Introduction to Adaptive Causal Bandits Algorithm

Yan-Yu Chen
Department of Statistics
University of California, Davis
Davis, CA 95616
ynychen@ucdavis.edu

Abstract

Multi-armed bandit problems have potentials to identify the best intervention in a sequence of repeated experiments. Initially, the minimax optimal performance is well-understood without the framework of counterfacturals. However, recent research demonstrated that causal bandit algorithms incorporating observed variables that d-separate the intervention from the outcome of interests can result in lower regret. In practice, it is desirable to have an algorithm that is agnostic to whether observed variables act as d-separators. This means that the algorithm should adapt and perform nearly as well as an algorithm that has oracle knowledge of the presence or absence of a d-separator. Importantly, the algorithm does not require any oracle knowledge of causal mechanism. To formalize and explore the notion of adaptivity, we introduce a recent algorithm (HAC-UCB) from Bilodeau et al. (2022) that (a) achieves optimal regret when a d-separator is observed; (b) incurs significantly smaller regret compared to recent causal bandit algorithms when the observed variables are not d-separators. The concept of adaptivity is further extended to other conditions, but unfortunately it fails to completely exploit the tools developed by recent research in causal inference. Therefore, we will discuss some improvement of adaptability for bandit problems.

1 Introduction

The task of learning the best intervention from observational data, without knowledge of the specific causal structure, is impossible according to Theorem 4.3.2 of Pearl (2009). Instead, the focus shifts to finding an efficient method for sequentially selecting interventions in i.i.d. repetitions of the environment. The main challenge lies in the inability to observe the counterfactual effects of interventions not chosen. Without any structural assumptions beyond i.i.d., one can learn the best intervention with high confidence by performing each intervention a sufficient number of times. However, performing each intervention a sufficient number of times is not an efficient way and most of time infeasible in practice.

The (multi-armed) bandit framework provides a natural context to address this problem. In this setting, the experimenter chooses actions over a sequence of interactions, based on past experiences, and observes the corresponding rewards. The goal is to achieve performance comparable to what would have been attained if the experimenter had always chosen the optimal action. Performance is measured by regret, the difference between the cumulative reward obtained by the experimenter and that of the optimal action. Regret captures the exploration-exploitation trade-off, where exploration involves choosing potentially suboptimal actions to learn their optimality, while exploitation entails selecting the action that appears to be the best based on empirical evidence. Other measures of performance, such as identifying the average treatment effect or the best action at the end of all interactions, fail to penalize suboptimal actions during exploration and are insufficient for studying the desired question.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

For an action set $\mathcal{A}$ and a time horizon T, the minimax optimal regret for bandit problems without any data assumptions (worst-case scenario) is $\mathcal{O}(\sqrt{|\mathcal{A}|T})$, and several algorithms achieve this bound. Recently, demonstrated that under additional causal structure, a new algorithm called C-UCB from Lu et al. (2020) can attain improved regret. Specifically, if the experimenter has access to a variable $Z \in \mathcal{Z}$ that d-separates (Pearl (2009)) the intervention and the reward, as well as the interventional distribution of Z for each action in $\mathcal{A}$, C-UCB achieves regret $\mathcal{O}(\sqrt{|\mathcal{Z}|T})$. We will introduce d-separation formally later. For now, we can understand it as a desirable causal mechanism in the environment. However, as shown, when the d-separation assumption fails, the performance of C-UCB is significantly worse than that of UCB. This raises the question of strict adaptation. We seek that there exists an algorithm that combines the guarantee of C-UCB when Z is a d-separator and the guarantee of UCB in all other scenarios, without prior knowledge of whether Z is a d-separator.

Currently, Bilodeau et al. (2022) proposed hypothesis-tested adaptive causal UCB (HAC-UCB) that partially achieves this goal. HAC-UCB perforems hypothesis testing on each trial to decide whether the experimenter should swich form C-UCB to UCB. Another closest approach is the Corral algorithms, which employ online mirror descent to merge "base" bandit algorithms. However, Corral requires the stability of each base algorithm when operating on importance-weighted observations, which is not the case for C-UCB, as evidenced by simulations. This poses a challenge for adapting to causal structure using Corral-like techniques and prompts the exploration of novel methods to achieve adaptivity. Still, both of them do not completely exploit the tools developed by recent research in causal inference. The most critical reason that causal assumption-based approaches fail is the presence of unmeasured confounding. If there presents the unmeasured confounding in the environment, $\mathcal{Z}$ cannot be a d-separator. However, not all unmeasured confounding cannot be remedied. Recent research in causal inference has developed several tools to address the issue of presence of unmeasured confounding. Therefore, there is a space to improve the adaptivity for bandit problem.

Contributions. We review the essential concept of adaptive causal bandit algorithm: the conditionally benign property for environments. We introduce several causal/non-causal algorithms attempting to adapt the optimal minmax regret rate without the knowledge of the presence or absence of d-separator. We reproduce two regrets figures from Bilodeau et al. (2022), the conditionally benign environments and worst-case environment. Our numerical experiments confirm that HAC-UCB is surely state-of-art of the adaptive causal bandit algorithm nowadays. However, we provide some possible improvement of adaptability for HAC-UCB in the discussion. Such improvement makes algorithms adapt to the environment in certain case even when the conditionally benign property fails.

2 Preliminaries

2.1 d-separation and Conditionally Benign Property

For clarity, we introduce the concept of d-separation from Pearl (2009). We can express the causal mechanism among varialbes of interests in terms of a directed acyclic graph (DAG) by d-separation.

Definition 1. *Let A, Z, and Y be variables of interests, where $Z \in \mathcal{Z}$.*

(1) A and Y are d-connected if there is an unblocked path between theme.

(2) A and Y are d-connected, conditioned on a set $\mathcal{Z}$ of nodes, if there is a collider-free path between A and Y that traverses no member of $\mathcal{Z}$. If no such path exists, we say that A and Y are d-separated by $\mathcal{Z}$.

(3) If a collider is a member of the conditioning set $\mathcal{Z}$, or has a descendant in $\mathcal{Z}$, then it no longer blocks any path that traces this collider.

For example, in the case of $A \to Y$, A and Y are d-connected according to case (1). For the case (2), consider $A \to Z \to Y$. A and Y are d-separated by Z since when conditioning on Z, there is no path between A and Y. Finally, consider a similar DAG $A \to Z \leftarrow Y$. In this case A and Y are now d-connected by Z since Z is a collider. When conditioning on Z, there is new path open from A to Y through Z.

With the concept of d-separation, we are ready to state the conditional benign property.

Definition 2. *An environment is conditionally benign if there exists a random variable Z such that the conditional distribution of the reward given Z is unchanged for each action $a \in \mathcal{A}$.*

This definition does not require any causal terminology to define or use for regret bounds, but it is actually closely related to the causal setting. Shown by Bilodeau et al. (2022), for an environment equipped with action, post-action variables and outcome of interest ($\mathcal{A}$, $\mathcal{Z}$, Y), the conditional benign property is equivalently to say that there exists a $Z \in \mathcal{Z}$ being a d-separator when $\mathcal{A}$ is all interventions.

Conditionally benign property plays an essential role for causal-type algorithms to work. The most important example and counterexample of conditionally benign environment are $A \to Z \to Y$ and $A \to Z \to Y \leftarrow A$, respectively. It is clear that the conditional benign property holds in the first case since Z is a d-separator. For the second case, Z fails to d-separate A and Y since conditioning on Z, there is still a path between A and Y. Therefore, the conditional benign property fails. Furthermore, conditionally benign property is a slightly weaker assumption than d-separator environments as it allows the presence of null intervention.

2.2 Causal Bandit Problem

We explore a generalized extension of the traditional bandit setting from Lattimore et al. (2016), called causal bandit problems. In addition to observing the reward associated with the chosen action, the experimenter also has access to additional variables after making their action selection. We refer to this as the post-action context. It is important to note that this differs from the contextual bandit problem, where the experimenter has access to side-information or contextual variables before making their action choice.

Let Y_t be the outcome of interest, A_t be the action or intervention, and Z_t be the post-action context variable. For each round $t = 1, \ldots, T$, the experimenter selects $A_t \in \mathcal{A}$ from a policy π and simultaneously the environment samples $\{(Z_t(a), Y_t(a)) : a \in \mathcal{A}\}$, where $Z_t(a)$ and $Y_t(a)$ are counterfactuals of Z_t and Y_t when action $A_t = a$. The experimenter only observes $(Z_t(A_t), Y_t(A_t))$ and receives reward $Y_t(A_t)$. The performance of a policy is quantified by the regret $R(T) = T \cdot \max_{a\in\mathcal{A}} \mathbb{E}[Y(a)] - \mathbb{E}_\pi \sum_{t=1}^{T} \mathbb{E}[Y(A_t)]$. Conceptually, The experimenter seeks the most effective intervention $A \in \mathcal{A}$ for $A \to Y$ to minimize the regret $R(T)$ with the presence of post-action variables Z

It is worth to note that, in general, $\mathbb{E}[Y(a)] \neq \mathbb{E}[Y|A = a]$ but $\mathbb{E}[Y(a)] = \mathbb{E}[\mathbb{E}[Y|pa(Y), A = a]]$, where $pa(Y)$ is the set of parents of Y in DAG. Therefore, when there exists unmeasured confounding in the environment i.e., $\mathcal{Z} \neq pa(Y)$, statistics related to causal quantities based on observed trials are no longer consistent, and thus leads to a significant bias.

2.3 Algorithms

We now introduce additional notation to define the three main algorithms of interest in this work: UCB, C-UCB, and HAC-UCB. Let $\delta \in (0, 1)$ be the confidence parameter. First, let $\mathbb{T}_t^{\mathcal{A}}(a) = 1 \vee \sum_{s=1}^{t} I(A_s = a)$, $\hat{\mu}_t^{\mathcal{A}}(a) = \mathbb{T}_t^{\mathcal{A}}(a)^{-1} \sum_{s=1}^{t} Y_s(A_s) I(A_s = a)$, $\mathrm{UCB}_t^{\mathcal{A}}(a) = \hat{\mu}_t^{\mathcal{A}}(a) + \sqrt{\log(2/\delta)/(2\mathbb{T}_t^{\mathcal{A}}(a))}$. We define the policy of UCB, by $A_{t+1} = \arg\max_{a\in\mathcal{A}} \mathrm{UCB}_t^{\mathcal{A}}(a)$. It is well-known that without any assumption of environment, the regret of UCB is $R(T) \leq 2|\mathcal{A}| + 4\sqrt{2|\mathcal{A}|T \log T}$.

Next, let $\mathbb{T}_t^{\mathcal{Z}}(z) = 1 \vee \sum_{s=1}^{t} I(Z_s(A_s) = z)$, $\hat{\mu}_t^{\mathcal{Z}}(z) = \mathbb{T}_t^{\mathcal{Z}}(z)^{-1} \sum_{s=1}^{t} Y_s(A_s) I(Z_s(A_s) = z)$, $\mathrm{UCB}_t^{\mathcal{Z}}(z) = \hat{\mu}_t^{\mathcal{Z}}(z) + \sqrt{\log(2/\delta)/(2\mathbb{T}_t^{\mathcal{Z}}(z))}$, $\widetilde{\mathrm{UCB}}_t(a) = \sum_{z\in\mathcal{Z}} \mathrm{UCB}_t^{\mathcal{Z}}(z) \mathbb{P}_{\tilde{\nu}_a}(Z = z)$, and $\tilde{\nu}_a$ is an estimate of the interventional marginal distribution ν_a of Z when $A = a$. We define the policy of C-UCB by $A_{t+1} = \arg\max_{a\in\mathcal{A}} \widetilde{\mathrm{UCB}}_t(a)$. Bilodeau et al. (2022) showed that the regret of C-UCB $R(T)$ is lower when $\tilde{\nu}_a$ is close to ν_a and conditionally benign property holds; however, $R(T)$ is linear in T, when conditionally benign property fails even if $\tilde{\nu}_a = \nu_a$.

Finally, we state the idea of HAC-UCB and describe it precisely in the pseudocode from Bilodeau et al. (2022) in Figure 1. Heuristically, the HAC-UCB algorithm initially goes through an exploration phase to ensure that the estimate $\tilde{\nu}_a$ is sufficiently accurate. If it is not, the maximum likelihood estimate (MLE) of the marginals is used as a replacement. After this exploration phase, HAC-

Algorithm 1: HAC-UCB($\mathcal{A}$, $\mathcal{Z}$, T, $\tilde{\nu}(Z)$)

do Play each $a \in \mathcal{A}$ for $\lceil 4\sqrt{T} \,/\, |\mathcal{A}| \,\rceil$ rounds, and let $\hat{\nu}_a(Z)$ be the MLE of $\nu_a(Z)$
if $\sup_{a\in\mathcal{A}} \sum_{z\in\mathcal{Z}} \left| \mathbb{P}_{\tilde{\nu}_a}[Z=z] - \mathbb{P}_{\hat{\nu}_a}[Z=z] \right| > 2T^{-1/4}\sqrt{|\mathcal{A}|\,|\mathcal{Z}|\log T}$
  replace $\tilde{\nu}(Z) \longleftarrow \hat{\nu}(Z)$
do Play each $a \in \mathcal{A}$ for $\lceil \sqrt{T} \,/\, |\mathcal{A}| \,\rceil$ rounds
set flag $=$ True
while $t \le T$
  if flag
    `/* Check if either of the two conditions fail */`
    set $\mathrm{D}^{\mathcal{A}}_{t-1}(a) = \mathrm{UCB}^{\mathcal{A}}_{t-1}(a) - \widetilde{\mathrm{UCB}}_{t-1}(a) + \frac{\sqrt{|\mathcal{A}||\mathcal{Z}|\log T}}{T^{1/4}}$
    for $a \in \mathcal{A}$ **do**
      if not $-2\sum_{z\in\mathcal{Z}} \sqrt{\frac{\log T}{\mathbb{T}^{\mathcal{Z}}_{t-1}(z)}}\,\mathbb{P}_{\tilde{\nu}_a}[Z=z] \le \mathrm{D}^{\mathcal{A}}_{t-1}(a) \le 2\sqrt{\frac{\log T}{\mathbb{T}^{\mathcal{A}}_{t-1}(a)}} + 2\frac{\sqrt{|\mathcal{A}||\mathcal{Z}|\log T}}{T^{1/4}}$
        set flag $=$ False; **break**
    `/* If conditions pass, play C-UCB */`
    if flag
      set $A^{\mathrm{HAC}}_t = A^{\mathrm{C}}_t$;
    else
      set $A^{\mathrm{HAC}}_t = A^{\mathrm{UCB}}_t$;
  else
    `/* If conditions ever fail, play UCB forever */`
    set $A^{\mathrm{HAC}}_t = A^{\mathrm{UCB}}_t$;

Figure 1: Pseudo code of HAC-UCB

UCB optimistically plays the C-UCB algorithm until there is enough evidence indicating that the environment is not conditionally benign.

The switch from C-UCB to UCB is determined by a hypothesis test conducted at each round. This test utilizes confidence intervals that would hold if the environment were conditionally benign. When there is a disagreement between the arm mean estimates of UCB and C-UCB, it provides evidence that the environment is not conditionally benign. The decision to switch is made when the size of the disagreement is sufficiently large compared to the size of the confidence intervals themselves.

It is shown that with high probability, this hypothesis test will not trigger a switch in a conditionally benign environment. Furthermore, if the environment is not conditionally benign, the test sufficiently limits the regret incurred by C-UCB. Overall, this heuristic approach aims to ensure accurate estimation during the exploration phase, followed by an adaptive switch between C-UCB and UCB based on evidence of the environment's conditionally benign nature or lack thereof. By doing so, the regrest of HAC-UCB $R(T)$ always achieves sublinear; as good as C-UCB when $\tilde{\nu}_a$ is close to ν_a and conditionally benign property holds.

3 Implementation

In this section, we reproduce the two regret plots for comparison in Bilodeau et al. (2022). Each plots take about 3 minutes to generate. We compare the empirical performance of several algorithms in two key settings: conditionally benign environment $A \to Z \to Y$ and worst-case environment $A \to Z \to Y \leftarrow A$. We evaluate HAC-UCB, against UCB, C-UCB, C-UCB-2, and Corral. For all algorithms, we use the optimal parameters as prescribed by existing theory. To isolate the impact of the conditionally benign property, we set the $\tilde{\nu}_a(Z) = \nu_a(Z)$. The results of this section present representative simulations, demonstrating the following empirical findings: (a) In conditionally benign environments, both HAC-UCB and C-UCB exhibit improved performance compared to UCB, Corral, and C-UCB-2. The latter three algorithms all experience regret growth proportional to $|\mathcal{A}|T$, whereas HAC-UCB and C-UCB achieve better performance.(b) In worst-case environments, both C-UCB and C-UCB-2 experience linear regret, while HAC-UCB successfully transitions to incur sublinear regret, enabling it to compete with Corral and UCB.

The Python codes for the algorithms and the experiment are submitted in Canvas.

3.1 Conditionally Benign Environment

We consider the case, $A \to Z \to Y$. Let $\Delta = \sqrt{|\mathcal{A}|(\log T)/T}$ be the fixed effect. Let $\mathcal{Z} = \{0, 1\}$ and $Y \mid Z \sim \mathrm{Ber}(1/2 + (1 - Z)\Delta)$. Let $\mathbb{P}_{\nu_1}[Z = 0] = 1 - \varepsilon$; $\mathbb{P}_{\nu_a}[Z = 0] = \varepsilon$ for all other $a \in \mathcal{A}\backslash\{1\}$, where ε is selected to a extremely small number. In this case, we can calculate the outcome under the optimal action. That is, $\max_{a \in \mathcal{A}} \mathbb{E}[Y(a)] = \mathbb{E}[\mathbb{E}[Y|Z, A = 1]] = 1/2+(1-\varepsilon)\Delta$. We repeat the entire experiments over 300 epochs and visualize it in Figure 2. We find that C-UCB and HAC-UCB perform similarly, both achieving much smaller regret. In the meanwhile, UCB grows at the worst-case rate.

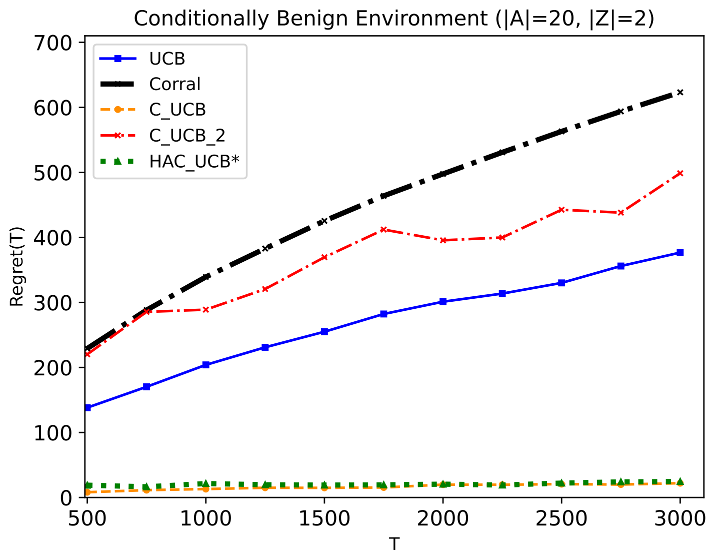

Figure 2: Regret when the conditionally benign property holds

3.2 Worst Case Environment

We consider the case, $A \to Z \to Y \leftarrow X \leftarrow A$. Let $X = I(a > 10)$ be the unmeasured confounder. Let $\mathcal{Z} = \{0, 1\}$ and $Y|X, Z \sim Ber(p)$, where $p = 0.5/6$ if $(X, Z) = (0, 0)$; $p = 5/6$ if $(X, Z) = (1, 0)$; $p = 5.5/6$ if $(X, Z) = (0, 1)$; $p = 1/6$ if $(X, Z) = (1, 1)$. Let $\mathbb{P}_{\nu_a}[Z = 0] = 6/8$ for $a = 1, \ldots, 10$; $\mathbb{P}_{\nu_a}[Z = 0] = 7/8$ for the other $a = 11, \ldots, 20$.

In this case, we can also calculate the outcome under the optimal action. That is, $\max_{a \in \mathcal{A}} \mathbb{E}[Y(a)] = \mathbb{E}[\mathbb{E}[Y|Z, X, A = a]] = 34/48$ for $a = 1, \ldots, 10$. We repeat the entire experiments over 300 epochs and visualize it in Figure 3. We find C-UCB incurs linear regret (worst-case rate), and HAC-UCB achieves sublinear regret, although worse than UCB as expected.

4 Discussion

4.1 Impossibility of Strict Adaptivity

We now answer the question of strict adaptivity (adaptive minimax optimality) in the beginning. It is impossible for any algorithm to always realize the benefits of the conditionally benign property while also recovering the worst-case rate $\mathcal{O}(\sqrt{|\mathcal{A}|T})$, even when the algorithm has access to the true marginals. Lu et al. (2020) shown that any algorithm that does not take advantage of causal structure cannot be adaptively minimax optimal; Bilodeau et al. (2022) further extend it to the case: even algorithms that use $\mathcal{Z}$ and $\tilde{\nu}_a = \nu_a$ cannot be adaptively minimax optimal. Therefore, it is still remains an open problem to find the maximal adaptability of bandit algorithms and develop the optimal algorithm in the sense of adaptivity.

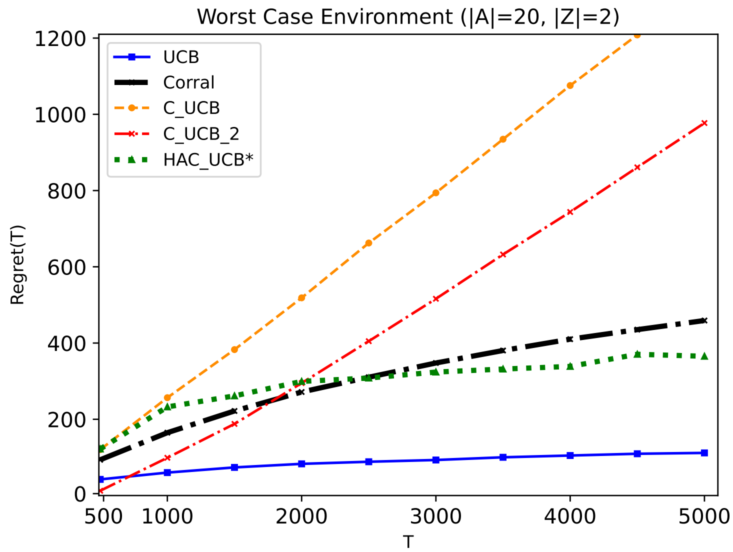

Figure 3: Regret when the conditionally benign property fails

4.2 Improvement of Adaptability

Here we propose a possible improvement of HAC-UCB to extend the adaptability beyond the conditionally benign environments. Consider the case $A \to Z \to Y$ with the presence of unmeasured confounding $Z \leftarrow U \to Y$. The conditional begin property fails in this environment. However, it is a typical confounding case in causal inference. In this case, we can apply the instrumental variable (IV) estimation to correct the misleading effect due to the confounding as A now plays a role of an IV. Several researches has explore IV with contextual bandit problem. For instance, see Kallus (2017) and Zhang et al. (2022) for IV bandit algorithms. With the modification from IV, we can extend the ideas of HAC-UCB for improvement. Whenever we have enough evidence of disagreement between C-UCB and UCB, we further check if there is enough evidence that the IV bandit algorithms disagree with UCB neither. If so, we switch to UCB. Otherwise, we switch to the IV bandit algorithms to enjoy the benefit of lower regret rate from the causal-type algorithms since it indicates that the environment is the confounding case we considered. By doing so, we extend the adaptability beyond the conditionally benign environment. Therefore, we improve the adaptability of HAC-UCB.

References

Bilodeau, B., Wang, L., and Roy, D. M. (2022). Adaptively exploiting d-separators with causal bandits. *Advances in Neural Information Processing Systems*.

Kallus, N. (2017). Instrument-armed bandits. *Proceedings of Algorithmic Learning Theory*.

Lattimore, F., Lattimore, T., and Reid, M. D. (2016). Causal bandits: Learning good interventions via causal inference. *Advances in Neural Information Processing Systems*.

Lu, Y., Meisami, A., Tewar, A., and Yan, Z. (2020). Regret analysis of bandit problems with causal background knowledge. *In Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*.

Pearl, J. (2009). Causality. *Cambridge University Press*.

Zhang, J., Chen, Y., and Singh, A. (2022). Causal bandits: Online decision-making in endogenous settings. In *NeurIPS 2022 Workshop on Causality for Real-world Impact*.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No] We do not see any negative societal impacts of this project.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] All computations were done on CPU on a personal laptop computer.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]