P3O: Policy-on Policy-off Policy Optimization Supplementary Material

Rasool Fakoor
Amazon Web Services
fakoor@amazon.com

Pratik Chaudhari
Amazon Web Services
prtic@amazon.com

Alexander J. Smola
Amazon Web Services
smola@amazon.com

A Hyper-parameters for all experiments

Table 1: **A2C hyper-parameters on Atari benchmark**

Hyper-parameters	Value
Architecture	conv (32-8 $\times$ 8-4) conv (64-4 $\times$ 4-2) conv (64-3 $\times$ 1-1) FC (512)
Learning rate	7×10^{-4}
Number of environments	16
Number of steps per iteration	5
Entropy regularization (α)	0.01
Discount factor (γ)	0.99
Value loss Coefficient	0.5
Gradient norm clipping coefficient	0.5
Random Seeds	$\{0 \ldots 2\}$

Table 2: **ACER hyper-parameters on Atari benchmark**

Hyper-parameters	Value
Architecture	Same as A2C
Replay Buffer size	5×10^4
Learning rate	7×10^{-4}
Number of environments	16
Number of steps per iteration	20
Entropy regularization (α)	0.01
Number of training epochs per update	4
Discount factor (γ)	0.99
Value loss Coefficient	0.5
importance weight clipping factor	10
Gradient norm clipping coefficient	0.5
Momentum factor in the Polyak	0.99
Max. KL between old & updated policy	1
Use Trust region	True
Random Seeds	$\{0 \ldots 2\}$

Table 3: **PPO hyper-parameters on Atari benchmark**

Hyper-parameters	Value
Architecture	Same as A2C
Learning rate	7×10^{-4}
Number of environments	8
Number of steps per iteration	128
Entropy regularization (α)	0.01
Number of training epochs per update	4
Discount factor (γ)	0.99
Value loss Coefficient	0.5
Gradient norm clipping coefficient	0.5
Advantage estimation discounting factor (τ)	0.95
Random Seeds	$\{0 \ldots 2\}$

Table 4: **P3O hyper-parameters on Atari benchmark**

Hyper-parameters	Value
Architecture	Same as A2C
Learning rate	7×10^{-4}
Replay Buffer size	5×10^4
Number of environments	16
Number of steps per iteration	16
Entropy regularization (α)	0.01
Off policy updates per iteration (ξ)	Poisson(2)
Burn-in period	15×10^3
Samples from replay buffer	6
Discount factor (γ)	0.99
Value loss Coefficient	0.5
Gradient norm clipping coefficient	0.5
Advantage estimation discounting factor (τ)	0.95
Random Seeds	$\{0 \ldots 2\}$

B Comparisons with baseline algorithms

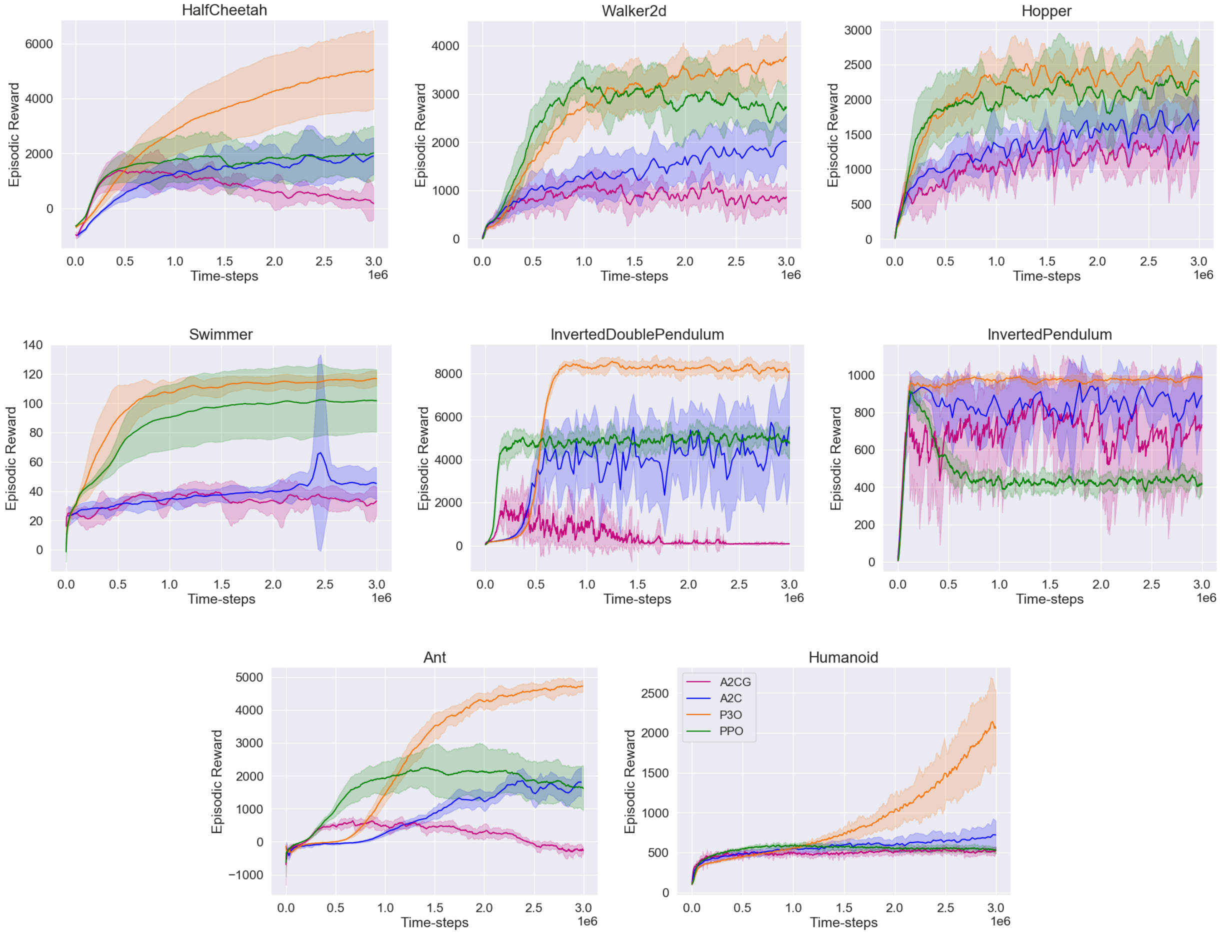

Figure 1: **Training curves of A2C (blue), A2CG [A2C with GAE] (magenta), PPO (green) and P3O (orange) on 8 MuJoCo environments.**

Table 5: **P3O hyper-parameters for MuJoCo tasks**

Hyper-parameters	Value
Architecture	FC(100) - FC(100)
Learning rate	3×10^{-4}
Replay Buffer size	5×10^{3}
Number of environments	2
Number of steps per iteration	64
Entropy regularization (α)	0.0
Off policy updates per iteration (ξ)	Poisson(3)
Burn-in period	2500
Number of samples from replay buffer	15
Discount factor (γ)	0.99
Value loss Coefficient	0.5
Gradient norm clipping coefficient	0.5
Advantage estimation discounting factor (τ)	0.95
Random Seeds	$\{0 \dots 9\}$

Table 6: **A2C (and A2C with GAE) hyper-parameters on MuJoCo tasks**

Hyper-parameters	Value
Architecture	FC(64) - FC(64)
Learning rate	13×10^{-3}
Number of environments	8
Number of steps per iteration	32
Entropy regularization (α)	0.0
Discount factor (γ)	0.99
Value loss Coefficient	0.5
Gradient norm clipping coefficient	0.5
Random Seeds	$\{0 \dots 9\}$

Table 7: **PPO hyper-parameters on MuJoCo tasks**

Hyper-parameters	Value
Architecture	FC(64) - FC(64)
Learning rate	3×10^{-4}
Number of environments	1
Number of steps per iteration	2048
Entropy regularization (α)	0.0
Number of training epochs per update	10
Discount factor (γ)	0.99
Value loss Coefficient	0.5
Gradient norm clipping coefficient	0.5
Advantage estimation discounting factor (τ)	0.95
Random Seeds	$\{0 \ldots 9\}$

Table 8: **Returns on MuJoCo continuous-control tasks after 3M time-steps of training and 10 random seeds**.

Games	A2CG	A2C	PPO	P3O
Half-Cheetah	181.46	1907.42	2022.14	**5051.58**
Walker	855.62	2015.15	2727.93	**3770.86**
Hopper	1377.07	1708.22	2245.03	**2334.32**
Swimmer	33.33	45.27	101.71	**116.87**
Inverted Double Pendulum	90.09	5510.71	4750.69	**8114.05**
Inverted Pendulum	733.34	889.61	414.49	**985.14**
Ant	-253.54	1811.29	1615.55	**4727.34**
Humanoid	530.12	720.38	530.13	**2057.17**

Table 9: **Returns of agents on 49 Atari-2600 games after 28M timesteps (112M frames) of training.**

Games	A2C	ACER	PPO	P3O
Alien	1425.00	2436.20	2260.43	**3124.80**
Amidar	439.43	1393.24	1062.73	**1787.40**
Assault	3897.73	**6996.46**	5941.23	6222.27
Asterix	12272.50	24414.00	7574.33	**25997.00**
Asteroids	2052.27	1874.83	2147.33	**2483.30**
Atlantis	2847251.67	2832752.33	2647593.67	**3077883.00**
BankHeist	910.43	**1281.60**	1236.90	864.03
BattleZone	6250.00	10726.67	**22856.67**	12793.33
BeamRider	5149.29	6486.07	3834.01	**11163.49**
Bowling	24.19	**38.61**	31.75	27.04
Boxing	0.21	99.33	98.06	**99.44**
Breakout	403.25	**474.81**	328.80	351.81
Centipede	3722.24	6755.41	4530.21	**8615.36**
ChopperCommand	1389.67	**10376.00**	9504.33	8878.33
CrazyClimber	111418.67	136527.67	118501.00	**168115.00**
DemonAttack	65766.90	181679.27	37026.17	**331454.95**
DoubleDunk	−17.86	−8.37	−6.29	**−3.83**
Enduro	0.00	0.00	**1092.52**	0.00
FishingDerby	29.54	45.74	29.34	**52.07**
Freeway	0.00	0.00	**32.83**	0.00
Frostbite	269.87	304.23	**1266.73**	312.13
Gopher	3923.13	**99855.53**	6451.07	29603.60
Gravitar	377.33	387.00	**1042.67**	987.50
IceHockey	−6.39	−3.97	−5.11	**−3.50**
Jamesbond	453.83	457.50	**683.67**	475.00
Kangaroo	507.33	1524.67	11583.67	**13360.67**
Krull	8935.40	**9115.73**	8718.40	7812.03
KungFuMaster	25395.00	30002.33	34292.00	**46761.67**
MontezumaRevenge	0.00	0.00	0.00	**805.33**
MsPacman	2220.63	4892.33	3502.20	**7516.21**
NameThisGame	5977.63	**15640.83**	6011.03	9232.70
Pitfall	−65.50	−7.64	**−1.94**	−7.40
Pong	20.21	20.80	20.69	**20.95**
PrivateEye	49.24	**99.00**	97.33	92.61
Qbert	16289.08	22051.67	21830.17	**27619.33**
Riverraid	9680.33	**17794.03**	11841.03	13966.67
RoadRunner	35918.33	40428.67	50663.33	**58728.00**
Robotank	4.30	4.89	18.54	**33.69**
Seaquest	1485.33	1739.87	**1953.53**	1851.87
SpaceInvaders	1894.02	**3140.17**	2124.57	2699.33
StarGunner	55469.33	**65005.00**	63375.67	63905.00
Tennis	−22.22	−11.26	−6.72	**−5.27**
TimePilot	3359.00	7012.00	7535.67	**10789.00**
Tutankham	105.28	**291.09**	206.42	268.24
UpNDown	30932.20	159642.17	173208.13	**279107.53**
Venture	0.00	0.00	**0.00**	0.00
VideoPinball	21061.76	373803.36	220680.47	**377935.99**
WizardOfWor	1256.33	2973.00	5744.67	**10637.33**
Zaxxon	17.00	89.33	8872.67	**16801.33**

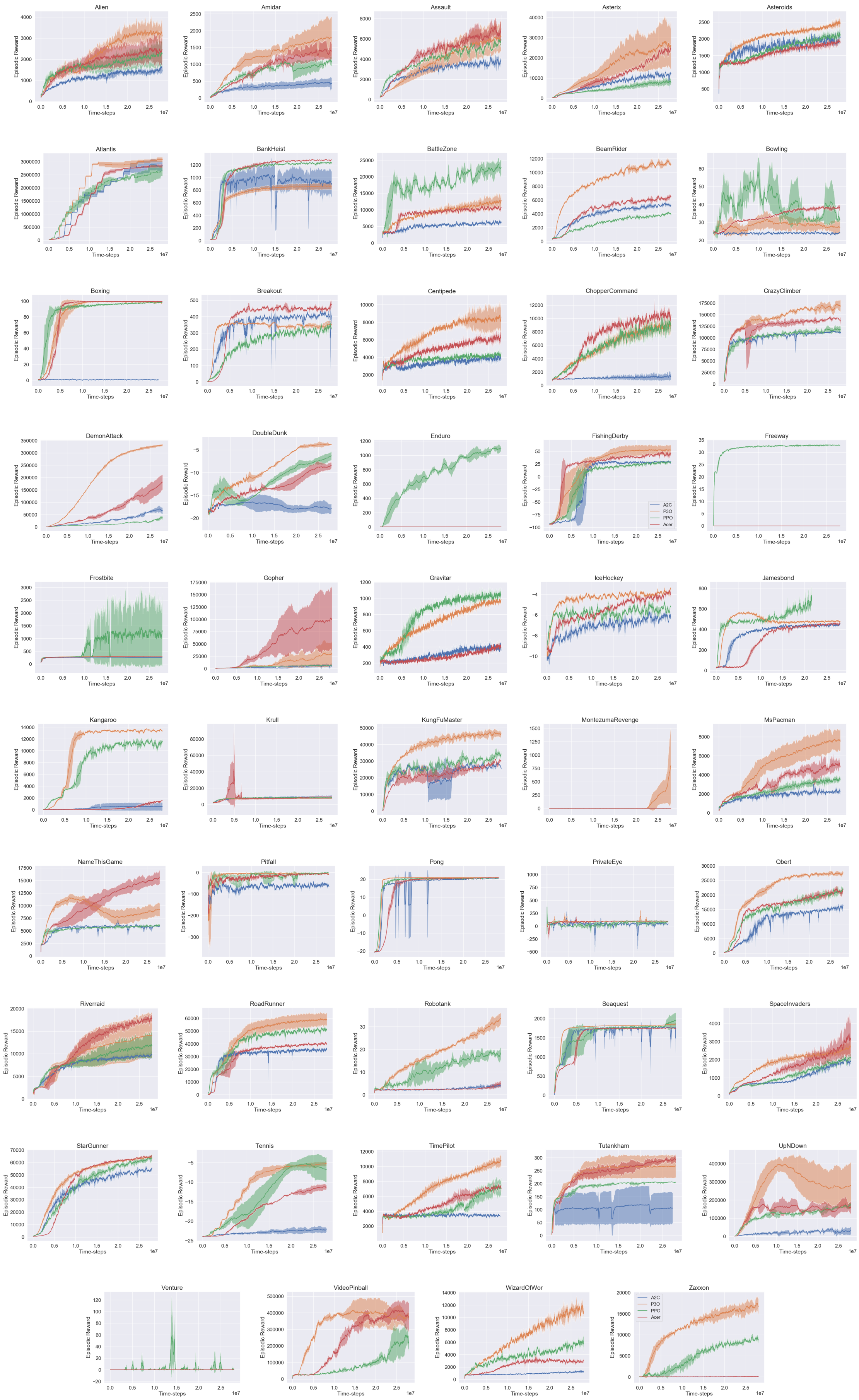

Figure 2: **Training curves of A2C (blue), ACER (red), PPO (green) and P3O (orange) on all 49 Atari games.**