Improving On-Policy Learning with Statistical Reward Accumulation

Yubin Deng[1], Ke Yu[1], Dahua Lin[1], Xiaoou Tang[1], Chen Change Loy[2]

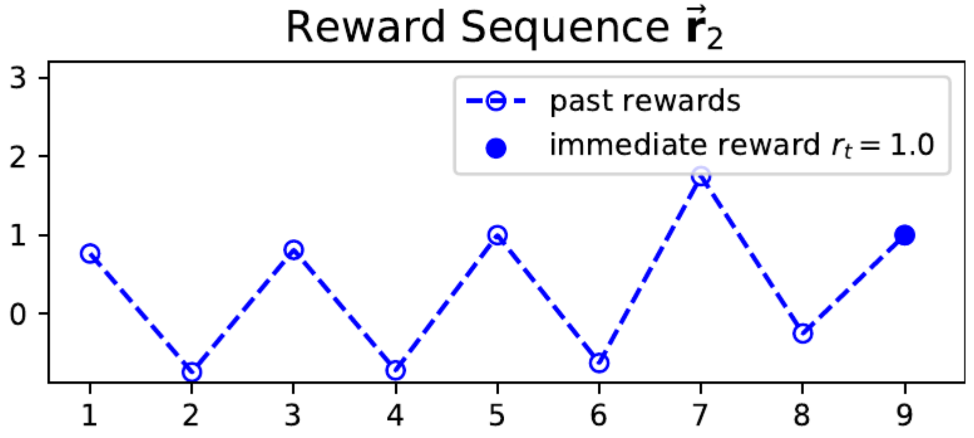
[1]CUHK-SenseTime Joint Laboratory, The Chinese University of Hong Kong

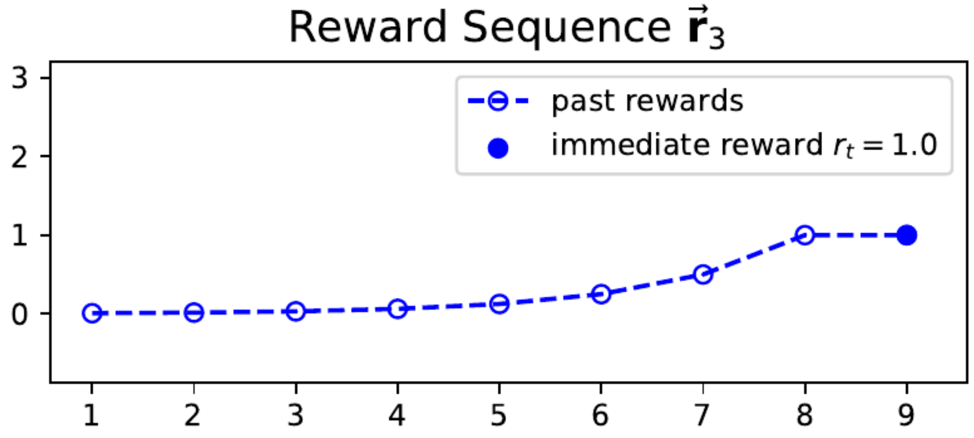
[2]SenseTime-NTU Joint AI Research Centre, Nanyang Technological University

Motivation

- **Better Reward Characterization?**
  - How high the immediate reward is
  - How varied the past rewards were

$$Sharpe\ Ratio = \frac{\mathbb{E}(r)}{\sigma(r)}$$





Our Approach

- **New Characterization**
  - How high the immediate reward is:

$$\mathcal{R}_H = e^{\frac{1}{T}\ln\frac{\mathcal{R}_T}{\mathcal{R}_0}} - 1 = \frac{\mathcal{R}_T^{1/T} - \mathcal{R}_0^{1/T}}{\mathcal{R}_0^{1/T}}$$

  - How varied the past rewards were:

$$\omega = 1 - \left[\frac{\sigma(\delta_{\mathcal{R}})}{\sigma_{max}}\right]^{\tau}$$

- **Variability-Weighted Reward (VWR)**

$$r^{vwr} = \mathcal{R}_H \times (1 - \left[\frac{\sigma(\delta_{\mathcal{R}})}{\sigma_{max}}\right]^{\tau})$$

Our Approach

- $\mathcal{R}_H$: How high the immediate reward is

$$\vec{\mathbf{r}} = [r_{t-(T-1)}, \cdots, r_{t-2}, r_{t-1}, r_t]$$

$$\vec{\mathbf{d}} = [r_{t-(T-1)}, r_{t-(T-2)} - r_{t-(T-1)} \cdots, r_t - r_{t-1}]$$

$$\vec{\mathbf{f}} = [f_1, f_2, \cdots, f_t] = [d_t, d_{t-1}, \cdots, d_{t-(T-1)}]$$

$$\vec{\mathcal{R}} = [\mathcal{R}_0, \mathcal{R}_1, \cdots, \mathcal{R}_T] = \frac{1}{T+1}[f_0, f_0 + f_1, \cdots, \sum_{i=0}^{T} f_i]$$

$$\mathcal{R}_H = \frac{\mathcal{R}_T^{1/T} - \mathcal{R}_0^{1/T}}{\mathcal{R}_0^{1/T}} \quad \text{where } \mathcal{R}_T - \mathcal{R}_0 = \frac{1}{T+1} r_t$$

$d_n = r_n - r_{n-1}$

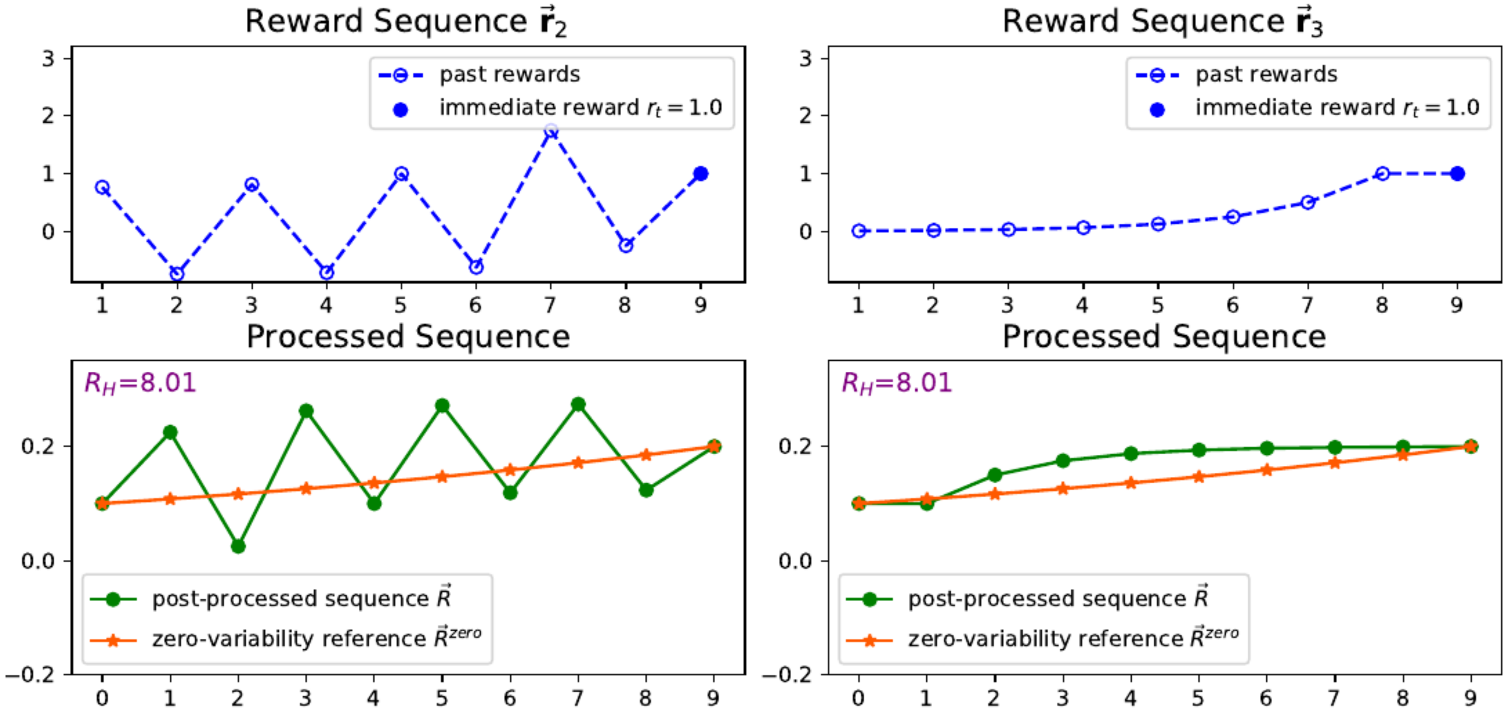
flip

normalized cumulative sum

$f_0 = 1$

Our Approach

- An example: $\mathcal{R}_H = \frac{\mathcal{R}_T^{1/T} - \mathcal{R}_0^{1/T}}{\mathcal{R}_0^{1/T}} = 8.01$



The green curve is $\vec{\mathcal{R}}$

Our Approach

- ω: How <u>varied</u> the past rewards were

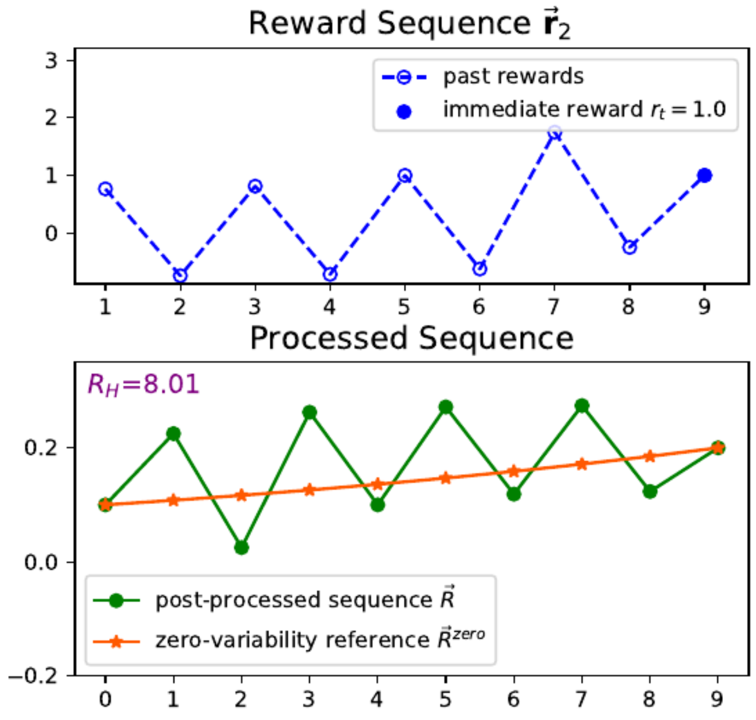
$$\vec{\mathcal{R}}^{zero} = \mathcal{R}_0\left[e^{0\times\tilde{\mathcal{R}}}, e^{1\times\tilde{\mathcal{R}}}, \cdots, e^{T\times\tilde{\mathcal{R}}}\right] \quad \text{with } \tilde{\mathcal{R}} = \frac{1}{T}\ln\frac{\mathcal{R}_T}{\mathcal{R}_0}$$

$$\delta_{\mathcal{R}}(n) = \frac{\mathcal{R}_n - \mathcal{R}_n^{zero}}{\mathcal{R}_n^{zero}}$$

$$\omega = 1 - \left[\frac{\sigma(\delta_{\mathcal{R}})}{\sigma_{max}}\right]^{\tau}$$

The green curve is $\vec{\mathcal{R}}$

The orange curve is $\vec{\mathcal{R}}^{zero}$

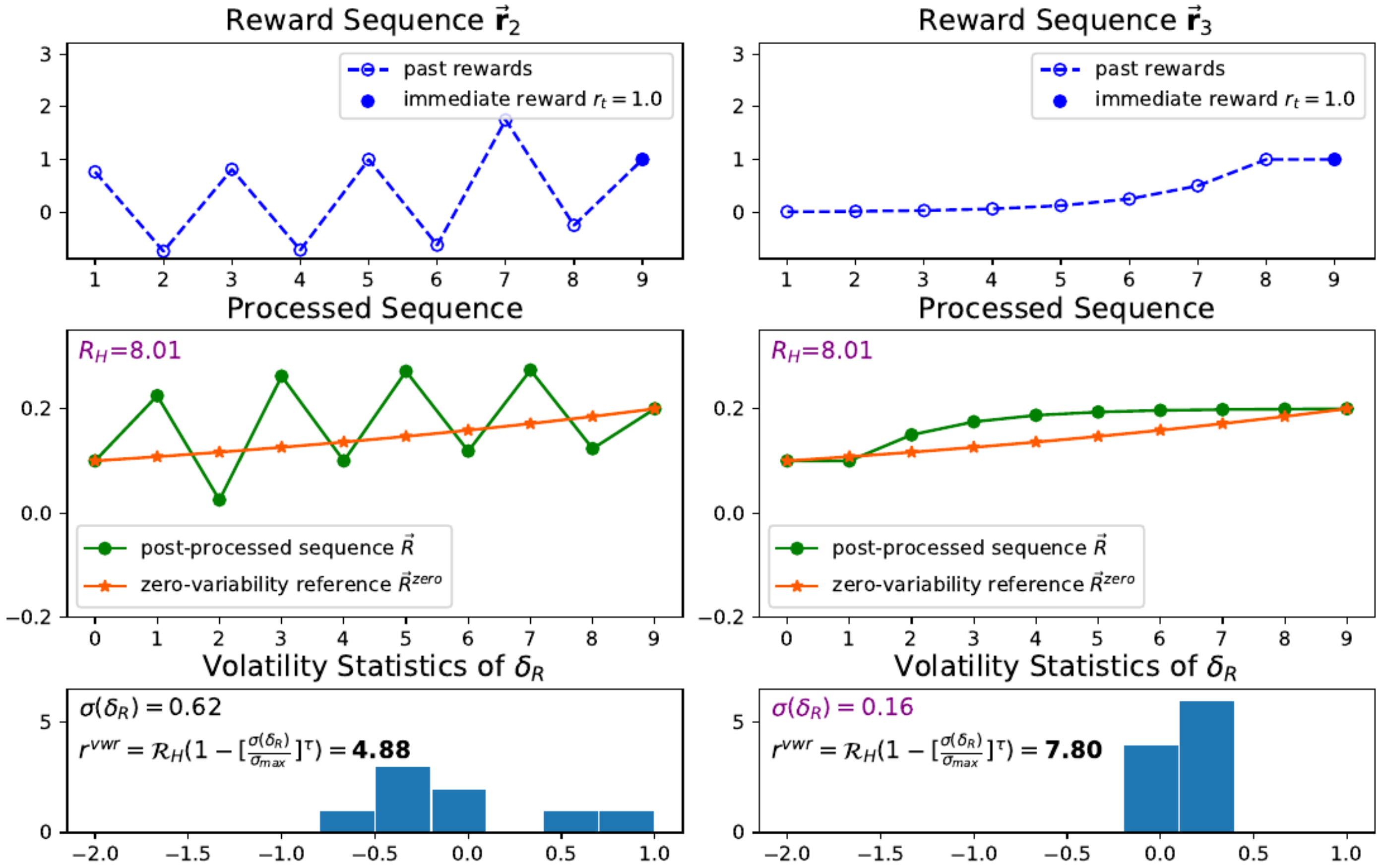


Our Approach

- Variability-Weighted Reward (VWR)

$$r^{vwr} = \begin{cases} \mathcal{R}_H\left(1 - \left[\frac{\sigma(\delta_{\mathcal{R}})}{\sigma_{max}}\right]^{\tau}\right) & \text{if } \sigma(\delta_{\mathcal{R}}) < \sigma_{max}, \mathcal{R}_T > 0 \\ 0 & \text{otherwise} \end{cases}$$

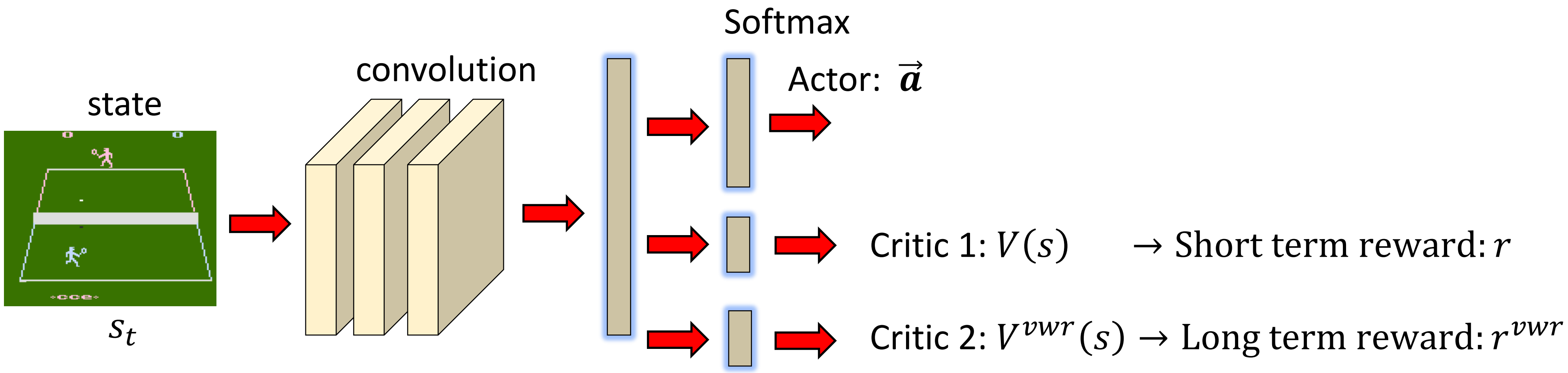
$r^{vwr} = 4.88$



$\boldsymbol{r^{vwr} = 7.80}$

Our Approach

- Advantage Actor Multi-Critic (A2MC)



$$r^{vwr} = \mathcal{R}_H \left(1 - \left[\frac{\sigma(\delta_{\mathcal{R}})}{\sigma_{max}}\right]^{\tau}\right)$$

Our Approach

- Hot-Wire Exploration

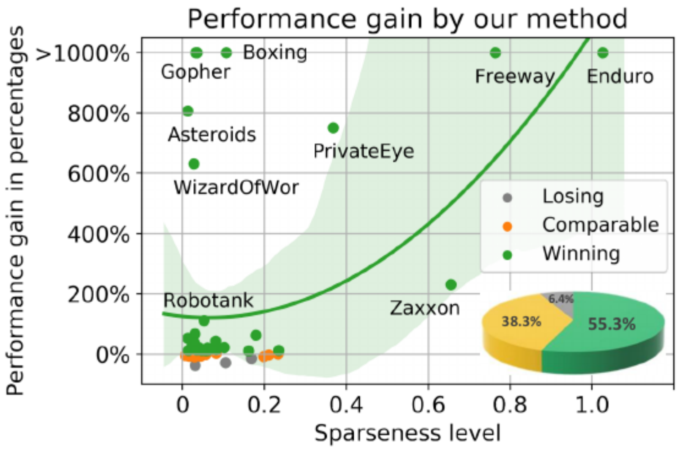
$$a_{t+k} = \begin{cases} \text{a random action identical for all k} & \text{prob} = \epsilon \\ \pi(a_{t+k}|s_{t+k}) \ \text{ for } k = 0, ..., N-1 & \text{prob} = 1 - \epsilon \end{cases}$$

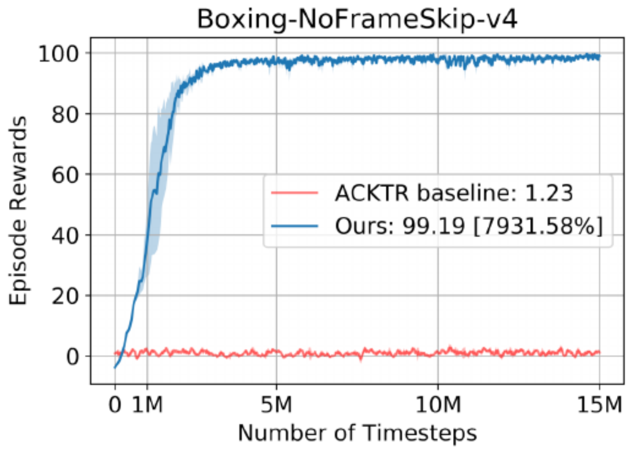
Experiments

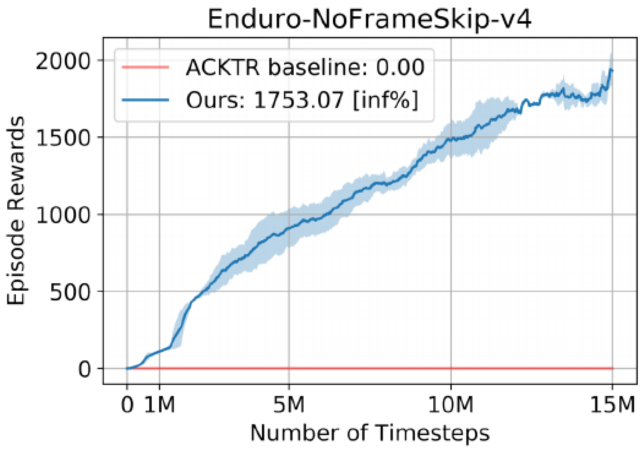
A2MC vs. ACKTR

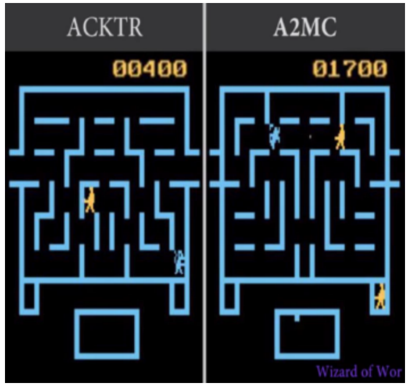
- Atari

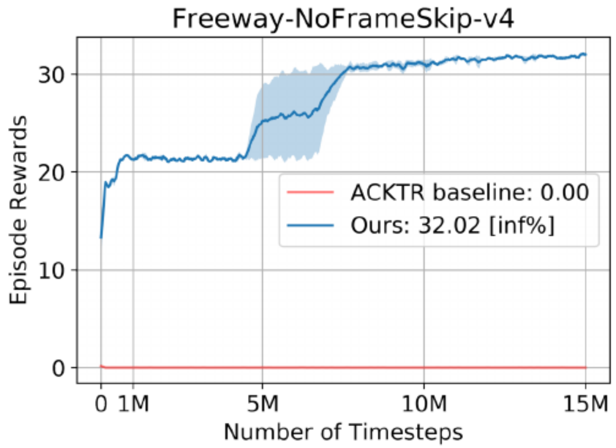
Win: 55.3%
Fair: 38.3%
Lose: 6.4%

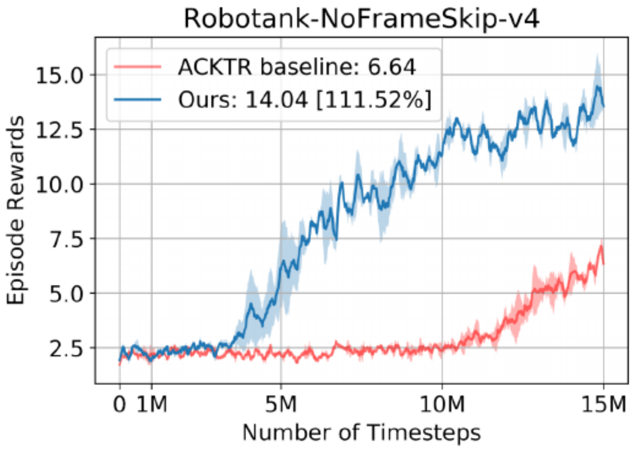


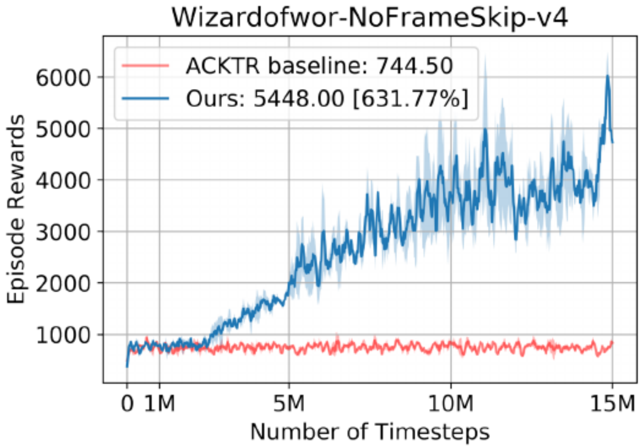










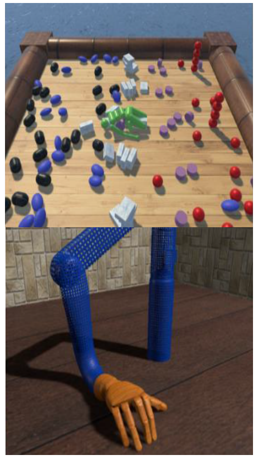


Experiments

- Atari: A2MC has a human-level performance rate of 74.5% (38 out of 51 games) in the Atari benchmarks, compared to 63.6% reached by ACKTR.

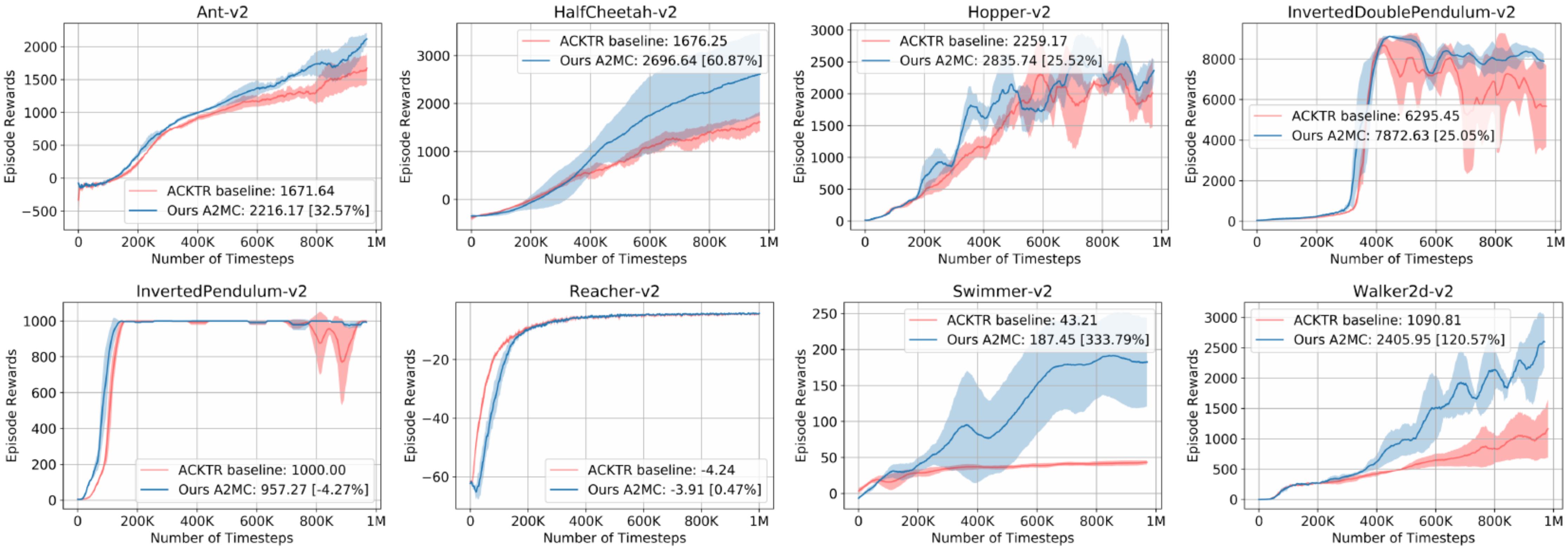
		ACKTR		A2MC	
Domain	Human	Rewards	Eps	Rewards	Eps
Asteroids	47388.7	34171.0	N/A	**830232.5**	**11314**
Beamrider	5775.0	13581.4	3279	13564.3	**3012**
Boxing	12.1	1.5	N/A	**99.1**	**158**
Breakout	31.8	**735.7**	4097	411.4	**3664**
Double Dunk	-16.4	-0.5	742	**21.3**	**544**
Enduro	860.5	0.0	N/A	**3492.2**	**730**
Freeway	29.6	0.0	N/A	**32.7**	**1058**
Pong	9.3	20.9	904	19.4	**804**
Q-bert	13455.0	21500.3	**6422**	**25229.0**	7259
Robotank	11.9	16.5	-	**25.7**	4158
Seaquest	20182.0	1776.0	N/A	**1798.6**	N/A
Space Invaders	1652.0	**19723.0**	14696	11774.0	**11064**
Wizard of Wor	4756.5	702	N/A	**7471.0**	**8119**

Experiments



- MuJoCo

A2MC vs. ACKTR



Experiments

- MuJoCo

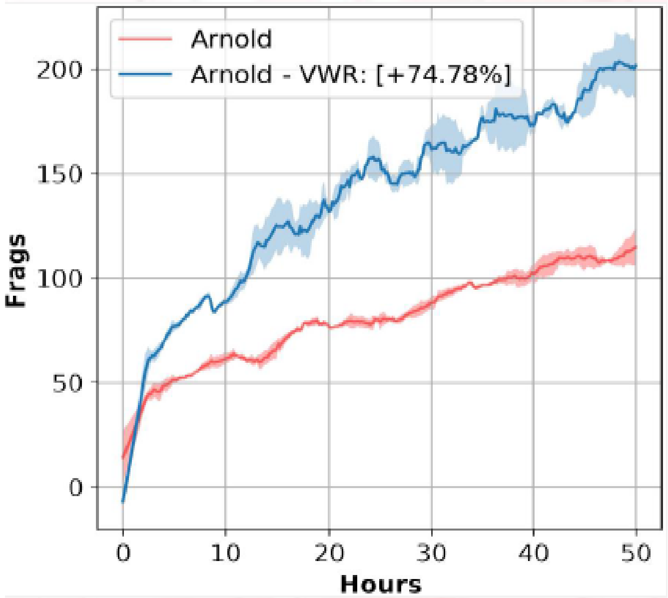
GAMES	ACKTR	Our A2MC		PPO		PPO+LIRPG	Our MC-PPO	
Ant	1671.6	**2216.1**	**(32.5%)**	411.4	(± 107.7)	~ -50	**618.9**	(**50.4**%)
HalfCheetah	1676.2	**2696.6**	**(60.8%)**	1433.7	(± 83.9)	~ 2000	**2473.4**	(**72.5**%)
Hopper	2259.1	**2835.7**	**(25.5%)**	2055.8	(± 274.6)	~ 2200	**3131.3**	(**52.3**%)
Inv. D-Pendulum	6295.4	**7872.6**	**(25.0%)**	4454.1	(± 1098.1)	N/A	**7648.7**	(**71.7**%)
Inv. Pendulum	1000.0	957.2	(-4.2%)	839.7	(± 127.1)	N/A	777.4	(-7.4%)
Reacher	-4.2	-3.9	(0.4%)	-5.47	(± 0.3)	N/A	-10.3	(-8.5%)
Swimmer	43.2	**187.4**	**(333.7%)**	79.1	(± 31.2)	N/A	**102.9**	(**30.2**%)
Walker2d	1090.8	**2405.9**	**(120.5%)**	2300.8	(± 397.6)	~ 2100	**3718.1**	(**61.6**%)
Win — Fair — Lose	N/A	6 — 2 — 0		N/A		N/A	6 — 2 — 0	

Experiments

- FPS Game DOOM

After 24 hours	Arnold	Arnold + VWR
Kills	105	**183**
Frags	87	**173**
K/D Ratio	1.48	**2.08**

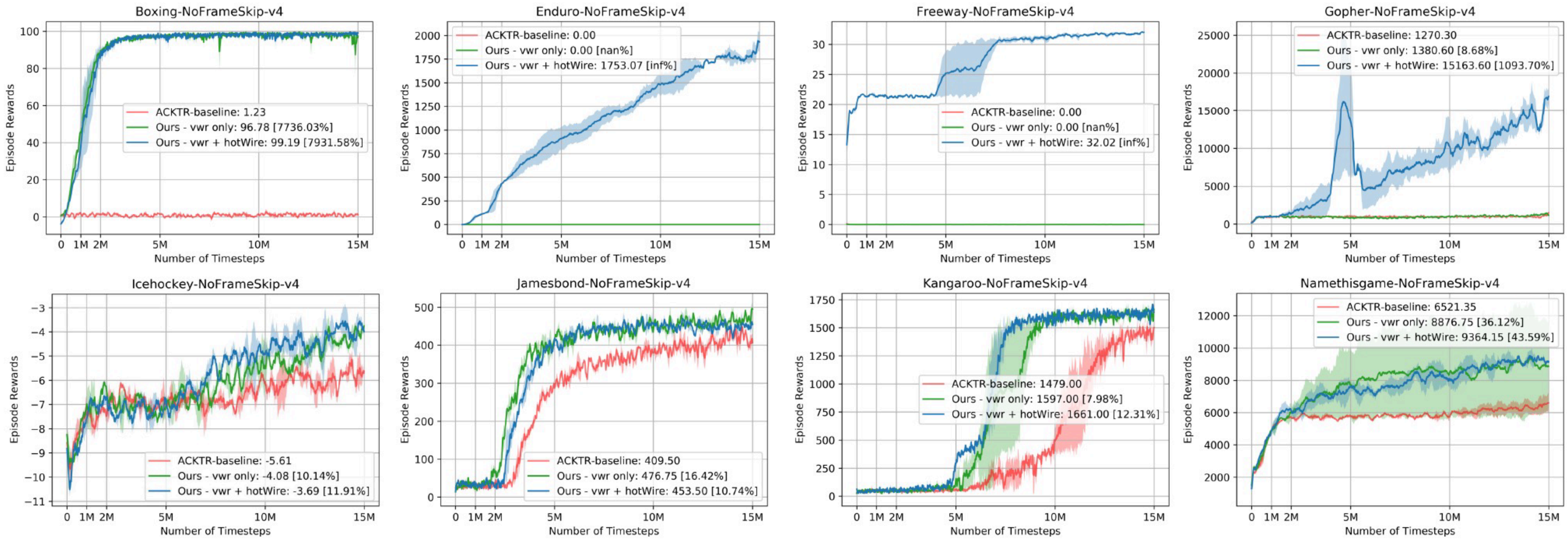
After 50 hours		
Kills	116	**224**
Frags	113	**223**
K/D Ratio	2.00	**2.65**



Experiments

— ACKTR
— A2MC (- hotwire)
— A2MC

- Ablation Study on Hot-Wire Exploration



Demo

- https://youtu.be/zBmpf3Yz8tc

Summary

- We introduce an effective auxiliary reward signal (VWR) that considers both the current reward and the volatility of past rewards.
- The original and auxiliary rewards are trained in a multi-critic manner.
- Extensive experiments in discrete and continuous domains validate the effectiveness of our approach.

Thanks!
Q&A

Project Page

Improving On-Policy Learning with Statistical Reward Accumulation

Yubin Deng, Ke Yu, Dahua Lin, Xiaoou Tang, Chen Change Loy