Adaptive Active Learning for Regression via Reinforcement Learning

Weighted Improved Greedy Sampling (WiGS)

Simon Nguyen, Troy Russo, Kentaro Hoffman, and Tyler McCormick

Department of Statistics, University of Washington

Introduction: The Active Learning Bottleneck

- **Problem:** High cost of acquiring labeled data in regression tasks (e.g., robotics, drug discovery, environmental science).
- **Goal:** Select the most informative points from a large unlabeled pool to minimize labeling costs.
- **Core Challenge:** Determining "informativeness" involves balancing two competing objectives:
  1. **Exploration:** Spanning the feature space ($\mathcal{X}$).
  2. **Investigation:** Reducing uncertainty in the output space ($\mathcal{Y}$).

Background: Greedy Sampling Heuristics

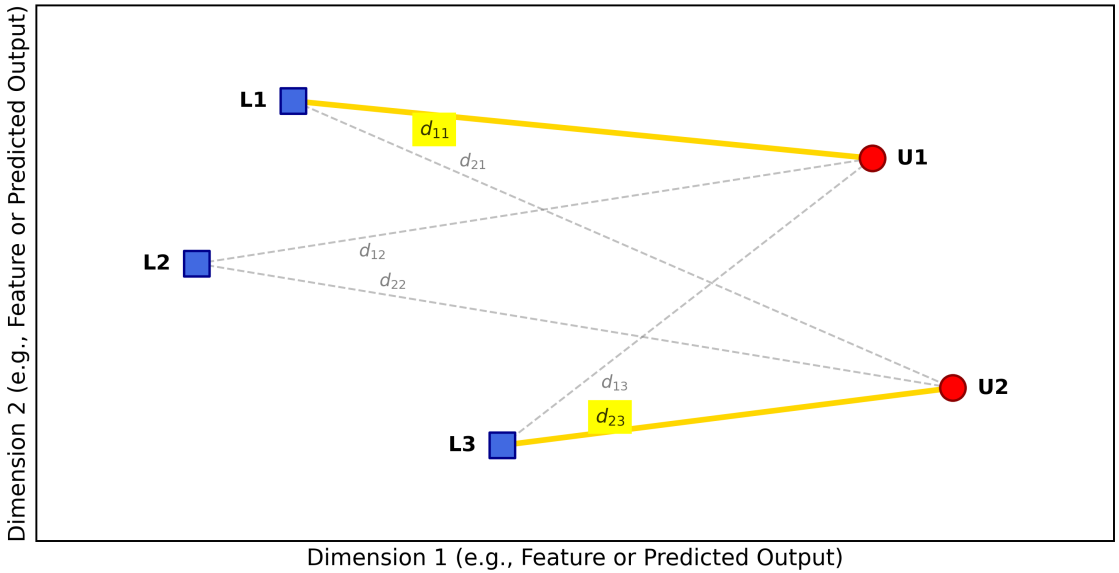


Figure: The "Furthest Nearest Neighbor" Principle

- **GSx (Exploration):** Selects points furthest from labeled data in *feature* space.
- **GSy (Investigation):** Selects points where the model prediction is furthest from known labels.

The State-of-the-Art Baseline: iGS

- **improved Greedy Sampling (iGS)** combines these using a *multiplicative*

The Problem: Asymptotic Blindness

The "Trap" Scenario:

- Consider a region with **high feature density** ($d^x \approx 0$) but **high model error** ($d^y \approx 1$).
- **Multiplicative Failure:**

$$\text{Score} = 0 \times 1 = 0$$

- The iGS algorithm becomes "blind" to high-error points simply because they are in dense regions.

Hypothesis: The optimal balance is *not* fixed but dynamic, varying with data distribution and learning stage.

Proposed Framework: Weighted improved Greedy Sampling

WiGS Recasts the Criterion as Additive:

- We introduce a dynamic weight $w_x^{(t)}$ to explicitly control the trade-off.
- Scores are min-max normalized (d') to be on comparable scales.

The WiGS Score

$$s_n^{WiGS} = \min_m \left(w_x^{(t)} d'_{nm,x} + (1 - w_x^{(t)}) d'_{nm,y} \right)$$

- If $w_x \to 0$, the agent can focus purely on error, allowing it to see into "traps" that iGS ignores.
- **Challenge:** How to optimally select $w_x^{(t)}$?

Strategy 1: Heuristic Weighting

Before using learning, we explore simple heuristics for $w_x^{(t)}$:

- **Static Weights:**
  - Fix w_x (e.g., 0.25 for investigation focus, 0.75 for exploration focus).
- **Time-Decay Weights:**
  - Hypothesis: Exploration is critical early; investigation is critical later.
  - **Linear Decay:** $w_x^{(t)} = 1 - (c \cdot t/T)$
  - **Exponential Decay:** $w_x^{(t)} = \exp(-c \cdot t/T)$

Strategy 2: Adaptive Weighting via RL

"Learning to Active Learn"

- We frame the weight selection as a sequential decision problem.
- An **Agent** observes the state of the learning process and outputs the optimal weight $w_x^{(t)}$.

Two RL Approaches:

1. **Discrete:** Multi-Armed Bandits (WiGS-MAB). Selects from a set $\{0.25, 0.5, 0.75\}$ using UCB1.
2. **Continuous:** Soft Actor-Critic (WiGS-SAC). Outputs exact continuous values.

The MDP Formulation (WiGS-SAC)

To avoid "double-dipping," the agent learns solely from the training set D_{tr}.

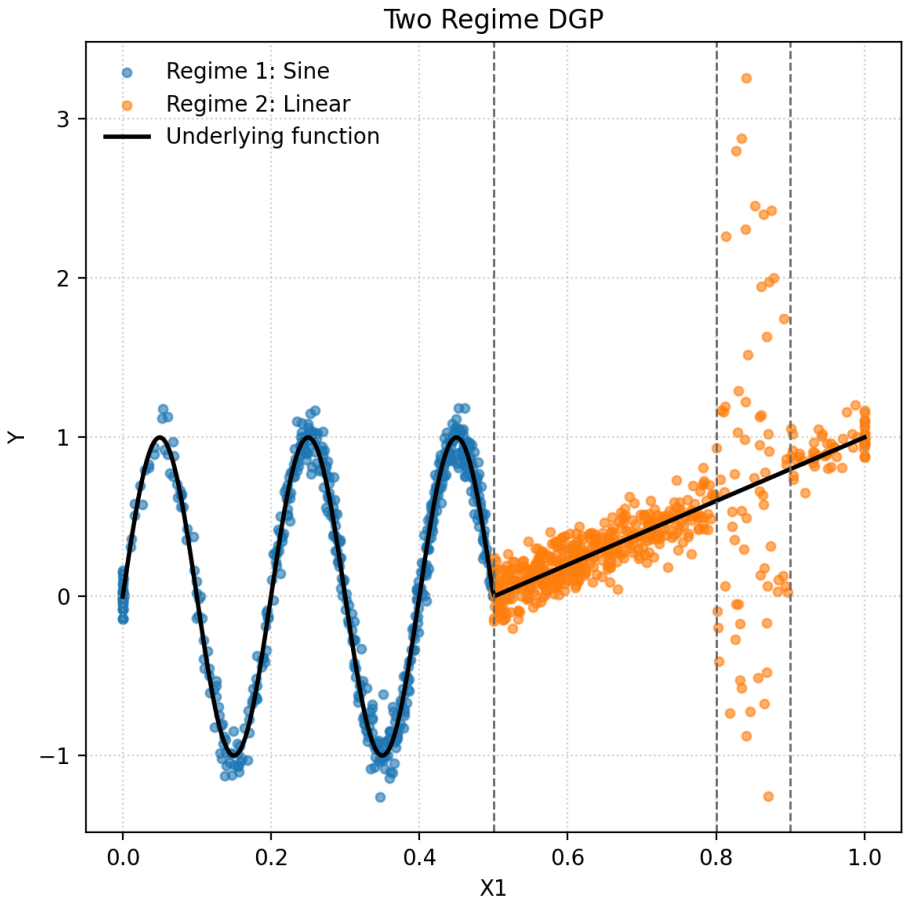
- **State (s_t):**
  - Current iteration t.
  - Distributional statistics of D_{tr} (mean, std dev).
  - Model performance (K-Fold CV RMSE on D_{tr}).
- **Action (a_t):**
  - Continuous weight $w_x^{(t)} \in [0, 1]$.
- **Reward (r_t):**
  - Improvement in generalization performance estimated via Cross-Validation.

$$r_t = CV_{RMSE}^{(t-1)} - CV_{RMSE}^{(t)}$$

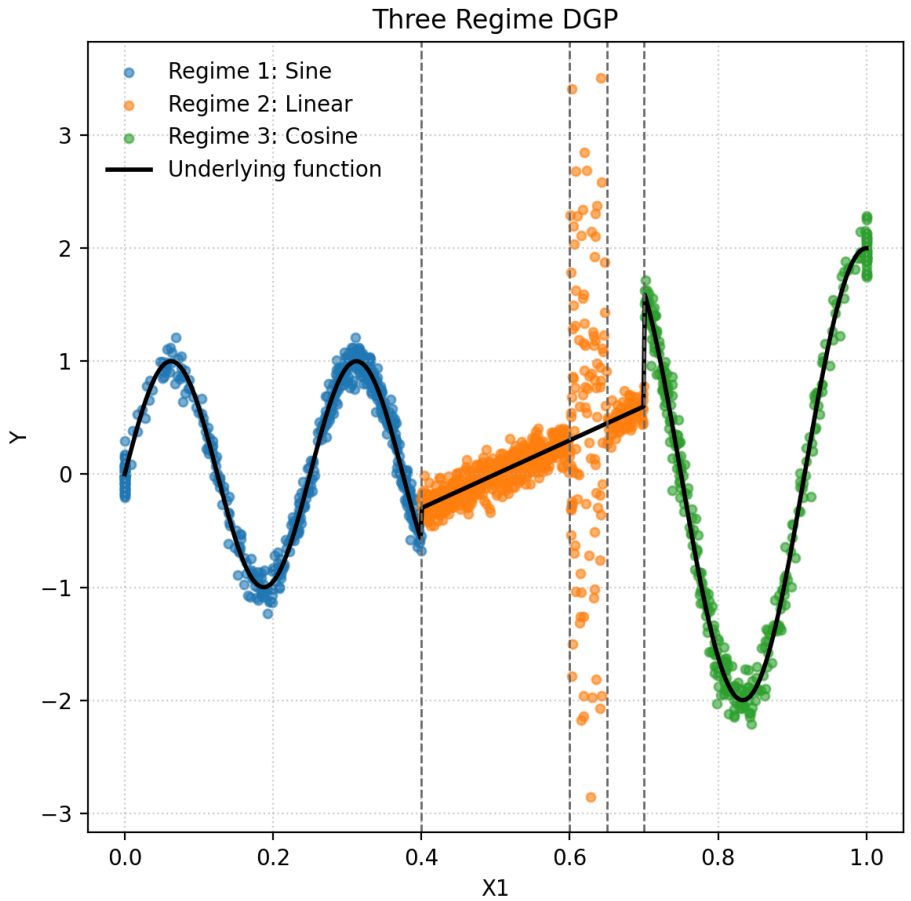
Synthetic Experiments: Stress Testing

Designed to test "Asymptotic Blindness" using Traps.

- **Two-Regime:** Sine wave (complex) + Linear (simple).
- **Three-Regime:** Includes an "Extreme Noise Desert" ($\sigma = 1.5$).



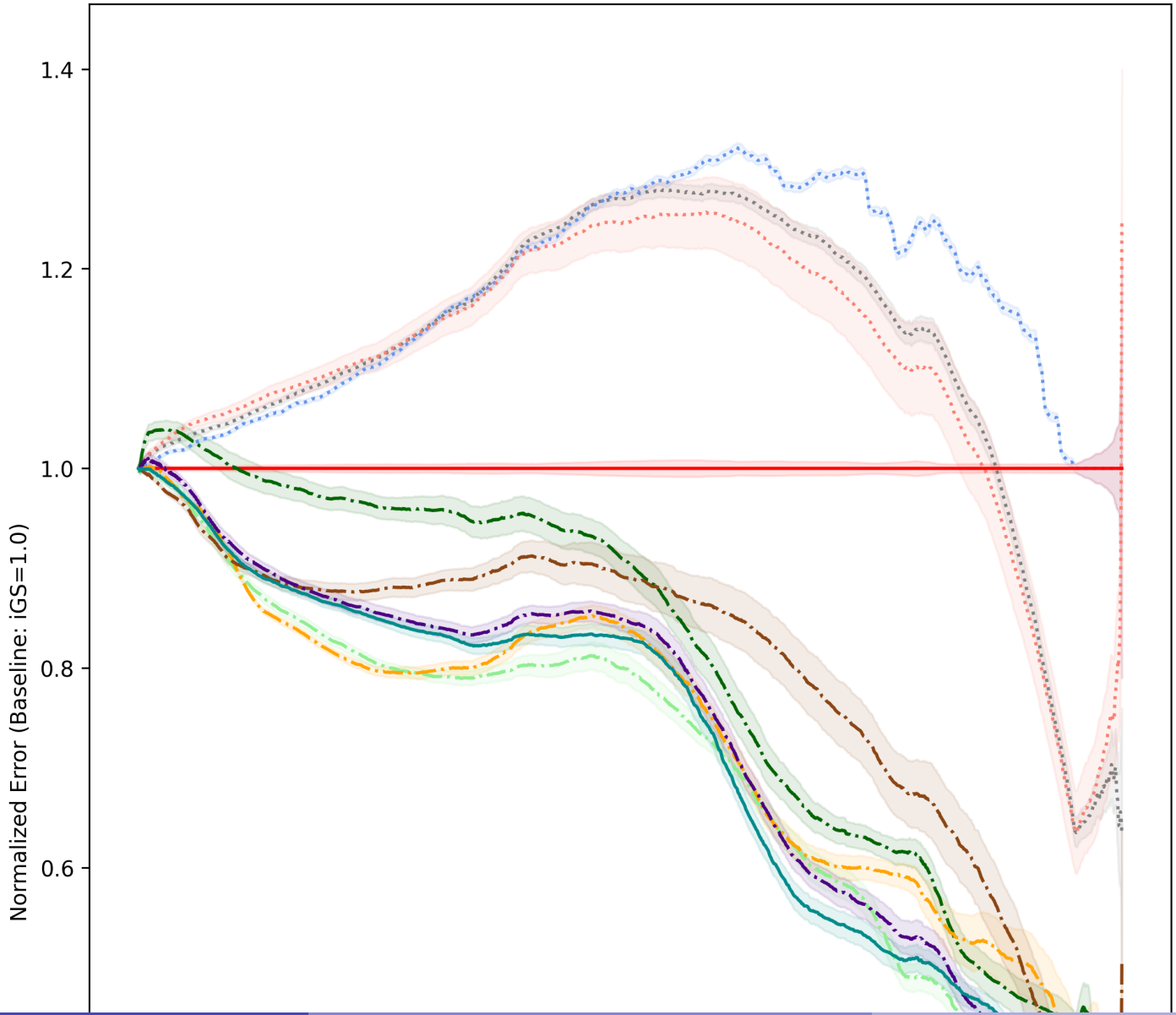
Two-Regime DGP



Three-Regime DGP

Results: Performance on Synthetic Data

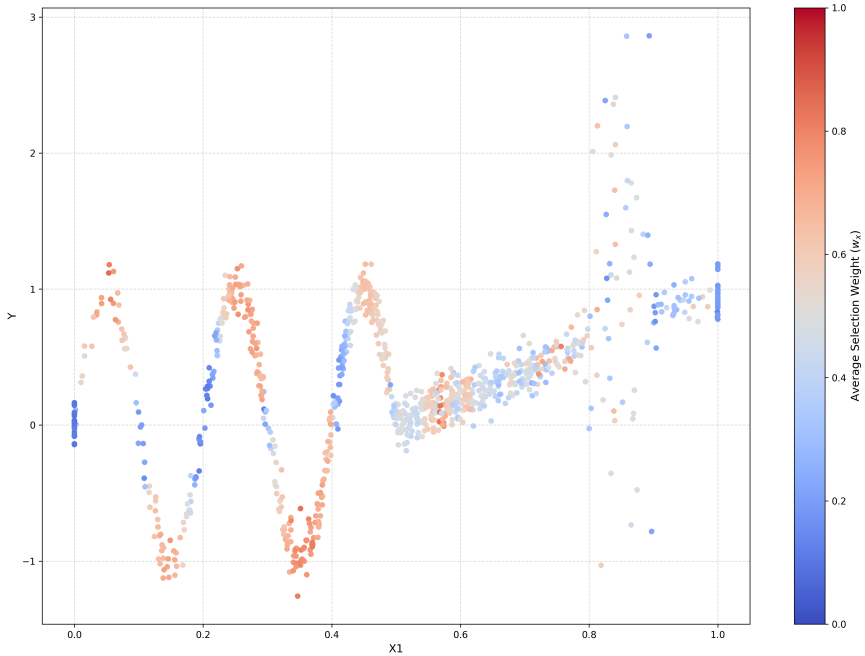
Evaluation: Full-Pool RMSE relative to iGS baseline (Red Line at 1.0).



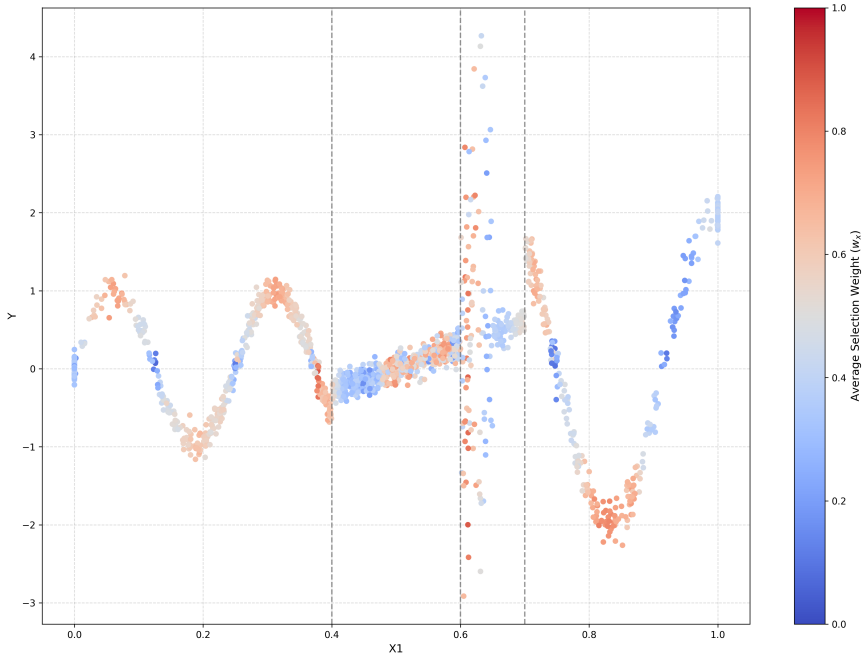
Analysis: What did the Agent Learn?

Spatial Adaptation: The agent learned a region-specific policy.

- **High Curvature:** High Exploration ($w_x \to 1$, Red).
- **Linear Slopes/Traps:** High Investigation ($w_x \to 0$, Blue).



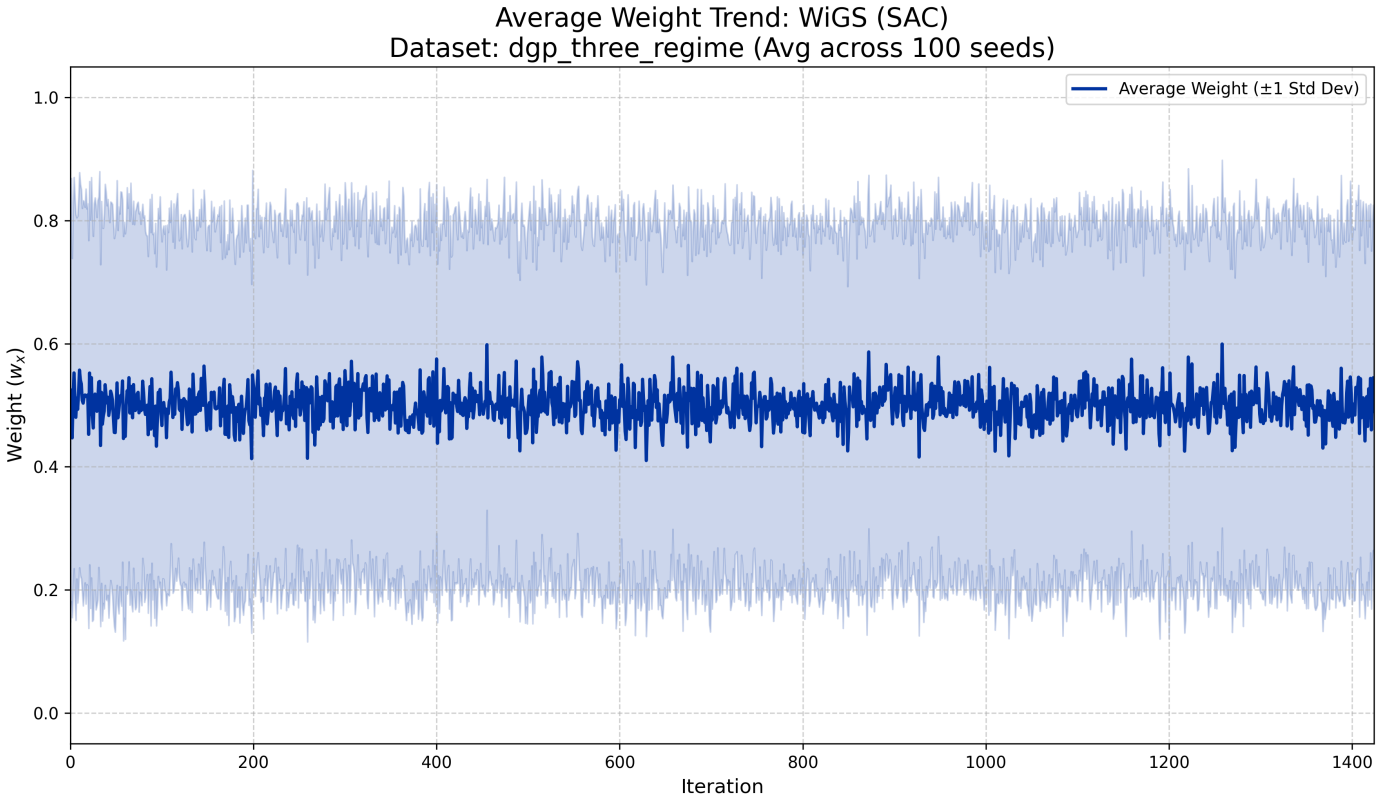
Two-Regime Policy



Three-Regime Policy

Analysis: Temporal Adaptation

Temporal Evolution:



- The agent does **not** converge to a single static number.
- High variance (shaded region) implies the optimal trade-off is a dynamic decision made at every step.

Experiments on Real-World Benchmarks

Tested on 18 public regression datasets (UCI, Kaggle, etc.).

- **Robustness:** WiGS-SAC consistently matches or outperforms iGS.
- **Complex Domains:** Significant error reduction in complex datasets like `wine_red` and `wine_white` ($\approx 50\%$ reduction).
- **Simple Domains:** In simpler tasks (e.g., `yacht`), performance converges to the baseline (safe behavior).

Conclusion

1. **WiGS Framework:** An additive, weighted formulation resolves the "Asymptotic Blindness" of multiplicative baselines.
2. **RL for Active Learning:** We successfully framed the exploration-investigation trade-off as a continuous control problem using Soft Actor-Critic.
3. **Results:**
   - The adaptive agent learns sophisticated, non-stationary policies.
   - Robust performance across 18 benchmarks and 2 adversarial synthetic environments.

Code available at GitHub:
https://github.com/thatswhatsimonsaid/WeightedGreedySampling