Reward Inflation: A Healthy Stimulus for Reinforcement Learning

Ganghun Lee[1], Minji Kim[1], Minsu Lee[2], and Byoung-Tak Zhang[1]
[1]Seoul National University [2]Sungshin Women's University

Sydney Poster Session 6 Thu 10 Dec, 17:00, Hall 1-4

Project Page

NEURAL INFORMATION PROCESSING SYSTEMS

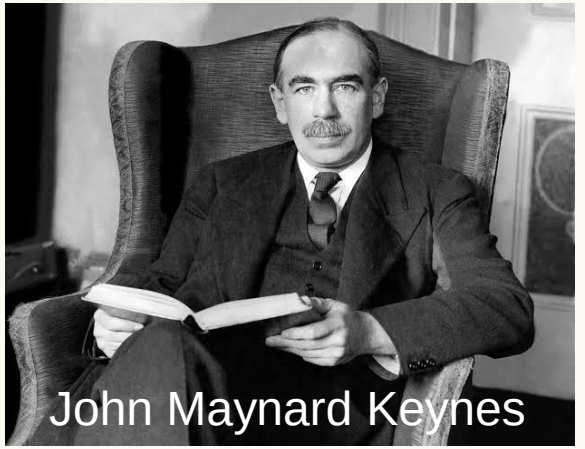
AI Institute | Seoul National University

Motivation

- Moderate inflation could fuel the economy.



≈



Can moderate reward inflation help drive learning?

Method

- **Reward inflation**
  - Steady and exponential growth of reward scale $I(g)$.

g: global training step

Inflation level (reward scale)

$$I(g) := \prod_{k=1}^{g} (1 + \tilde{\rho}(k)), \qquad r'_t(g) := r_t \cdot I(g)$$

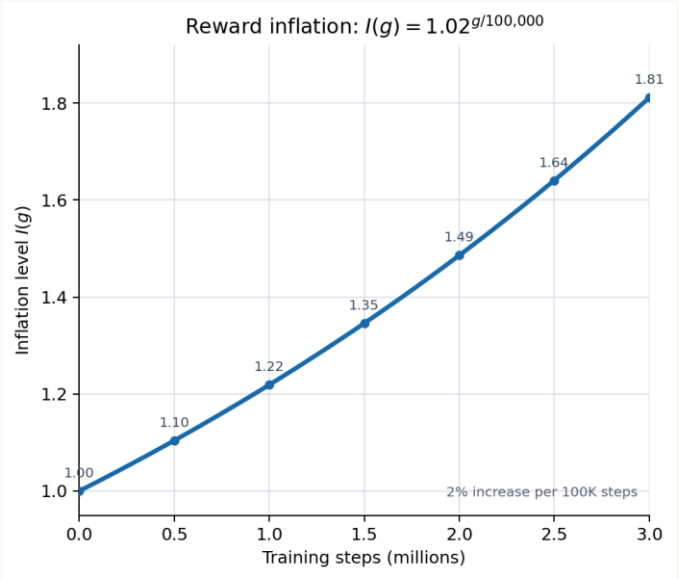
Reward × Inflation level

Inflation rate (stepwise)

$$\tilde{\rho}(g) := (1 + \rho_Y(g; Y))^{\frac{1}{Y}} - 1$$

Inflation rate (per Y steps)





2% inflation per 100K steps

Method

- **Fed**: adaptive reward inflation

  Central bank controls the **interest rate** based on the **state of the economy**.

  Fed controls the **inflation rate** based on the **state of the learning**.



Rule of Thumb:

- **Raise** inflation rate if gradient norm is **decreasing**. Stagnation
- **Reduce** inflation rate if gradient norm is **increasing**. Overheating

Using gradient norm of actor/critic $\mathcal{G}(\phi) = \|\nabla_\phi J(\phi)\|_2$

Fed implementation:

- Inspired by the **Taylor Rule** (originally for interest rate) with a **target inflation**.

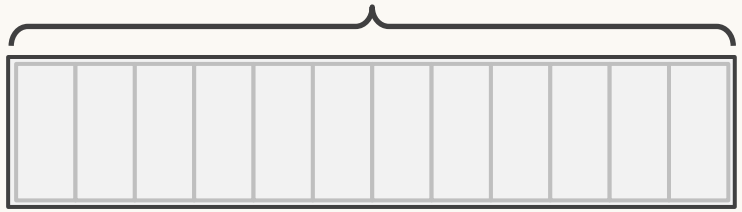
Theoretical effects of reward inflation

- Assumptions

$$r \times I(g), \qquad I(g) > 0$$

No optimality change

$$\bar{I} \coloneqq \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}[I(g)]$$



Limited replay buffer

Theoretical effects of reward inflation

- Effect one: **Recency weighting**

$\Delta\theta(g)$: parameter update of the sample at g
based on TD learning

$$\Delta\theta(g) = \underbrace{\eta\, I(g)\, \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}[r\, \nabla_\theta Q_\theta(s,a)]}_{\text{reward-driven component}} + \underbrace{\eta\, \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left(\gamma \max_{a'} Q(s',a') - Q_\theta(s,a)\right) \nabla_\theta Q_\theta(s,a)\right]}_{\text{bootstrapping component}}.$$

g_1: Old sample
g_2: Recent sample

$$\left\|\Delta\theta(g_2) - \Delta\theta(g_1)\right\| = \left|I(g_2) - I(g_1)\right| \cdot \eta \left\|\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}[r\, \nabla_\theta Q_\theta(s,a)]\right\|.$$

Inflation gap

> Reward inflation **emphasizes** the **recent samples**.

Theoretical effects of reward inflation

- Effect two: **Gradient norm scaling**

Q scale increases with inflation level: $Q_\theta(s,a) \approx \bar{I} \cdot Q^*(s,a)$



Critic gradient $\nabla_\theta L(\theta) = -\mathbb{E}_\mathcal{D}\left[\left(y - Q_\theta(s,a)\right)\nabla_\theta Q_\theta(s,a)\right] \implies \|\nabla_\theta L(\theta)\| \propto \bar{I}^2.$

Actor gradient $\|\nabla_\phi J(\phi)\| \propto \bar{I}.$

> Reward inflation **scales** the **gradient norms**.

Main result

0% 2% Fed

Task	0%	2%	Fed (2%)	Task	0%	2%	Fed (2%)
Alien	1250±352	**1513**±124	1482±96	Frostbite	2732±364	**3377**±543	3172±579
Amidar	296±33	**468**±64	456±83	Gopher	12595±3870	**16160**±4948	15954±5352
Assault	3654±252	**3814**±449	3401±396	Hero	**2758**±1779	2731±624	2549±215
Asterix	22178±7214	**31424**±5245	29492±3913	Jamesbond	486±63	**601**±154	567±120
BankHeist	650±251	776±189	**797**±109	Kangaroo	7821±1762	10885±2742	**11983**±2016
BattleZone	22659±1074	22195±4406	**24501**±945	Krull	7584±2605	7969±2602	**8799**±419
BeamRider	6906±662	**8682**±904	7908±759	KungFu	3853±4542	8257±8977	**12075**±11337
Bowling	22.46±0.63	**23.55**±0.90	23.30±0.50	MsPacman	2218±141	**2365**±102	2181±135
Boxing	91.37±2.06	94.93±1.02	**95.44**±1.00	Pong	20.40±0.1	**20.58**±0.09	20.30±0.40
Breakout	367±19	**375**±14	**375**±21	PrivateEye	**621**±1229	63±117	81±143
ChopperCmd	1178±508	1492±658	**1553**±1301	Qbert	11754±1849	12317±1376	**12831**±1756
CrazyClimbe	93919±8113	103214±10735	**105703**±10671	Riverraid	8228±641	**8699**±1122	8350±840
DemonAttack	32572±9621	50007±6358	**50466**±6846	RoadRunner	34017±1172	35502±6092	**40112**±3054
Enduro	1275±108	1542±165	**1553**±150	Seaquest	4074±632	**5138**±1024	4635±1048
Freeway	23.44±3.58	**24.81**±4.80	23.53±3.73	SpaceInvade	1349±106	**1565**±148	1483±222

ALE (DQN)

+11.53%, +12.59%

2% inflation, Fed

Method	HalfCheetah	Walker2d	Humanoid	Swimmer	Hopper	Ant	Pusher	Reacher	NS
-2%	14020±1082	4342±461	5300±176	92±43	2421±460	3929±1090	-24.25±4.10	-3.46±0.23	36.8
0%	15241±734	4399±527	5826±313	116±36	2869±299	4763±1642	-23.28±2.89	**-3.41**±0.23†	68.6
1%	**15707**±715†	5007±560	5859±363	121±25	2821±406	5245±1312†	-22.71±2.87	-3.42±0.23	79.8
2%	15607±372	5102±520†	5861±199	136±11†	3040±499	4346±1267	-22.07±1.58†	-3.42±0.24	81.3†
3%	15091±927	4690±624	5873±306†	126±11	3035±445	4392±1298	-22.58±1.58	-3.44±0.24	71.5
5%	14484±996	4615±518	5545±492	131±12	3161±346	4795±1089	-22.20±0.89	-3.45±0.23	71.5
10%	14186±1200	4167±650	5209±1241	129±19	**3166**±288†	4248±1219	-24.60±1.47	-3.54±0.33	46.6
Fed (2%)	15621±991	**5395**±452	**6048**±436	**143**±13	3105±293	5180±1352	**-21.99**±1.23	-3.43±0.24	**92.8**

MuJoCo (SAC)

+18.51%, +35.21%

2% inflation, Fed (NS)

Main result

-2% 0% 10%

Method	HalfCheetah	Walker2d	Humanoid	Swimmer	Hopper	Ant	Pusher	Reacher	NS
-2%	14020±1082	4342±461	5300±176	92±43	2421±460	3929±1090	-24.25±4.10	-3.46±0.23	36.8
0%	15241±734	4399±527	5826±313	116±36	2869±299	4763±1642	-23.28±2.89	**-3.41**±0.23†	68.6
1%	**15707**±715†	5007±560	5859±363	121±25	2821±406	5245±1312†	-22.71±2.87	-3.42±0.23	79.8
2%	15607±372	5102±520†	5861±199	136±11†	3040±499	4346±1267	-22.07±1.58†	-3.42±0.24	81.3†
3%	15091±927	4690±624	5873±306†	126±11	3035±445	4392±1298	-22.58±1.58	-3.44±0.24	71.5
5%	14484±996	4615±518	5545±492	131±12	3161±346	4795±1089	-22.20±0.89	-3.45±0.23	71.5
10%	14186±1200	4167±650	5209±1241	129±19	**3166**±288†	4248±1219	-24.60±1.47	-3.54±0.33	46.6
Fed (2%)	15621±991	**5395**±452	**6048**±436	**143**±13	3105±293	5180±1352	**-21.99**±1.23	-3.43±0.24	**92.8**
2% (lr)	14566±1006	4922±515	5417±298	77±41	2899±359	4458±1888	-23.95±2.88	-3.42±0.24	59.2
PER	14734±1934	4797±592	4542±2290	57±23	2683±529	**5762**±979	-26.41±2.98	-3.43±0.24	48.7
RN	15022±436	5036±466	5578±285	118±13	3149±272	5017±1666	-29.17±13.37	-3.46±0.24	64.1
RTN	11877±369	4454±571	4451±633	87±21	2742±463	4046±1603	-23.41±2.05	-3.45±0.13	32.1
RWS	15482±687	5000±546	5730±362	104±40	2923±469	5466±1318	-23.32±3.55	-3.43±0.20	76.8

MuJoCo (SAC)

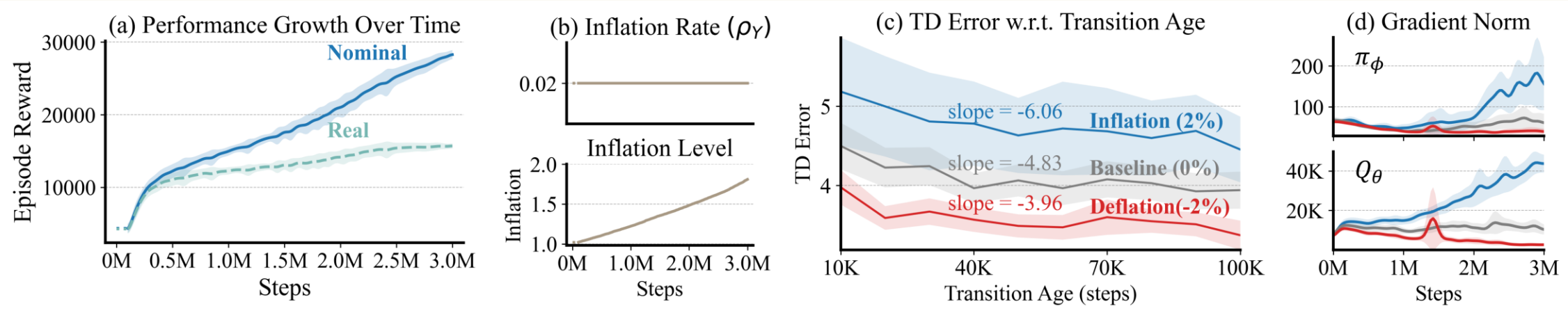
-46.43%, **-35.16%**

2% deflation, 10% inflation

Modest reward inflation can benefit RL.
However, **deflation** and **excessive inflation** generally hurt.

Empirical observations

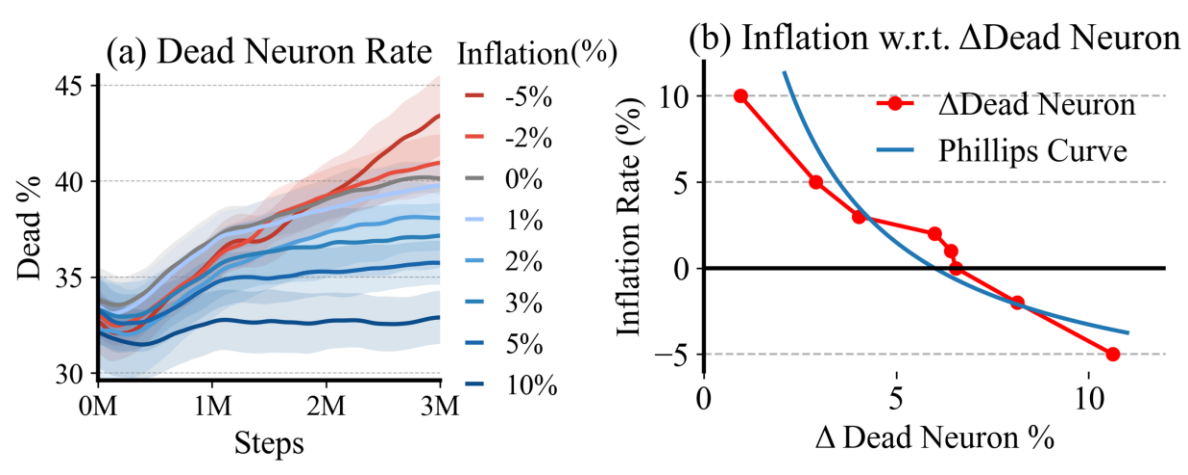
- Empirical cues for theoretical claims

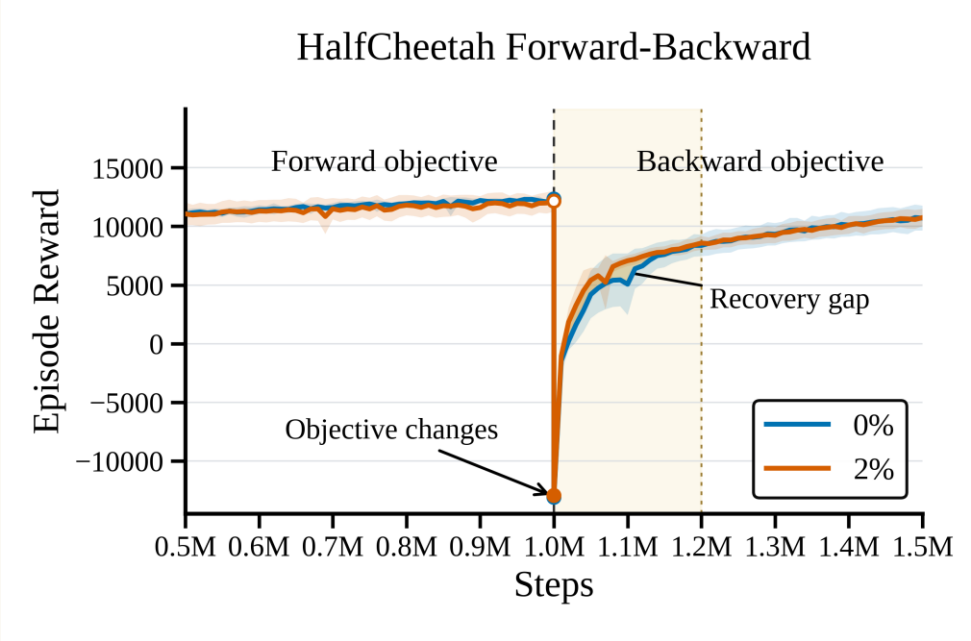


Recency weighting: Reward inflation scales the TD error slope by sample recency.
Gradient norm scaling: Reward inflation scales the gradient norm of actor/critic.

Empirical observations

- Effects on **neural plasticity**





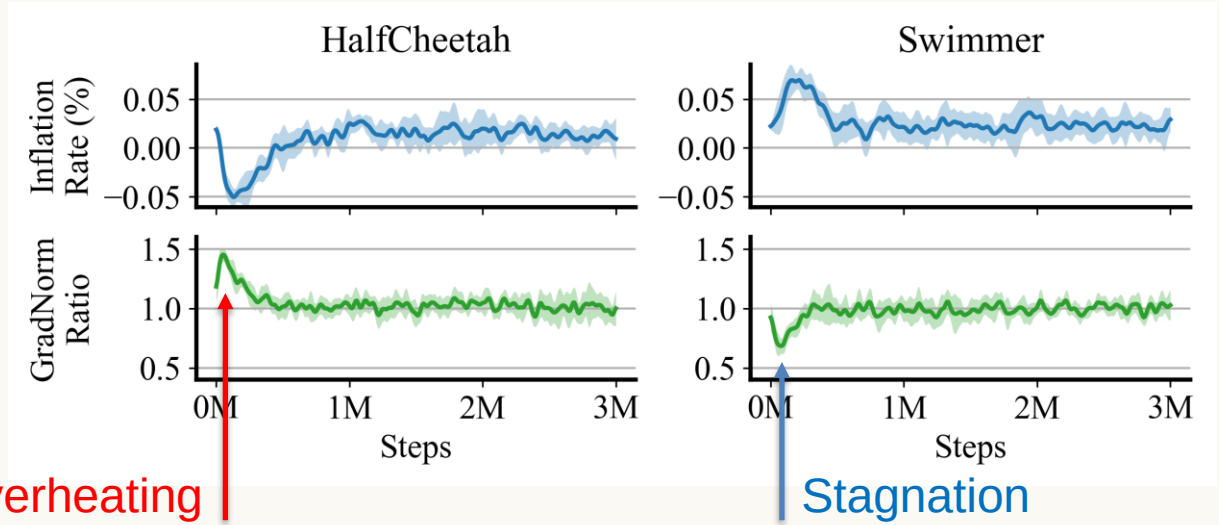
Plasticity curve: Higher inflation slows the increase in dormant neurons.
However, less loss of plasticity does not mean better performance.

What else we found

- Opposite behaviors in Fed



Fed



HalfCheetah's gradnorm peaks → Fed reduces inflation rate → gradient norm attenuated

Swimmer's gradnorm bottoms out → Fed raises inflation rate → gradient norm recovered

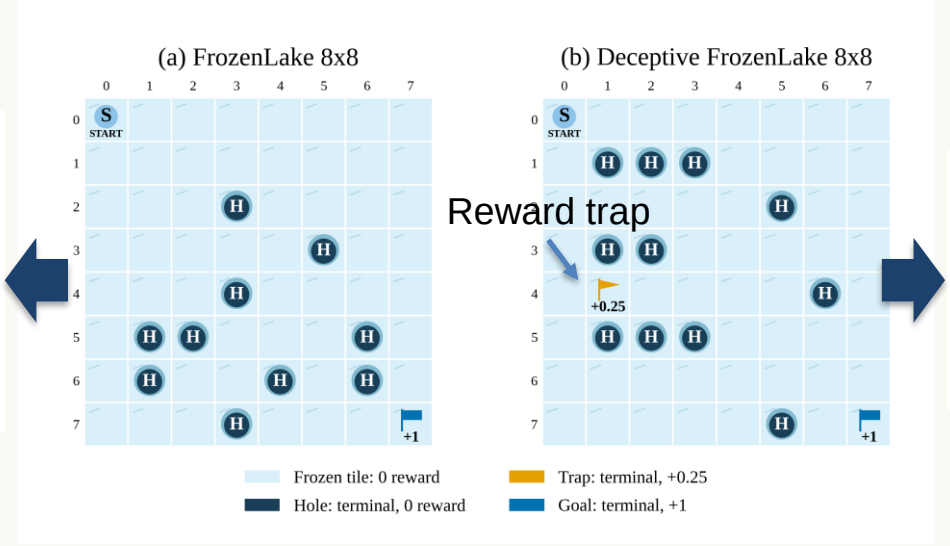
> Fed **monitors gradient norm** and **intervenes by adjusting inflation rate**.

What else we found

- **Speed-risk trade-off** in reward inflation
  - Tabular-Q learning

Inflation	TD-error slope	Success rate	AUC
0%	-2.039×10^{-4}	0.478 ± 0.132	0.407 ± 0.014
1%	-7.645×10^{-4}	0.483 ± 0.166	0.408 ± 0.015
5%	-3.220×10^{-3}	0.520 ± 0.109	0.422 ± 0.015
10%	-6.464×10^{-3}	0.507 ± 0.109	0.431 ± 0.018
50%	-2.600×10^{-2}	0.584 ± 0.047	0.475 ± 0.015
100%	-3.753×10^{-2}	0.585 ± 0.030	0.482 ± 0.015
500%	-8.895×10^{-2}	0.586 ± 0.033	$\mathbf{0.484} \pm 0.014$
1000%	-1.231×10^{-1}	$\mathbf{0.594} \pm 0.031$	$\mathbf{0.484} \pm 0.018$
5000%	-1.930×10^{-1}	0.574 ± 0.035	0.474 ± 0.019

Aggressive inflation boosted learning



Trapped rate

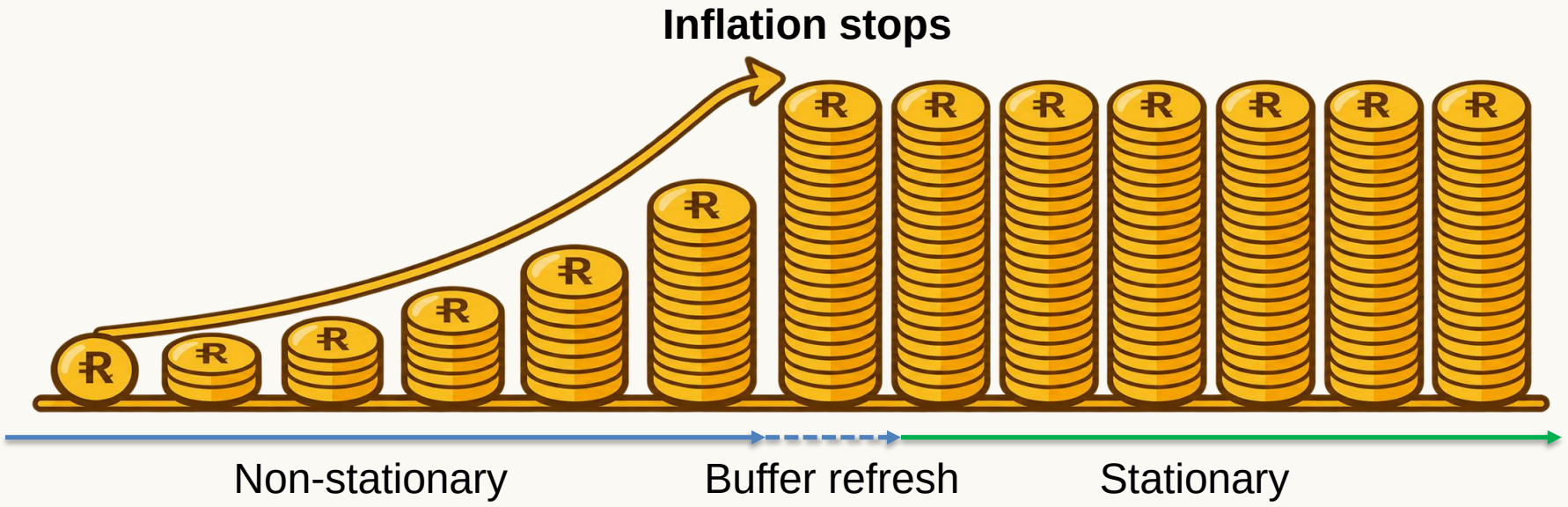
Inflation	Low exploration	Higher exploration
0%	18.8%	0.0%
1%	15.6%	0.0%
5%	31.2%	0.0%
10%	28.1%	0.0%
50%	93.8%	59.4%
100%	100.0%	96.9%

Aggressive inflation made reward traps easier to fall into

Reward inflation may need **a larger exploration budget** for **deceptive environments**.

What else we found

- Convergence discussion



Ongoing inflation can induce non-stationary learning.
Stopping inflation restore stationarity allowing standard convergence conditions.

Limitations

1. Inflation-rate sensitivity
2. Fed heuristics
3. Evaluation scope

Takeaways

- Modest reward inflation in RL can benefit performance and plasticity.
- More theoretically grounded methods and broader experiments are needed.
- The code can be found in https://github.com/northadventure/RewardInflation.

Economic growth and RL may share a similar dynamic.

Reward Inflation: A Healthy Stimulus for Reinforcement Learning

zxc8594@snu.ac.kr

Sydney Poster Session 6 Thu 10 Dec, 17:00, Hall 1-4

Project Page