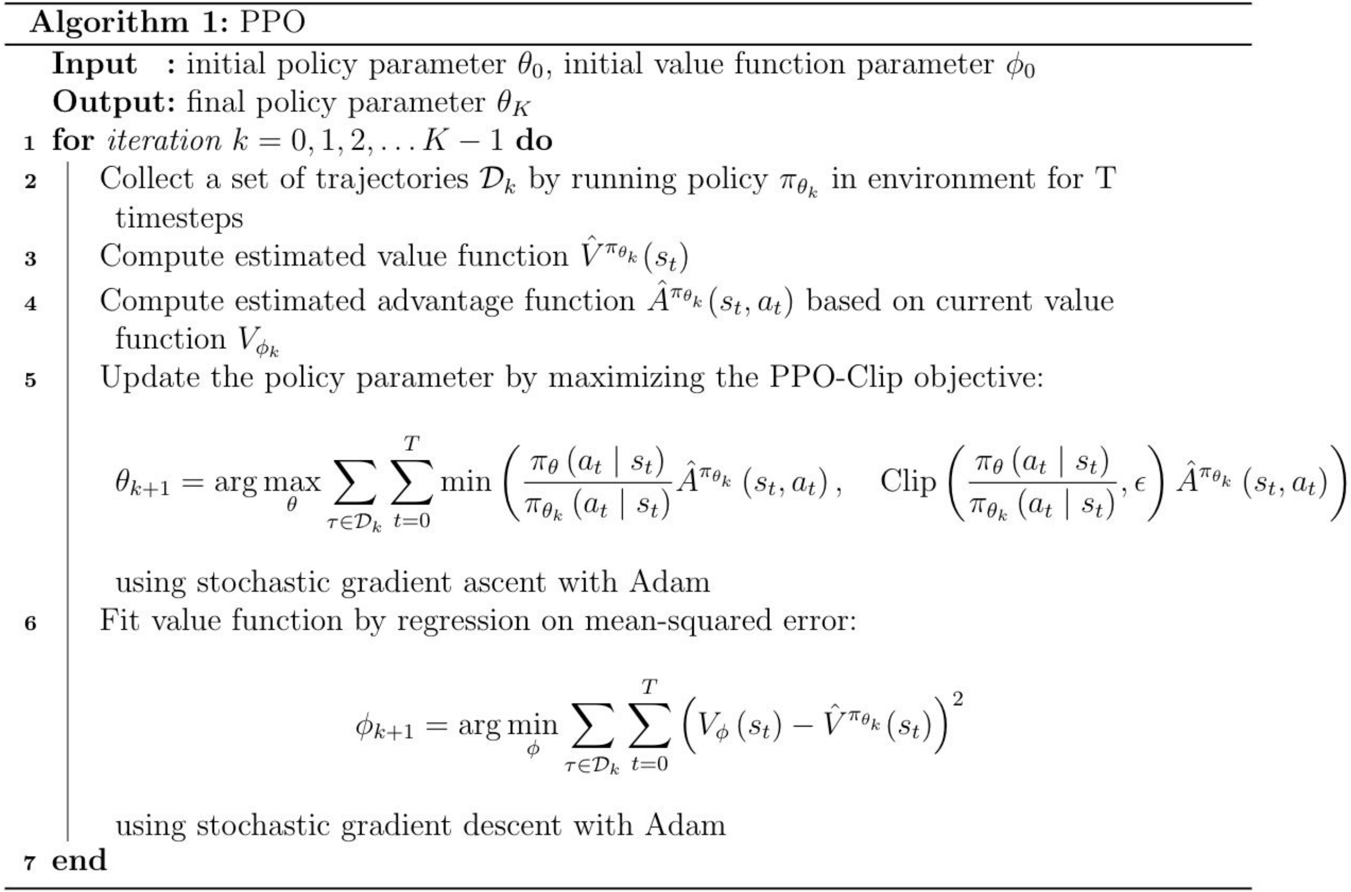
Inventory Control with Multiple Suppliers

Hemeng Li
Ruifan Yang

PPO



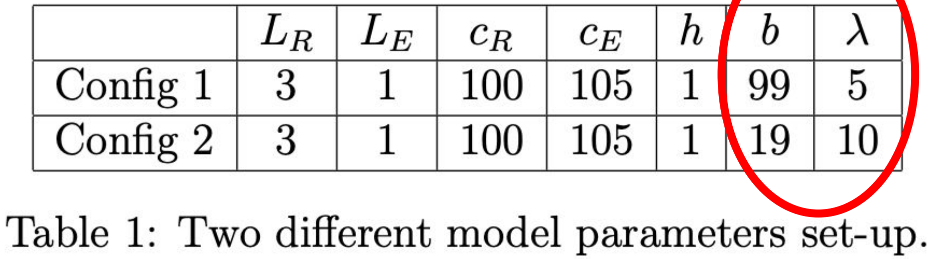
```
Algorithm 1: PPO
Input   : initial policy parameter $\theta_0$, initial value function parameter $\phi_0$
Output: final policy parameter $\theta_K$
for iteration $k = 0, 1, 2, \ldots K-1$ do
    Collect a set of trajectories $\mathcal{D}_k$ by running policy $\pi_{\theta_k}$ in environment for T timesteps
    Compute estimated value function $\hat{V}^{\pi_{\theta_k}}(s_t)$
    Compute estimated advantage function $\hat{A}^{\pi_{\theta_k}}(s_t, a_t)$ based on current value function $V_{\phi_k}$
    Update the policy parameter by maximizing the PPO-Clip objective:
        $$\theta_{k+1} = \arg\max_{\theta} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^{T} \min\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_k}\left(a_t \mid s_t\right)} \hat{A}^{\pi_{\theta_k}}\left(s_t, a_t\right), \quad \operatorname{Clip}\left(\frac{\pi_\theta\left(a_t \mid s_t\right)}{\pi_{\theta_k}\left(a_t \mid s_t\right)}, \epsilon\right) \hat{A}^{\pi_{\theta_k}}\left(s_t, a_t\right)\right)$$
    using stochastic gradient ascent with Adam
    Fit value function by regression on mean-squared error:
        $$\phi_{k+1} = \arg\min_{\phi} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^{T} \left(V_\phi\left(s_t\right) - \hat{V}^{\pi_{\theta_k}}\left(s_t\right)\right)^2$$
    using stochastic gradient descent with Adam
end
```

Implementation details

- Two separate neural networks for actor and critic
- Initial Policy: train initial actor with a TBS policy
- Value function estimation: TD(1)
- Normalized Advantage function
- Maximization/Minimization using Adam

Experimental Results

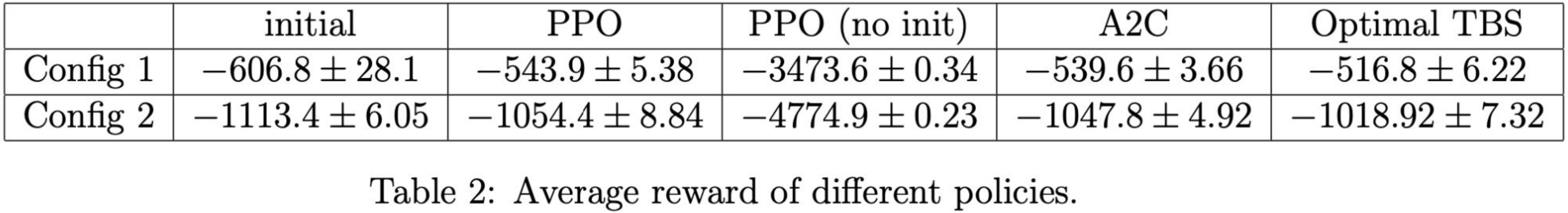
Experiment Setup:



	L_R	L_E	c_R	c_E	h	b	λ
Config 1	3	1	100	105	1	99	5
Config 2	3	1	100	105	1	19	10

Table 1: Two different model parameters set-up.

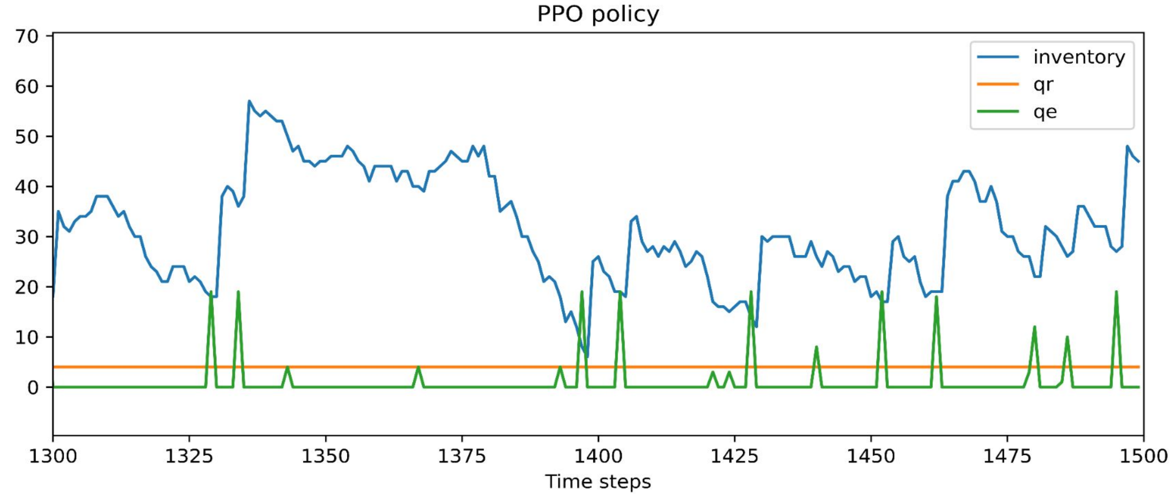
Numerical results and comparison with other group:

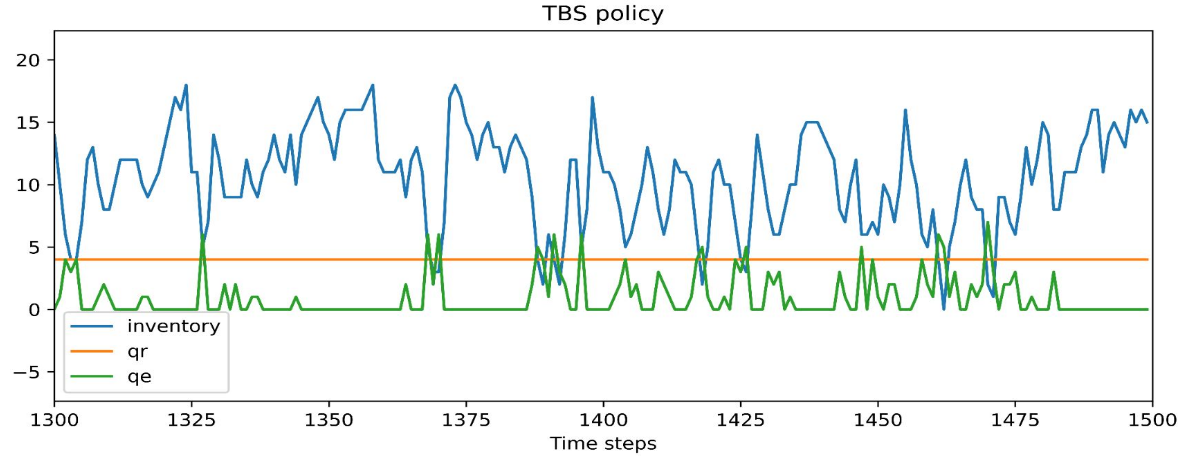


	initial	PPO	PPO (no init)	A2C	Optimal TBS
Config 1	-606.8 ± 28.1	-543.9 ± 5.38	-3473.6 ± 0.34	-539.6 ± 3.66	-516.8 ± 6.22
Config 2	-1113.4 ± 6.05	-1054.4 ± 8.84	-4774.9 ± 0.23	-1047.8 ± 4.92	-1018.92 ± 7.32

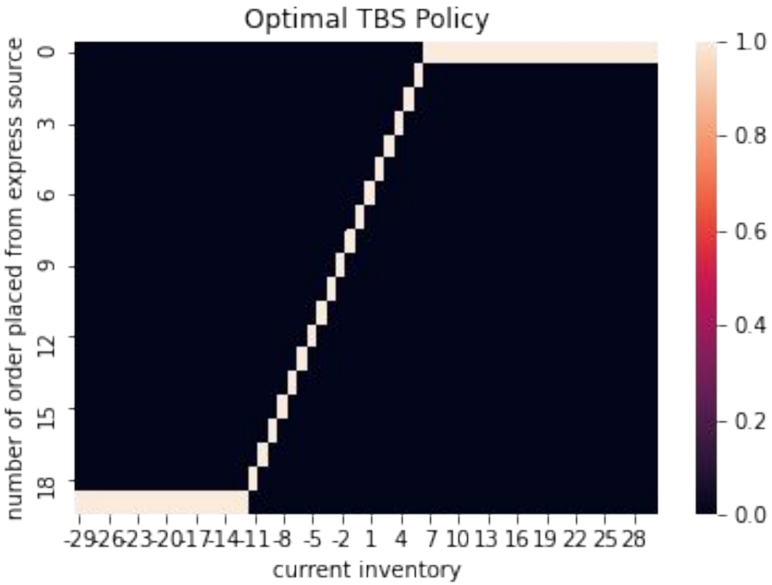
Table 2: Average reward of different policies.

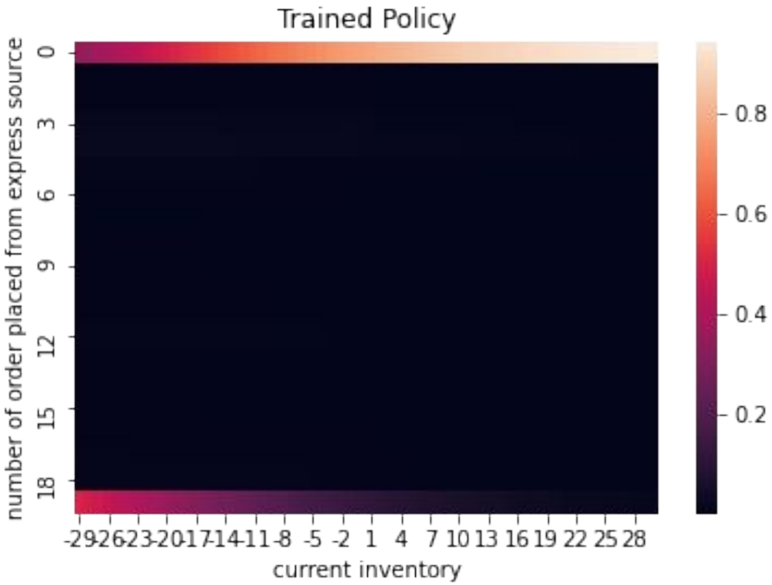
Policy Visualization





Policy Visualization





Future Work

- Explore more complex NN structure
- Use multiple actors
- Reward normalization
- Adam annealing