Gardant Biweekly Report 4

Han Xiao, Kexuan Ma, Haomin Mo

November 7, 2024

Summary

(a) What Has Been Done

1. **Linear Function**

   1) Original linear function

   2) Piecewise-linear function

   3) Additon of oscillation

   4) A weighted linear combination of context features

2. **Statistical Tests Between DR Estimators and Ordinary Estimators**

   Performed paired t-tests and bootstrap tests on DR Estimator and Estimator without DR. The result showed that we only have one entry with linear function 3 and alpha 50 reached the 10% significance level, other entries have no significance.

3. **Statistical Tests Between with and without Exploration Policies DR Estimators**

   Performed paired t-tests and bootstrap tests on the DR estimators of policies with and without exploration. The results show no significant difference between the DR rewards for policies with and without exploration.

4. **Literature Review on Simulation Process in the Paper**

   Review on the paper **"To Update or Not to Update? Delayed Nonparametric Bandits with Randomized Allocation"**, including 4 linear reward functions and the delay reward setting.

(b) Any Questions or Issues

1. Given the results, where no significant difference was observed in DR rewards, does this suggest that the DR estimator maintains robustness and consistency across different exploration parameters?

Detail

1. Linear Function

- **Original linear function**

  This function uses a simple linear combination of sampled context features, combined with non-stationary components such as drift and seasonality. When the action is optimal, the mean reward is higher, and the variance is lower.

- **Piecewise-linear function**

  The piecewise-linear function adds conditional adjustments to the context factor, creating a reward structure that changes based on the value of the context.

  Specifically, the context factor undergoes transformation within predefined intervals, leading to abrupt variations in the reward. This approach enables the function to capture scenarios where different context levels lead to distinct effects, making rewards more sensitive to variations in context.

$$\text{context\_factor} = \begin{cases} 1, & \text{if context\_factor} < 0.5 \\ -10 \cdot \text{context\_factor} + 4, & \text{if } 0.5 \leq \text{context\_factor} < 0.6 \\ 0, & \text{if context\_factor} \geq 0.6 \end{cases}$$

- **Additon of oscillation**

  This variant adds an oscillatory component to the reward structure. The oscillation, which varies periodically over time, simulates environmental cycles or recurring trends. By incorporating oscillation, the function adapts to scenarios where actions yield rewards that vary cyclically, adding a layer of time-based complexity to the reward dynamics.

- **A weighted linear combination of context features**

  This function assigns specific weights to each sampled context feature, creating a weighted sum. It also includes scaled-down seasonality, with each context feature contributing differently to the final reward.

  The weighted approach allows for a more nuanced representation of the reward, as each feature's effect is differentiated, aligning more closely with environments where certain factors are more influential than others. This function is particularly suited for cases requiring a flexible and detailed relationship between context and reward.

2. Statistical Tests Between DR Estimators & Non-DR Estimators

To evaluate the significance between the DR (Doubly Robust) estimators and non-DR estimators, we selected the following statistical tests:

- **Paired t-test**: This test was used to compare the mean rewards between the DR and non-DR estimators to determine if there was a significant difference.

- **Bootstrap Test**: A non-parametric test conducted with 10,000 bootstrap samples to assess the significance of the observed mean difference in rewards.

Comparison Logic

Our objective was to examine the rewards under four different reward functions: `lin1`, `lin2`, `lin3`, and `lin4`. For each reward function, we compared the DR rewards and non-DR rewards. The statistical tests were applied to each pair of rewards across the four reward functions to check for any significant differences.

Results and Analysis

The following table summarizes the average rewards for both DR and non-DR estimators, as well as the results from the paired t-test and bootstrap test.

Reward Type	Alpha	Avg DR Rewards	Avg Non-DR Rewards	t-stat	t-test p-value	Bootstrap p-value
lin1	1.0	1.173692	1.286153	-0.816596	0.416120	0.5017
lin1	10.0	1.173692	1.186552	-0.097091	0.922850	0.5017
lin1	50.0	1.173692	1.214703	-0.296791	0.767248	0.5004
lin2	1.0	2.037807	1.994189	0.564542	0.573662	0.4947
lin2	10.0	2.038871	1.981871	0.723120	0.471311	0.5010
lin2	50.0	2.040583	1.944418	1.207884	0.229969	0.5040
lin3	1.0	0.650627	0.709812	-0.224772	0.822620	0.4992
lin3	10.0	0.650627	0.708549	-0.220033	0.826299	0.4875
lin3	50.0	0.650627	0.787066	-0.523644	0.601696	0.5079
lin4	1.0	1.608151	1.720932	-1.305946	0.194597	0.5057
lin4	10.0	1.608091	1.711212	-1.231751	0.220960	0.4938
lin4	50.0	1.608354	1.653525	-0.535966	0.593184	0.4919

Table 1: Statistical Test Results for DR vs. Non-DR Rewards with Different Reward Functions

From the table above, we observe the following:

- Across all `lin1`, `lin2`, `lin3`, and `lin4` reward functions, the average rewards for both DR and non-DR estimators show some differences, but these differences were not statistically significant.
- The **paired t-test** results consistently yield non-significant p-values (all above 0.1) except for only one entry with lin3 and alpha value of 50. This indicates that there is no statistically significant difference in the mean rewards between the DR and non-DR estimators for any of the reward functions. In some cases, the t-statistics were low or near zero, reinforcing the observation that the rewards were similar.
- The **bootstrap test** results similarly returned p-values close to 0.5 or above, demonstrating that the observed mean differences between the DR and non-DR rewards were not significant. This supports the findings from the paired t-test by providing a non-parametric validation.

Conclusion

The statistical tests reveal that the DR and non-DR estimators performed similarly across the different reward functions tested. The results suggest that neither method had a significant advantage in terms of average rewards, indicating robustness in the performance of both estimators under these test conditions.

3. Statistical Tests Between Policies DR Estimators

To evaluate the significance between the policies' DR estimators with and without exploration, we selected the following statistical tests:

- **Paired t-test**: This test was used to compare the mean DR rewards between the exploration and no-exploration policies to determine if there was a significant difference.
- **Bootstrap Test**: A non-parametric test conducted with 10,000 bootstrap samples to assess the significance of the observed mean difference in DR rewards.

Comparison Logic

Our objective was to examine the DR rewards under four different reward functions: `lin1`, `lin2`, `lin3`, and `lin4`. For each reward function, we compared the DR rewards of the policy with exploration (using a selected α value) and without exploration ($\alpha = 0.01$). The statistical tests were applied to each pair of DR rewards across the four reward functions to check for any significant differences.

Results and Analysis

The following figures and tables illustrate the results of the comparisons for each reward function. The average DR rewards for both with- and without-exploration policies are shown, along with the results of the statistical tests (paired t-test and bootstrap test).

	Reward Type	Alpha Pair	Avg DR Rewards (No Exploration)	Avg DR Rewards (With Exploration)	Paired t-stat	Paired p-value	Observed Mean Diff	Bootstrap p-value
0	lin1	0.01 vs 5.0	1.105564	1.105564	NaN	NaN	0.000000e+00	1.0000
1	lin1	0.01 vs 20.0	1.105564	1.105564	NaN	NaN	0.000000e+00	1.0000
2	lin1	0.01 vs 50.0	1.105564	1.105564	1.000000	0.319748	-2.220446e-18	0.7457
3	lin2	0.01 vs 5.0	2.003172	2.003172	NaN	NaN	0.000000e+00	1.0000
4	lin2	0.01 vs 20.0	2.003172	2.003172	NaN	NaN	0.000000e+00	1.0000
5	lin2	0.01 vs 50.0	2.003172	2.003172	NaN	NaN	0.000000e+00	1.0000
6	lin3	0.01 vs 5.0	0.849734	0.849734	NaN	NaN	0.000000e+00	1.0000
7	lin3	0.01 vs 20.0	0.849734	0.849734	0.331847	0.740706	-2.220446e-18	0.5731
8	lin3	0.01 vs 50.0	0.849734	0.849734	0.779334	0.437642	-5.551115e-18	0.5327
9	lin4	0.01 vs 5.0	1.550043	1.550043	NaN	NaN	0.000000e+00	1.0000
10	lin4	0.01 vs 20.0	1.550043	1.550043	NaN	NaN	0.000000e+00	1.0000
11	lin4	0.01 vs 50.0	1.550043	1.550043	NaN	NaN	0.000000e+00	1.0000

Figure 1: Statistical Test Results for DR Rewards with Different Reward Functions

From the table above, we observe that the DR rewards for the policies with and without exploration are nearly identical across all four reward functions (`lin1`, `lin2`, `lin3`, `lin4`). Specifically:

- In all cases, the average DR rewards are the same for both exploration and no-exploration settings, indicating that **the exploration parameter (α) had no observable effect on the DR reward values**.
- The **paired t-test** results show non-significant p-values (where valid), with some comparisons returning NaN values due to the identical DR rewards, making it impossible to calculate a meaningful t-statistic.
- The **bootstrap test** results consistently return p-values close to or equal to 1.0000, further supporting the lack of any significant difference between the DR rewards for policies with and without exploration.

These results suggest that exploration had minimal or no effect on the DR rewards across these reward functions. Thus, both policies (with and without exploration) performed similarly in terms of the DR rewards, demonstrating that the DR estimator is robust and consistent under these test conditions, regardless of the exploration parameter.

4. Literature Review on Simulation Process in the Paper "To Update or Not to Update? Delayed Nonparametric Bandits with Randomized Allocation"

Here are 4 reward function setups to evaluate the adaptability and robustness of different strategies in various delayed reward scenarios.

4.1 Reward Function Definitions

Setup 1: Well-Separated Sinusoidal Functions Use two well-separated sinusoidal functions that represent different reward signals. One function is a shifted version of the other. The reward functions $g_1(x)$ and $g_2(x)$ are defined as follows:

$$g_1(x) = -2\sin(20\pi x) + 3,$$

$$g_2(x) = -2\sin(20\pi x) + 2,$$

where $x \in [0, 1]$. These functions are combined with an additional context variable x_2 to form the reward functions:

$$f_1(x_1, x_2) = g_1(x_1) \cdot x_2,$$
$$f_2(x_1, x_2) = g_2(x_1) \cdot x_2.$$

Setup 2: Piecewise Linear Functions Involves three piecewise linear functions $g_1(x)$, $g_2(x)$, and $g_3(x)$ that are well-separated over different regions in the covariate space. The piecewise functions are defined as follows:

$$g_1(x) = \begin{cases} 1, & 0 \le x < 0.5, \\ -10x + 6, & 0.5 \le x < 0.6, \\ 0, & x \ge 0.6, \end{cases}$$

$$g_2(x) = \begin{cases} 0, & 0 \le x < 0.5, \\ 10x - 5, & 0.5 \le x < 0.6, \\ 1, & x \ge 0.6, \end{cases}$$

$$g_3(x) = \begin{cases} 0, & 0 \le x < 0.3, \\ 20x - 6, & 0.3 \le x < 0.4, \\ 2, & 0.4 \le x < 0.6, \\ -20x + 14, & 0.6 \le x < 0.7, \\ 0, & x \ge 0.7. \end{cases}$$

These functions are used to define the reward functions as follows:

$$f_1(x_1, x_2) = x_2 g_1(x_1),$$

$$f_2(x_1, x_2) = x_2 g_2(x_1),$$
$$f_3(x_1, x_2) = x_2 g_3(x_1).$$

Setup 3: High-Frequency Sinusoidal Functions Consider two sinusoidal functions with high-frequency oscillations, representing a scenario where the best action alternates rapidly. The reward functions $g_1(x)$ and $g_2(x)$ are defined as:

$$g_1(x) = 2\cos(5\pi x) + 2,$$

$$g_2(x) = -2\sin(5\pi x) + 2,$$

where $x \in [0, 1]$.

Setup 4: Dominant Arm with High Regret Region One reward function dominates over the majority of the covariate space, except for a small region where it incurs a significantly higher regret. The reward functions $g_1(x)$ and $g_2(x)$ are defined as:

$$g_1(x) = 1, \quad \text{for all } x \in [0, 1],$$

$$g_2(x) = \begin{cases} 0, & 0 \le x < 0.5, \\ 100000x - 50000, & 0.5 \le x \le 0.502, \\ 200, & 0.502 < x < 0.503, \\ -100000x + 50500, & 0.503 \le x \le 0.505, \\ 0, & 0.505 < x \le 1. \end{cases}$$

4.2 Delay Reward Settings

The delay reward scenarios in this simulation introduce different levels of feedback delay to better understand how strategies adapt under non-instantaneous reward conditions. Four delay settings are utilized:

- **No Delay**: Rewards are observed instantaneously.
- **Delay 1**: A geometric delay with a 0.3 probability of observing the reward immediately, modeling a low probability of instant feedback.
- **Delay 2**: Every 5th reward is unobserved, and other rewards follow a geometric delay with $p = 0.3$, representing a moderate delay scenario.
- **Delay 3**: Each reward has a 0.7 probability of delay, with the delay time following a half-normal distribution (scale parameter $\sigma = 1500$), simulating a high degree of delay variability.
- **Delay 4**: Rewards are observed only at progressively longer intervals within four sequential segments, creating a severe delay environment with limited feedback.

These delay scenarios are designed to challenge the estimation accuracy of strategies under increasingly delayed reward conditions, thus evaluating their performance in adapting to non-instantaneous feedback.