Proximal Policy Optimization Algorithms

Руслан Ахтариев
НИУ ВШЭ

12 декабря 2019 г.

Policy Gradient Methods

$$L^{PG}(\theta) = \hat{\mathbb{E}}_t \left(\log \pi_\theta(a_t \mid s_t) \hat{A}_t \right)$$

- π_θ — стохастическая стратегия (политика)
- a_t — действие совершаемое в момент t
- s_t — состояние среды в момент t
- $\hat{A}_t$ — ожидаемая награда в момент t

$$\hat{g} = \hat{\mathbb{E}}_t \left(\nabla_\theta \log \pi_\theta(a_t \mid s_t) \hat{A}_t \right)$$

Trust Region Methods

maximize $\hat{\mathbb{E}}_t \left(\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)} \hat{A}_t \right)$

subject to $\hat{\mathbb{E}}_t \left[KL \left(\pi_\theta(\cdot \mid s_t), \pi_{\theta_{old}}(\cdot \mid s_t) \right) \right] \leqslant \delta$

$$\hat{\mathbb{E}}_t \left[\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)} \hat{A}_t - \beta KL \left(\pi_\theta(\cdot \mid s_t), \pi_{\theta_{old}}(\cdot \mid s_t) \right) \right]$$

Clipped Surrogate Objective

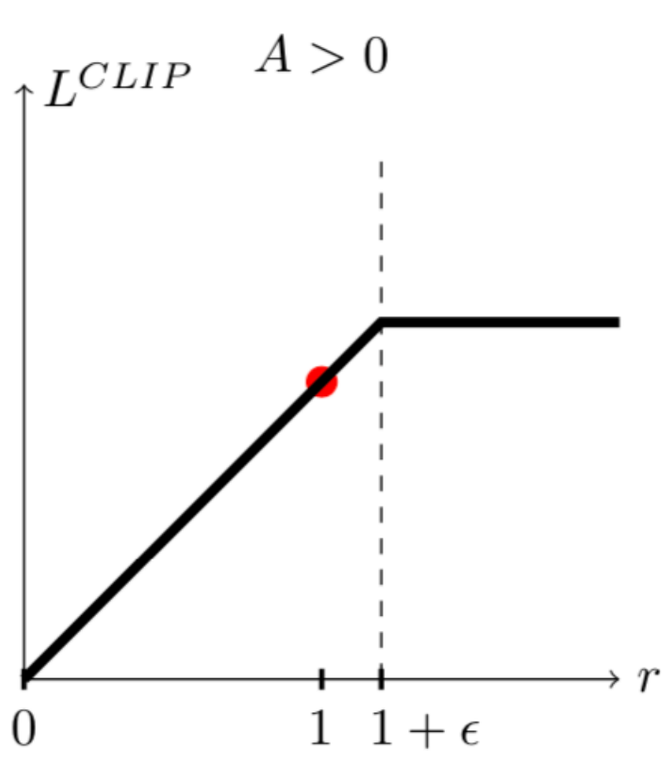
$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)}, r_t(\theta_{old}) = 1$$

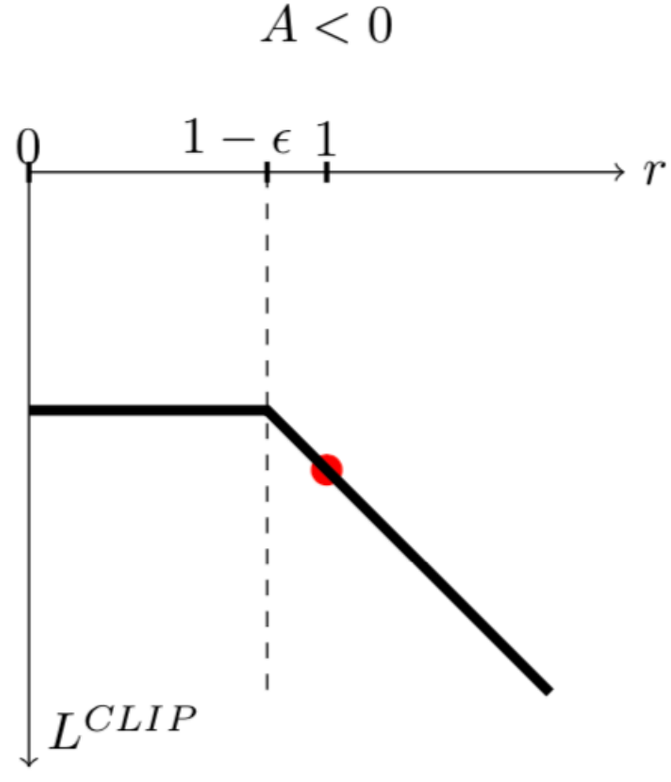
$$L^{CPI}(\theta) = \hat{\mathbb{E}}_t\left(\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)}\hat{A}_t\right) = \hat{\mathbb{E}}_t\left(r_t(\theta)\hat{A}_t\right)$$

Это приводит к слишком большим обновлениям, постараемся не отдалять $r_t(\theta)$ от единицы

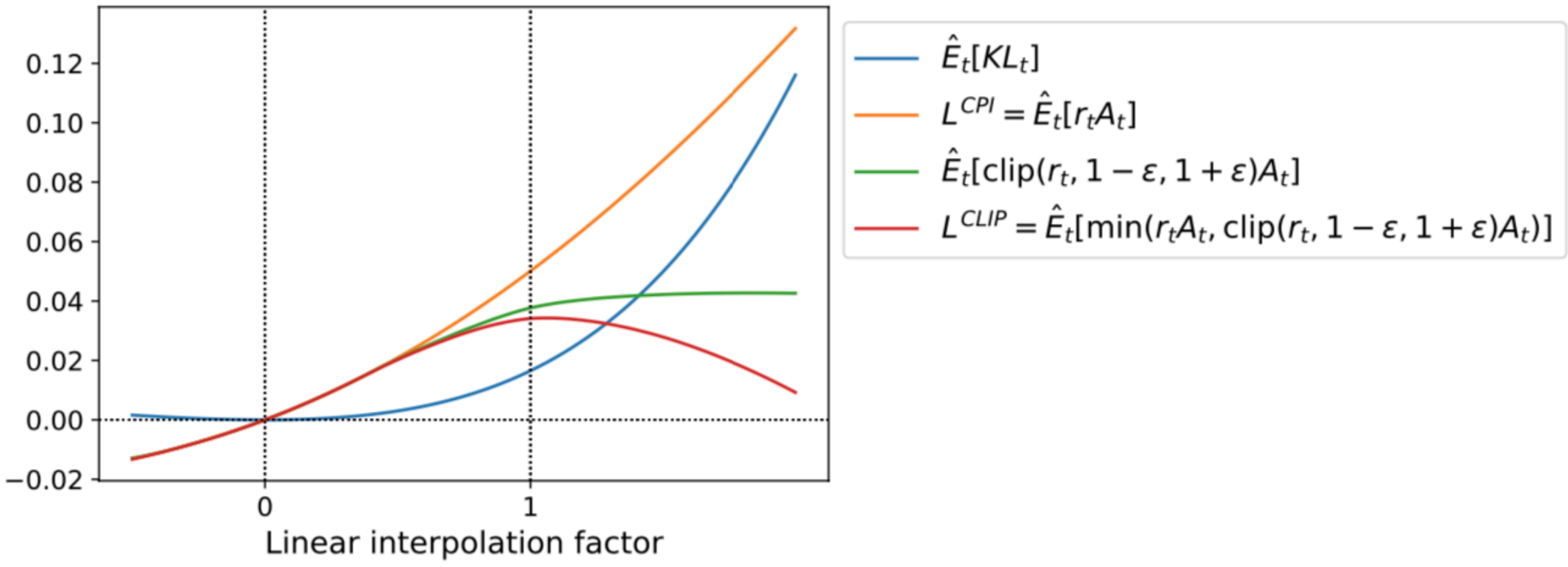
Clipped Surrogate Objective

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





Clipped Surrogate Objective



Adaptive KL Penalty Coefficient

$$L^{KLPEN}(\theta) = \hat{\mathbb{E}}_t \left[\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)} \hat{A}_t - \beta KL\left(\pi_\theta(\cdot \mid s_t), \pi_{\theta_{old}}(\cdot \mid s_t)\right) \right]$$

$$d = \hat{\mathbb{E}}_t \left[KL\left(\pi_\theta(\cdot \mid s_t), \pi_{\theta_{old}}(\cdot \mid s_t)\right) \right]$$

- $d < d_{targ}/1.5, \;\; \beta \leftarrow \beta/2$
- $d < d_{targ} \cdot 1.5, \;\; \beta \leftarrow \beta \cdot 2$

Algorithm

$$L^{CLIP+VF+S}(\theta) = \hat{\mathbb{E}}_t \left[L^{CLIP}(\theta) - c_1 L^{VF}(\theta) - c_2 S[\pi_\theta](s_t) \right]$$

- c_1, c_2 — *коэффициенты*
- S — *бонус за exploration*
- $L^{VF} = (V_\theta(s_t) - V_t^{targ})^2$ — *лос функции, которая предугадывает выигрыш*

Algorithm

Запускаем нашу стратегию на T шагов (T сильно меньше одного эпизода), вычисляем $\hat{A}_t$, используем полученные семплы для обновления.

$$\hat{A}_t = -V(s_t) + r_t + \gamma r_{t+1} + \ldots + \gamma^{T-t+1} r_{T-1} + \gamma^{T-t} V(s_T)$$

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$$

$$\hat{A}_t = \delta_t + \gamma \delta_{t+1} + \ldots + \gamma^{T-t+1} \delta_{T-1}$$

Algorithm

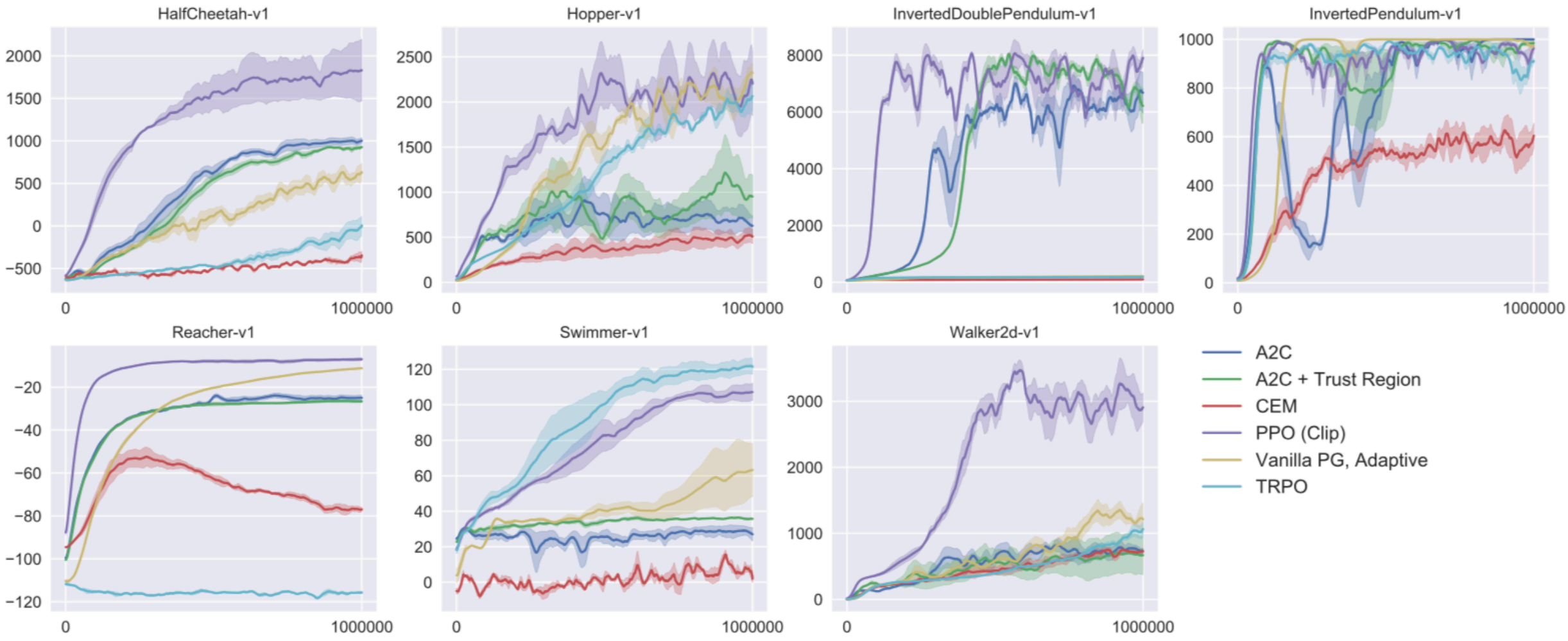
Algorithm 1 PPO, Actor-Critic Style

for iteration=$1, 2, \ldots$ **do**
  for actor=$1, 2, \ldots, N$ **do**
    Run policy $\pi_{\theta_{\text{old}}}$ in environment for T timesteps
    Compute advantage estimates $\hat{A}_1, \ldots, \hat{A}_T$
  end for
  Optimize surrogate L wrt θ, with K epochs and minibatch size $M \leq NT$
  $\theta_{\text{old}} \leftarrow \theta$
end for

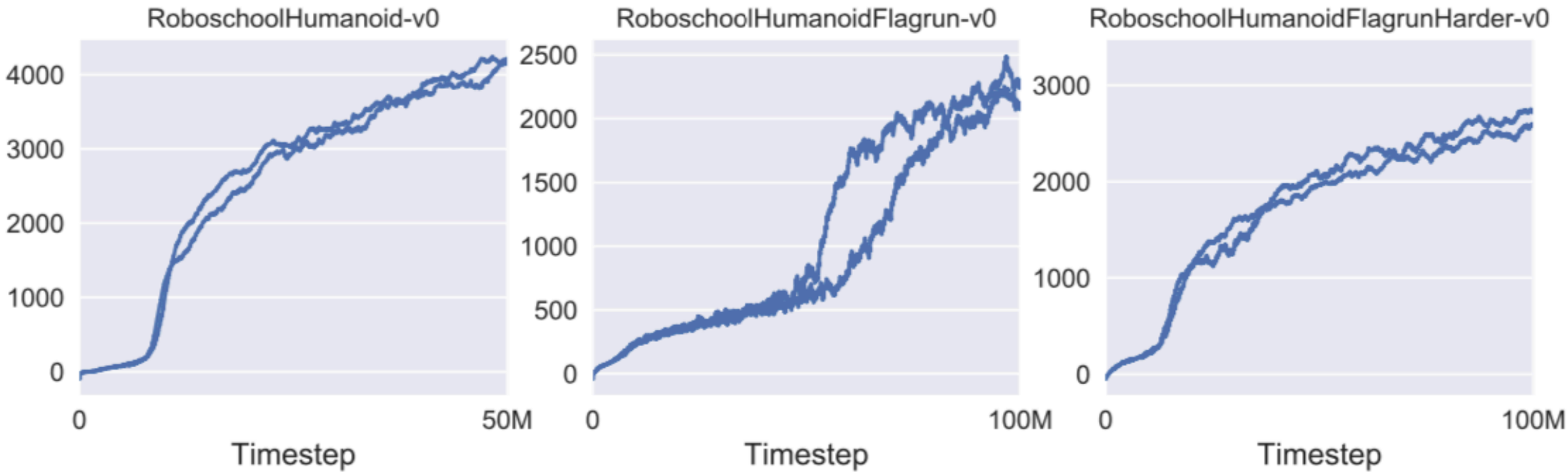
Results

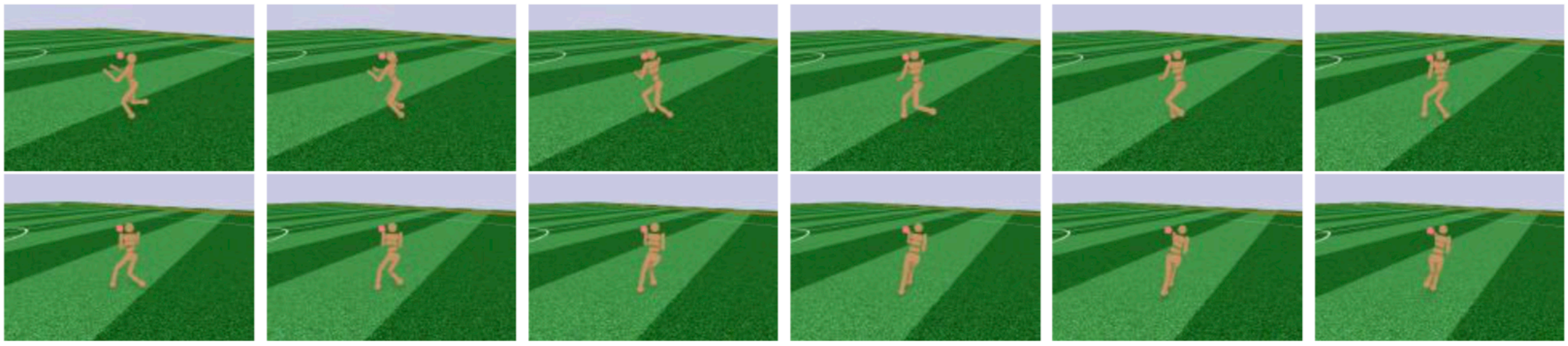
algorithm	avg. normalized score
No clipping or penalty	-0.39
Clipping, $\epsilon = 0.1$	0.76
Clipping, $\epsilon = 0.2$	**0.82**
Clipping, $\epsilon = 0.3$	0.70
Adaptive KL $d_{\text{targ}} = 0.003$	0.68
Adaptive KL $d_{\text{targ}} = 0.01$	0.74
Adaptive KL $d_{\text{targ}} = 0.03$	0.71
Fixed KL, $\beta = 0.3$	0.62
Fixed KL, $\beta = 1.$	0.71
Fixed KL, $\beta = 3.$	0.72
Fixed KL, $\beta = 10.$	0.69

Results


HalfCheetah-v1
Hopper-v1
InvertedDoublePendulum-v1
InvertedPendulum-v1
Reacher-v1
Swimmer-v1
Walker2d-v1
A2C
A2C + Trust Region
CEM
PPO (Clip)
Vanilla PG, Adaptive
TRPO

Results


RoboschoolHumanoid-v0
4000
3000
2000
1000
0
0
50M
Timestep
RoboschoolHumanoidFlagrun-v0
2500
2000
1500
1000
500
0
0
100M
Timestep
RoboschoolHumanoidFlagrunHarder-v0
3000
2000
1000
0
0
100M
Timestep



Questions

- *Какую целевую функцию предлагается оптимизировать в методе TRPO?*
- *Итоговая функция PPO, расписать её компоненты*
- *Опишите алгоритм обучения с PPO*