Gardant B Report 3

Summary

1. What we have done

- Literature Review of ”Doubly Robust Policy Evaluation and Learning”
- LinUCB Combined with Doubly Robust

The Doubly Robust (DR) method enhances reward estimation in policy evaluation by combining observed rewards from historical data with predicted rewards from a model. This integration allows for more accurate estimates of outcomes under new policies, even in scenarios with sparse data.

When applied to LinUCB, the DR method improves action selection by balancing exploration and exploitation through confidence bounds. It adjusts observed rewards using propensity scores to align with LinUCB's policy.

- Different values for alpha in LinUCB (Exploration & Exploitation)

We plotted the cumulative reward for LinearUCB with tanh function, the result states that we still need to do exploration, since only exploitation (low alpha) will decrease the cumulative reward overall.

2. Any Questions or Issues

- Correctness of the ”Doubly Robust Estimator Formular” in attachment

In the attahment it shows the formular of Doubly Robust Estimator:

$$\hat{Q}(C, A_{\text{LinUCB}}) = \frac{1}{N} \sum_{i=1}^{N} \left[R_i \cdot \frac{\hat{\pi}(A_{\text{LinUCB}}|C_i)}{\pi(A_i|C_i)} + \hat{R}(C_i, A_{\text{LinUCB}}) \cdot \frac{\pi(A_i|C_i)}{\hat{\pi}(A_{\text{LinUCB}}|C_i)} \right]$$

However, as we verify the formular from the paper, it should be:

$$\hat{V}_{\text{DR}}^{\pi} = \frac{1}{|S|} \sum_{(x,h,a,r_a) \in S} \Bigg[\frac{(r_a - \hat{\varrho}_a(x)) \mathbf{I}(\pi(x) = a)}{\hat{p}(a \mid x, h)} + \hat{\varrho}_{\pi(x)}(x) \Bigg]. \quad (1)$$

Therefore, we rewrite it as:

$$\hat{Q}(C, A\text{LinUCB}) = \frac{1}{N} \sum_{i=1}^{N} \left[\hat{R}(C_i, A_{\text{LinUCB}}) + \frac{\pi(A_{\text{LinUCB}}|C_i)}{\hat{\pi}(A_i|C_i)} \cdot \Big(R_i - \hat{R}(C_i, A_i) \Big) \right]$$

We would like to confirm if this formula is now correct.

3. The plan until next meeting

- Refine selected models by fine-tuning them on the Statlog dataset.
- Evaluate performance using offline policy evaluation and delayed rewards.

Details

1. Literature Review of Doubly Robust Estimator

In policy evaluation, there are two mainstream approaches preceding the development of the Doubly Robust Estimator (DRE)

1. Direct Method

$$\hat{V}_{\mathrm{DM}}^{\pi} = \frac{1}{|S|} \sum_{x \in S} \hat{q}_{\pi(x)}(x)$$

where:

– $\hat{q}_a(x)$ is the estimated expected reward for action a given context x.

– $\hat{q}_{\pi(x)}(x)$ is the estimated expected reward for the action chosen by policy π in context x.

Cons：

- Section bias: without considering the distribution shift between the historical policy and the target policy.
- DM relies entirely on the accuracy of the reward model. If the reward model is not accurate, DM will produce a biased estimate.

2. Inverse Propensity Score

$$\hat{V}_{\mathrm{IPS}}^{\pi} = \frac{1}{|S|} \sum_{(x,a,r) \in S} \frac{r \cdot I(\pi(x) = a)}{p(a|x)}$$

where

- S is the sample set from the dataset.
- r is the observed reward for taking action (a) in context (x)
- $I(\pi(x) = a)$ is an indicator function
- $p(a|x)$ is the propensity score

Pros:

- Unbiasedness: Under accurate propensity score estimation

Cons:

- High variances: When some actions have very low propensity scores, the weighting factors in IPS can be very large, resulting in high variance and instability in IPS estimates

3. Doubly Robust Estimate

Doubly Robust (DR) method is a statistical technique used in policy evaluation to improve the accuracy of estimating rewards (outcomes) for different actions. It combines two sources of information:

- Observed Rewards from Historical Data: This is the actual reward (outcome) observed when a particular action was taken in a certain context.
- Predicted Rewards from a Model: This is an estimate of the reward generated by a predictive model based on the context and action.

Combine DM and IPS, results in DRE:

$$\hat{V}_{\mathrm{DR}}^{\pi} = \frac{1}{|S|} \sum_{(x,a,r) \in S} \left(\frac{r - \hat{q}_a(x)}{p(a|x)} I(\pi(x) = a) + \hat{q}_{\pi(x)}(x) \right)$$

By combining these two components, the DR method helps to create a more accurate estimate of what would happen if a new policy were applied, even in

situations where data is sparse or the historical policy differs significantly from the new policy.If at least one of the propensity score and reward model is accurate, the estimate is unbiased.

2. Combine DR with LinUCB

- Action Selection by LinUCB

LinUCB selects the optimal action A_{LinUCB} based on confidence bounds. Given a new context C , LinUCB uses its linear model to estimate the reward for each action and chooses the one with the highest upper confidence bound, balancing exploration and exploitation.

- Predicted Reward (Direct Model)

For the selected action A_{LinUCB} , a direct model predicts the reward $\hat{R}(C, A_{\text{LinUCB}})$. This model estimates the expected reward for A_{LinUCB} in the current context, complementing historical observations with model–based estimates.

- Observed Reward with Propensity Score Adjustment

The observed reward R from historical data is adjusted using the propensity scores—probabilities of taking actions in specific contexts under different policies. The observed action's propensity score $\pi(A|C)$ from historical data and LinUCB's propensity score $\hat{\pi}(A_{\text{LinUCB}}|C)$ for the chosen action are used to reweight the observed reward, aligning it with LinUCB's selection policy.

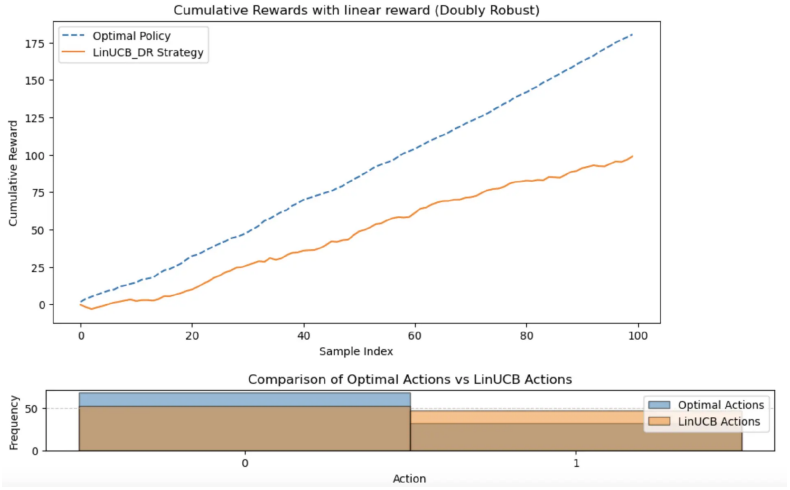
- Calculating the DR Estimator

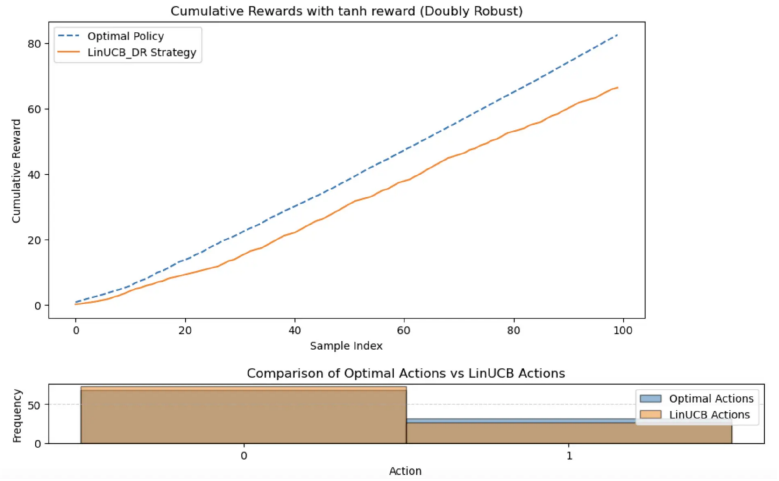
The DR estimator for LinUCB's expected reward is calculated as:

$$DR\_reward = \frac{1}{N}\sum_{1}^{N}\frac{\hat{\pi}(A_{\text{LinUCB}}|C)}{\pi(A|C)} \times (R_i - \hat{R}(C, A_{\text{LinUCB}})) + \hat{R}(C, A_{\text{LinUCB}})$$

Results

The application of the Doubly Robust (DR) method in LinUCB yields notable results.



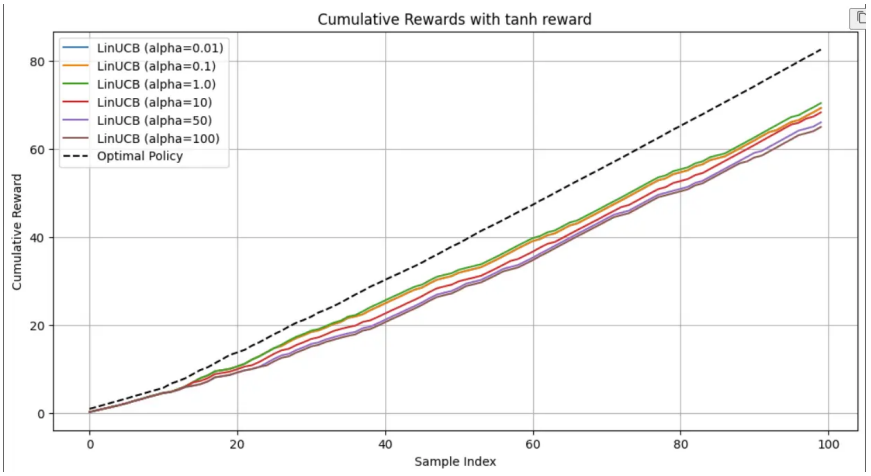


- Improved Accuracy: DR effectively reduces bias and variance, leading to more accurate reward estimates, particularly in sparse data scenarios.
- Enhanced Robustness: Its two-part structure, combining observed and predicted rewards, provides reliable estimates even when predictive models contain errors.

However, the observed results may still diverge from actual data due to several factors:

- Data Quality Issues: Real-world data can be noisy or sparse, impacting the model's ability to learn effectively.
- Dynamic Environments: Changes in the environment can affect the applicability of learned policies, leading to potential discrepancies between estimated and actual rewards.
- Risk of Overfitting: While DR enhances robustness, there is a risk of overfitting, especially in complex models or small sample sizes.

3. Exploration and Exploitation by adjusting alpha



From our last week's discussion with Marco, we performed the cumulative reward plot for the tanh reward which is the best non-linear reward function we discovered last week. While Low alpha emphasizes exploitation, using known information to maximize immediate rewards, High alpha emphasizes exploration, seeking new actions to potentially discover higher rewards.

As we can see from the plot, The optimal alpha values (around 1.0 and 10 in the plot) seem to balance exploration and exploitation effectively, as these curves are closest to the optimal policy. Therefore we need to perform some aspects of the exploration in this scenario to make the reward optimal.