Hierarchical Reinforcement Learning with Uncertainty-Guided Diffusional Subgoals

Vivienne Wang Tinghuai Wang Joni Pajarinen

Aalto University

June 14, 2025

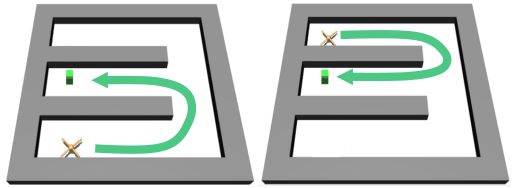
Aalto University
School of Electrical
Engineering

ICML
International Conference
On Machine Learning

Motivation Problem in HRL

Hierarchical Reinforcement Learning

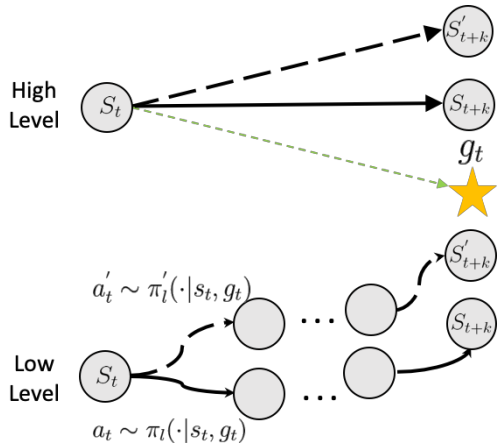
- Decomposes complex tasks into manageable sub-problems.
- High-level policy sets subgoals for low-level policy.



Example: Ant navigating a W-maze.

Key Challenge in Goal-Conditioned HRL

- **Non-stationarity**: Low-level policy constantly changes during training.
- This makes it difficult for high-level policy to generate effective subgoals.
- High-level needs to:
  - Adapt to evolving low-level skills.
  - Capture a complex subgoal distribution.
  - Account for uncertainty in its estimates.



Illustrating non-stationarity.

Our Approach: HIDI's Core Contributions

$\rightsquigarrow$ 1. Generative Subgoals

- Learns diverse, state-aware subgoal possibilities.
- Adapts to changing low-level abilities.

$\mathcal{GP}$ 2. Principled Uncertainty

- Gaussian Process guides with learned knowledge.
- Quantifies confidence in subgoal choices.
- Steers towards reliable, feasible paths.

ϵ-mix 3. Smart Selection Strategy

- Blends diffusion's creativity with GP's certainty.
- Balances exploration of new ideas with exploitation of known good ones.

How HIDI Works: Diffusional Subgoals

High-Level Policy as a Conditional Diffusion Model

- Reverse diffusion process generates subgoal $\boldsymbol{g}$ given state $\boldsymbol{s}$:

$$\pi^h_{\theta_h}(\boldsymbol{g}|\boldsymbol{s})p_{\theta_h}(\boldsymbol{g}^{0:N}|\boldsymbol{s}) = \mathcal{N}(\boldsymbol{g}^N; \boldsymbol{0}, \boldsymbol{I}) \prod_{i=1}^{N} p_{\theta_h}(\boldsymbol{g}^{i-1}|\boldsymbol{g}^i, \boldsymbol{s}).$$

- Subgoals $\boldsymbol{g}^{i-1}$ iteratively refined from noise $\boldsymbol{g}^N$ using a learned noise predictor $\boldsymbol{\epsilon}_{\theta_h}$:

$$\boldsymbol{g}^{i-1} = \frac{1}{\sqrt{\alpha_i}} \left(\boldsymbol{g}^i - \frac{\beta_i}{1-\bar{\alpha}_i} \boldsymbol{\epsilon}_{\theta_h}(\boldsymbol{g}^i, s, i) \right) + \sqrt{\beta_i}\boldsymbol{\epsilon}.$$

Combined Learning Objective

- The subgoal generator is trained to minimize:

$$\mathcal{L}_d(\theta_h) = \underbrace{\mathcal{L}_{dm}(\theta_h)}_{\text{Diffusion Loss}} + \psi \underbrace{\mathcal{L}_{gp}(\theta_h, \theta_{gp})}_{\text{GP Prior Loss}} + \eta \underbrace{\mathcal{L}_{dpg}(\theta_h).}_{\text{RL (DPG) Loss}}$$

- $\mathcal{L}_{dm}$: Matches relabeled "optimal" subgoals from experience.
- $\mathcal{L}_{gp}$: Regularizes towards GP's view of good subgoals.
- $\mathcal{L}_{dpg}$: Maximizes expected high-level Q-values (task reward).

How HIDI Works: Uncertainty Selection

Uncertainty Modeling with Gaussian Process (GP) Prior

- A (sparse) GP models $p(\mathbf{g}|\mathbf{s};\theta_{gp})$.
- Provides predictive mean $\mu_*(\mathbf{s}_*)$ and variance $\sigma_*^2(\mathbf{s}_*)$ for a new state $\mathbf{s}_*$.
- **Regularizes** diffusion: Guides ϵ_{θ_h} towards feasible subgoals.
- **Quantifies uncertainty**: Informs about reliability of subgoal regions.

Inducing States Informed Subgoal Selection

- Hybrid strategy to select subgoal $\mathbf{g}_*$ at state $\mathbf{s}_*$:

$$\mathbf{g}_* = \begin{cases} \boldsymbol{\mu}_*(\mathbf{s}_*), & \text{with probability } \varepsilon \text{ (Exploit GP certainty)}, \\ \mathbf{g} \sim \pi_{\theta_h}(\mathbf{g} \,|\, \mathbf{s}_*), & \text{with probability } 1-\varepsilon \\ & \text{(Leverage diffusion variety)}. \end{cases}$$

- Balances structured, data-driven guidance (GP mean) with flexible, adaptive generation (diffusion).

Experimental Results: Performance Comparison

Evaluated on challenging MuJoCo continuous control tasks. Baselines: HLPS, SAGA, HIGL, HRAC, HIRO.

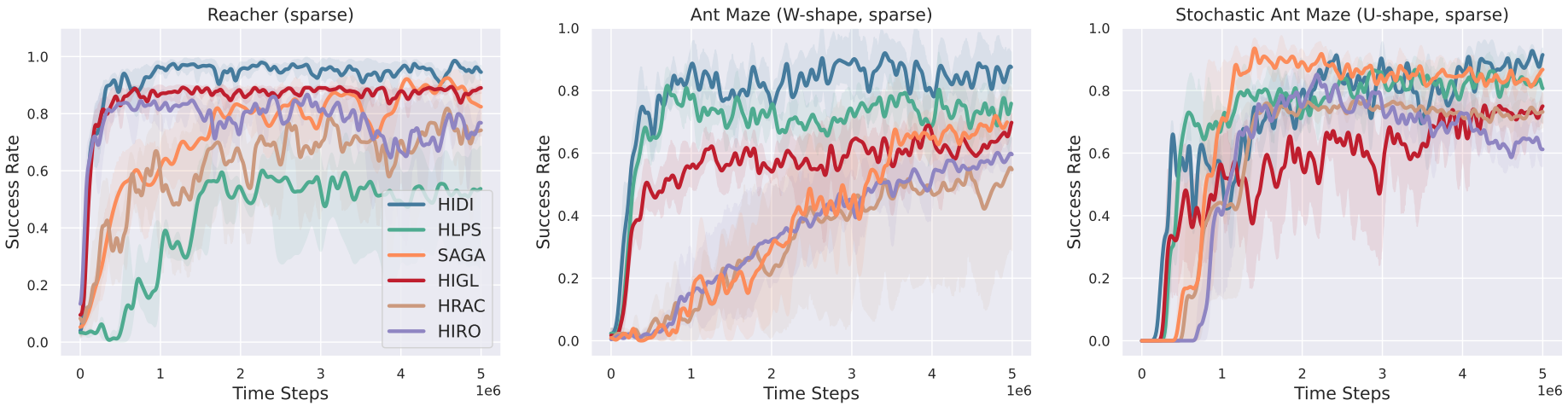


Figure: Learning curves on Reacher (Left), AntMaze W-Sparse (Center), AntMaze U-Stochastic (Right).

Key Observation: HIDI demonstrates better sample efficiency, higher asymptotic performance and robustness in stochastic environments.

Subgoal Quality & Ablation Insights

Subgoal Quality

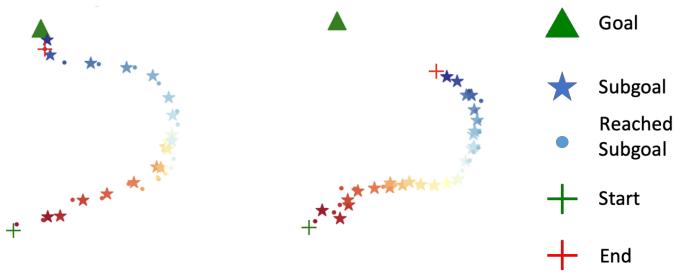


Figure: *

Generated (blue) vs. Reached (red) subgoals. HIDI (left) generates more achievable subgoals than baselines like HRAC (right). AntMaze (W-shape).

HIDI generates reachable subgoals, providing stable low-level learning signals.

Ablation Studies

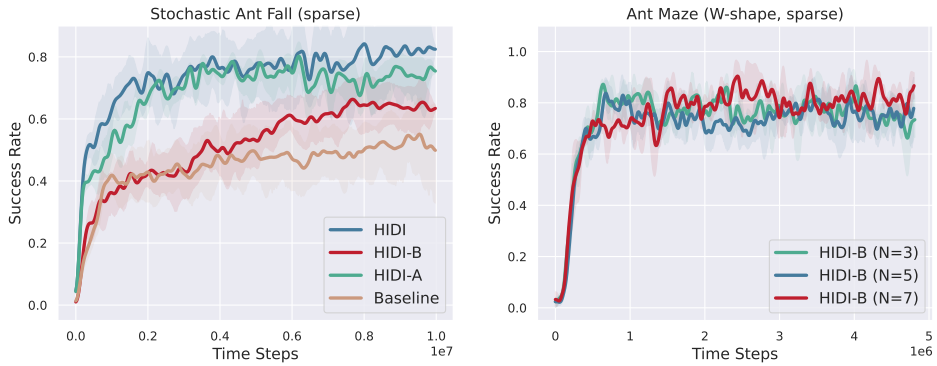


Figure: *

Left: HIDI vs. variants (-A: no selection, -B: no selection & no GP). Right: Effect of diffusion steps N.

- **Diffusional Subgoals**: + 15% perf.
- **GP Regularization**: + 15-16% perf. & sample eff.
- **Subgoal Selection**: + 7-8% perf.
- Optimal diffusion steps $N = 5$.

Conclusion & Key Takeaways

We introduced a novel HRL framework

- Employs a **conditional diffusion model** for expressive subgoal generation.
- Leverages a **GP prior** to regularize learning and explicitly quantify uncertainty.
- Uses a **subgoal selection strategy** combining GP's mean and diffusion model's samples for robust, adaptive decision-making.

Impact:

- HIDI demonstrates significant improvements in both **sample efficiency** and **asymptotic performance** on challenging continuous control benchmarks.
- Highlights the benefits of modeling complex subgoal distributions and incorporating principled uncertainty quantification in HRL.

Thank You!