We thank the reviewers for their thoughtful feedback. We believe that a few misreadings of our work made some of the evaluations overly harsh and would ask reviewers to reconsider our paper in the light of clarifications provided below.

**1. Why Online learning? (R2)** It is acknowledged in the literature (Chakraborty and Murphy, 2014; Chakraborty and Moodie, 2013) that most of learning methods in DTRs focus on the batch (offline) settings, and "development of statistically sound estimation and inference techniques for" the online reinforcement learning (RL) in DTRs "is an important research direction." We answers this open problem by proposing the first online RL algorithm in DTRs with provably theoretical guarantees. The applications of online RL in health care are motivated by the increasing "use of sophisticated mobile devices" which enables continuous monitoring and intervention on the fly (Chakraborty and Murphy, 2014). For example, a physician could prescribe a set of safe treatments and use online methods to find the optimal combination for a patient with chronic conditions. Broadly speaking, when the combination of the observational data (e.g., $P(\boldsymbol{v})$) and causal knowledge (causal diagram $G$) does not ensure the identifiability of the causal effects ($E_{\boldsymbol{\pi}}[Y]$) in DTRs, the state of art methods would suggest running randomized clinical trials (RCTs) such as SMART (Murphy, 2003, 2005), and solve for the optimal policy using the standard offline methods (e.g., Q-learning). As the reviewer (R2) suggested, running online experiments in the real environment could be dangerous and expensive. Our results are the first adaptive randomization algorithm that could identify the optimal policy in DTRs while achieving near-optimal bounds on the cost of experimentation (regret). For more discussions on adaptive randomization and online RL, see (Gittins, 1979; Rosenberger and Lachin, 2015).

**2. High-dimensional State-Action Space (R1, R2)** For experimental studies (e.g., RCTs) in DTRs, issues of sample complexity could arise when the state-action space $|\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|$ is high-dimensional. It was believed in the DTR literature that adaptive randomization procedures (e.g., online RL) could circumvent this issue (Cheung, Chakraborty, and Davidson, 2015). Our analysis reveals that this is not the case. In particular, we present first results (Thms. 1-3) analyzing the information complexity of experimental studies in DTRs. We show that a regret bound of $\Omega(\sqrt{|\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|T})$ is inevitable for any randomization procedure, regardless of how sophisticated it might be. This suggests that we should explore other methods to improve the learning convergence. For example, one could exploit the parametric assumptions of the underlying functions (e.g., linear). Our approach takes another route and tries to leverage the abundant, observational data that are available prior to the experimental studies. Specifically, we consider the non-identifiable settings where system dynamics (causal effects $P_{\bar{\boldsymbol{x}}_k}(s_{k+1}|\bar{\boldsymbol{s}}_k)$ and $E_{\bar{\boldsymbol{x}}_K}[Y|\bar{\boldsymbol{s}}_K]$) could not be uniquely determined due to unobserved confounding in canonical DTR models defined in Def. 1 (e.g., Fig. 1(a)). We derive informative bounds about the system dynamics in DTRs from the confounded observational data and incorporate the derived bounds into the online algorithm in an elegant way. We show that this novel approach combining both the online RL and offline bounding could significantly improve the learning performance of online learners for a large family of DTR instances.

**3. Identification Conditions (R2)** Our online algorithms (Sec. 2.1) are developed under the conditions of sequential back-door in DTRs (e.g., Fig. 1(b)); while bounding results in Sec. 3.1 are applicable to DTRs with arbitrary unobserved confounding (Fig. 1(a)). Reviewer 2 (R2) seems to be somewhat confused with our identification conditions, and might mistook Fig. 1(a) as the basis of causal assumptions used for online RL methods in Sec. 2.1. As R2 suggested, randomized experiments and online RL are similar in nature (Bareinboim, Forney, and Pearl, 2015). Since each candidate policy $\boldsymbol{\pi}$ does not take $\boldsymbol{U}$ as input, the sequential backdoor holds in the samples collected by the online RL algorithm in Alg. 1. Causal quantities $P_{\bar{\boldsymbol{x}}_k}(s_{k+1}|\bar{\boldsymbol{s}}_k)$ and $E_{\bar{\boldsymbol{x}}_K}[Y|\bar{\boldsymbol{s}}_K]$ are thus immediately identifiable, and estimable from experimental data. For instance, in Fig. 1(b), $S_1, S_2$ block all back-door paths from $X_1, X_2$ to $Y$. The causal effect $E_{x_1,x_2}[Y|s_1,s_2]$ could be estimated as $E[Y|x_1,x_2,s_1,s_2]$. Given these clarifications, we would like to respectfully ask R2 to re-evaluate our online RL algorithm (Alg. 1) and theoretical regret analysis (Thms. 1-3).

**R2:** (1) Regarding the factorization of distribution $P_{\boldsymbol{\pi}}(\bar{\boldsymbol{x}}_K, \bar{\boldsymbol{s}}_K, y)$ (below Line 113, Page 3), the exogenous $\boldsymbol{U}$ are subsumed in the product of causal quantities $P_{\bar{\boldsymbol{x}}_k}(s_{k+1}|\bar{\boldsymbol{s}}_k)$ and $E_{\bar{\boldsymbol{x}}_K}[Y|\bar{\boldsymbol{s}}_K]$. Specifically, we average $\boldsymbol{U}$ over distribution $P(\boldsymbol{u})$, and factorize the resultant causal effect $P_{\bar{\boldsymbol{x}}_K}(y, \bar{\boldsymbol{s}}_K)$ following the basic definition of conditional probabilities and exclusion restrictions. (2) Bounds in Thm. 2 is a decreasing function relative to $\Delta_{\boldsymbol{\pi}}$ and is maximized when $\Delta_{\boldsymbol{\pi}}$ is the smallest, i.e., $\boldsymbol{\pi}$ is the second best policy. (3) The consistency axiom suggests that counterfactual probabilities $P(\bar{\boldsymbol{s}}_{k+1_{\bar{\boldsymbol{x}}_k}}|\bar{\boldsymbol{x}}_k) = P(\bar{\boldsymbol{s}}_{k+1}|\bar{\boldsymbol{x}}_k)$. However, Lem. 1 is concerned with gap between causal quantities $P_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_{k+1}) - P_{\bar{\boldsymbol{x}}_k}(\bar{\boldsymbol{s}}_k)$, which generally do not equate to the bound $P(\bar{\boldsymbol{s}}_{k+1}, \bar{\boldsymbol{x}}_k) - P(\bar{\boldsymbol{s}}_k, \bar{\boldsymbol{x}}_k)$. (4) $\Gamma(s_1)$ is well defined since we define $\overline{X}_k$ as an empty set when $k < 1$. $\Gamma$ is a function over state-action space for all horizon $k = 1, \ldots, K$. (5) In Corol. 1, the upper bound must be no larger than 1 since $E[Y|\bar{\boldsymbol{s}}_K, \bar{\boldsymbol{x}}_K] \leq 1$. (6) Thm. 6 is *stronger* than standard non-identifiability proof since for the constructed DTRs $M_1, M_2$, not only their transitional probabilities $P_{\bar{\boldsymbol{x}}_k}(s_{k+1}|\bar{\boldsymbol{s}}_k)$ have to be different, but they also need to *match exactly* the lower and upper bounds in Thm. 5. The bounds in Thm. 5 are thus tight given the confounded observational data. To the best of our knowledge, we are not aware of any other tightness result regarding DTRs in the literature.

**R1, R3:** We really appreciate the reviewers for the helpful suggestions and references. We will incorporate these changes in the camera-ready version of the paper if accepted.