Dense and Diverse Goal Coverage in Multi-Goal Reinforcement Learning

Sagalpreet Singh, Rishi Saket, Aravindan Raghuveer

RLC 2026

Google DeepMind | {sagalpreet, rishisaket, araghuveer}@google.com

Problem Setting: High Return is Not Enough—We Need Goal Diversity

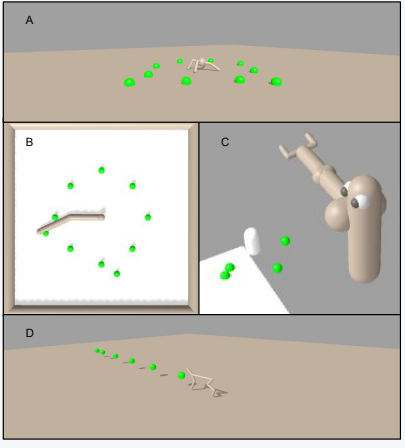
Multi-Goal RL Setup

- MDP state space $S = S^+ \cup S^-$ with sparse reward $R(s) = \mathbb{I}(s \in S^+)$.
- Goal manifold S^+ is **unknown upfront** (goals discovered only upon visiting).

Limitations of Prior Approaches

- **Standard RL (Return Max):** Exploits 1-2 easy goals $\rightarrow$ policy collapse over S^+.
- **Exploration RL (SAC/MaxEnt, RND/Counts):** Visits all under-explored states, including non-goals $S^- \rightarrow$ **wastes steps & lowers return**.
- **State Marginal Matching (SMM) / GFlowNets:** Require target distribution upfront or DAG structure.

Goal-Rich Environments



MuJoCo environments with multiple goal regions (green spheres).

Core Insight: A Concave Objective Balancing Return & Goal Diversity

Goal Distribution vs Non-Goal Distribution

Maximize return while spreading state distribution $d[\overline{\pi}](s)$ **uniformly across goal states** S^+, assigning **zero mass** to non-goals S^-.

Gini-Regularized Optimization Objective

We define the objective $Z(\overline{\pi})$ over policy mixture simplex $\overline{\pi} \in \overline{\Pi} = \Delta(\Pi)$:

$$\max_{\overline{\pi} \in \overline{\Pi}} Z(\overline{\pi}) := \underbrace{J_\gamma(\overline{\pi})}_{\text{Expected Return}} + \underbrace{\mathcal{I}^{S^+}(\overline{\pi})}_{\text{Gini Diversity on } S^+} = \sum_{s \in S^+} \left(d[\overline{\pi}](s) - \frac{1}{2} d[\overline{\pi}](s)^2 \right)$$

Key Property: Strict Concavity

$Z(\overline{\pi})$ is **strictly concave** over policy mixture simplex $\overline{\Pi}$, enabling efficient Frank-Wolfe optimization!

Why Policy Mixtures?

A single policy rarely covers disjoint goals. A **mixture of specialized policies** $\overline{\pi} = \sum \lambda_k \mu_k$ naturally covers S^+.

The Algorithm: Dense & Diverse Goal Coverage (DDGC)

Frank-Wolfe Framework with Dynamic Custom Rewards

Iteratively construct policy mixture $\overline{\pi}_K = (1 - \lambda_K)\overline{\pi}_{K-1} + \lambda_K \mu_K$ with $\lambda_k = \frac{2}{k+1}$:

1. **Direction Step (Linearizing Z):** Computes dynamic custom reward $r_k(s)$ based on past mixture frequency $d[\overline{\pi}_{k-1}](s)$:

$$r_k(s) = \begin{cases} \mathbf{1} - \mathbf{d}[\overline{\pi}_{\mathbf{k-1}}](\mathbf{s}) & \text{for } s \in S^+ \\ \mathbf{0} & \text{for } s \in S^- \end{cases}$$

Intuition: Over-visited goals get penalized ($r_k \to 0$); under-visited goals yield high reward ($r_k \to 1$).

2. **Batch Offline RL Subroutine (FAC):** Fits action-value Q-function and updates policy μ_k using Fitted Actor-Critic on accumulated replay buffer $\mathcal{D}_k$.
3. **Mixture Update:** Update mixture $\overline{\pi}_k \leftarrow (1 - \lambda_k)\overline{\pi}_{k-1} + \lambda_k \mu_k$.

Theoretical Analysis: Fast $\mathcal{O}(1/K)$ Convergence Guarantees

Main Theoretical Guarantee

Under standard concentrability and Bellman error assumptions, the sub-optimality gap $Z(\overline{\pi}^*) - Z(\overline{\pi}_K)$ for mixture $\overline{\pi}_K$ is provably bounded by:

$$Z(\overline{\pi}^*) - Z(\overline{\pi}_K) = \underbrace{\mathcal{O}\left(\frac{1}{K}\right)}_{\text{Frank-Wolfe Rate}} + \underbrace{\mathcal{O}(\gamma^H + \gamma^{N_{\text{FQI}}})}_{\text{Horizon \& Optimization}} + \underbrace{\mathcal{O}(\epsilon_{\text{approx}})}_{\text{RL Function Approx Error}} + \underbrace{\mathcal{O}\left(\frac{1}{\sqrt{N_T}}\right)}_{\text{Trajectory Sampling Error}}$$

Proof Key Step 1: Optimization

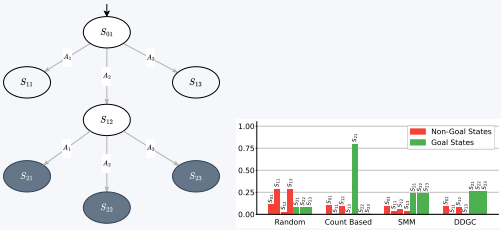
Curvature bound $C_Z \leq 1$ guarantees $\mathcal{O}(1/K)$ convergence rate of Frank-Wolfe in policy mixture space.

Proof Key Step 2: Sample Efficiency

Concentration bounds (Hoeffding) + FQI error analysis bound errors from batch offline RL updates.

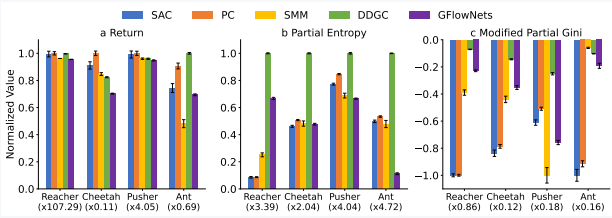
Empirical Results: High Return + Uniform Goal Coverage

Controlled Synthetic MDP



- **Count-RL:** Collapses to single path (S_{21}).
- **SMM:** Wastes probability mass on non-goals (S^-).
- **DDGC (Ours): Uniform mass on goals, 0 mass on non-goals!**

MuJoCo Benchmarks (Brax)



- **Return:** DDGC matches top-performing SAC baseline.
- **Diversity (Gini & Entropy):** DDGC achieves **2-3x higher goal coverage** than SAC, RND, SMM, and GFlowNets.

Summary & Visit Our Poster!

TL;DR: Key Takeaways

1. **Problem:** Standard RL ignores goal diversity; existing exploration wastes steps visiting non-goals.
2. **Formulation:** Multi-Goal RL objective balancing return and Gini goal diversity over policy mixtures.
3. **Algorithm:** DDGC uses Frank-Wolfe with dynamic custom rewards to iteratively train policy mixtures.
4. **Results:** Fast $\mathcal{O}(1/K)$ convergence, near-optimal returns, and superior goal coverage across benchmarks.

Poster & Presentation Details

Poster #44 3:00 PM – 6:00 PM

Code & Paper:

`github.com/google-deepmind/ddgc`

Contact:

`sagalpreet@google.com`