Expected Return causes Mode collapse in RL

Fix this by Inverse Probability Scaling of Rewards

Abhijeet Sinha, Sundari Elango, Dianbo Liu

NUS National University of Singapore

ICML International Conference On Machine Learning

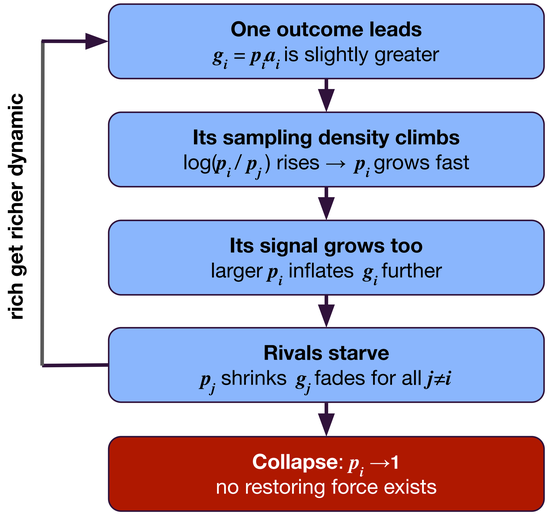
THE PROBLEM

Many RL problems have **multiple equally good outcomes** like diverse molecules, diverse proofs, diverse hypotheses not a single best answer.
Standard RL policies collapse onto a **narrow subset** of outcomes, even when many others are valid and equally rewarding which is usually blamed on weak exploration or regularization.

WHAT CAUSES THIS?

$$J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[r(\phi(\tau))] = \sum_{o\in\mathcal{O}} p_\theta(o)\, r(o)$$

$$\frac{d}{dt}\log\frac{p_i(t)}{p_j(t)} = p_i(t)\, a_i(t) \;-\; p_j(t)\, a_j(t)$$



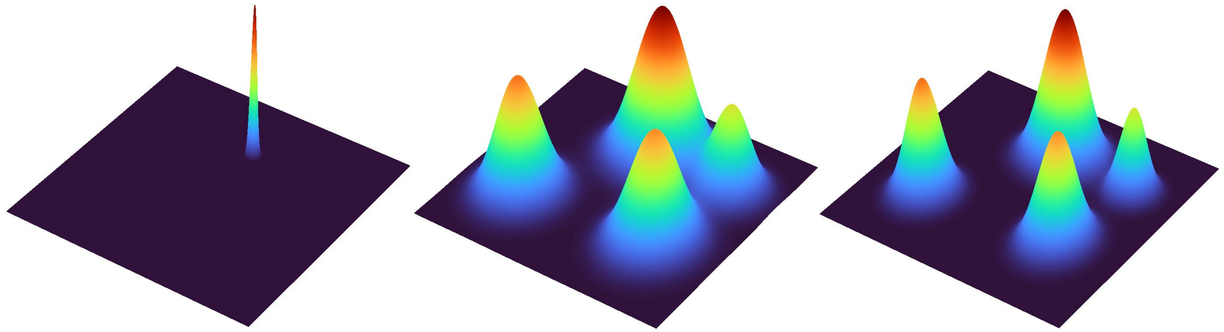
IPS: Inverse Probability Scaling

$$\tilde{r}_\theta(o) \;:=\; \frac{r(o)}{p_\theta(o)}$$

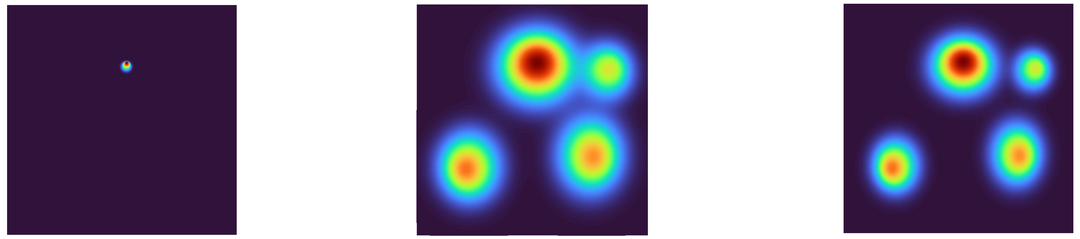
$$J_{\mathrm{IPS}}(\theta) \;=\; \mathbb{E}_{o\sim p_\pi}[\tilde{r}(o)] = \mathbb{E}_{o\sim p_\pi}\left[\frac{r(o)}{p_\theta(o)}\right]$$

$$\frac{d}{dt}\log\frac{p_i(t)}{p_j(t)} = \big(r(i) - r(j)\big) - \big(p_i(t) - p_j(t)\big)\sum_{k=1}^{K} r(k)$$

$$p_i^\star \;=\; \frac{r(i)}{\sum_{k=1}^{K} r(k)}$$



Regular RL Sampling | **Reward** | **IPS - RL Sampling**



IPS is a drop-in reward replacement fully compatible with regular RL algorithms, such as GRPO and REINFORCE, that don't require a separate critic model.
However, this method cannot replace algorithms like PPO that are reliant on a separate critic model.

RESULTS

HYPOSPACE

Method	Causal	3D	Boolean
GRPO	16.0%	26.9%	9.5%
FlowRL	12.8%	26.9%	10.7%
IPS-GRPO	**43.9%**	**90.0%**	**60.7%**

DRUG DISCOVERY

	Threshold	Algorithm	Yield	OB100
SYNTH	0.80	REINVENT	5695	605
		IPS-GRPO	**7371**	**554**
	0.85	REINVENT	418	2934
		IPS-GRPO	**1580**	**1178**
ALL-AMIDE	0.80	REINVENT	1	>5000
		IPS-GRPO	**1478**	**773**
	0.85	REINVENT	0	>5000
		IPS-GRPO	**20**	**>5000**