Google DeepMind

Curiosity in Hindsight: Intrinsic Exploration in Stochastic Environments

Daniel Jarrett, Corentin Tallec, Florent Altché, Thomas Mesnard, Rémi Munos, Michal Valko

LINK TO PAPER

LINK TO PAPER

Introduction

How to explore the world when external rewards are sparse or absent?

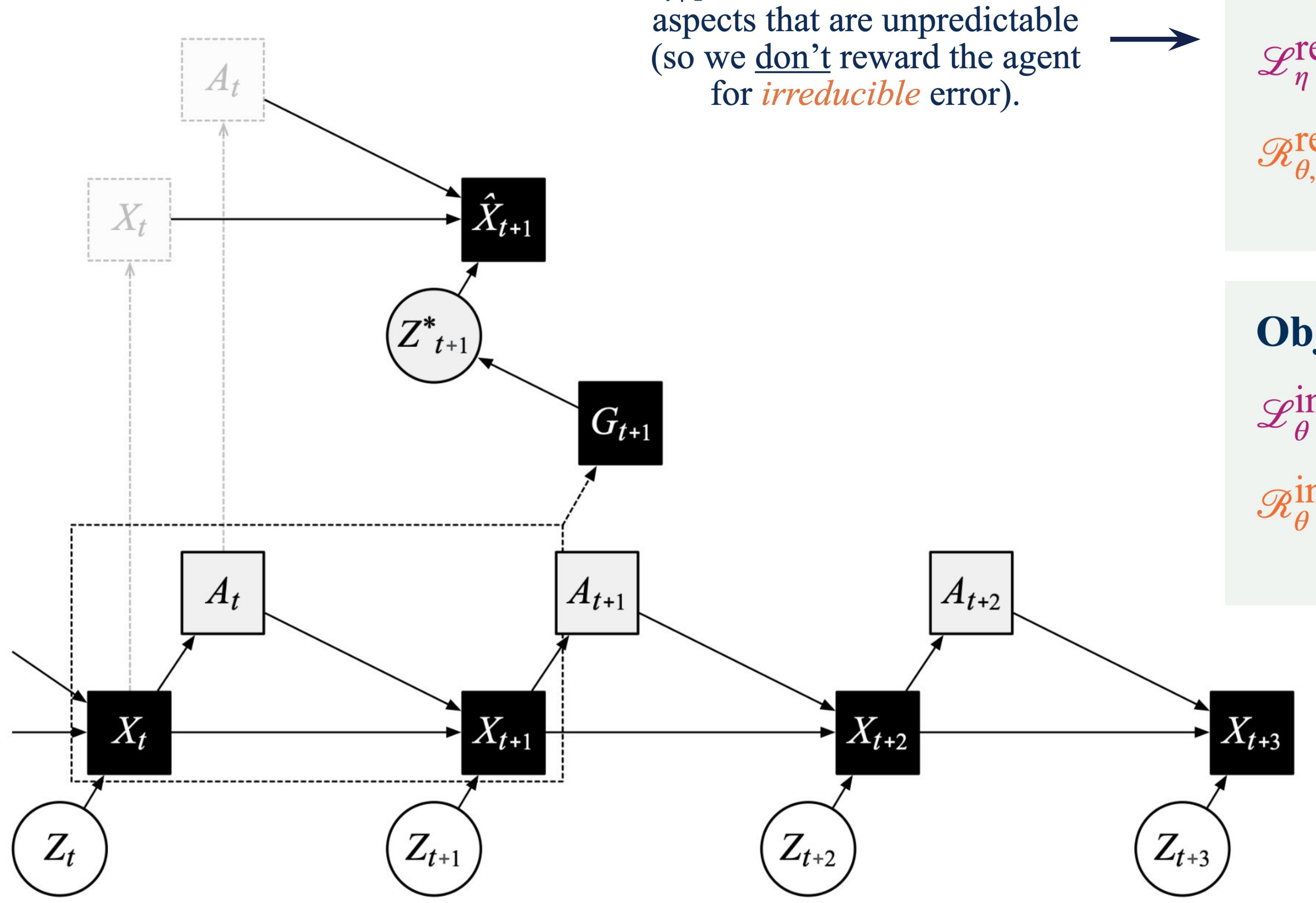
- **Curiosity-Driven Exploration**: Prioritize exploring (and learning) what is not yet understood, where "understanding" := ability to **predict** outcome.
- **Problem of Stochasticity**: Predictive error-based exploration agents are often "stuck" around **high-entropy** elements in the state-action space.
- **Example (Noisy TV Problem)**: Since the agent never knows what the next state is, they will stand in front of it, hoping to learn what is never learnable.

Key Idea: Disentangle (irreducible) "**noise**" from (reducible) "**novelty**", and only use novelty to guide exploration. How?

- Learn **representations of the future** capturing precisely the **unpredictable** aspects of each outcome (no more, no less).
- Use this as additional input when making predictions, such that intrinsic rewards only reflect the **predictable** aspects of world dynamics.

Two Models + Two Objectives

Learn a **Reconstructor** f_η for $f(X_t, A_t, Z_{t+1})$ and a **Generator** p_θ for $p_\eta(Z_{t+1} \,|\, X_t, A_t, X_{t+1})$.



Z_{t+1} should capture *at least* all aspects that are unpredictable (so we don't reward the agent for *irreducible* error).

Objective 1 (Reconstruction)

$$\mathscr{L}^{\text{rec.}}_{\eta}(x_t, a_t, z_{t+1}, x_{t+1}) := \left\| x_{t+1} - f_\eta(x_t, a_t, z_{t+1}) \right\|_2^2$$

$$\mathscr{R}^{\text{rec.}}_{\theta,\eta}(x_t, a_t) := \mathbb{E}_{\substack{X_{t+1} \sim \tau(\cdot \,|\, x_t, a_t) \\ Z_{t+1} \sim p_\theta(\cdot \,|\, x_t, a_t, X_{t+1})}} \mathscr{L}^{\text{rec.}}_{\eta}(x_t, a_t, z_{t+1}, x_{t+1})$$

Objective 2 (Invariance)

$$\mathscr{L}^{\text{inv.}}_{\theta}(x_t, a_t, z_{t+1}) := \log \frac{p_\theta(z_{t+1} \,|\, x_t, a_t)}{p_\theta(z_{t+1})}$$

$$\mathscr{R}^{\text{inv.}}_{\theta}(x_t, a_t) := \mathbb{E}_{\substack{X_{t+1} \sim \tau(\cdot \,|\, x_t, a_t) \\ Z_{t+1} \sim p_\theta(\cdot \,|\, x_t, a_t, X_{t+1})}} \mathscr{L}^{\text{inv.}}_{\theta}(x_t, a_t, z_{t+1})$$

Z_{t+1} should capture *at most* all aspects that are unpredictable (so we do reward the agent for *reducible* error).

implement as

Implementation

Ask a **Critic** to g_ν to maximize a contrastive loss.

Objective 3 (Contrastive Learning)

$$\mathscr{L}^{K,\text{con.}}_{\theta,\nu}(x_t, a_t, z_{t+1}) := \mathbb{E}_{\substack{(X_t^1,\ldots,X_t^{K-1}) \sim \prod_{i=1}^{K-1} \rho_\pi \\ (A_t^1,\ldots,A_t^{K-1}) \sim \prod_{i=1}^{K-1} \pi(\cdot \,|\, X_t^i) \\ (X_{t+1}^1,\ldots,X_{t+1}^{K-1}) \sim \prod_{i=1}^{K-1} \tau(\cdot \,|\, X_t^i, A_t^i) \\ (Z_{t+1}^1,\ldots,Z_{t+1}^{K-1}) \sim \prod_{i=1}^{K-1} p_\theta(\cdot \,|\, X_t^i, A_t^i, X_{t+1}^i)}} \log \frac{e^{g_\nu(x_t, a_t, z_{t+1})}}{\frac{1}{K}\left(e^{g_\nu(x_t, a_t, z_{t+1})} + \sum_{i=1}^{K-1} e^{g_\nu(x_t, a_t, Z_{t+1}^i)} \right)}$$

$$\mathscr{R}^{K,\text{con.}}_{\theta,\nu}(x_t, a_t) := \mathbb{E}_{\substack{X_{t+1} \sim \tau(\cdot \,|\, x_t, a_t) \\ Z_{t+1} \sim p_\theta(\cdot \,|\, x_t, a_t, X_{t+1})}} \mathscr{L}^{K,\text{con.}}_{\theta,\nu}(x_t, a_t, z_{t+1})$$

In Practice: (1) batch size $K < \infty$, (2) ν is not fully optimized, (3) λ is a hyperparameter.

The intrinsic reward is now:

$$\mathscr{R}^{K}_{\theta,\eta,\nu}(x_t, a_t) := \frac{1}{\lambda} \mathscr{R}^{\text{rec.}}_{\theta,\eta}(x_t, a_t) + \mathscr{R}^{K,\text{con.}}_{\theta,\nu}(x_t, a_t)$$

and the agent performs:

$$\underset{\pi}{\overset{\text{(policy)}}{\text{maximize}}} \;\; \underset{\theta,\eta}{\overset{\text{(model)}}{\min}} \; \max_{\nu} \; \mathbb{E}_{\substack{X_t \sim \rho_\pi \\ A_t \sim \pi(\cdot \,|\, X_t)}} \mathscr{R}^{K}_{\theta,\eta,\nu}(X_t, A_t)$$

Overall, simple *drop-in modification* on top of any choice of curiosity-driven exploration.

Curiosity

Definition 1 (Curiosity): Define the *intrinsic reward*

$$\mathscr{R}_\eta(x_t, a_t) := -\,\mathbb{E}_{X_{t+1} \sim \tau(\cdot \,|\, x_t, a_t)} \log \tau_\eta(X_{t+1} \,|\, x_t, a_t)$$

The agent performs

$$\underset{\pi}{\overset{\text{(policy)}}{\text{maximize}}} \;\; \underset{\eta}{\overset{\text{(model)}}{\min}} \; \mathbb{E}_{\substack{X_t \sim \rho_\pi \\ A_t \sim \pi(\cdot \,|\, X_t)}} \mathscr{R}_\eta(X_t, A_t)$$

Curiosity in Hindsight

Definition 2 (Curiosity in Hindsight): Define the *hindsight intrinsic reward*

$$\mathscr{R}_{\theta,\eta}(x_t, a_t) := \frac{1}{\lambda} \mathscr{R}^{\text{rec.}}_{\theta,\eta}(x_t, a_t) + \mathscr{R}^{\text{inv.}}_{\theta}(x_t, a_t)$$

The agent performs

$$\underset{\pi}{\overset{\text{(policy)}}{\text{maximize}}} \;\; \underset{\theta,\eta}{\overset{\text{(model)}}{\min}} \; \mathbb{E}_{\substack{X_t \sim \rho_\pi \\ A_t \sim \pi(\cdot \,|\, X_t)}} \mathscr{R}_{\theta,\eta}(x_t, a_t)$$

Illustrative Example