Exhibit Hall 1 # 311

Flipping Coins to Estimate Pseudocounts for Exploration in Reinforcement Learning

Sam Lobel*, Akhil Bagaria* & George Konidaris
Brown University

BROWN

Exploration is Key to Scaling Reinforcement Learning

- One of the biggest open problems in RL
- Bonus-based exploration via novelty search

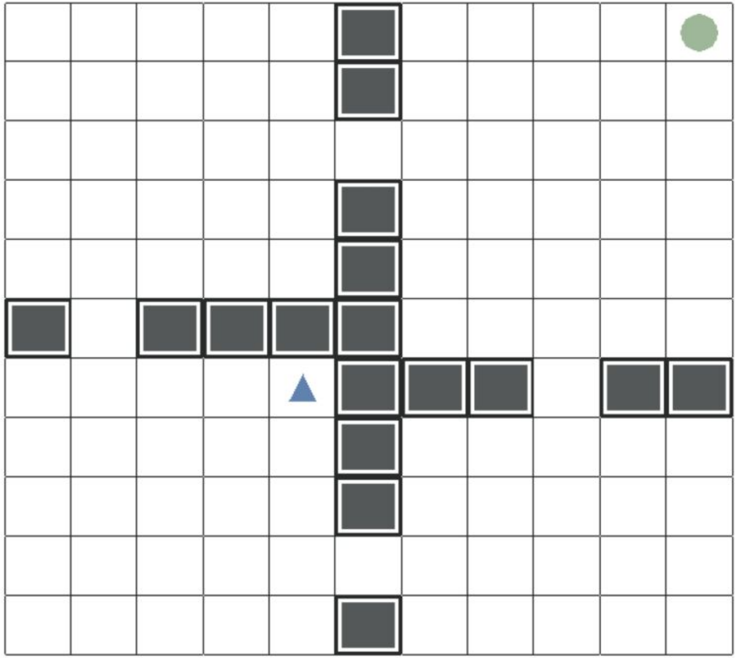
Sparse Reward

Open Ended

Optimal Exploration in Tabular Domains

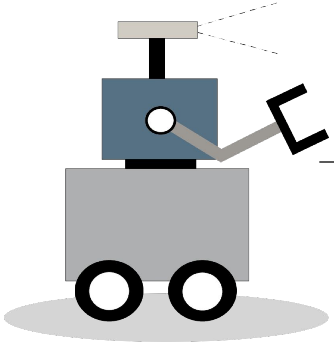
$$R(s,a) = R_e(s,a) + \mathcal{B}(s,a)$$

$$\mathcal{B}(s,a) \propto \sqrt{\frac{1}{\mathcal{N}(s,a)}}$$

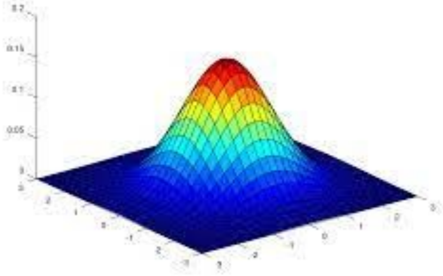


Count-based bonus leads to optimal policies in poly time

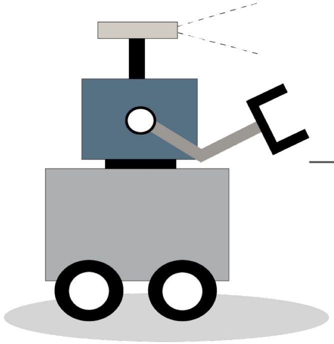
Current Approach to Pseudocounts: Density Modeling



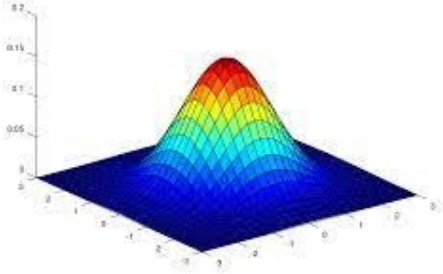
$$\xrightarrow{p_\phi(s)}$$



$$\boxed{n(s) = \text{Algebra}\Big(p_{\phi'}(s), p_\phi(s)\Big)}$$



$$\xrightarrow{p_{\phi'}(s)}$$



Current Approach to Pseudocounts: Density Modeling

Existing methods track **changes in probability density** and place strong **restrictions** on the density models:

- **Training**: fully online, learning positive, update on a state once
- **Architectural**: normalized probability (no GANs, VAEs etc)

Count-based exploration can be improved by computing pseudocounts directly
& under a less restrictive setting

Using Randomness to Extract Counts

Counts *naturally emerge* from coin-flip distributions made on state visitations

$$c_i \sim \{\text{-}1, 1\}^d$$

$f^*_\theta($  $) = \quad 1 \qquad 0 \quad \xrightarrow{\frac{1}{d}E\left[||v||^2\right]} \quad \frac{1}{2}$

An Objective Function for Tracking Counts

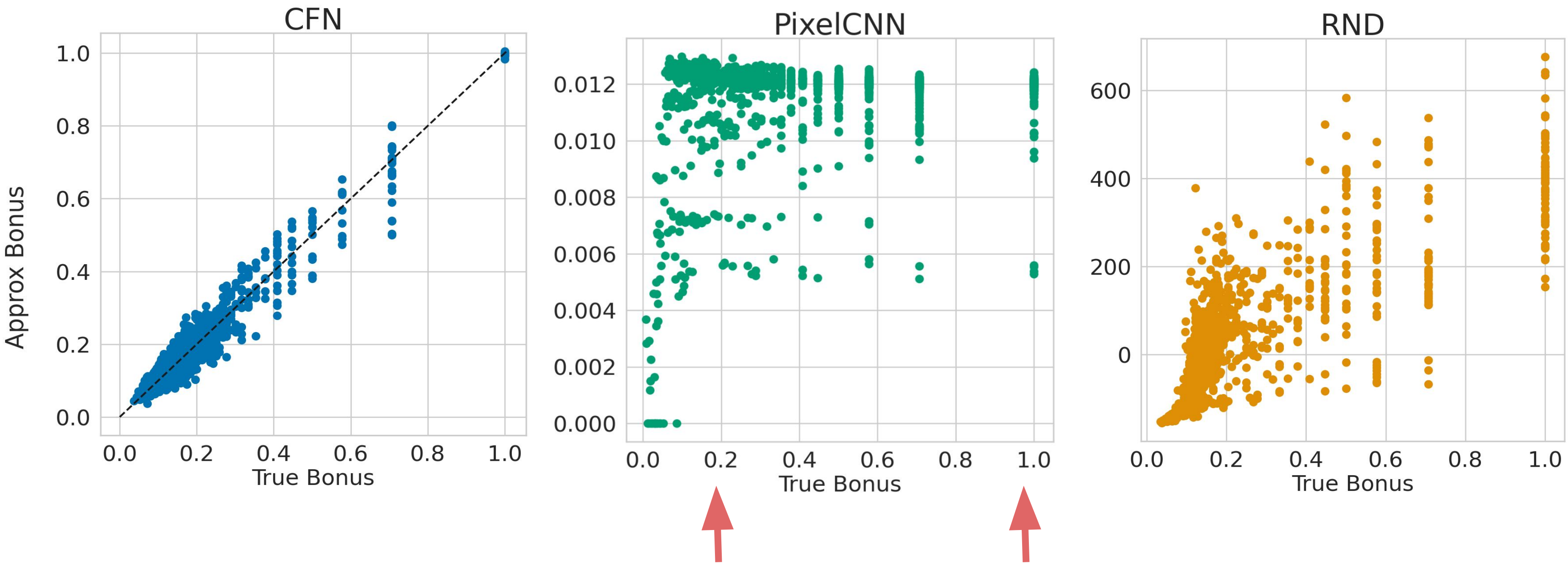
- MSE rephrases averaging as an optimization problem
- Func approximator to map states to coin-flip vectors

$$f^*(s) = \arg\min_f \sum_{i=1}^{m} \|v_i - f(s_i)\|^2 = \arg\min_f \sum_{i=1}^{m} \sum_{j=1}^{d} (v_{ij} - f(s_i)_j)^2$$

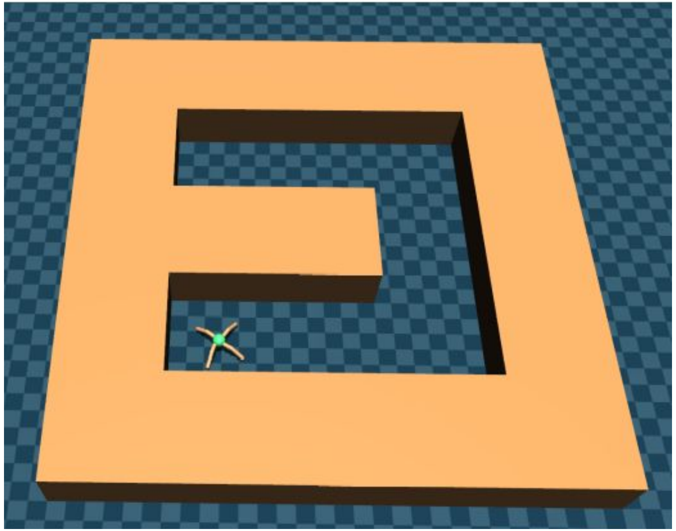
$$\|f^*(s)\| \approx \sqrt{\frac{1}{n}}$$

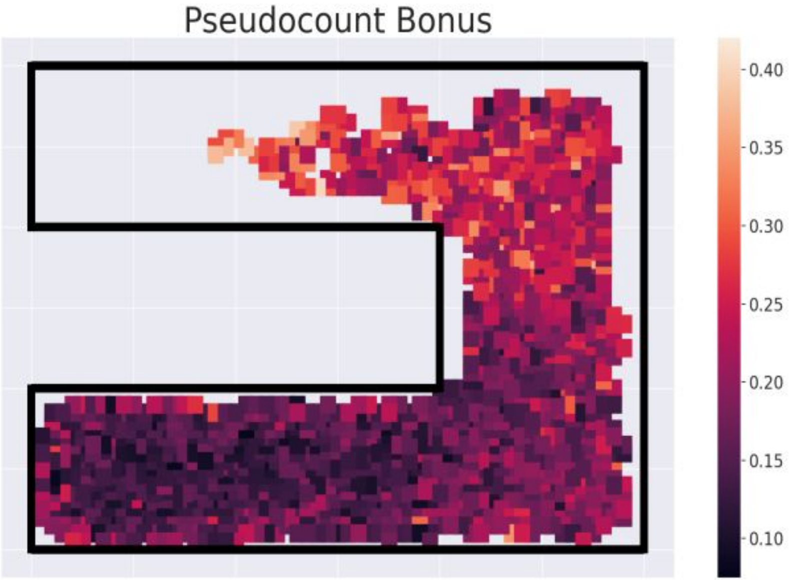
- Standard supervised learning (regression) objective

How Accurate are the Pseudocounts?



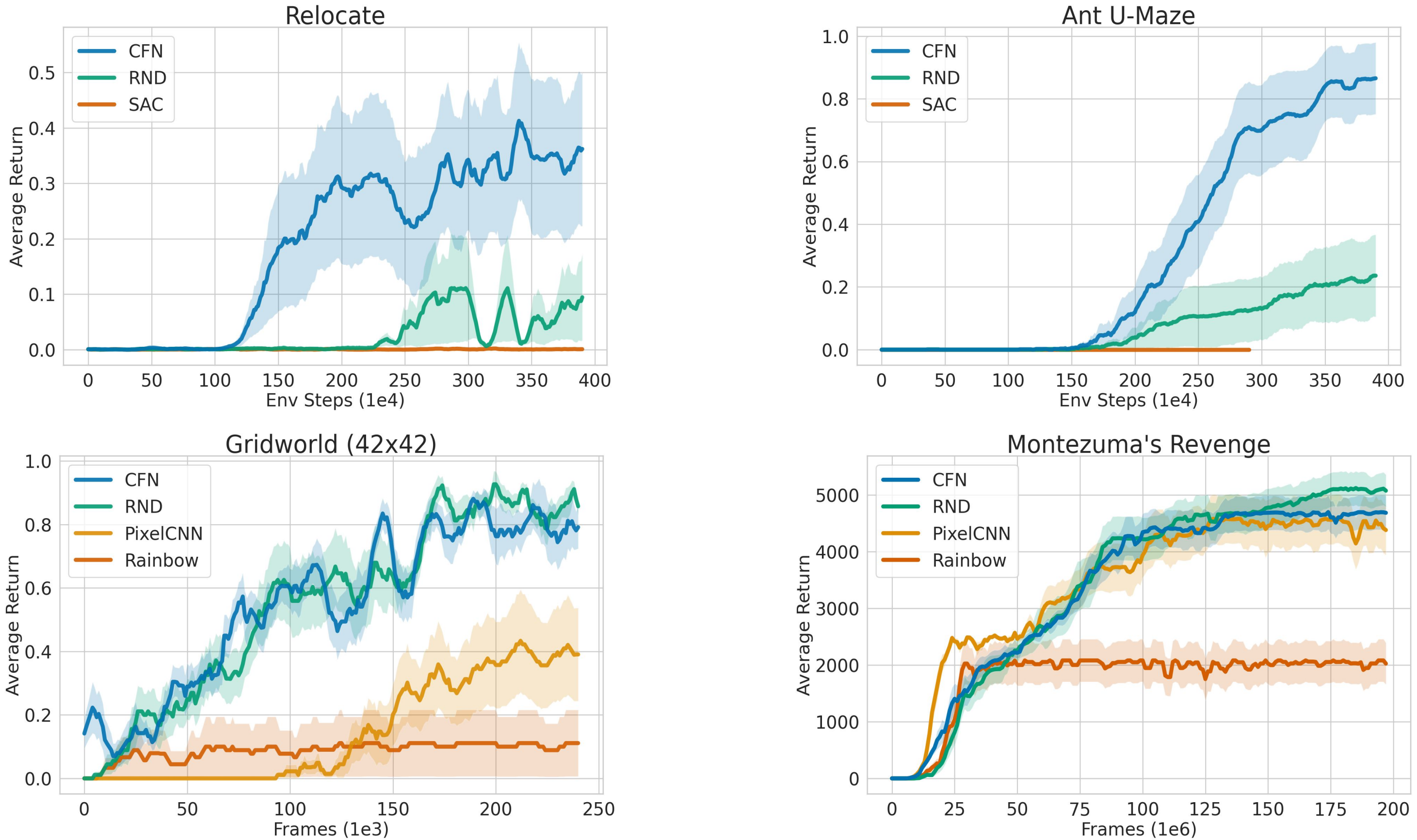
What does the Exploration Bonus look like?



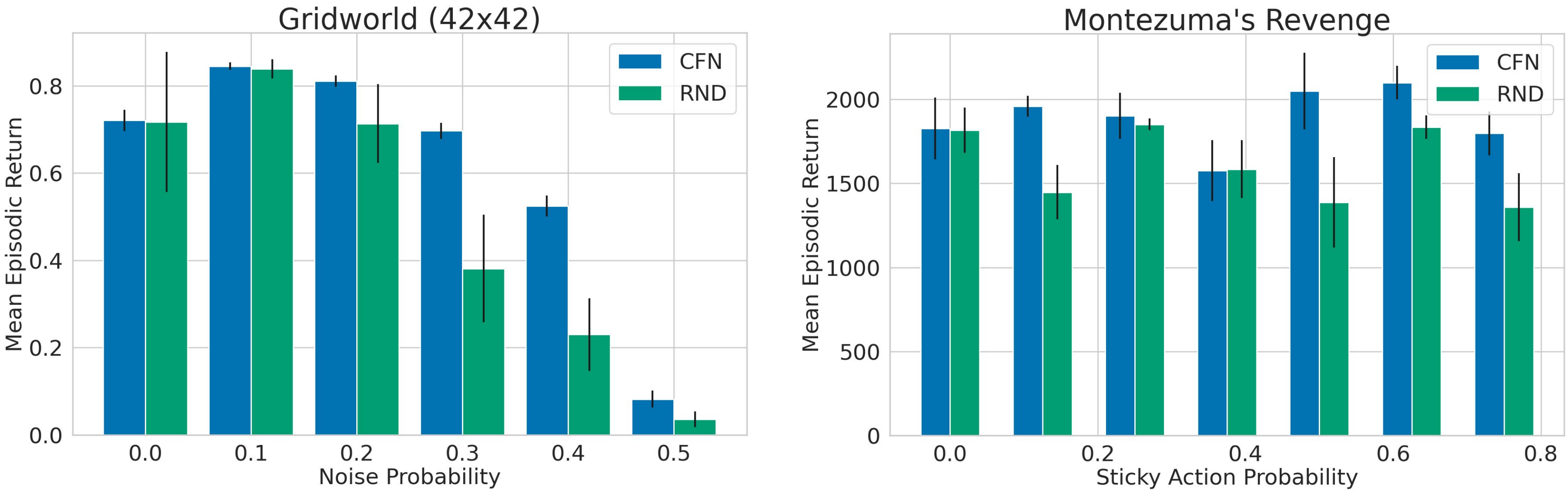


Exploration bonus attracts the agent to the frontier

Better Bonus Leads to Better RL



Pseudocounts are More Robust to Stochasticity



Gradually decaying bonus is crucial in stochastic domains

Conclusion

- Pseudocounts **without** density modeling
- Counts **emerge** from the **sampling distribution** of Bernoulli trials
- Standard **supervised learning objective**; use favorite **DL tricks** and **representations**
- CFN gets **accurate counts**
- Better **RL performance** than existing methods

BROWN

Exhibit Hall 1 # 311

AI