Designing Time Series Experiments in A/B Testing with Transformer Reinforcement Learning

Xiangkun Wu

Joint work with **Qianglin Wen, Yingying Zhang, Hongtu Zhu,Ting Li, and Chengchun Shi**

March 22, 2026

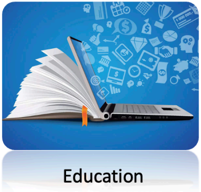
The first three authors contribute equally to the paper,and the last two authors are the corresponding authors.

Outline

- Background and Motivation
- Proposed Method
- More Explanations for this Method
- Experimental Results

Background

- A/B testing : A/B testing is a method to compare two methods and measure which performs better .
- Average Treatment Effect (ATE): measures the average difference in outcomes between the treatment group and the control group.
- Time series Experiment: Experimental units are sequentially exposed to different interventions (Treatment or Control) over a period of time, resulting in data with intrinsic temporal dependencies.

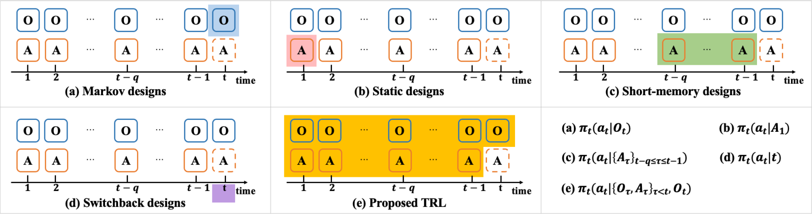


Challenges

- Carryover Effects: Current actions (e.g., dispatch policies) have delayed impacts on future outcomes by altering system states.
- Small Treatment Effects: Improvements are often very modest, typically ranging from only 0.5% to 2%
- Limited Duration: Experiments are usually restricted to a few weeks, resulting in small sample sizes for ATE estimation.

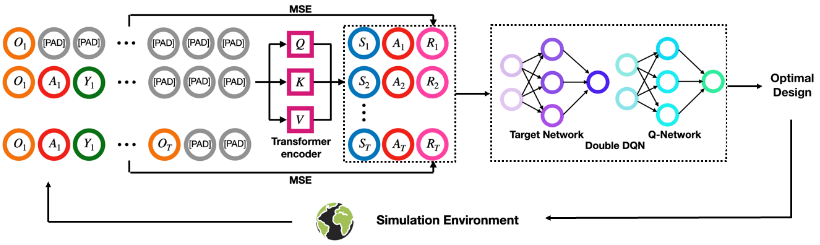
Motivation

- Information Loss: Existing designs (e.g., Markov or Switchback) often restrict treatment allocation to depend only on the initial action, current observation, or a short history.
- Restrictive Assumptions: Current methods rely on strong parametric assumptions (e.g., MDP, $ARMA$, or linearity) to simplify MSE optimization. These often fail in complex, non-linear real-world environments.



Proposed Method

- **Transformer Encoder**: Utilizes masked self-attention to process variable-length sequences, capturing and leveraging the **full historical information** H_{t-1}.
- **Double DQN Agent**: The state S_t is input into a Double Deep Q-Network to output an **optimal allocation strategy**, deciding the treatment A_t for the current time step.



Why this Method?

- **Impossibility theorem**: Suppose we set $\widehat{ATE}$ to the double robust estimator. Then there exist data generating processes $\{\mathcal{P}_t\}_t$ under which the optimal policy π that minimizes $\mathrm{Var}(\pi)$ depends on the entire past history for all $1 \leq t \leq T$, and this optimal policy is unique.
- **A complex optimization problem**: Owing to RL's effectiveness in policy optimization, RL has been adopted as a computational tool to tackle complex combinatorial optimization problems.

The Specific Process

- *State*: We define the state S_t as the full history $\{O_1, A_1, Y_1, \ldots, O_{t-1}, A_{t-1}, Y_{t-1}, O_t\}$ up to time point t, in order to capture all potential temporal dependencies that might influence the optimal treatment allocation at that time;
- *Action*: The action is the same to A_t, determining which policy (standard or new) to implement at time t;
- *Reward*: The reward at each time t is set to

$$R_t = -\alpha^{T-t}\big[\widehat{\mathrm{ATE}}(t) - \mathrm{ATE}_{mc}\big]^2, \tag{1}$$

- Transformer-based DDQN. We develop a variant of double deep Q-network that leverages transformer architectures to learn the optimal policy. Specifically, we define the Q-function as

$$Q_t(S_t, A_t) = \mathbb{E}_{\pi^{opt}}\left[\sum_{k=t}^{T} \gamma^{k-t} R_{k-t} \,\middle|\, S_t, A_t\right],$$

Numerical Experiments

- **TRL**: the proposed transformer RL design.
- **TMDP**/**NMDP**: Designs proposed by tailored for MDPs. These designs switch treatments on a daily basis and assign the same treatment within each day.
- **Switchback designs** proposed by (**HW**), (**BSZ**),(**XCT**), and(**WSY**). These designs switch treatments every few time intervals. WSY uses fixed switching intervals, whereas HW, BSZ, and XCT consider regular switchback designs with random switching intervals. The difference among HW, BSZ, and XCT lies in the ATE estimators they employ.

Simulations

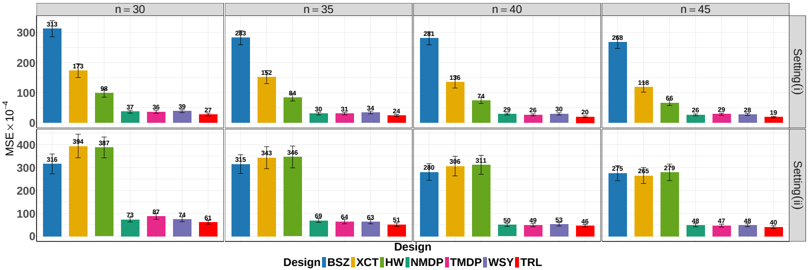


Figure 1: Barplots of empirical MSEs under different designs with their confidence intervals in the synthetic environment, across varying variances (Setting (i)) and transition structures (Setting (ii)).

Real data-based simulator

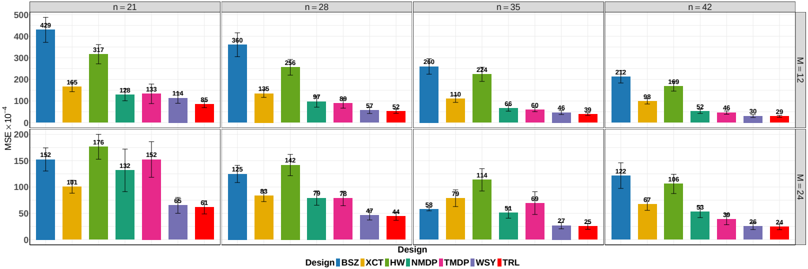


Figure 2: Barplots of the empirical MSEs under different designs in the real-data-based simulation, with a 5% performance improvement from the new policy.

Public dispatch simulator

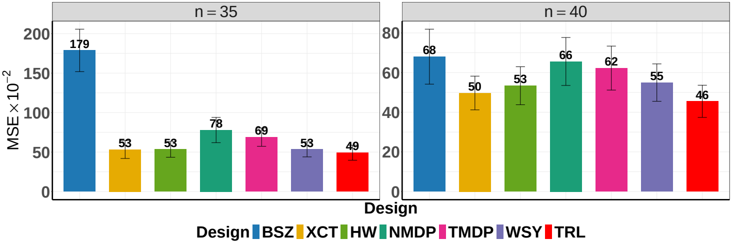


Figure 3: Barplots of empirical MSEs under different designs with their confidence intervals in the dispatch environment, across different days.

Thank you!