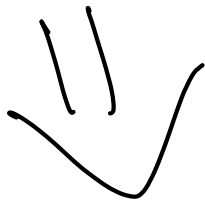
$\Rightarrow$  statespace remains same

$\Rightarrow$  How sparse reward is created and used in policy gradient algorithm

$\Rightarrow$

---

ALGORITHM 1 STARTS BELOW



①  $\Theta = \Theta_I \Leftarrow \text{Initialized}$

$$\text{policy-loss} = - \sum_{t=0}^{T=1344} \log(\pi_{\Theta_I}(s_t)) R_t$$

where,  
at  $t = 1344$ ,  $R_t = 0$  if,  $t \% 50 \neq 0$

$\Rightarrow$  The negative sign is to indicate  
we need to carry out gradient  
ascent instead of descent.

After training,  $\Theta = \Theta_T$

(2) Approach 2: Initialize  $\Theta^P = \Theta^P_I$ .

Prettraining:

$T=672$  pump.

$$\text{policy-loss} = - \sum_{t=1}^T \log(P_{\Theta^P}(\leq t)) \cdot R_t$$

After prettraining:  $\Theta = \Theta^P_{P.T}$

Fine tuning:

$$\begin{aligned} \text{policy-loss} \\ = - \sum_{t=1}^{T=1344} \log(P_{\Theta^P}(\leq t)) R_t \end{aligned}$$

where,  
 $R_t = 0$  if  $t/50! = 0$

After Fine tuning:

$$\Theta = \Theta^P_F$$

### ③ Approach 3:

Here we pretrain using an environment model and finetune the environment model on bestest-hydronic.

We then use the finetuned environment model to train the RL agent

Initialize:  $\Theta^E = \Theta^I$

Pretraining:

$$\begin{aligned} & (E_{\Theta^E}(s_t, a_t)[0] - s_{t+1})^2 \\ & + (E_{\Theta^E}(s_t, a_t)[1] - R_t)^2 \end{aligned}$$

After pretraining:  $\Theta^E = \Theta^{E}_{P.T}$

Finetuning  $T=1344$

$$\text{env\_loss} = \sum_{t=0}^T \left( E_{\Theta^{E}_{P.T}}(s_t, a_t)[0] - s_{t+1} \right)^2 + \left( E_{\Theta^{E}_{P.T}}(s_t, a_t)[1] - R_t \right)^2$$

$\Rightarrow$  only calculate loss when  $t \% 50 \neq 0$

After Finetuning:  $\Theta^E = \Theta^E_F$

Now training the policy gradient

$$\text{policy\_loss} = - \sum_{t=1}^{T=1344} \log(P_{\theta_I}^p(s_t)) \cdot R_t$$

where,

If  $t \cdot 0.50 = 0 \Rightarrow$  Use the actual environment

$t \cdot 0.50 \neq 0 \Rightarrow$  Use the surrogate environment.