Mediator-Based Reward Design in Online Contextual Bandit

Anonymous Authors[1]

Abstract

In reinforcement learning, different reward functions may lead to the same optimal policy, while some reward functions can be substantially easier to learn. This paper proposes a framework that constructs surrogate rewards based on mediators between actions and rewards, informed by expert-provided causal directed acyclic graphs (DAGs). These DAGs encode domain knowledge from scientists. We show that our surrogate reward is unbiased and has reduced variance compared to the original reward when the mediator fully captures all causal pathways from the action to the reward. We further introduce an online reward-design agent that adaptively learns a surrogate reward in an unknown environment. We show that this reward-design agent can improve the regret guarantees of an online contextual bandit algorithm. Furthermore, our framework highlights improvement even without the surrogacy assumption, when total horizon is small relative to the error term induced by surrogacy violations. We complement the theoretical analysis with simulation studies with HeartSteps V1 dataset.

1. Introduction

Reinforcement learning (RL), in which an agent sequentially interacts with an unknown environment and learns to maximize cumulative rewards, is widely used for solving decision-making problems. The reward function is the agent's learning signal, and prior literature (Laud, 2004; Eschmann, 2021) has shown that a well-designed reward can preserve the optimal policy while substantially improving learning efficiency. This paper focuses on a central reward-design challenge: how to reduce variance when the original reward is **noisy**. The problem arises naturally in applications such as mobile health (Trella et al., 2023; Ghosh et al., 2024) and advertising (Li et al., 2010; Schwartz et al., 2017), where rewards are defined through highly noisy human behaviors. Figure 1 illustrates a smoking-cessation example (Battalio et al., 2021), where a mobile app delivers digital interventions to prompt users' stress-management behavior. These interventions reduce stress levels and thereby decrease subsequent smoking behavior, and the action-to-stress pathway is often less noisy than the action-to-smoking pathway.

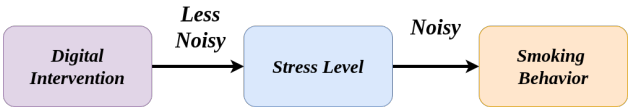


Figure 1. A noisy reward example in mobile health, where a digital intervention reduces smoking behavior by prompting stress management behavior, whose effect is mediated by stress level.

In domains like behavior psychology, the domain knowledge is often available in the form of causal directed acyclic graphs (DAGs) representing causal relations between variables. These causal relations are usually the domain experts' knowledge of the major causal effects between variables in the system (Gopnik et al., 2004), and they may be included in the design of RL algorithms to improve the learning efficiency (Deng et al., 2023; Bareinboim et al., 2024), especially in scientific fields (Yang et al., 2024; Gao et al., 2025). This paper aims to design surrogate rewards based on the **mediators**, the variables that mediate causal paths between actions and rewards. For example, the *stress level* is the mediator between digital interventions and smoking behavior in Figure 1.

The mediator-based surrogate index was introduced by Athey et al. (2019) to estimate the average treatment effect under the surrogacy assumption, with the primary outcome unobserved in the experimental sample and the treatment variable unobserved in the observational sample. In policy learning, Yang et al. (2024) imputed missing long-term outcomes using the mediator-based surrogate index and maximized the imputed outcomes via off-policy optimization in a batched bandit setting. Additional related-work background is deferred to Appendix A. Our work differs from the above by adaptively learning the surrogate rewards online. **A key feature of our work is its modularity**—we develop an online reward design agent that can interface with any existing

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

online bandit oracle, which is an agent making decisions in bandit environments; thus separating the reward design from the decision-making process. A technical challenge of this general framework is the non-stationarity in the surrogate reward due to the reward-learning process. We propose to use an **adversarial bandit oracle** to address this challenge.

Contributions. We make four contributions. First, we propose a framework for mediator-based reward design in contextual bandits and show that, under the surrogacy assumption, the resulting surrogate reward is unbiased with strictly lower variance. Second, we develop an online reward-design agent that adaptively learns the target mediator-based reward and interfaces with any online bandit oracle. Third, we prove that, when paired with an adversarial oracle to handle reward non-stationarity, the method achieves tighter regret bounds than LinUCB-style stochastic linear contextual bandits, with or without exact surrogacy. Fourth, we validate these gains empirically in simulation studies.

2. Problem Setup

Notation. We use I_d and 0_d to denote the identity matrix and the null matrix of dimension $d \times d$ respectively. We use $\mathbf{0}$ to denote the null vector. We use $\|\cdot\|$ to denote the L_2-norm. For a symmetric matrix A, we use $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ to denote the smallest and the greatest eigenvalue of A respectively, and $\|\mathbf{x}\|_A$ to denote $\sqrt{\mathbf{x}^\top A\mathbf{x}}$ for a vector $\mathbf{x}$. A random vector $\mathbf{x} \sim SG_d(\sigma)$ is a d-dimensional sub-Gaussian variable such that, for any $\mathbf{u} \in \mathbb{R}^d$ with $\|\mathbf{u}\| = 1$, $\mathbb{E}[\exp(\mathbf{u}^\top \mathbf{x}t)] \leq \exp(\sigma^2 t^2/2), \forall t \in \mathbb{R}$.

Contextual bandit with mediators. We consider a contextual bandit problem where the agent at each step $t = 1, \ldots, T$ observes context $\boldsymbol{S}_t \in \mathcal{S} \subset \mathbb{R}^{d_S}$, based on which it chooses an action $A_t \in \mathcal{A}$. The environment reveals the mediator $\boldsymbol{M}_t \in \mathbb{R}^{d_M}$ and the reward $R_t \in \mathbb{R}$ jointly from a distribution $P(\cdot \mid \boldsymbol{S}_t, A_t)$, where each $P(\cdot \mid \boldsymbol{s}, a)$ is a probability distribution over $\mathbb{R}^{d_M+1}$. The agent aims to optimize the cumulative reward $\sum_{t=1}^T R_t$. We define the mean reward function $R(\boldsymbol{s}, a) = \mathbb{E}[R_t \mid A_t = a, \boldsymbol{S}_t = \boldsymbol{s}]$.

Mediator-based surrogate reward. Figure 2 demonstrates a typical DAG for contextual bandit problems. When the dashed line is absent, $\boldsymbol{M}_t$ blocks all causal paths from A_t to R_t. In the causal inference literature, this is called surrogacy (Assumption 1). In our theoretical and numerical analysis, we will **allow weak violation of the surrogacy assumption** to accommodate different real world scenarios. Although it is implicitly assumed in general, we emphasize that **causal sufficiency** is assumed in our DAG, which means that for a set of variables V, every common cause of any pair of variables in V is also in V (Spirtes et al., 2001).

Assumption 1 (Surrogacy, (Prentice, 1989))**.** *The reward is independent of the action given the mediator and the state, i.e.* $R_t \perp A_t \mid \boldsymbol{M}_t, \boldsymbol{S}_t$.

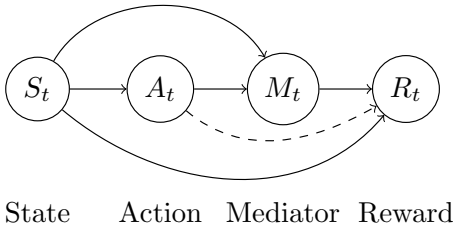


Figure 2. A DAG for contextual linear bandit with mediators. The surrogacy assumption is violated **when the dashed line between R and A exists**.

We aim to design a surrogate reward $\tilde{R}_t = f(\boldsymbol{M}_t, \boldsymbol{S}_t)$ for some function $f : \mathbb{R}^{d_M} \times \mathcal{S} \mapsto \mathbb{R}$. We show in Proposition 1 that under surrogacy, $\tilde{R}_t = \mathbb{E}[R_t \mid \boldsymbol{M}_t, \boldsymbol{S}_t]$ is an unbiased surrogate reward having lower variance than the true reward.

Proposition 1 (Unbiasedness under surrogacy)**.** *Define* $\tilde{R}^*(\boldsymbol{m}, \boldsymbol{s}) = \mathbb{E}[R_t \mid \boldsymbol{M}_t = \boldsymbol{m}, \boldsymbol{S}_t = \boldsymbol{s}]$*, the mean function of* $\tilde{R}_t$ *given* $\boldsymbol{M}_t = \boldsymbol{m}$ *and* $\boldsymbol{S}_t = \boldsymbol{s}$*. If Assumption 1 holds, we have*

$$\begin{aligned}
&\mathbb{E}[R_t \mid A_t = a, \boldsymbol{S}_t = \boldsymbol{s}] \\
=&\mathbb{E}[\tilde{R}^*(\boldsymbol{M}_t, \boldsymbol{S}_t) \mid A_t = a, \boldsymbol{S}_t = \boldsymbol{s}], \\
&\mathrm{Var}(R_t \mid A_t = a, \boldsymbol{S}_t = \boldsymbol{s}) \\
\geq&\mathrm{Var}(\tilde{R}^*(\boldsymbol{M}_t, \boldsymbol{S}_t) \mid A_t = a, \boldsymbol{S}_t = \boldsymbol{s})
\end{aligned}$$

for any $a \in \mathcal{A}$ *and* $\boldsymbol{s} \in \mathcal{S}$*. The inequality strictly holds if and only if there exists some subset* $\mathcal{M} \subseteq supp(\boldsymbol{M}_t | A_t = a, \boldsymbol{S}_t = \boldsymbol{s})$ *with* $P(\mathcal{M}) > 0$*, such that* $\mathrm{Var}(R_t | \boldsymbol{M}_t = \boldsymbol{m}, A_t = a, \boldsymbol{S}_t = \boldsymbol{s}) > 0$ *for all* $\boldsymbol{m} \in \mathcal{M}$*.*

The proof relies on the law of total variance and is deferred to Appendix B.

A linear working model. Throughout the paper, we make an additional linear assumption about the reward and mediator generating processes:

$$\begin{aligned}
R_t &= \boldsymbol{\theta}_{A_t}^\top \boldsymbol{S}_t + \boldsymbol{\theta}_S^\top \boldsymbol{S}_t + \boldsymbol{\theta}_M^\top \boldsymbol{M}_t + \epsilon_t, \\
\boldsymbol{M}_t &= \boldsymbol{\Gamma}_{A_t} \boldsymbol{S}_t + \boldsymbol{\omega}_t,
\end{aligned} \tag{1}$$

where $\boldsymbol{\theta}_a \in \mathbb{R}^{d_S}$, $\boldsymbol{\Gamma}_a \in \mathbb{R}^{d_M \times d_S}$ for each $a \in \mathcal{A}$, $\boldsymbol{\theta}_S \in \mathbb{R}^{d_S}, \boldsymbol{\theta}_M \in \mathbb{R}^{d_M}$, and $\epsilon_t \sim SG_1(\sigma_\epsilon)$, $\boldsymbol{\omega}_t \sim SG_{d_M}(\sigma_\omega)$. Note that we have $\boldsymbol{\theta}_{A_t}$ here to **allow violation of the surrogacy assumption**, which holds when $\boldsymbol{\theta}_a \equiv \mathbf{0}$ for all $a \in \mathcal{A}$.

Policy and regret. We define a policy π as a mapping from context to action, i.e., $\pi : \mathcal{S} \mapsto \mathcal{A}$. The performance of an online algorithm is measured by regret, which is defined as the difference between the expected cumulative rewards generated by the optimal policy and the expected cumulative rewards generated by the algorithm. Assuming $\pi^*(\boldsymbol{s}) := \arg\max_a R(\boldsymbol{s}, a)$ is the optimal policy, we define the regret as

$$\text{Regret}_T := \mathbb{E}\left[\sum_{t=1}^{T} \left(R(\boldsymbol{S}_t, \pi^*(\boldsymbol{S}_t)) - R(\boldsymbol{S}_t, A_t)\right)\right].$$

3. Mediator-based Online Reward Learning

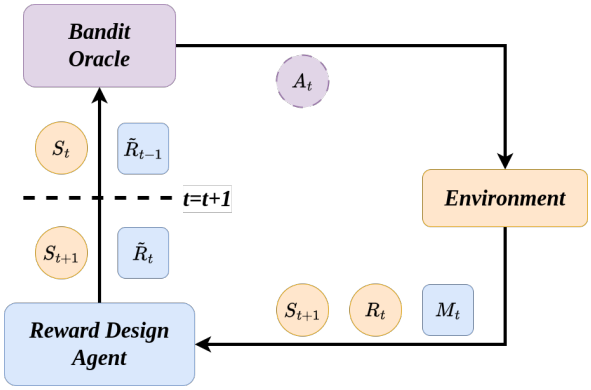


Figure 3. At each decision time t, the oracle observes the state $\boldsymbol{S}_t$, chooses an action $A_t \in \mathcal{A}$, and receives the surrogate reward $\tilde{R}_t$. Variables in **orange (circle)** are observed variables in the standard bandit setting, and those in **blue (rounded square)** are the additional variables introduced in this framework for surrogate reward design. Action is in **purple (dashed circle)** to show that it's dependent on the oracle.

Our proposed solution has two critical components, an **online reward design agent** that adaptively learns a reward mapping based on $\boldsymbol{S}_t$ and $\boldsymbol{M}_t$, and an online bandit oracle. The interaction between the online reward design agent and the online bandit oracle is characterized in Figure 3.

3.1. Online Reward Design Agent

The **online reward design agent** is learning the function $\tilde{R}^*(\boldsymbol{m}, \boldsymbol{s}) = \mathbb{E}[R_t | \boldsymbol{M}_t = \boldsymbol{m}, \boldsymbol{S}_t = \boldsymbol{s}] = \boldsymbol{s}^\top \boldsymbol{\theta}_S + \boldsymbol{m}^\top \boldsymbol{\theta}_M$, where $\boldsymbol{\theta}_S$ and $\boldsymbol{\theta}_M$ are the coefficients of the linear model in Eq. (1).

We use online ridge regression to learn the coefficients of the linear model. At each decision time t, the estimators are

$$\begin{bmatrix} \hat{\boldsymbol{\theta}}_{S,t} \\ \hat{\boldsymbol{\theta}}_{M,t} \end{bmatrix} = \left(\sum_{\tau=1}^{t-1} \boldsymbol{X}_\tau \boldsymbol{X}_\tau^\top + \lambda \mathbf{I}_{d_S+d_M}\right)^{-1} \sum_{\tau=1}^{t-1} \boldsymbol{X}_\tau R_\tau, \tag{2}$$

where λ is a tuning parameter and $\boldsymbol{X}_t = (\boldsymbol{S}_t^\top, \boldsymbol{M}_t^\top)^\top$ denotes the regression covariates. The overall process is described in Algorithm 1.

Algorithm 1: Mediator-based Online Reward Learning

Input: online bandit oracle $\mathcal{O}$; dataset for reward design $\mathcal{D}_0 = \{\}$; ridge regularization parameter λ.

1: Initialize $\pi_1(\cdot|s)$ uniform over $\mathcal{A}$ for all s.
2: **for** $t = 1, \ldots, T$ **do**
3: Observe current state $\boldsymbol{S}_t$
4: Sample $A_t \sim \pi_t(\boldsymbol{S}_t)$, and environment generates $(R_t, \boldsymbol{M}_t) \sim P(\cdot \mid \boldsymbol{S}_t, A_t)$
5: Add observation for reward design $\mathcal{D}_t = \mathcal{D}_{t-1} \cup \{(\boldsymbol{S}_t, \boldsymbol{M}_t, R_t)\}$
6: Run ridge regression in Eq. (2) on $\mathcal{D}_t$ to get estimators $\hat{\boldsymbol{\theta}}_{S,t}$ and $\hat{\boldsymbol{\theta}}_{M,t}$
7: Construct reward $\tilde{R}_t = \hat{\boldsymbol{\theta}}_{S,t}^\top \boldsymbol{S}_t + \hat{\boldsymbol{\theta}}_{M,t}^\top \boldsymbol{M}_t$
8: Update the oracle $\mathcal{O}$ with $(\boldsymbol{S}_t, A_t, \tilde{R}_t)$ and obtain π_{t+1}
9: **end for**

3.2. The Need for an Adversarial Bandit Oracle

Online learning of the surrogate reward induces non-stationarity in the reward signal, so we must use an adversarial bandit oracle rather than a stochastic one.

Assumption 2 (Variance-adaptive adversarial bandit regret guarantee)**.** *We assume that the adversarial bandit oracle has upper bound on the following regret, with a high probability 1-δ,*

$$\text{AdvReg}_T = \tilde{\mathcal{O}}\left(\sigma_R^2 \cdot \textit{poly}(d_S, |\mathcal{A}|, \log(T/\delta)) \cdot T^{1-\alpha}\right),$$

where $\alpha \in [0, \frac{1}{2}]$, poly stands for polynomial functions, $\tilde{\mathcal{O}}$ hides the logarithmic terms, and $\sigma_R^2 := \sup_{\boldsymbol{s},a} \text{Var}[R_t \mid \boldsymbol{S}_t = \boldsymbol{s}, A_t = a]$ is the variance of the reward signal.

3.3. Regret Analysis

We first introduce an additional regularity assumption concerning the boundedness of the coefficients in Eq. (1).

Assumption 3 (Bounded coefficients)**.** *We assume that there exist constants $\mathcal{E}, C > 0$ s.t.*

$$\forall t \in \{1, 2, \cdots, T\}, \|\boldsymbol{S}_t\| \le 1 \textit{ almost surely},$$
$$\max_{a \in \mathcal{A}} \|\boldsymbol{\Gamma}_a\| \le C, \max_{a \in \mathcal{A}} \|\boldsymbol{\theta}_a\| \le \mathcal{E}C,$$
$$\|\boldsymbol{\theta}_M\| \le C.$$

*Note that **surrogate error** $\mathcal{E}$ controls the ratio of the effect of A_t on R_t not captured by $\boldsymbol{M}_t$ to that captured by $\boldsymbol{M}_t$. A greater $\mathcal{E}$ implies more severe violation of the surrogacy assumption. $\mathcal{E} = 0$ implies surrogacy (Assumption 1).*

We now present the high-probability regret bound for Algorithm 1 when the adversarial online bandit oracle is chosen

as *RealLinExp3* (Neu & Olkhovskaya, 2020). The standard proof is extended to incorporate the noise of the reward.

Theorem 1 (Regret bound for Algorithm 1)**.** *When Assumption 3 holds, the regret of Algorithm 1 is bounded with a high probability by*

$$\tilde{\mathcal{O}}\left(\sigma_\omega\sqrt{d_S|\mathcal{A}|T}+\sigma_\epsilon\sqrt{(d_S+d_M)T}+\mathcal{E}T\right). \quad (3)$$

Now we compare our regret bound with the original bound (omitting the $\tilde{\mathcal{O}}$ notation). The minimax regret bound for stochastic contextual linear bandits is $\sigma_\omega\sqrt{d_S|\mathcal{A}|T}+\sigma_\epsilon\sqrt{d_S|\mathcal{A}|T}$. To interpret the improvement, we note that our regret bound in Eq. (3) has a reduction in a multiplicative factor of $\sqrt{|\mathcal{A}|}$ at the price of an additional $\sqrt{d_M}$ term. This improvement is due to the fact that the ridge regression estimator pools information across all arms to learn the effect of the mediator on the reward, while naive contextual linear bandits with a discrete action set estimates the coefficient of each arm separately. Making this improvement requires an additional regression based on the mediator, thus leading to the additional $\sqrt{d_M}$ term. The linear term $\mathcal{E}T$ only appears if Assumption 1 is violated; this term comes from the bias in the surrogate reward due to the violation. Accordingly, even when there is a violation of Assumption 1, our method has a better regret bound when

$$\sqrt{T}\le\frac{\sigma_\epsilon}{\mathcal{E}}(\sqrt{d_S|\mathcal{A}|}-\sqrt{d_S+d_M}). \quad (4)$$

The improvement is significant in the following cases:

1. the mediator is less noisy than the original reward, i.e., $\sigma_\omega \ll \sigma_\epsilon$;
2. the context dimension is substantially larger than the mediator dimension, i.e., $d_S \gg d_M$;
3. the action space $\mathcal{A}$ is relatively large; and
4. the uncaptured direct effect $\mathcal{E}$ is small.

The full proof of Theorem 1 is deferred to Appendix C.

4. Simulation Study

We evaluate our framework on the HeartSteps V1 mobile health dataset, with synthetic-data results moved to Appendix D and environment details deferred to Appendix E. We compare our reward-designed algorithms (R-LinExp3, R-LinUCB) against the base oracles (LinExp3, LinUCB) using expected reward difference relative to LinUCB, in an environment calibrated to the HeartSteps V1 mobile health study (Klasnja et al., 2019), where every 30 minutes an agent decides whether to prompt physical activity based on observed user states (e.g., weather, location, and recent activity summaries). The mediator and reward are functions of short- and long-term physical activity, respectively (details in Appendix E.2).

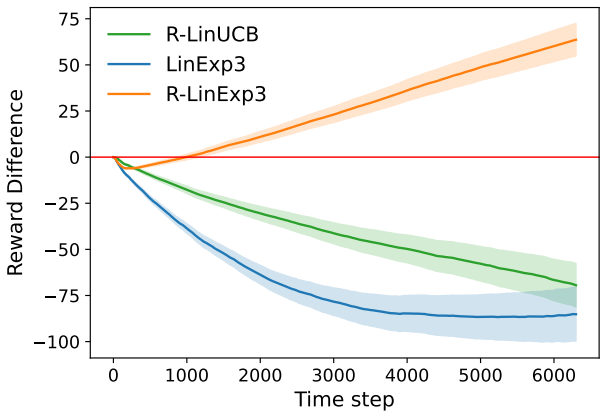


Figure 4. Real-world (HeartSteps V1) performance. The axes have the same meaning as the synthetic-data plots in Appendix D. Error bars represent standard errors over 500 independent runs.

Figure 4 summarizes the reward difference relative to the LinUCB baseline over time. Consistent with the synthetic experiments moved to Appendix D, we observe an initial "warm-up" period during which reward-designed agents may underperform. After this transient phase, R-LinExp3 achieves higher rewards than LinUCB and maintains a clear advantage over R-LinUCB, supporting our motivation for adopting an adversarial bandit oracle to accommodate the nonstationarity induced by online reward learning.

Moreover, the reward-designed variants (R-LinExp3 and R-LinUCB) exhibit, on average, 70.4% of the standard error of LinExp3 across time. This empirical reduction in variability supports the claim that incorporating mediator-based structural information decreases the effective noise in the learning signal.

Finally, we do not observe the long-horizon performance deterioration of reward-designed agents predicted by the discussion following Eq. (3), because the real-world dataset contains only 6,394 decision points, the horizon is insufficiently large for the bias term to dominate.

5. Conclusion

In this paper, we propose a flexible framework for low-noise surrogate rewards design, exploiting the structural information provided by causal DAGs. We analyze the regret bound of our framework when an adversarial bandit oracle is used. From both theoretical and applicative views, we show significant improvements when high-quality (low-noise, high surrogacy) mediators are available. This framework accommodates real-world applications by harnessing the power of existing online-learning algorithms and domain knowledge.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Online least squares estimation with self-normalized processes: An application to bandit problems. *arXiv preprint arXiv:1102.2670*, 2011.

Athey, S., Chetty, R., Imbens, G. W., and Kang, H. The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical report, National Bureau of Economic Research, 2019.

Banihashem, K., Hajiaghayi, M., Shin, S., and Springer, M. An improved relaxation for oracle-efficient adversarial contextual bandits. *Advances in Neural Information Processing Systems*, 36, 2024.

Bareinboim, E., Zhang, J., and Lee, S. An introduction to causal reinforcement learning. Technical Report R-65, Causal Artificial Intelligence Lab, Columbia University, 12 2024.

Baron, R. M. and Kenny, D. A. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173, 1986.

Battalio, S. L., Conroy, D. E., Dempsey, W., Liao, P., Menictas, M., Murphy, S., Nahum-Shani, I., Qian, T., Kumar, S., and Spring, B. Sense2stop: a micro-randomized trial using wearable sensors to optimize a just-in-time-adaptive stress management intervention for smoking relapse prevention. *Contemporary Clinical Trials*, 109: 106534, 2021.

Deng, Z., Jiang, J., Long, G., and Zhang, C. Causal reinforcement learning: A survey. *Transactions on Machine Learning Research*, 2023.

Digitale, J. C., Martin, J. N., and Glymour, M. M. Tutorial on directed acyclic graphs. *Journal of Clinical Epidemiology*, 142:264–267, 2022. ISSN 0895-4356. doi: https://doi.org/10.1016/j.jclinepi.2021.08.001. URL https://www.sciencedirect.com/science/article/pii/S0895435621002407.

Eschmann, J. Reward function design in reinforcement learning. *Reinforcement Learning Algorithms: Analysis and Applications*, pp. 25–33, 2021.

Gao, D., Lai, H.-Y., Klasnja, P., and Murphy, S. Harnessing causality in reinforcement learning with bagged decision times. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, pp. 658–666. PMLR, 2025.

Ghosh, S., Kim, R., Chhabria, P., Dwivedi, R., Klasnja, P., Liao, P., Zhang, K., and Murphy, S. Did we personalize? assessing personalization by an online reinforcement learning algorithm using resampling. *Machine Learning*, pp. 1–37, 2024.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., and Danks, D. A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, 111(1):3, 2004.

Gupta, A., Pacchiano, A., Zhai, Y., Kakade, S., and Levine, S. Unpacking reward shaping: Understanding the benefits of reward engineering on sample complexity. *Advances in Neural Information Processing Systems*, 35:15281–15295, 2022.

Hare, J. Dealing with sparse rewards in reinforcement learning, 2019. URL https://arxiv.org/abs/1910.09281.

He, J., Zhou, D., Zhang, T., and Gu, Q. Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. *Advances in neural information processing systems*, 35:34614–34625, 2022.

Hirtz, T., Tian, H., Yang, Y., and Ren, T.-L. Unsupervised reward engineering for reinforcement learning controlled manufacturing. *Journal of Intelligent Manufacturing*, pp. 1–14, 2024.

Ibrahim, S., Mostafa, M., Jnadi, A., Salloum, H., and Osinenko, P. Comprehensive overview of reward engineering and shaping in advancing reinforcement learning applications. *IEEE Access*, 12:175473–175500, 2024. doi: 10.1109/ACCESS.2024.3504735.

Icarte, R. T., Klassen, T. Q., Valenzano, R., and McIlraith, S. A. Reward machines: Exploiting reward function structure in reinforcement learning. *Journal of Artificial Intelligence Research*, 73:173–208, 2022.

Klasnja, P., Smith, S., Seewald, N. J., Lee, A., Hall, K., Luers, B., Hekler, E. B., and Murphy, S. A. Efficacy of contextually tailored suggestions for physical activity: a micro-randomized optimization trial of heartsteps. *Annals of Behavioral Medicine*, 53(6):573–582, 2019.

Kuroki, Y., Rumi, A., Tsuchiya, T., Vitale, F., and Cesa-Bianchi, N. Best-of-both-worlds algorithms for linear contextual bandits. In Dasgupta, S., Mandt, S., and Li, Y. (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1216–1224. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/kuroki24a.html.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Laud, A. D. *Theory and application of reward shaping in reinforcement learning*. University of Illinois at Urbana-Champaign, 2004.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

Luis, J. J. G. Policy gradient rl algorithms as directed acyclic graphs. *arXiv preprint arXiv:2012.07763*, 2020.

Mohan, A., Zhang, A., and Lindauer, M. Structure in deep reinforcement learning: A survey and open problems. *Journal of Artificial Intelligence Research*, 79:1167–1236, 2024.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with noisy labels. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/3871bd64012152bfb53fdf04b401193f-Paper.pdf.

Neu, G. and Olkhovskaya, J. Efficient and robust algorithms for adversarial linear contextual bandits. In *Conference on Learning Theory*, pp. 3049–3068. PMLR, 2020.

Ng, A., Harada, D., and Russell, S. J. Policy invariance under reward transformations: Theory and application to reward shaping. In *International Conference on Machine Learning*, 1999. URL https://api.semanticscholar.org/CorpusID:5730166.

Prentice, R. L. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine*, 8(4): 431–440, 1989.

Rakhlin, A. and Sridharan, K. Bistro: An efficient relaxation-based method for contextual bandits. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1977–1985, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/rakhlin16.html.

Schwartz, E. M., Bradlow, E. T., and Fader, P. S. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4): 500–522, 2017.

Scott, C., Blanchard, G., and Handy, G. Classification with asymmetric label noise: Consistency and maximal denoising. In Shalev-Shwartz, S. and Steinwart, I. (eds.), *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pp. 489–511, Princeton, NJ, USA, 12–14 Jun 2013. PMLR. URL https://proceedings.mlr.press/v30/Scott13.html.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, prediction, and search*. MIT press, 2001.

Syrgkanis, V., Luo, H., Krishnamurthy, A., and Schapire, R. E. Improved regret bounds for oracle-based adversarial contextual bandits. *Advances in Neural Information Processing Systems*, 29, 2016.

Thost, V. and Chen, J. Directed acyclic graph neural networks, 2021. URL https://arxiv.org/abs/2101.07965.

Trella, A. L., Zhang, K. W., Nahum-Shani, I., Shetty, V., Doshi-Velez, F., and Murphy, S. A. Reward design for an online reinforcement learning algorithm supporting oral self-care. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 15724–15730, 2023.

Tropp, J. A. et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Wang, J., Liu, Y., and Li, B. Reinforcement learning with perturbed rewards. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 6202–6209, 2020.

Yang, J., Eckles, D., Dhillon, P., and Aral, S. Targeting for long-term outcomes. *Management Science*, 70(6): 3841–3855, 2024.

A. Related Work

Reward Design. The term **reward engineering** is used interchangeably with **reward design** sometimes (Hirtz et al., 2024), where both terms refer to the creation and modification of reward functions to align the learned policy with the goal of the task. However, in other literature (Gupta et al., 2022; Ibrahim et al., 2024), **reward engineering** refers to creating a reward function that aligns with the human intent. Our method differs from the latter setting as we directly observe rewards. **Reward shaping** involves modifying the reward function to guide exploration or reduce variance in the learning process, without changing the optimal policy (Laud, 2004). Existing methods such as potential-based reward shaping (Ng et al., 1999) and dense reward functions that allocate a sparse reward into multiple steps (Hare, 2019; Eschmann, 2021) guide exploration but rarely address high-noise scenarios, which is the central issue tackled in this paper. Our work uniquely focuses on mediator-based surrogate rewards to reduce variance from immediate noisy rewards.

Reward Design for Noisy Data. Prior work (Natarajan et al., 2013; Scott et al., 2013; Wang et al., 2020) has studied the definition of unbiased surrogate rewards, where the knowledge of the **noise** is used to recover the true loss. But these works are mainly concerned about eliminating bias instead of **variance reduction**. What's more, the structural knowledge in the generation process of the reward has not garnered adequate attention.

Reward Design with Structural Knowledge. Mohan et al. (2024) provide a comprehensive survey on accelerating the learning of RL through structural knowledge. Our proposed method broadly falls into the range of incorporating side information. Icarte et al. (2022) proposed a reward shaping utilizing structural information but provided no theoretical explanation and posed requirements about the reward function specification. On the other hand, we propose to exploit a common format for depicting structural information in causal inference, Directed Acyclic Graphs.

Causal Directed Acyclic Graphs and Mediator. Causal Directed Acyclic Graphs (DAGs) are constructed to depict prior knowledge about specific causal systems (Digitale et al., 2022). DAGs have been used in causal problems to inform study design, statistical analysis, and algorithm improvement (Luis, 2020; Thost & Chen, 2021). Mediator, as defined by Baron & Kenny (1986), is a variable that captures most part of relation from an independent variable to its dependent variable. A more straightforward definition is given in Figure 5.

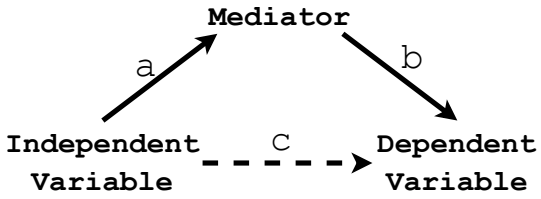


Figure 5. A variable functions as a mediator when it meets the following conditions: (a) variations in levels of the independent variable significantly account for variations in the presumed mediator (i.e., Path a), (b) variations in the mediator significantly account for variations in the dependent variable (i.e., Path b), and (c) when Paths a and b are controlled, a previously significant relation between the independent and dependent variables is no longer significant, with the strongest demonstration of mediation occurring when Path c is zero. (Baron & Kenny, 1986)

B. Proof of Proposition 1

For any $a \in \mathcal{A}$ and $\boldsymbol{s} \in \mathcal{S}$,

$$
\begin{aligned}
&\mathbb{E}[\tilde{R}^*(\boldsymbol{M}_t, \boldsymbol{s}) \mid A_t = a, \boldsymbol{S}_t = \boldsymbol{s}] \\
=&\mathbb{E}[\mathbb{E}[R_t \mid \boldsymbol{M}_t, \boldsymbol{S}_t = \boldsymbol{s}] \mid A_t = a, \boldsymbol{S}_t = \boldsymbol{s}] \\
=&\mathbb{E}[\mathbb{E}[R_t \mid \boldsymbol{M}_t, \boldsymbol{S}_t = \boldsymbol{s}, A_t = a] \mid A_t = a, \boldsymbol{S}_t = \boldsymbol{s}] && \text{(Assumption 1)} \\
=&\mathbb{E}[R_t \mid \boldsymbol{S}_t = \boldsymbol{s}, A_t = a]. && \text{(Law of total expectation)}
\end{aligned}
$$

The variance reduction follows from the law of total variance

$$\begin{aligned}
&\mathrm{Var}(R_t \mid A_t = a, \boldsymbol{S}_t = \boldsymbol{s}) \\
=&\,\mathrm{Var}(\mathbb{E}[R_t \mid \boldsymbol{M}_t, A_t = a, \boldsymbol{S}_t = \boldsymbol{s}] \mid A_t = a, \boldsymbol{S}_t = \boldsymbol{s}) \\
+&\,\mathbb{E}[\mathrm{Var}(R_t \mid \boldsymbol{M}_t, A_t = a, \boldsymbol{S}_t = \boldsymbol{s}) \mid A_t = a, \boldsymbol{S}_t = \boldsymbol{s}] \qquad \text{(Law of total variance)} \\
\geq&\,\mathrm{Var}(\mathbb{E}[R_t \mid \boldsymbol{M}_t, A_t = a, \boldsymbol{S}_t = \boldsymbol{s}] \mid A_t = a, \boldsymbol{S}_t = \boldsymbol{s}) \\
=&\,\mathrm{Var}(\tilde{R}^*(\boldsymbol{M}_t, \boldsymbol{s}) \mid A_t = a, \boldsymbol{S}_t = \boldsymbol{s}).
\end{aligned}$$

If there exists $\boldsymbol{m} \in supp(\boldsymbol{M}_t | A_t = a, \boldsymbol{S}_t = \boldsymbol{s})$ such that $\mathrm{Var}(R_t | \boldsymbol{M}_t = \boldsymbol{m}, A_t = a, \boldsymbol{S}_t = \boldsymbol{s}) > 0$, then $\mathbb{E}[\mathrm{Var}(R_t \mid \boldsymbol{M}_t, A_t = a, \boldsymbol{S}_t = \boldsymbol{s}) \mid A_t = a, \boldsymbol{S}_t = \boldsymbol{s}] > 0$ and thus the inequality strictly holds.

C. Proof of Theorem 1

Proof Sketches. The regret in our framework can be attributed to two components: (1) The surrogate reward error incurred by the reward design agent, which accounts for both the cost of collecting structural information and the inaccuracy of the mediators. We bound the information cost using properties of the ℓ^2-regularized least-squares estimator (Abbasi-Yadkori et al., 2011; Tropp et al., 2015), and we capture the mediator's inaccuracy with a linear term $\mathcal{E}T$. (2) The standard regret arising from the exploration–exploitation trade-off in the bandit setting, for which extensive analyses exist in the online learning literature (Neu & Olkhovskaya, 2020).

Due to the sub-Gaussianity of ϵ_t and ω_t, $\forall \delta \in (0, 1)$,

$$\begin{aligned}
&\mathbb{P}(|\epsilon_t| \geq z) \leq 2\exp\left(-\frac{z^2}{2\sigma_\epsilon^2}\right), \forall z \in \mathbb{R}^+ \\
\Longrightarrow\ &\mathbb{P}\left(\max_{t=1,\cdots,T} |\epsilon_t| \leq \sigma_\epsilon \sqrt{2\log(2T/\delta)}\right) \geq 1 - \delta. \\
&\mathbb{P}\left(\max_{\|\mathbf{u}\|\leq 1} |\mathbf{u}^\top \omega_t| \leq \sigma_\omega \sqrt{2\log(2d_M/\delta)}\right) \geq 1 - \delta.
\end{aligned}$$

To simplify notation we denote

$$R_\omega := \sigma_\omega \sqrt{2\log(2d_M/\delta)} = \tilde{O}(\sigma_\omega).$$

C.1. Concentration for Sample Covariance Matrix

In this subsection, we introduce and prove a proposition to be used later. For the purpose of analyzing the growth rate of the design matrix, we restrict attention to a stationary phase of the interaction. In particular, we assume that the policy is fixed throughout this analysis (i.e., $\pi_t \equiv \pi$) and that the context sequence $\{S_t\}_{t\geq 1}$ is i.i.d. Any finite initial exploration phase does not affect the asymptotic order and is therefore omitted from this analysis. Let $d := d_S + d_M$.

Proposition 2 (Sample Covariance Matrix Concentration)**.** *In a contextual linear bandit problem, suppose action $a \in \mathcal{A}$ is chosen $T_a \in \mathbb{N}$ times up to time T. So we obtain $|\mathcal{A}|$ groups of observations X. The group of a contains T_a observations, whose i-th observation is denoted as $X_i^{(a)}$. For any $\lambda > 0$,*

$$\lambda_{\min}\Big(\lambda I + \sum_{t=1}^{T} X_t X_t^\top\Big) = \tilde{\Omega}(T).$$

Proof. Let Σ_S and Σ_ω be the covariance matrix of states $\boldsymbol{S}$ and mediator noises ω. We denote

$$\begin{aligned}
V_T &:= \lambda I + \sum_{t=1}^{T} X_t X_t^\top, U_a := \begin{bmatrix} I_{d_S} \\ \boldsymbol{\Gamma}_a \end{bmatrix} \in \mathbb{R}^{d \times d_S}, \\
\Sigma_a &:= U_a \Sigma_S U_a^\top + \mathrm{diag}\{0_{d_S}, \Sigma_\omega\}.
\end{aligned}$$

Under the fixed policy π and i.i.d. contexts, the sequence $\{X_t\}_{t=1}^T$ is i.i.d. (with sub-Gaussian tails). Let $\Sigma := \mathbb{E}[X_t X_t^\top]$ denote the population second-moment matrix induced by the context distribution and the fixed policy π. Assume $\lambda_0 :=$

$\lambda_{\min}(\Sigma) > 0$. Moreover, since $M_t = \Gamma_{A_t} S_t + \omega_t$ and ω_t is independent sub-Gaussian noise, X_t is a sub-Gaussian random vector. In particular, there exists $\sigma > 0$ such that $X_t \sim SG_d(\sigma)$. One convenient sufficient bound is to define σ^2 as

$$\lambda_{\max}(\Sigma) \le \max\left\{\max_{a\in\mathcal{A}} \lambda_{\max}(U_a \Sigma_S U_a^\top),\ \lambda_{\max}(\Sigma_\omega)\right\}.$$

We know that

$$\lambda_{\min}(V_T) \ge T\lambda_{\min}\underbrace{\left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t^\top\right)}_{\text{denoted as } \hat{\Sigma}}.$$

By standard concentration for sample covariance matrices of i.i.d. sub-Gaussian vectors (Tropp et al., 2015), we have with high probability

$$|\lambda_{\min}(\hat{\Sigma}) - \lambda_{\min}(\Sigma)| \le \rho(\hat{\Sigma} - \Sigma) = \tilde{\Theta}\left(\sigma^2\sqrt{d/T}\right),$$

which implies

$$\begin{aligned}
&\lambda_{\min}(\hat{\Sigma}) \ge \lambda_0 - \tilde{\Theta}\left(\sigma^2\sqrt{d/T}\right).\\
\Longrightarrow\ &\lambda_{\min}(V_T) \ge T\Big(\lambda_0 - \tilde{\Theta}\big(\sigma^2\sqrt{d/T}\big)\Big) = \tilde{\Omega}(T).
\end{aligned}$$

□

C.2. Surrogate Reward Error

In this section, we bound the regret incurred by using the surrogate reward designed by the reward design agent. We begin by analyzing the general properties of the estimator given the true covariance matrix, and then apply Proposition 2 to estimate the sample covariance matrix in this environment.

We use the following notation:

1. $$\begin{aligned}R(\boldsymbol{s}, a) :=& \mathbb{E}[R_t \mid A_t = a, \boldsymbol{S}_t = \boldsymbol{s}]\\ =& \left(\boldsymbol{\theta}_S^\top + \boldsymbol{\theta}_M^\top \boldsymbol{\Gamma}_a + \boldsymbol{\theta}_a^\top\right)\boldsymbol{s}\end{aligned}$$

   is the expectation of true reward for the whole problem, which cannot be known exactly because our observation is corrupted by noise.

2. $r_t(\boldsymbol{s}, a) := \hat{\boldsymbol{\theta}}_{M,t}^\top \boldsymbol{\Gamma}_a \boldsymbol{s} + \hat{\boldsymbol{\theta}}_{M,t}^\top \omega_t$ is an estimation of the contribution from action a to $R(\boldsymbol{s}, a)$ at decision time t. This is the surrogate reward we feed to the oracle since it only determines the selection of action. And we have $|r_t(\boldsymbol{s}, a)| \le C^2 + CR_\omega$ according to Assumption 3.

We first analyze the error incurred when using r_t to replace the true contribution of a in R. We define

$$\begin{aligned}
e_t(\boldsymbol{s}, a) &= \left(\boldsymbol{\theta}_M^\top \boldsymbol{\Gamma}_a + \boldsymbol{\theta}_a^\top\right)\boldsymbol{s} - \mathbb{E}[r_t(\boldsymbol{s}, a)]\\
&= \left(\boldsymbol{\theta}_M - \hat{\boldsymbol{\theta}}_{M,t}\right)^\top \boldsymbol{\Gamma}_a \boldsymbol{s} + \boldsymbol{\theta}_a^\top \boldsymbol{s}\\
&\le \sup_{\boldsymbol{s},a} \|\boldsymbol{\Gamma}_a \boldsymbol{s}\| \cdot \|\boldsymbol{\theta}_M - \hat{\boldsymbol{\theta}}_{M,t}\| + \|\boldsymbol{\theta}_a\|\|\boldsymbol{s}\|\\
&\le C\|\boldsymbol{\theta}_M - \hat{\boldsymbol{\theta}}_{M,t}\| + \mathcal{E}C. \qquad \text{(Assumption 3)}
\end{aligned}$$

We assume that after a sufficient time t, the policy converges to π, and we denote

$$\begin{aligned}
\bar{\boldsymbol{\theta}}_a &:= \mathbb{E}[\boldsymbol{\theta}_a \mid \pi],\\
X_t &:= \begin{bmatrix}\boldsymbol{S}_t\\ \boldsymbol{M}_t\end{bmatrix} \in \mathbb{R}^d.
\end{aligned}$$

Then for $R_t = \left[(\boldsymbol{\theta}_S + \boldsymbol{\theta}_{a_t})^\top \boldsymbol{S}_t + \boldsymbol{\theta}_M^\top \boldsymbol{M}_t\right] + \epsilon_t$, there exists $\boldsymbol{\theta}^* := (\boldsymbol{\theta}_S + \bar{\boldsymbol{\theta}}_a, \boldsymbol{\theta}_M)$ such that

$$\xi_t := R_t - \boldsymbol{\theta}^{*\top} X_t = \epsilon_t + (\boldsymbol{\theta}_a - \bar{\boldsymbol{\theta}}_a)^\top S_t.$$

The sequence $\{\xi_t\}$ can be verified to be a martingale difference process and is sub-Gaussian (Appendix G). And $\|X_t\|$ is bounded because $\|\boldsymbol{S}_t\|$ is bounded. Let $\|X_t\| \le L$.

Abbasi-Yadkori et al. (2011, Corollary 10) proved that, when using the l^2-regularized least-squares parameter estimate $\hat{\boldsymbol{\theta}}_T$ with regularization coefficient $\lambda = 1/T$, with probability $1-\delta$,

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_{V_T} \le \sigma_\epsilon \sqrt{d \log(1 + T^2 L) - d \log \delta} + \|\boldsymbol{\theta}^*\|/\sqrt{T},$$

Using Proposition 2, $\lambda_{\min}(V_T) \ge \tilde{\Theta}(T)$. We know that

$$\begin{aligned}
\|\boldsymbol{\theta}_M - \hat{\boldsymbol{\theta}}_{M,T}\| \le & \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| \\
\le & \frac{1}{\sqrt{\lambda_{\min}(V_T)}} \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_{V_T} \\
= & \tilde{\mathcal{O}}(\frac{\sigma_\epsilon \sqrt{d}}{\sqrt{T}}).
\end{aligned}$$

Summing over $t = 1, 2, \cdots, T$, we obtain

$$\begin{aligned}
& \mathbb{E}\left[\sum_{t=1}^{T} e_t(\boldsymbol{s}, a_t)\right] \\
\le & C \cdot \mathbb{E}\left[\sum_{t=1}^{T} \left\|\hat{\boldsymbol{\theta}}_{M,t} - \boldsymbol{\theta}_M\right\|\right] + \mathcal{E}CT \\
= & C \cdot (\sigma_\epsilon \sqrt{d}) \tilde{\mathcal{O}}(\sqrt{T}) + \mathcal{E}CT.
\end{aligned}$$

C.3. True Regret

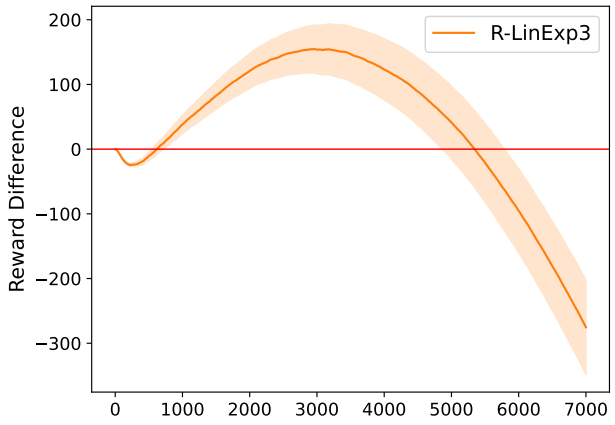
Note that the Assumption 2 implies that $\pi^*(s) = \arg\max_{a \in \mathcal{A}} R(\boldsymbol{s}, a)$ also satisfies that

$$\begin{aligned}
& \mathbb{E}\left[\sum_{t=1}^{T} \left(r_t\left(\boldsymbol{S}_t, \pi^*\left(\boldsymbol{S}_t\right)\right) - r_t\left(\boldsymbol{S}_t, A_t\right)\right)\right] \\
= & \tilde{\mathcal{O}}\left(\text{poly}(d_S, |\mathcal{A}|, \log(T/\delta)) \cdot T^{1-\alpha}\right).
\end{aligned}$$

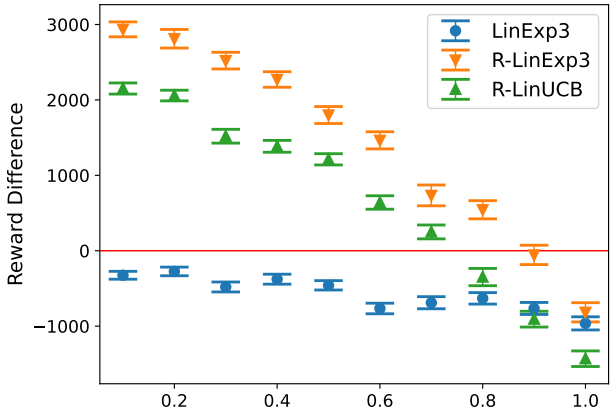
To proceed, we analyze the true regret w.r.t $r^*(a, s)$. We have

$$\begin{aligned}
& \mathbb{E}\left[\sum_{t=1}^{T} \left(R\left(\boldsymbol{S}_t, \pi^*\left(\boldsymbol{S}_t\right)\right) - R\left(\boldsymbol{S}_t, A_t\right)\right)\right] \\
\le & \mathbb{E}\left[\sum_{t=1}^{T} \left(r_t\left(\boldsymbol{S}_t, \pi^*\left(\boldsymbol{S}_t\right)\right) - r_t\left(\boldsymbol{S}_t, A_t\right)\right)\right] \\
& + 2 \sup_{\boldsymbol{s}, a} \mathbb{E}\left[\sum_{t=1}^{T} e_t(\boldsymbol{s}, a)\right] \\
= & (C^2 + C R_\omega) \cdot \tilde{\mathcal{O}}\left(\text{poly}(d_S, |\mathcal{A}|, \log(T/\delta)) \cdot T^{1-\alpha}\right) \\
& + C \cdot (\sigma_\epsilon \sqrt{d}) \tilde{\mathcal{O}}(\sqrt{T}) + \mathcal{E}CT. \qquad \text{(Assumption 2 for the first term)}
\end{aligned}$$

We elaborate the usage of Assumption 2: Assumption 2 is applied to the surrogate r_t. Let π_{adv} be optimal under r_t. The true-reward R optimizer π is no better than π_{adv} on r_t, so the guarantee (against π_{adv}) also holds for π. And as established in Neu & Olkhovskaya (2020, Theorem 2), the polynomial term and exponent α are $\sqrt{d_S |\mathcal{A}| \log T}$ and $\frac{1}{2}$, respectively. Plugging these into the regret bound yields the final result.



(a) Reward difference over time T with $\sigma_\omega = 0.1$ and $\mathcal{E} = 2$.



(b) Reward difference versus surrogate error $\mathcal{E}$ with $\sigma_\omega = 0.01$.

Figure 6. Synthetic-data comparison of R-LinExp3, R-LinUCB, and LinExp3 relative to LinUCB. Error bars represent standard errors over 700 independent runs.

D. Additional Synthetic Experiments

We include synthetic-data validation here to preserve space in the main text. The environment follows Eq. 1; complete generation details are given in Appendix E. We compare the same four algorithms as in the main text: LinExp3 and LinUCB, together with their reward-designed counterparts R-LinExp3 and R-LinUCB.

Figure 6a shows how performance evolves with the horizon T. The reward-designed methods exhibit a short warm-up period because the surrogate reward must first be learned online, so both R-LinExp3 and R-LinUCB may initially lag behind their base counterparts. Once this transient phase passes, R-LinExp3 becomes the strongest method and maintains a positive reward gap relative to LinUCB. This is consistent with Eq. 3: the variance reduction from mediator-based reward design eventually outweighs the extra estimation error introduced by learning the surrogate.

The same panel also highlights why an adversarial oracle is preferable after reward design. Even though the underlying synthetic environment is stochastic, the learned surrogate reward changes over time, which makes the effective learning problem faced by the bandit oracle nonstationary. R-LinExp3 handles this nonstationarity more robustly than R-LinUCB and consistently outperforms it after the warm-up period. By contrast, the blue curve remains mostly below zero, indicating that vanilla LinExp3 underperforms LinUCB when no reward design is applied, as expected in a stochastic environment.

Figure 6b examines the effect of surrogate error $\mathcal{E}$, which measures the action-to-reward effect not captured by the mediator relative to the captured effect. As $\mathcal{E}$ increases, both reward-designed methods deteriorate, matching the dependence on surrogate error predicted by Eq. 3. Still, the degradation is gradual rather than abrupt: R-LinExp3 continues to outperform LinUCB even when $\mathcal{E} = 0.8$, where the mediator captures only about $1/(1 + 0.8) \approx 56\%$ of the causal effect. This robustness suggests that the method remains useful even when the mediator is informative but not perfectly surrogate.

Overall, the synthetic results support three messages from the theory: reward design is most beneficial after an initial learning phase, its gains decrease smoothly as the surrogacy assumption is violated, and pairing online reward design with an adversarial contextual-bandit oracle is important in practice because the surrogate-learning step induces nonstationarity.

E. Detail of Environment Design

E.1. Synthetic Data

In the simulated bandit environment, we set the number of arms to $|\mathcal{A}| = 20$, with feature dimensions $d_S = 10$ and $d_M = 5$. The state vectors S_t are sampled uniformly from the unit ball. All coefficient vectors are drawn from normal distributions and rescaled to satisfy the norm constraint specified in Assumption 3, with norm bound $C = 2$. The surrogate error $\mathcal{E}$ varies by experimental setting. Noise terms are defined as $\omega_t \sim \mathcal{N}(\mathbf{0}, \sigma_\omega^2 I_{d_M})$ and $\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$, where $\sigma_\epsilon = 10$ and σ_ω is case-specific. Each experimental run consists of 7,000 steps and is repeated 700 times.

E.2. HeartSteps V1

We construct a generative contextual bandit environment calibrated to the HeartSteps V1 mobile health study (Klasnja et al., 2019). At each decision time t, the context S_t is formed by replaying observed covariates from the HeartSteps dataset (e.g., recent activity summaries, weather/temperature, and location indicators), with total dimension $d_S = 7$. The action set is binary, $A_t \in \{0, 1\}$, corresponding to sending vs. not sending an intervention. Conditioned on (S_t, A_t), we generate the mediator M_t and reward R_t from linear-Gaussian models fit offline, as in Eq. (1).

HeartSteps V1 Dataset. The dataset contains 37 users and a total of 6,394 decision points (min 73, max 202, mean 172.8 per user). Each decision point includes covariates such as `sum_step.log`, `jbsteps30pre.log`, `dec.temperature`, `sd_steps60`, `dosage`, and `home.location`. The original action indicator (send vs. not send) defines the treatment and is also used to construct `dosage` during preprocessing, but it is not treated as an additional state coordinate.

Generative model. At time t, we form a 7-dimensional feature vector x_t with an intercept,

$$[1,\ \text{sum\_step.log},\ \text{jbsteps30pre.log},\ \text{dec.temperature},\ \mathbb{I}\{\text{sd\_steps60}\},\ \mathbb{I}\{\text{home.location}\},\ \text{dosage}],$$

where the dosage coordinate is formed as an exponentially-discounted sum of the user's recent treatment history: for decision index t, $\text{dosage}_t := \sum_{k=1}^{K} \gamma^k \mathbb{I}\{A_{t-k} = 1\}$ with discount $\gamma = 0.95$ and a finite window $K = 40$ past decision points (computed from the logged HeartSteps actions during preprocessing), and we optionally perturb it by additive Gaussian noise.

For all users, we fit linear-Gaussian models for the mediator and the reward, yielding action-specific coefficients $\theta_a^{(M)}$ and $\theta_a^{(R)}$ (for $a \in \{0, 1\}$) and a mediator-effect coefficient β_m. We generate

$$m_t = x_t^\top \theta_{A_t}^{(M)} + \xi_t, \qquad \xi_t \sim \mathcal{N}(0, \sigma_m^2),$$

and

$$r_t = x_t^\top \theta_{A_t}^{(R)} + \beta_m m_t + \eta_t, \qquad \eta_t \sim \mathcal{N}(0, \sigma_r^2),$$

where the target (ground-truth) used is the log-transformed short-window post-decision step count as the mediator (10-minute steps, `jbsteps10.log`) and the log-transformed 30-minute post-decision step count as the reward (`jbsteps30.log`), both derived from HeartSteps V1 activity data.

Experiment protocol and evaluation. For each replication, we shuffle users at random and evaluate all algorithms on the same replayed context stream for comparability. User contexts are replayed sequentially, and episodes terminate after a fixed horizon. We report learning curves of cumulative *expected* reward and reward differences relative to the LinUCB baseline; uncertainty bands correspond to standard errors across replications.

F. Practicality of Assumption 2

To demonstrate the practicality of the assumption about the oracle, we note that there exist adversarial bandit algorithms (Rakhlin & Sridharan, 2016; Neu & Olkhovskaya, 2020; He et al., 2022) that already consider the case where the mean reward is arbitrarily chosen by an adversary.

Besides, algorithms that assume a strict linear reward can be a candidate for our oracle (Syrgkanis et al., 2016; Kuroki et al., 2024; Banihashem et al., 2024). These algorithms are developed for the case of a deterministic reward without variance adaptivity. Here we introduce a reduction method proposed by Lattimore & Szepesvári, which absorbs the noise term as an additional dimension of the adversarial reward vector, to fit these algorithms into our framework.

Assume that in a stochastic linear environment, at each decision time t,

1. The agent observes the state S_t and chooses an action $a_t \in \mathcal{A}$ accordingly.
2. The agent receives reward $r_t := \langle S_t, \theta_{a_t} \rangle + \epsilon_t$, where ϵ_t is a centered, bounded noise.

One can frame a stochastic linear environment as an adversarial linear one, by the following steps.

We define

$$\psi : \mathbb{R}^{d_S} \longrightarrow \mathbb{R}^{d_S+1},$$
$$s \mapsto (s, 1),$$

where the additional dimension allows us to incorporate the noise ϵ_t as a part of the parameter vector θ'_a of action a provided by the adversary, by letting $\theta'_{a_t} := (\theta_{a_t}, \epsilon_t)$. Thus $r_t := \langle S_t, \theta_{a_t} \rangle + \epsilon_t = \langle \psi(S_t), \theta'_{a_t} \rangle$. Note that action set $\mathcal{A}$ is not changed, while the adversarial environment chooses action vector θ_a in a higher-dimensional set.

G. property of $\{\xi_t\}$

In this section, we verify that the noise sequence $\{\xi_t\}$ used in the self-normalized concentration result of Abbasi-Yadkori et al. (2011) is (i) a martingale difference sequence and (ii) conditionally sub-Gaussian.

Recall from Section C that

$$\xi_t := R_t - \boldsymbol{\theta}^{*\top} X_t = \epsilon_t + (\boldsymbol{\theta}_{a_t} - \bar{\boldsymbol{\theta}}_{a_t})^\top \boldsymbol{S}_t, \tag{5}$$

where $\bar{\boldsymbol{\theta}}_a := \mathbb{E}[\boldsymbol{\theta}_a \mid \pi]$ under the stationary policy π. Let $\mathcal{F}_t$ be the natural filtration generated by the history up to time t,

$$\mathcal{F}_t := \sigma(\boldsymbol{S}_1, A_1, R_1, \ldots, \boldsymbol{S}_t, A_t, R_t),$$

and note that $\boldsymbol{S}_t, A_t$ are $\mathcal{F}_{t-1}$-measurable.

Martingale Difference Process. We show that $\{\xi_t\}$ is a martingale difference sequence with respect to $\{\mathcal{F}_t\}$, i.e., $\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}] = 0$ for all t. By Eq. (1), ϵ_t is mean-zero and independent of $\mathcal{F}_{t-1}$, hence $\mathbb{E}[\epsilon_t \mid \mathcal{F}_{t-1}] = 0$. For the second term in Eq. (5), conditioning on $\mathcal{F}_{t-1}$ fixes $(\boldsymbol{S}_t, A_t) = (\boldsymbol{s}, a)$, and under the stationarity assumption (policy has converged to π) the action-specific random effect satisfies $\mathbb{E}[\boldsymbol{\theta}_a - \bar{\boldsymbol{\theta}}_a \mid \pi] = 0$. Therefore,

$$\begin{aligned}
\mathbb{E}\big[(\boldsymbol{\theta}_{a_t} - \bar{\boldsymbol{\theta}}_{a_t})^\top \boldsymbol{S}_t \mid \mathcal{F}_{t-1}\big] &= \mathbb{E}\big[(\boldsymbol{\theta}_a - \bar{\boldsymbol{\theta}}_a)^\top \boldsymbol{s} \mid \mathcal{F}_{t-1}\big] \\
&= \boldsymbol{s}^\top \mathbb{E}\big[\boldsymbol{\theta}_a - \bar{\boldsymbol{\theta}}_a \mid \pi\big] \\
&= 0,
\end{aligned}$$

which implies $\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}] = 0$.

Sub-Gaussianity. We next verify conditional sub-Gaussianity. By Assumption 3, we have $\|\boldsymbol{\theta}_a\| \le \mathcal{E}C$ and $\|\boldsymbol{S}_t\| \le 1$; hence

$$\begin{aligned}
\big|(\boldsymbol{\theta}_{a_t} - \bar{\boldsymbol{\theta}}_{a_t})^\top \boldsymbol{S}_t\big| &\le \|\boldsymbol{\theta}_{a_t} - \bar{\boldsymbol{\theta}}_{a_t}\| \, \|\boldsymbol{S}_t\| \\
&\le (\|\boldsymbol{\theta}_{a_t}\| + \|\bar{\boldsymbol{\theta}}_{a_t}\|) \cdot 1 \le 2\mathcal{E}C.
\end{aligned}$$

Thus $Z_t := (\boldsymbol{\theta}_{a_t} - \bar{\boldsymbol{\theta}}_{a_t})^\top \boldsymbol{S}_t$ is conditionally sub-Gaussian with proxy at most $2\mathcal{E}C$ (a bounded, mean-zero random variable is sub-Gaussian). Moreover, ϵ_t is conditionally σ_ϵ-sub-Gaussian (Eq. 1). Since the sum of conditionally sub-Gaussian random variables is conditionally sub-Gaussian, it follows that, for all t,

$$\xi_t = \epsilon_t + Z_t \quad \text{is conditionally sub-Gaussian with parameter} \quad \sigma_\xi := \sqrt{\sigma_\epsilon^2 + (2\mathcal{E}C)^2}.$$

In particular, $\{\xi_t\}$ satisfies the noise condition required by Abbasi-Yadkori et al. (2011).