# Incentivizing Exploration in Reinforcement Learning


Max Simchowitz (MSR) msimchow@mit.edu

Alex Slivkins (MSR) slivkins@microsoft.com

(work conducted during internship at MSR)


**Episodic Tabular MDPs:** state **x**, reward **r**, step **h**, action **a** horizon **H**, #states = **S**, #actions = **A**,

$$\textbf{value}(\pi;\mu) := \mathsf{E}^{\pi}_{\mu}\left[\sum_{h=1}^{H}\mathbf{r}_h\right] = \mathsf{E}^{\pi}_{\mu}\left[\sum_{h=1}^{H}\mathsf{r}_{\mu}(\mathbf{x}_h,\mathbf{a}_h,h)\right].$$

policy · MDP instance · mean rewards

expectation under transition $x_{h+1} \sim \mathsf{p}_{\mu}(\cdot \mid x_h, a_h)$

**Typical Goal:** find $\pi$ such that $\textbf{value}(\pi;\mu) \geq \max_{\pi'} \textbf{value}(\pi';\mu) - \epsilon$.

(requires learning **rewards + transitions**)

## Episodic RL with Strategic Agents

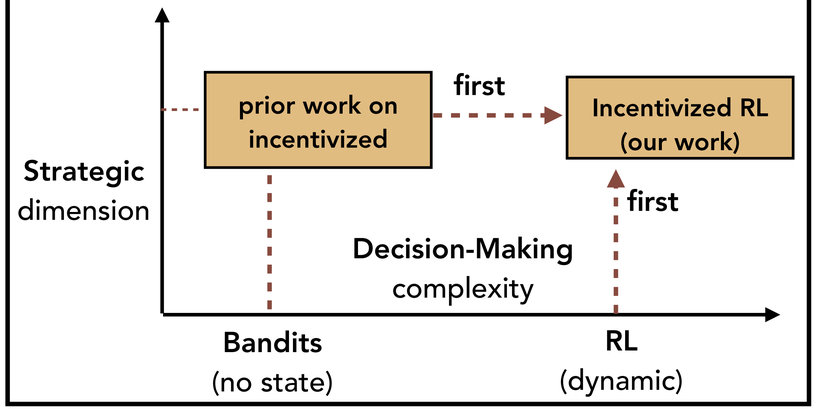
1. the principal chooses a signal $\sigma_k$;
2. agent $k$ arrives, observes $k$ and $\sigma_k$, and chooses a policy $\pi_k$;
3. this policy is executed in the MDP;
4. the agent receives reward $\sum_{h\in[H]} r_{k,h}$;
   the principal observes the resultant trajectory $\tau_k = (\, x_{k;h},\, a_{k;h},\, r_{k;h},\, h\,)_{h\in[H]}$.

Agents select **Bayes-greedy policy** under **shared prior**, with knowledge of principal's signal-generating mechanism

$$\pi_k \in \arg\max_{\pi\in\Pi_{\text{mkv}}} \mathbb{E}\left[\,\textbf{value}(\pi;\mu_\star) \mid \sigma_k\,\right], \qquad \mu \sim \mathbf{p}$$

**New Goal: Design a signal-revealing mechanism to to `explore MDP' by learning rewards + transitions**

## Where we fit in




## Results Summary

**Our result:** Sample complexity for ***'natural exploration objective'***
a. qualitatively matches prior art for simpler **Incentivized Exploration** settings (bandits)
b. **exponentially** improves over naive reduction from bandits

(Slightly More) **Formally…**

**Defn.** $(\rho, N)$-exploration: Visit all "$\rho$-reachable" triples $(x, a, h)$ at least $N$ times

some policy visits that triple with prob. at least **ρ**

**Thm.** We perform $(\rho, N)$-exploration in $K \approx \tilde{O}(N)\cdot C_{\rho}^{-SAH}\cdot \text{poly}(S, A, H, 1/\rho)$ episodes

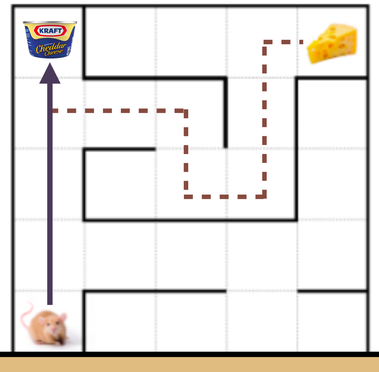
prior-dependent constant typically scaling as 1/**ρ**

Exponential dependence in **SAH** is large for **non-strategic RL**, but qualitatively matches exp(# arms) bounds for **bandits** (no state)

**Cor.** $\epsilon$-suboptimal policy $\pi$ for $K \approx \epsilon^{-O(SAH)}$.

## Pitfalls of Full-History Revelation

Fake "Cheese"
Prior Reward: 50
Realized Reward: **15**



Real Cheese
Prior Reward: 10
Realized Reward: **100**

Agent **exploits** (and only gets) the **fake cheese.**

Can happen even with **constant probability** under **well-specified** priors.

## The Mechanism

'Implement' classical RL Algorithm (E3) using a novel signaling policy we call 'Hidden Hallucination'

1. **Rarely** signal a 'hallucinated' history which makes unexplored states look **as if they have small reward**
2. **Key Challenge:** getting the agent to 'trust' the hallucinated history (*rarity helps!*).
3. **Principal Learns!** While compatible with agent beliefs

**Some Key Objects:**

1. a *ledger* $\lambda$ reveals (partial) history of trajectories
2. *honest ledger* $\lambda_{\text{hon};\ell}$ reveals all history from exploration phases $\ell' < \ell$
3. *censored ledger* $\lambda_{\text{hon};\ell}$ reveals all transition history from exploration phases $\ell' < \ell$, but no reward data
4. hallucinated ledger $\lambda_{\text{hal};\ell}$ - synthetic history created by changing *reward data* in honest ledger.
5. *punishing-event* $\mathcal{E}_{\text{pun},\ell}$ on which all rewards at under-explored triples $(x, a, h)$ look "small'

**Algorithm 1** `HiddenHallucination`

**Input:** phase length $n_{\text{ph}}$, target #samples $n_{\text{lrn}}$, punishment parameter $\varepsilon_{\text{pun}} > 0$
**for** each phase $\ell = 1, 2, \ldots$ **do**
  $\textbf{phase}_\ell = (\ell-1)n_{\text{ph}} + [n_{\text{ph}}] \subset \mathbb{N}$ `% the next` $n_{\text{ph}}$ `episodes`
  Draw "hallucination episode" $k^{\text{hal}}_\ell$ uniformly from $\textbf{phase}_\ell$
  **for** each episode $k \in \textbf{phase}_\ell$ **do**
    **if** $k = k^{\text{hal}}_\ell$ **then** `% hallucination episode`
      `% censored ledger` $\lambda_{\text{cens};\ell}$, `punish-event` $\mathcal{E}_{\text{pun},\ell}$
      Define distribution $\mathbf{p}_{\text{hal},\ell}$ over MDP models by
        $\mathbf{p}_{\text{hal},\ell}(\cdot) := \mathbb{P}\left[\,\mu_\star \in \cdot \mid \lambda_{\text{cens};\ell},\ \mathcal{E}_{\text{pun},\ell}\,\right]$
      Draw hallucinated MDP $\mu_{\text{hal};\ell}$ at random from $\mathbf{p}_{\text{hal},\ell}$
      For each fully-explored $(x, a, h)$ triple, `% hallucinate rewards`
        each time this triple appears in the ledger,
        draw its reward as prescribed by $\mu_{\text{hal};\ell}$.
      Reveal *hallucinated ledger* $\lambda_{\text{hal};\ell}$ formed by
        inserting the hallucinated rewards into censored ledger $\lambda_{\text{cens};\ell}$.
    **else** `% exploitation`
      Reveal honest ledger: $\lambda_k \leftarrow \lambda_{\text{hon};\ell}$.
    Observe the trajectory $\tau_k$ from this episode.

**Some Key Ideas in Analysis:**
1. **'Canonical Probabilities'** that 'ignore' the algorithm
2. **Ledger Hygiene**: Ledgers that enable canonical posteriors
3. A general purpose one-step '**hidden hallucination**' lemma to capture one round of the algorithm.