Optimizing Pessimism in Dynamic Treatment Regimes: A Bayesian Learning Approach

Yunzhe Zhou
UC Berkeley

Zhengling Qi
GWU

Chengchun Shi
LSE

Lexin Li
UC Berkeley

Abstract

In this article, we propose a novel pessimism-based Bayesian learning method for optimal dynamic treatment regimes in the offline setting. When the coverage condition does not hold, which is common for offline data, the existing solutions would produce sub-optimal policies. The pessimism principle addresses this issue by discouraging recommendation of actions that are less explored conditioning on the state. However, nearly all pessimism-based methods rely on a key hyper-parameter that quantifies the degree of pessimism, and the performance of the methods can be highly sensitive to the choice of this parameter. We propose to integrate the pessimism principle with Thompson sampling and Bayesian machine learning for optimizing the degree of pessimism. We derive a credible set whose boundary uniformly lower bounds the optimal Q-function, and thus we do not require additional tuning of the degree of pessimism. We develop a general Bayesian learning method that works with a range of models, from Bayesian linear basis model to Bayesian neural network model. We develop the computational algorithm based on variational inference, which is highly efficient and scalable. We establish the theoretical guarantees of the proposed method, and show empirically that it outperforms the existing state-of-the-art solutions through both simulations and a real data example.

1 INTRODUCTION

Due to heterogeneity in patients' responses to the treatment, one-size-fits-all strategy may no longer be optimal (Jiang et al., 2017). Precision medicine aims to identify the most effective treatment strategy based on individual patient information. For example, for many complex diseases, such as cancer, mental disorders and diabetes, patients are usually treated at multiple stages over time based on their evolving treatment and clinical covariates (Sinyor et al., 2010; Maahs et al., 2012). Dynamic treatment regimes (DTRs) provide a useful framework of leveraging data to learn the optimal treatment strategy by incorporating heterogeneity across patients and time (Murphy, 2003). Formally, a DTR is a sequence of decision rules, where each rule takes the patient's past information as input, and outputs the treatment assignment. An optimal DTR is the one that maximizes patient's expected clinical outcomes. DTRs generally follow an *online* learning paradigm, where the process involves repeatedly collecting patient's response to the assigned treatment. In medical studies, however, it is often impractical to constantly collect such interactive information. This prompts us to study the problem of learning optimal DTRs in an *offline* setting, where the data have already been pre-collected. In this article, we propose a novel Bayesian learning approach using a pessimistic-type Thompson sampling for finding DTRs.

1.1 Related Work

Statistical methods for DTRs. There is a vast literature on statistical methods for finding optimal DTRs, which, broadly speaking, includes Q-learning, A-learning and value search methods. See Tsiatis et al. (2019); Kosorok and Laber (2019) for an overview. See also Robins (2004); Qian and Murphy (2011); Zhang et al. (2013); Chakraborty and Murphy (2014); Zhao et al. (2015); Chen et al. (2016); Shi et al. (2018a,b); Qi et al. (2020); Chen et al. (2020); Zhang (2020); Cai et al. (2021); Qiu et al. (2021); Zhou et al. (2021); Qi et al. (2022); Tan et al. (2022), and the references therein. However, most existing methods rely on a positivity assumption in the offline data, which essentially requires the probability of each treatment assignment at each stage is uniformly bounded away from zero. In the observational data, such an assumption could easily fail, as certain treatments are prohibited in some scenarios. Therefore, applying these methods may produce sub-

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

optimal DTRs.

Offline reinforcement learning (RL). Built on Markov decision process (MDP), Offline RL learns an optimal policy from historical data without any online interaction (Prudencio et al., 2022). It is thus highly relevant for precision medicine type applications. However, many RL algorithms rely on a crucial coverage assumption, which requires the offline data distribution to provide a good coverage over the state-action distribution induced by all candidate policies. This assumption may be too restrictive and may not hold in observational studies. To address this challenge, the pessimism principle has been adopted that discourages recommending actions that are less explored conditioning on the state. The solutions in this family can be roughly classified into two categories, including model-based algorithms (see e.g., Kidambi et al., 2020; Yu et al., 2020; Uehara and Sun, 2021; Yin et al., 2021), and model-free algorithms (see e.g., Fujimoto et al., 2019; Kumar et al., 2019; Wu et al., 2019; Buckman et al., 2020; Kumar et al., 2020; Rezaeifar et al., 2021; Jin et al., 2021; Xie et al., 2021; Zanette et al., 2021; Bai et al., 2022; Fu et al., 2022). The main idea of the model-based solutions is to penalize the reward or transition function whose state-action pair is rarely seen in the offline data, whereas the main idea of the model-free ones is to learn a conservative Q-function that lower bounds the oracle Q-function. Nevertheless, most of these solutions either require a well-specified parametric model, or rely on a key hyperparameter to quantify the degree of pessimism. It is noteworthy that the performance of those solutions can be highly sensitive to the choice of the hyperparameter; see Section 2.2 for more illustration. In addition, many algorithms are developed in the context of long or infinite-horizon Markov decision process. Their generalizations to medical applications with non-Markovian and finite-horizon systems remain unknown. Finally, we note that there is concurrent work by Jeunen and Goethals (2021) that adopts a Bayesian framework for offline contextual bandit. However, their method requires linear function approximations, and cannot handle complex nonlinear systems, nor more general sequential decision making.

Thompson sampling. Thompson sampling (TS) is a popular Bayesian approach proposed by Thompson (1933) that randomly draws each arm according to its probability of being optimal, so to balance the exploration-exploitation trade-off in the online contextual bandit problems. It has demonstrated a competitive performance in empirical applications. For instance, Chapelle and Li (2011) showed that TS outperforms the upper confidence bound (UCB) algorithm in both synthetic and real data applications of advertisement and news article recommendation. The success of TS can be attributed to the Bayesian framework it adopts. In particular, the prior distribution serves as a regularizer to prevent overfitting, which implicitly discourages exploitation. In addition, actions are selected randomly at each time step according to the posterior distribution, which explicitly encourages exploration and is useful in settings with delayed feedback (Chapelle and Li, 2011).

Bayesian machine learning. Bayesian machine learning (BML) is a paradigm for constructing machine learning models based on the Bayes theorem, and has been successfully deployed in a wide range of applications (see, e.g., Seeger, 2006, for a review). Popular BML methods include Bayesian linear basis model (Smith, 1973), variational autoencoder (Kingma and Welling, 2013), Bayesian random forests (Quadrianto and Ghahramani, 2014), Bayesian neural network (Blundell et al., 2015), among many others. An appealing feature of BML is that, through posterior sampling, the uncertainty quantification is straightforward. In contrast, the frequentist methods for uncertainty quantification that are based on asymptotic theories can be highly challenging with complex machine learning models, whereas those based on bootstrap can be computationally intensive with large datasets.

1.2 Our Proposal and Contributions

In this article, we propose a novel pessimism-based Bayesian learning approach for offline optimal dynamic treatment regimes. We integrate the pessimism principle and Thompson sampling with the Bayesian machine learning framework. In particular, we derive an explicit and uniform uncertainty quantification of the Q-function estimator given the data, which in turn offers an alternative way of constructing confidence interval without having to specify a parametric model or tune the degree of pessimism, as required by nearly all existing pessimism-based offline RL and DTR algorithms. Compared to the RL and DTR algorithms without pessimism, our method yields a better decision rule when the coverage condition is seriously violated, and a comparable result when the coverage approximately holds. Compared to the RL and DTR algorithms adopting pessimism, our method achieves a more consistent and competitive performance. Theoretically, we show that the regret of the proposed method depends only on the estimation error of the *optimal* action's Q-estimator, and we provide the explicit form of its upper bound in a special case of parametric model. The resulting bound is much narrower than the regret of the standard Q-learning algorithm that depends on the uniform estimation error of the Q-estimator at *each* action. Methodologically, our approach is fairly general, and works with a range of different BML models, from simple Bayesian linear basis model to more complex Bayesian neural network model. Scientifically, our proposal offers a viable solution to a critical problem in precision medicine that can assist patients to achieve the best individualized treatment strategy. Finally, computationally, our algorithm is efficient and scalable to large datasets, as it adopts a variational inference approach to approximate the posterior distribution, and does not require computationally

intensive posterior sampling method such as Markov chain Monte Carlo (Geman and Geman, 1984).

Our method shares a similar spirit as TS, in that we also adopt a Bayesian framework for uncertainty quantification and exploration-exploitation trade-off. We also remark that, although the concepts of pessimism, TS and BML are not completely new, how to integrate them properly is highly nontrivial, and is the main contribution of this article. First of all, in the online setting, TS randomizes over actions to address the exploration-exploitation dilemma. However, randomization contradicts the pessimistic principle in the offline setting. To address this issue, we borrow the idea from the Bayesian UCB method (Kaufmann et al., 2012) for online bandits and generalize it to offline sequential decision making. Second, although posterior sampling allows one to conveniently quantify the *pointwise* uncertainty of the estimated Q-function at a given individual state-action pair, it remains challenging to lower bound the Q-function *uniformly* for any state-action pair. Developing a uniform credible set is crucial for implementing the pessimism principle. Our proposal provides an effective solution with a uniform uncertainty quantification.

2 PRELIMINARIES

2.1 Bayesian Machine Learning

Let $p(o|w)$ denote a machine learning model indexed by w that parameterizes the probability mass or density function of some random variable O, and let $\mathcal{D}_n = \{o_i\}_{i=1}^n$ denote a set of i.i.d. random samples. BML treats w as a random quantity, and learns the entire posterior distribution $p(w|\mathcal{D}_n)$ of w given the data $\mathcal{D}_n$ based on the Bayes rule, by combining the likelihood function $p(\mathcal{D}_n|w)$ and a prior distribution $p(w)$ that reflects prior knowledge about w. Once the posterior distribution of w is learned, a commonly used point estimator for w is the posterior mean denoted by $\widehat{w} = \mathbb{E}(w \mid \mathcal{D}_n)$. One can then make the prediction by using $\widehat{w}$ and the likelihood function. Alternatively, one can also make the prediction by using the posterior mean of the model output. We next consider two specific examples.

Bayesian Linear Basis Model (BLBM). BLBM is an extension of the classical Bayesian linear model (Lindley and Smith, 1972), and models the distribution of a response Y given $X = x$ as $y_i = w^T\phi(x_i) = \sum_{j=1}^K w_j\phi_j(x_i) + \epsilon_i$, where $\phi(x) = \{\phi_1(x), \cdots, \phi_K(x)\}^\top$ is a set of K basis functions, $w = (w_1, \ldots, w_n)^T$ is the weight vector, and the error ϵ_i follows a Gaussian distribution. Since the posterior distribution can be explicitly derived, BLBM is easy to implement in practice. However, it might suffer from potential model misspecification in high-dimensional complex problems.

Bayesian Neural Network (BNN). BNN learns the posterior distribution of the weight parameter w in a neural network. However, exact Bayesian inference is generally intractable due to the extremely complex model structure. Blundell et al. (2015) proposed to approximate the exact posterior distribution $p(w|\mathcal{D}_n)$ by a variational distribution $q(w|\theta)$ whose functional form is pre-specified, and then estimate θ by minimizing the Kullback-Leibler (KL) divergence, $\mathrm{KL}[q(w|\theta)||p(w|\mathcal{D}_n)]$. In practice, $q(w|\theta)$ can be set to a multivariate Gaussian distribution, and the parameters are updated based on Monte Carlo gradients. Blundell et al. (2015) developed an efficient computational algorithm, and showed BNN achieves a superior performance in numerous tasks.

2.2 The Pessimism Principle

In the offline setting, when the coverage condition is not met, the classical DTR and RL methods may yield sub-optimal policies. This is because some states and actions are less covered in the data, whose corresponding Q-values are difficult to learn, resulting in large variances and ultimately sub-optimal decisions. To address this issue, most existing offline RL methods adopt the pessimistic strategy, and derive the policies to avoid uncertain regions that are less covered in the data. Particularly, model-free offline RL methods learn a conservative Q-estimator that lower bounds the Q-function during the search of the optimal policy. We next briefly review a state-of-the-art solution of this type, the pessimistic value iteration method (PEVI) of Jin et al. (2021) based on linear models.

Consider a contextual bandit setting, where the offline data $\mathcal{D}_n$ consists of n i.i.d. realizations $\{s_i, a_i, r_i\}_{i=1}^n$ of the state, action and reward tuple $\{S, A, R\}$, where s_i collects the baseline covariates of the ith instance, a_i is the action received, and r_i is the corresponding reward. We assume R is uniformly bounded and a larger value of R indicates a better outcome. Denote the space of the covariates and actions by $\mathcal{S}$ and $\mathcal{A}$, respectively. In addition to estimating the conditional mean of the reward given the state-action pair, i.e., $Q(S, A) = \mathbb{E}(R|S, A)$, Jin et al. (2021) proposed to also learn a ξ-uncertainty quantifier Γ, such that the event

$$\Omega = \left\{|\widehat{Q}(s, a) - Q(s, a)| \leq \Gamma(s, a) \text{ for all } (s, a)\right\} \quad (1)$$

holds with probability at least $1 - \xi$ for any $\xi > 0$, where $\widehat{Q}$ is an estimator of Q. Instead of computing the greedy policy with respect to $\widehat{Q}$ as in the standard methods, they proposed to choose the greedy policy that maximizes the lower bound $\widehat{Q} - \Gamma$, and showed that the regret of the resulting policy is upper bounded by $\mathbb{E}[\Gamma(S, \pi^*(S))]$, where π^* is the true optimal policy. Note that this bound is much narrower than $\mathbb{E}[\max_a \Gamma(S, a)]$, i.e., the regret bound without taking pessimism into account. They further showed that the resulting policy is minimax optimal in linear finite horizon MDPs without the coverage assumption.

Despite its nice theoretical properties, it is challenging to

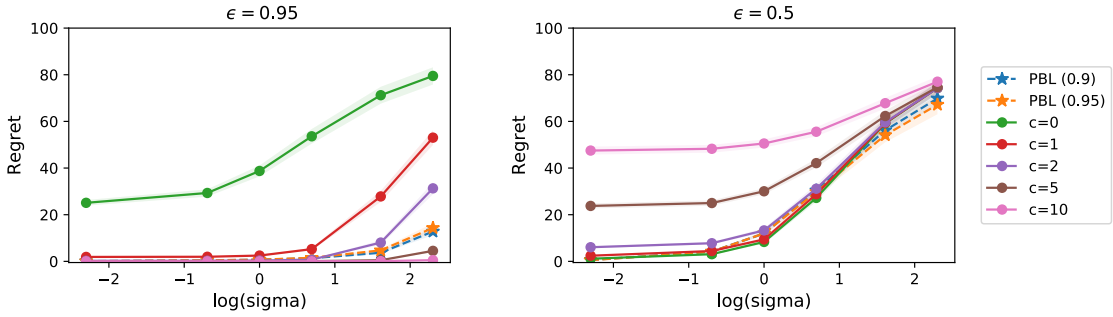


Figure 1: A toy example comparing the PEVI method of Jin et al. (2021) under different values of c and our proposed method PBL.

implement PEVI in practice due to the construction of a proper Γ that meets the requirement in (1). Jin et al. (2021) only developed a construction of Γ under a linear MDP model, and it cannot be easily generalized to more complex machine learning models. Even in the linear model case, their construction relies on a hyperparameter c, and the resulting policy can be highly sensitive to the choice of c. Actually, this is common for many pessimism-based RL methods, which often involve some hyperparameter to quantify the degree of pessimism, and the performances rely heavily on the tightness of this uncertainty quantifier. We consider the following toy example to elaborate.

A toy example. Suppose we model Q via a linear function: $f(s, a, w) = w^\top \phi(s, a)$, where $w \in \mathbb{R}^p$ is the coefficient of the linear basis function ϕ and is estimated by a ridge regression following Jin et al. (2021). They set

$$\Gamma(s, a) = cp[\phi(s, a)^T \Lambda^{-1} \phi(s, a)]^{1/2} \sqrt{\log(2dn/\xi)}, \quad (2)$$

for some constant $c > 0$, where $\Lambda = \sum_{i=1}^n \phi(s_i, a_i)\phi(s_i, a_i)^T + \lambda I$, λ is the ridge parameter, and I is the identity matrix. The choice of c in (2) is crucial for the performance, as a small c would fail to meet the requirement in 1 when the data coverage is inadequate, and a large c would over-penalize the Q-function when the coverage is sufficient. Figure 1 compares the regret of our method and PEVI, where there are two treatments $\{1, 2\}$ and a two-dimensional state $S = (S_1, S_2)$. The reward R is generated from a Gaussian distribution with mean $(0.8 + 0.2A)(S_1 + 2S_2)$ and variance σ^2, and the behavior is generated according to an ϵ-greedy policy that combines a uniformly random policy with a pretrained optimal policy. In this example, ϵ characterizes the level of the coverage, and we consider two levels $\epsilon = 0.95$ where sub-optimal actions are less explored, and $\epsilon = 0.5$ where the coverage holds. We vary the noise level σ, and compare our proposed method and PEVI under varying choices of $c = \{0, 1, 2, 5, 10\}$. It is seen that PEVI is highly sensitive to c under different values of ϵ and σ. By contrast, our proposed method takes a significance level as the input, which is fixed to 0.9 or 0.95 to ensure (1) holds with a large probability, and it achieves a much more stable performance.

3 BAYESIAN LEARNING WITH PESSIMISM

3.1 Basic Idea: Offline Contextual Bandit

As discussed earlier, the success of the pessimism-based methods relies crucially on the uniform uncertainty quantification of the Q-function estimation. Existing solutions require a hyperparameter to properly quantify the degree of pessimism, whereas the choice of such a parameter can be difficult. To address this challenge, and to make the pessimism approach more generally applicable in the offline setting, we propose a data-driven procedure and derive the uniform uncertainty quantification, without requiring specific models or tuning the degree of pessimism when searching for the optimal decision rules. We first illustrate our idea through a single-stage contextual bandit problem in this section, and discuss the dynamic setting of dynamic treatment regimes in the next section.

Suppose we observe the data $\mathcal{D}_n = \{s_i, a_i, r_i\}_{i=1}^n$. Motivated by Thompson sampling, we propose to model the conditional reward distribution given the state-action pair by $p(r|s, a, w)$, and estimate the model parameter $w \in \mathbb{R}^p$ under a Bayesian framework. Specifically, we first apply BML to obtain the posterior distribution $p(w|\mathcal{D}_n)$, and construct a credible set $\mathcal{W}$ given the posterior, such that $P(w \in \mathcal{W}|\mathcal{D}_n) \geq 1 - \alpha$, where $1 - \alpha \in (0, 1)$ is the user-specified coverage rate, which usually takes the fixed value of 0.9 or 0.95. Next, instead of choosing an action

that maximizes the conditional mean function

$$f(s,a,w)=\int_r p(r|s,a,w)dr,$$

with a randomly drawn w as in the online setting, we construct the lower bound of the credible set for $f(s,a,w)$, denoted by $f_L(s,a)$, by solving the following chance constraint optimization problem,

$$\begin{aligned}&\underset{w\in\mathcal{W}}{\text{minimize}}\ f(s,a,w),\\ &\text{subject to}\quad \mathbb{P}(w\in\mathcal{W}|\mathcal{D}_n)\geq 1-\alpha.\end{aligned}\tag{3}$$

Although our credible set is constructed using Bayesian inference, Proposition 1 in Section 4 guarantees that, by the Bernstein-von Mises theorem, the solution $f_L(s,a)$ to (3) provides a valid asymptotic lower bound for $Q(s,a)$ uniformly over $(s,a)\in\mathcal{S}\times\mathcal{A}$ from a frequentist perspective.

Note that the optimization in (3) may not be straightforward. First, it requires to specify the credible set $\mathcal{W}$ that satisfies the coverage constraint. This can be challenging for complex nonlinear models where the exact Bayesian inference is intractable. Second, it can be computationally difficult to optimize the objective function $f(s,a,w)$ with the inequality constraint.

To address the first challenge, we adopt the variational inference approach, and parameterize the posterior function using a Gaussian distribution $\mathcal{N}(\widehat{w},\widehat{\Sigma})$. The Gaussian model is correctly specified for BLBM, and provides a valid approximation for a large number of nonlinear models (Wang and Blei, 2019). Under the Gaussian approximation, the posterior distribution of $(w-\widehat{w})^\top\widehat{\Sigma}^{-1}(w-\widehat{w})$ follows a χ^2 distribution with the degree of freedom p, based on which we can easily construct $\mathcal{W}$.

To address the second challenge, we note that it is relatively straightforward to evaluate the objective function at feasible points. Therefore, we propose a sampling-based algorithm that first randomly collects N samples from the posterior distribution, denoted as $\{w_1,\ldots,w_N\}$. Among these sampling points, we compare the objective values that satisfy the quadratic constraint in (4), and select the smallest one, denoted by w^*. This yields the following optimization problem,

$$\begin{aligned}&\underset{j\in\{1,\cdots,N\}}{\text{minimize}}\ f(s,a,w_j),\\ &\text{subject to}\quad (w_j-\widehat{w})^\top\widehat{\Sigma}^{-1}(w_j-\widehat{w})\leq\chi^2_{1-\alpha}(p),\end{aligned}\tag{4}$$

where $\chi^2_{1-\alpha}(p)$ is the $(1-\alpha)$th quantile of the χ^2 distribution.

We denote the final solution by $\widehat{f}_L(s,a)=f(s,a,w^*)$. Proposition 2 in Section 4 shows that this solution $\widehat{f}_L(s,a)$ based on the Gaussian approximation and Monte Carlo sampling provides a valid uniform lower bound for Q-function estimation.

Algorithm 1 Pessimism-based Bayesian learning for offline contextual bandit.

Input: The observed data $\mathcal{D}_n=\{s_i,a_i,r_i\}_{i=1}^n$, and the significance level α.
Step 1: Fit BLBM or BNN on $\mathcal{D}_n$.
Step 2: Compute the posterior distribution $p(w|\mathcal{D}_n)$, with mean $\widehat{w}$ and covariance $\widehat{\Sigma}$ estimated by BLBM or BNN.
Step 3: Draw N random samples $\{w_i\}_{i=1}^N$ from $p(w|\mathcal{D}_n)$, and obtain the index set $J=\{j\in[N]\mid(w_j-\widehat{w})^\top\widehat{\Sigma}^{-1}(w_j-\widehat{w})\leq\chi^2_{1-\alpha}(p)\}$.
Step 4: Choose $j^*=\underset{j\in J}{\text{argmin}} f(s,a,w_j)$. Set $w^*=w_{j^*}$, and $\widehat{f}_L(s,a)=f(s,a,w^*)$.
Step 5: Compute the estimated optimal policy as $\widehat{\pi}(s)\in\text{argmax}_{a\in\mathcal{A}}\widehat{f}_L(s,a)$, for any $s\in\mathcal{S}$.
Output: A uniform lower bound $\widehat{f}_L$, and the estimated optimal policy $\widehat{\pi}$.

Finally, we output the greedy policy with respect to $\widehat{f}_L(s,a)$ as $\widehat{\pi}(s)\in\text{argmax}_{a\in\mathcal{A}}\widehat{f}_L(s,a)$ for any $s\in\mathcal{S}$. We summarize our procedure in Algorithm 1.

3.2 Dynamic Treatment Regimes

We next extend our method to the DTR problem, where the insufficient data coverage becomes more serious as the number of decision stages increases.

Suppose we observe the data $\mathcal{D}_n=\left\{s_i^{(1)},a_i^{(1)},s_i^{(2)},a_i^{(2)},\ldots,s_i^{(T)},a_i^{(T)},r_i^{(T)}\right\}_{i=1}^n$ consisting of i.i.d. realizations of state, action and reward tuples $\{S^{(1)},A^{(1)},S^{(2)},A^{(2)},\ldots,S^{(T)},A^{(T)},R^{(T)}\}$ at T stages. Denote $H^{(t)}=(S^{(1)},A^{(1)},\ldots,S^{(t)})$ as the history information up to the decision point t, and its realization for each instance as $h_i^{(t)}$ for $i=1,\ldots,n$. Here we only consider the sparse reward setting commonly seen in medical applications (Murphy, 2003), but our method can also be applied when there is an immediate reward at each decision point. We propose to incorporate our pessimism-based BML idea at each stage. Specifically, at the last stage, similar as in single-stage contextual bandit, we first construct the uniform lower confidence bound $\widehat{f}_L^{(T)}(h^{(T)},a^{(T)})$ for $\mathbb{E}(R^{(T)}|H^{(T)}=h^{(T)},A^{(T)}=a^{(T)})$. We then obtain the estimated optimal policy at the last stage as

$$\widehat{\pi}_T(h^{(T)})\in\text{argmax}_{a^{(T)}\in\mathcal{A}}\quad\widehat{f}_L^{(T)}(h^{(T)},a^{(T)}),$$

for every h_T. Next, to estimate the optimal policy for the $(T-1)$-stage, we employ dynamic programming, and construct the pseudo-reward for each instance at the $(T-1)$-stage as

$$r_i^{(T-1)}=\max_{a^{(T)}\in\mathcal{A}}\widehat{f}_L^{(T)}(h_i^{(T)},a^{(T)}),$$

Algorithm 2 Pessimism-based Bayesian learning for multi-stage dynamic treatment regimes.

Input: The observed data $\mathcal{D}_n = \{s_i^{(1)}, a_i^{(1)}, s_i^{(2)}, a_i^{(2)}, \ldots, s_i^{(T)}, a_i^{(T)}, r_i^{(2)}\}_{i=1}^n$, the length of horizon T, and the significance levels $\alpha_1, \ldots, \alpha_T$.
Initialize $\widehat{f}_L^{T+1} = 0$.
For $K = T, T-1, \cdots, 1$
Apply Algorithm 1 using the data rearranged as $\{s_i = h_i^{(K)}, a_i = a_i^{(K)}, r_i = r_i^{(K)}\}_{i=1}^n$, with the confidence level α_K to obtain $\widehat{\pi}_K$ with the estimated lower confidence bound $\widehat{f}_L^{(K)}$.
State: Construct the pseudo-reward, $r_i^{(K-1)} = \max_{a\in\mathcal{A}} \widehat{f}_L^{(K)}(h_i^{(K)}, a)$, for $i = 1, \ldots, n$.
End for.
Output: The estimated optimal policy $\{\widehat{\pi}_t\}_{1\le t\le T}$.

for $i = 1, \ldots, n$. We then apply Algorithm 1 again, using $\{h_i^{(T-1)}, a_i^{(T-1)}, r_i^{(T-1)}\}_{i=1}^n$ to construct the uniform lower confidence bound $\widehat{f}_L^{(T-1)}(h^{(T-1)}, a^{(T-1)})$ for

$$\mathbb{E}(\max_{a\in\mathcal{A}} \widehat{f}_L^{(T)}(H^{(T)}, a^{(T)})|H^{(T-1)} = h^{(T-1)}, A^{(T-1)} = a)$$

for every $h^{(T-1)}$ and a. We obtain the estimated optimal policy at the $(T-1)$-stage as

$$\widehat{\pi}_{T-1}(h^{(T-1)}) \in \text{argmax}_{a\in\mathcal{A}} \quad \widehat{f}_L^{(T-1)}(h^{(T-1)}, a),$$

for every $h^{(T-1)}$. We iterate the above process until the estimated optimal policy of the first stage is obtained. We summarize our proposed procedure in Algorithm 2. We also remark that dynamic treatment regimes differ from MDPs that impose the Markov assumption within each trajectory, in that, in our setting, the Markov assumption can be violated and the optimal treatment regime at each stage depends on the full data history.

4 THEORY

We next establish theoretical guarantees for our proposed method. We focus on the setting of offline contextual bandit of Section 3.1 here, and extend the results to the DRT setting in Appendix A.1.

We first list a set of regularity conditions. Recall that $p(r|s, a, w)$ corresponds to the model we impose for the conditional reward distribution given the state-action pair.

Assumption 1 *(i) The realization condition holds, i.e., there exists some w_0, such that $p(r|s, a, w_0)$ is the oracle conditional reward density function.*

(ii) The parameter space of ω is compact, and $p(r|s, a, w)$ is continuous and identifiable in ω.

(iii) $p(r|s, a, w)$ is differentiable in quadratic mean at the oracle parameter w_0 with a non-singular Fisher information matrix.

(iv) The prior measure of w is absolutely continuous in a neighborhood of w_0 with a continuous positive density at w_0.

Assumption 1 imposes the conditions on the parameter space and smoothness of the conditional density function, so that we can apply the Bernstein-von Mises theorem, and in turn establish the asymptotic equivalence between the derived credible interval and the confidence interval from the frequentist perspective. These conditions are all mild and standard in the literature (Kleijn and van der Vaart, 2012; Bickel and Kleijn, 2012; Kim, 2006).

We next obtain the following proposition.

Proposition 1 *Suppose Assumption 1 holds. Then,*

$$\liminf_{n\to\infty} \mathbb{P}\left(\cap_{(s,a)} \{f_L(s, a) \le Q(s, a)\}\right) \ge 1 - \alpha.$$

Note that $f_L(s, a)$ is the theoretical lower bound obtained by solving the exact optimization (3) and $Q(s, a) = f(s, a, w_0)$ with the oracle parameter w_0 under the realization condition in Assumption 1(i). Proposition 1 ensures that solving (3) is asymptotically equivalent to construct a valid and uniform lower bound, and thus the validity of using the Bayesian learning approach for quantifying the degree of pessimism.

Since the exact optimization (3) is difficult to solve, we next extend the above proposition to the case where Gaussian approximation and Monte Carlo sampling are applied to approximate the lower bound.

Proposition 2 *Suppose Assumption 1 holds. Then,*

$$\liminf_{n\to\infty} \liminf_{N\to\infty} \mathbb{P}\left(\cap_{(s,a)} \left\{\widehat{f}_L(s, a) \le Q(s, a)\right\}\right) \ge 1 - \alpha.$$

Note that $\widehat{f}_L(s, a)$ is the lower bound obtained by solving the surrogate optimization (4). Proposition 2 ensures that the resulting solution based on the Gaussian approximation and Monte Carlo sampling provides a valid uniform lower bound asymptotically.

Finally, we establish the theoretical guarantee that characterizes the average regret of the estimated optimal policy from Algorithm 1, i.e., the difference between the value function under the optimal policy and that under the estimated policy.

Theorem 1 *Suppose Assumption 1 holds. Then, as $n, N \to \infty$, with probability at least $1 - \alpha + o(1)$, the average regret is upper bounded by*

$$\mathbb{E}_{\pi^*}\left[\mathbb{E}(R|S, A) - \widehat{f}_L(S, A)\right].$$

Specifically, if we use BLBM with p basis functions for model fitting, then there exists some constant $\bar{c} > 0$, such that the average regret can be upper bounded by

$$\bar{c}p^{1/2}n^{-1/2}.$$

We note that the expectation $\mathbb{E}_{\pi^*}$ in the regret bound is taken with respect to the optimal policy π^*. In addition, the difference within the square brackets measures the estimation error of the Q-function. As such, the average regret of the proposed policy depends only on the estimation error of the optimal action's Q-estimator, instead of the uniform estimation error of the Q-estimator at each action. The latter can be much larger without the full coverage assumption. In the case of BLBM, N is not included in the upper bound since we can explicitly solve optimization (4) without Monte Carlo sampling.

We also remark that our theory can be extended to obtain finite-sample guarantee as well. As an example, Hipp and Michel (1976) showed that, under some regularity conditions, the Bernstein-von Mises approximation of the posterior distribution was of the order $n^{-1/2}$. Following similar arguments, we can further extend Theorem 1 to obtain a nonasymptotic probability bound. We provide more details in Corollary 1 in Appendix A.2.

5 SYNTHETIC DATA ANALYSIS

Simulation Setup. We conduct extensive numerical experiments to investigate the empirical performance of the proposed pessimism-based Bayesian learning method (PBL). We illustrate our method using both BLBM and BNN. We also compare with the method of Jin et al. (2021, PEVI), and a standard Q-learning method using BLBM or BNN but without pessimism (Non-Pessi).

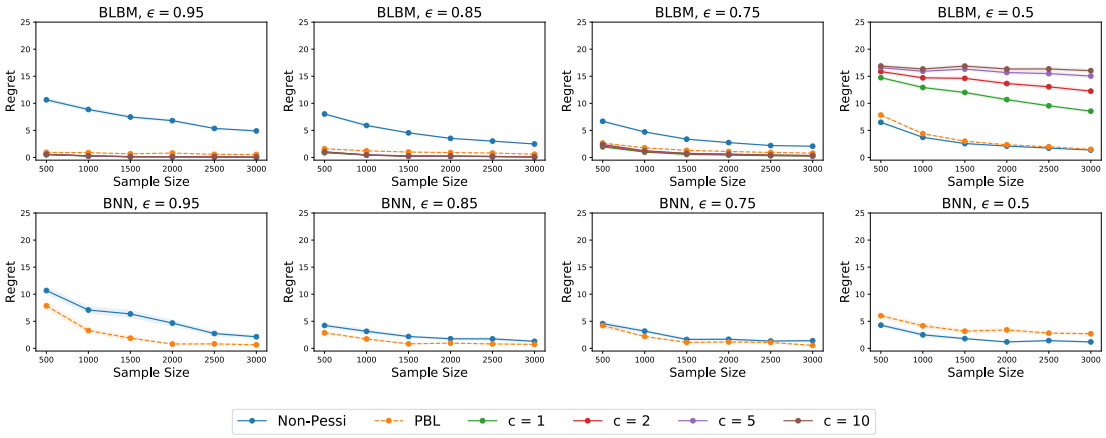
We consider both a single-stage contextual bandit setting and a two-stage DTR setting. In the two-stage setting, since a linear Q-function model is likely to be misspecified in backward induction, we did not implement BLBM under this setting. For both settings, we consider two data generating processes, with linear and nonlinear Friedman signals (see e.g., Zhao et al., 2017), respectively. We generate all actions by the ϵ-greedy policy, with $\epsilon \in \{0.95, 0.85, 0.75, 0.5\}$, and a smaller ϵ indicating a larger coverage over the state-action distribution. We choose the sample size n from $\{500, 1000, 1500, 2000, 2500, 3000\}$, and repeat each experiment 50 times. More details of the data generation and implementations are given in the Appendices C.1, C.2 and C.3. To implement PEVI, we choose the hyperparameter c from $\{1, 2, 5, 10\}$. We also conduct a sensitivity analysis by varying a number of parameters in Appendix C.4, including the number of Monte Carlo samples N, the number of ensembles M for the MC gradient computation in variational inference, and the significance level α. We find that our method is not overly sensitive to those parameters as long as they are in a reasonable range. We make our code publicly available at `https://github.com/yunzhe-zhou/PBL`.

Results. Figure 2 reports the results for the single-stage contextual bandit setting under the linear and nonlinear signals, whereas Figure A2 in Appendix C.5 reports the results for the two-stage DTR setting. It is clearly seen from these plots that both our proposed PBL and the PEVI method outperform the standard Q-learning method when $\epsilon \geq 0.75$ and the coverage assumption is seriously violated, demonstrating the advantage of the pessimism principle. Nevertheless, for the single-stage setting when $\epsilon = 0.5$ and the coverage is of less concern, PEVI over-penalizes the Q-function, leading to a large regret. By contrast, our proposed method performs comparably to the standard Q-learning algorithm in this setting. For the two-stage setting, our proposed method based on BNN outperforms PEVI in all cases. This is because PEVI uses a linear function approximation. The linearity assumption is likely to be violated in backward induction, leading to sub-optimal policies.

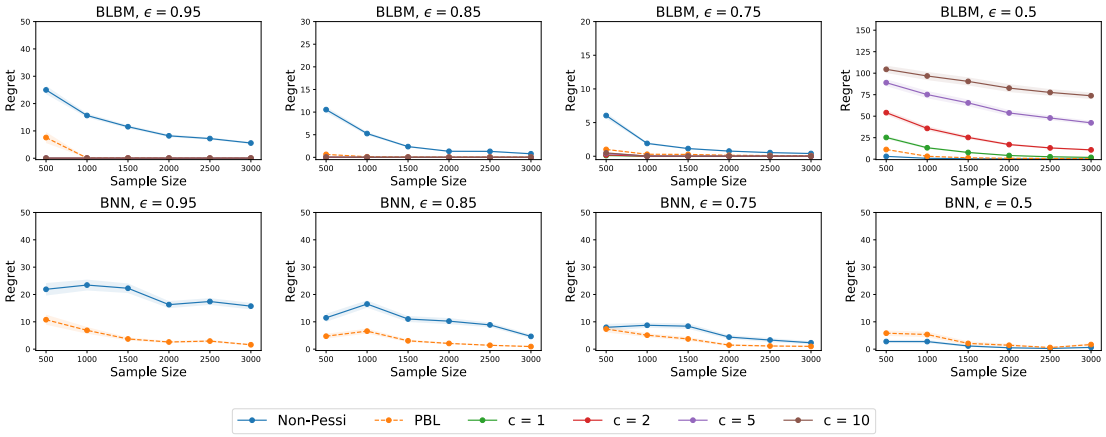
6 REAL DATA APPLICATION

We illustrate our method with the MIMIC-III v1.4 dataset that contains critical care data for over 40,000 patients from the Beth Israel Deaconess Medical Center between 2001 and 2012 (Johnson et al., 2016). Following the analysis of Raghu et al. (2017), we define a 5×5 action space by discretizing both medical interventions intravenous (IV) fluid and maximum vasopressor (VP) dosage into 5 levels. We define the reward as the negative value of the SOFA score that measures the organ failure of the patients (Lambden et al., 2019), so a larger reward is better. We consider the state space with 47 physiological features, including the demographics, lab values, vital signs, and intake and output events. We construct two datasets, one for single–stage contextual bandit, and the other for two-stage DTR. We randomly split each data into a training set and a testing set with equal sample size. We apply our algorithm to the training data and use BNN to fit the Q-function. We did not use BLBM or apply PEVI to this data, since the associations between the features and rewards are expected to be highly complex and nonlinear (Raghu et al., 2017). We compare our proposed PBL method with the standard Q-learning method based on BNN without using pessimism, as well as the conservative Q-learning (CQL) method implemented via the d3rlpy package at its default setting.

Figure 3 reports the frequencies of the assigned treatments, in terms of heatmaps, given by the physicians, the proposed PBL, and the standard Q-learning method (Non-Pessi). For each heatmap, the axis labels show different levels of each action, where 0 represents no drug is given, and a nonzero value corresponds to the dosage of the IV fluid or VP. It can



(a) Linear Signal



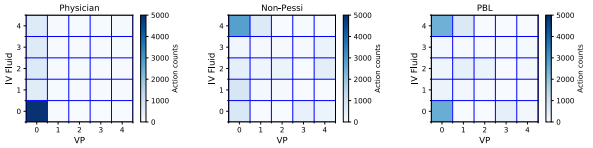
(b) Nonlinear Signal

Figure 2: The single-stage contextual bandit simulation with linear and nonlinear signals. The methods compared include the proposed PBL method, the PEVI method of Jin et al. (2021), and the standard Q-learning method without pessimism, each of which using BLBM and BNN.

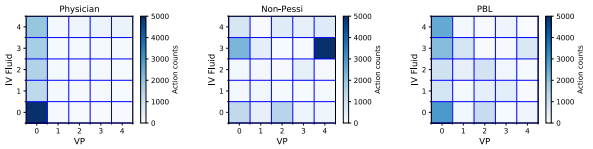
be seen from the plot that the physicians tended to prescribe no vasopressor to patients, but often considered IV fluids with various dosages. Meanwhile, the policy produced by our pessimism-based method tends to recommend treatment 0 or 4 for IV fluids and treatment 0 for VP, which is consistent with physicians' recommendations to some extent. Moreover, we use 5-fold cross-validation and apply the importance sampling method (Zhang et al., 2013) to evaluate the average reward under the estimated optimal policies produced by the proposed PBL, the standard Q-learning and CQL. Figure 4 reports the results. It is seen that PBL achieves the highest average award among the three methods, demonstrating the competitive performance of our method in this real data application.

7 DISCUSSIONS

In this article, we develop a novel pessimism-based Bayesian learning approach for offline optimal dynamic treatment regimes. We propose to combine the pessimism principle with Thompson sampling and Bayesian machine learning to optimize the degree of pessimism. Theoretically, we derive the upper bound for the regret of the proposed method, and obtain its explicit form in a specific case of a parametric model. Empirically, we develop a highly efficient and scalable computational algorithm based on variational inference. We also conduct extensive numerical experiments to illustrate the superior performance of our method. In terms of potential limitations of our proposed method, since it requires a large number of Monte Carlo samples, it can be computationally intensive, especially when the model dimension is large. How to further improve the computational efficiency warrants future research. In addition, it is of interest to extend our theoretical



(a) Single–stage decision making



(b) Two-stage decision making

Figure 3: Heatmap for the actions generated by the physician policy and the learned policy by the proposed PBL method and the standard Q-learning method without pessimism.

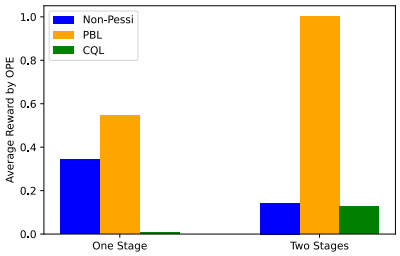


Figure 4: Average reward of proposed PBL method, the standard Q-learning method, and the conservative Q-learning evaluated by OPE.

results that take the model misspecification and approximation errors into consideration, and we leave it as future research.

Acknowledgements

Shi's research was partly supported by the EPSRC grant EP/W014971/1. Li's research was partly supported by the NSF grant CIF-2102227, and the NIH grants R01AG061303 and R01AG062542. The authors thank all the constructive comments from the referees and the area chair, which have led to a significant improvement of the earlier version of this paper.

References

Bai, C., Wang, L., Yang, Z., Deng, Z., Garg, A., Liu, P., and Wang, Z. (2022). Pessimistic bootstrapping for uncertainty-driven offline reinforcement learning. *arXiv preprint arXiv:2202.11566*.

Bickel, P. J. and Kleijn, B. J. (2012). The semiparametric bernstein–von mises theorem. *The Annals of Statistics*, 40(1):206–237.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR.

Buckman, J., Gelada, C., and Bellemare, M. G. (2020). The importance of pessimism in fixed-dataset policy optimization. *arXiv preprint arXiv:2009.06799*.

Cai, H., Shi, C., Song, R., and Lu, W. (2021). Jump interval-learning for individualized decision making. *arXiv preprint arXiv:2111.08885*.

Chakraborty, B. and Murphy, S. A. (2014). Dynamic treatment regimes. *Annual review of statistics and its application*, 1:447–464.

Chapelle, O. and Li, L. (2011). An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24.

Chen, G., Zeng, D., and Kosorok, M. R. (2016). Personalized dose finding using outcome weighted learning. *Journal of the American Statistical Association*, 111(516):1509–1521.

Chen, Y., Zeng, D., Xu, T., and Wang, Y. (2020). Representation learning for integrating multi-domain outcomes to optimize individualized treatment. *Advances in neural information processing systems*, 33:17976–17986.

Fu, Z., Qi, Z., Wang, Z., Yang, Z., Xu, Y., and Kosorok, M. R. (2022). Offline reinforcement learning with instrumental variables in confounded markov decision processes. *arXiv preprint arXiv:2209.08666*.

Fujimoto, S., Meger, D., and Precup, D. (2019). Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pages 2052–2062. PMLR.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741.

Hipp, C. and Michel, R. (1976). On the bernstein-v. mises approximation of posterior distributions. *The Annals of Statistics*, pages 972–980.

Jeunen, O. and Goethals, B. (2021). Pessimistic reward models for off-policy learning in recommendation. In *Fifteenth ACM Conference on Recommender Systems*, pages 63–74.

Jiang, R., Lu, W., Song, R., and Davidian, M. (2017). On estimation of optimal treatment regimes for maximizing t-year survival probability. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1165–1185.

Jin, Y., Yang, Z., and Wang, Z. (2021). Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Kaufmann, E., Cappé, O., and Garivier, A. (2012). On bayesian upper confidence bounds for bandit problems. In *Artificial intelligence and statistics*, pages 592–600. PMLR.

Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. (2020). Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33:21810–21823.

Kim, Y. (2006). The bernstein–von mises theorem for the proportional hazard model. *The Annals of Statistics*, 34(4):1678–1700.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kleijn, B. J. and van der Vaart, A. W. (2012). The bernstein-von-mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381.

Kosorok, M. R. and Laber, E. B. (2019). Precision medicine. *Annual review of statistics and its application*, 6:263–286.

Kumar, A., Fu, J., Soh, M., Tucker, G., and Levine, S. (2019). Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems*, 32.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. (2020). Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191.

Lambden, S., Laterre, P. F., Levy, M. M., and Francois, B. (2019). The sofa score—development, utility and challenges of accurate assessment in clinical trials. *Critical Care*, 23(1):1–9.

Lindley, D. V. and Smith, A. F. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(1):1–18.

Maahs, D. M., Mayer-Davis, E., Bishop, F. K., Wang, L., Mangan, M., and McMurray, R. G. (2012). Outpatient assessment of determinants of glucose excursions in adolescents with type 1 diabetes: proof of concept. *Diabetes technology & therapeutics*, 14(8):658–664.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355.

Prudencio, R. F., Maximo, M. R., and Colombini, E. L. (2022). A survey on offline reinforcement learning: Taxonomy, review, and open problems. *arXiv preprint arXiv:2203.01387*.

Qi, Z., Liu, D., Fu, H., and Liu, Y. (2020). Multi-armed angle-based direct learning for estimating optimal individualized treatment rules with various outcomes. *Journal of the American Statistical Association*, 115(530):678–691.

Qi, Z., Miao, R., and Zhang, X. (2022). Proximal learning for individualized treatment regimes under unmeasured confounding. *Journal of the American Statistical Association*, (just-accepted):1–33.

Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *Annals of statistics*, 39(2):1180.

Qiu, H., Carone, M., Sadikova, E., Petukhova, M., Kessler, R. C., and Luedtke, A. (2021). Optimal individualized decision rules using instrumental variable methods. *Journal of the American Statistical Association*, 116(533):174–191.

Quadrianto, N. and Ghahramani, Z. (2014). A very simple safe-bayesian random forest. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1297–1303.

Raghu, A., Komorowski, M., Ahmed, I., Celi, L., Szolovits, P., and Ghassemi, M. (2017). Deep reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1711.09602*.

Rezaeifar, S., Dadashi, R., Vieillard, N., Hussenot, L., Bachem, O., Pietquin, O., and Geist, M. (2021). Offline reinforcement learning as anti-exploration. *arXiv preprint arXiv:2106.06431*.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the second seattle Symposium in Biostatistics*, pages 189–326. Springer.

Seeger, M. (2006). Bayesian modelling in machine learning: A tutorial review.

Shi, C., Fan, A., Song, R., and Lu, W. (2018a). High-dimensional a-learning for optimal dynamic treatment regimes. *Annals of statistics*, 46(3):925.

Shi, C., Song, R., Lu, W., and Fu, B. (2018b). Maximin projection learning for optimal treatment decision with heterogeneous individualized treatment effects. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 80(4):681.

Sinyor, M., Schaffer, A., and Levitt, A. (2010). The sequenced treatment alternatives to relieve depression (STAR* D) trial: a review. *The Canadian Journal of Psychiatry*, 55(3):126–135.

Smith, A. F. (1973). A general bayesian linear model. *Journal of the Royal Statistical Society: Series B (Methodological)*, 35(1):67–75.

Tan, X., Qi, Z., Seymour, C. W., and Tang, L. (2022). Rise: Robust individualized decision learning with sensitive variables. *arXiv preprint arXiv:2211.06569*.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294.

Tsiatis, A. A., Davidian, M., Holloway, S. T., and Laber, E. B. (2019). *Dynamic treatment regimes: Statistical methods for precision medicine*. Chapman and Hall/CRC.

Uehara, M. and Sun, W. (2021). Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*.

Vaart, A. W. and Wellner, J. A. (1996). Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer.

Wang, Y. and Blei, D. M. (2019). Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, 114(527):1147–1161.

Wu, Y., Tucker, G., and Nachum, O. (2019). Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. (2021). Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34.

Yin, M., Bai, Y., and Wang, Y.-X. (2021). Near-optimal offline reinforcement learning via double variance reduction. *Advances in neural information processing systems*, 34.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J. Y., Levine, S., Finn, C., and Ma, T. (2020). Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142.

Zanette, A., Wainwright, M. J., and Brunskill, E. (2021). Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, 34.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, 100(3):681–694.

Zhang, J. (2020). Designing optimal dynamic treatment regimes: A causal reinforcement learning approach. In *International Conference on Machine Learning*, pages 11012–11022. PMLR.

Zhao, Q., Small, D. S., and Ertefaie, A. (2017). Selective inference for effect modification via the lasso. *arXiv preprint arXiv:1705.08020*.

Zhao, Y.-Q., Zeng, D., Laber, E. B., and Kosorok, M. R. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 110(510):583–598.

Zhou, W., Zhu, R., and Zeng, D. (2021). A parsimonious personalized dose-finding model via dimension reduction. *Biometrika*, 108(3):643–659.

A ADDITIONAL THEORETICAL RESULTS

A.1 Theory for Dynamic Treatment Regimes

We first generalize our theoretical guarantee for the DTR setting. We introduce the following notation. For any $K = 1, 2, \ldots, T-1$, let $f^{(K)}(h_K, a_K, w)$ denote the model at stage K used to fit the response of pseudo-reward $\max_{a^{(K+1)}} \widehat{f}_L^{(K+1)}(h^{(K+1)}, a^{(K+1)})$. For $K = T$, let $f^{(T)}(h_T, a_T, w)$ denote the model at final stage T for fitting $\mathbb{E}(R^{(T)}|H^{(T)} = h^{(T)}, A^{(T)} = a^{(T)})$. Denote $p^{(K)}(r|h_K, a_K, w)$ as the conditional density of the pseudo-reward given the history information under the model $f^{(K)}(h_K, a_K, w)$, such that

$$f^{(K)}(h_K, a_K, w) = \int_r p^{(K)}(r|h_K, a_K, w)dr,$$

We consider the following assumption for $K = 1, 2, \ldots, T$:

Assumption A1 *(i) The realization condition holds, i.e., there exists some $w_0^{(K)}$, such that $p^{(K)}(r|h_K, a_K, w_0^{(K)})$ is the oracle conditional pseudo-reward density function.*

(ii) The parameter space of ω is compact, and $p^{(K)}(r|h_K, a_K, w)$ is continuous and identifiable in ω.

(iii) $p^{(K)}(r|h_K, a_K, w)$ is differentiable in quadratic mean at the oracle parameter $w_0^{(K)}$ with a non-singular Fisher information matrix.

(iv) The prior measure of w is absolutely continuous in a neighborhood of $w_0^{(K)}$ with a continuous positive density at $w_0^{(K)}$.

Assumption A2 *Suppose the data used for fitting $\widehat{f}^{(K)}$ for $K = 1, \ldots T$ are independent.*

Assumption A1 is similar as Assumption 1, but extends to multiple stages. Assumption A2 imposes the cross-fitting condition to simplify our theoretical analysis. Without such an independence assumption, we need to impose certain entropy condition on the function class to prove Theorem A1 (Vaart and Wellner, 1996).

We next obtain the following proposition, and show that the $\widehat{f}_L^{(K)}(h_K, a_K)$ based on the Gaussian approximation and Monte Carlo sampling provides a valid uniform lower bound from the frequentist perspective.

Proposition A1 *Suppose Assumptions A1 and A2 hold. Then,*

$$\liminf_{\substack{n\to\infty \\ N\to\infty}} \mathbb{P}\left(\cap_{(h_K, a_K)} \left\{\widehat{f}_L^{(K)}(h_K, a_K) \leq Q^{(K)}(h_K, a_K)\right\}\right) \geq 1 - \alpha.$$

Finally, we establish the theoretical guarantee that characterizes the average regret of the estimated optimal policy from Algorithm 2.

Theorem A1 *Suppose Assumptions A1 and A2 sequentially hold for each stage K. Suppose the significance level $\alpha_K = \alpha/T$ is set following the Bonferroni correction. Then, as $n, N \to \infty$, with probability at least $1 - \alpha + o(1)$, the average regret is upper bounded by*

$$\sum_{K=1}^{T} \mathbb{E}_{\pi^*}\left[\mathbb{E}(R_K|H_K, A_K) - \widehat{f}_L^{(K)}(H_K, A_K)\right].$$

Specifically, if we use BLBM with p basis functions for model fitting, then there exists some constant $\bar{c} > 0$, such that the average regret can be upper bounded by

$$\bar{c}Tp^{1/2}n^{-1/2}.$$

A.2 Finite-sample Guarantee for Contextual Bandit

We next extend our theory to obtain the finite-sample guarantee for contextual bandit. As an example, following similar arguments in analyzing the Bernstein-von Mises approximation (Hipp and Michel, 1976), we can show that the statement of Theorem 1 holds with probability at least $1-\alpha-Cn^{-1/2}$, for some positive constant C that depends on the number of parameters p only. This yields a non-asymptotic probability upper bound as follows.

Corollary 1 *Suppose Assumption 1 holds. Then, as $N \to \infty$, with probability at least $1-\alpha-Cn^{-1/2}$, the average regret is upper bounded by*

$$\mathbb{E}_{\pi^*}\left[\mathbb{E}(R|S,A) - \widehat{f}_L(S,A)\right].$$

where C is a positive constant that depends on the number of parameters p.

B PROOFS

We provide the proofs for Proposition A1 and Theorem A1 in Appendix A.1. By setting $T=1$ as a special case, we obtain the proofs for Proposition 2 and Theorem 1 in Section 4.

B.1 Proof of Proposition A1

Denote $\mathcal{H}$ as the space for the history information. By definition that $f^{(K)}(h_K, a_K, w)$ is the model at stage K used to fit the response $\max_{a^{(K+1)}} \widehat{f}_L^{(K+1)}(h^{(K+1)}, a^{(K+1)})$, for any $K = 1, 2, \ldots, T-1$, and by (3), we know that, for any $(h_K, a_K) \in \mathcal{H} \times \mathcal{A}$ and $w \in \mathcal{W}$,

$$\mathbb{P}\left(\left\{w : \forall (h_K, a_K) \in \mathcal{H}\times\mathcal{A}, f_L^{(K)}(h_K,a_K) \le f^{(K)}(h_K,a_K,w)\right\} \mid \mathcal{D}_n\right) \ge \mathbb{P}(w \in \mathcal{W} \mid \mathcal{D}_n) \ge 1-\alpha.$$

For $Q^{(K)}(h_K, a_K) = f^{(K)}(h_K, a_K, w_0^{(K)})$, we obtain that

$$\mathbb{P}\left(\cap_{(h_K,a_K)}\left\{f_L^{(K)}(h_K,a_K) \le Q^{(K)}(h_K,a_K)\right\} \mid \mathcal{D}_n\right) \ge 1-\alpha.$$

Recall the assumptions that the parameter space is compact, and the likelihood function $p^{(K)}(r|h_K, a_K, w)$ is continuous and identifiable in w. Furthermore, recall that $p^{(K)}(r|h_K, a_K, w)$ is differentiable in quadratic mean at $w_0^{(K)}$ with non-singular Fisher information matrix, and the prior measure of w is absolutely continuous in a neighborhood of $w_0^{(K)}$ with a continuous positive density at $w_0^{(K)}$. Then, by Corollary 7 of Wang and Blei (2019), we have that

$$\liminf_{n\to\infty} \mathbb{P}\left(\cap_{(h_K,a_K)}\left\{f_L^{(K)}(h_K,a_K) \le Q^{(K)}(h_K,a_K)\right\}\right) \ge 1-\alpha.$$

Denote $w_{\text{opt}} = \underset{w\in\mathcal{W}}{\text{minimize}}\ f^{(K)}(h_K, a_K, w)$, subject to $\mathbb{P}(w \in \mathcal{W}|\mathcal{D}_n) \ge 1-\alpha$, which is the solution to the optimization problem in (3) for a given a_K and h_K. Let $\delta > 0$ be a positive constant, such that $|w_{\text{opt},j} - w_j| \le \delta$ for $j \in \{1, 2, \ldots, p\}$, and the value of δ will be determined later. Since f is Lipschitz continuous in w, there exists a constant $L > 0$, such that, for $\forall (h_K, a_K) \in \mathcal{H} \times \mathcal{A}$,

$$|f^{(K)}(h_K, a_K, w_{\text{opt}}) - f^{(K)}(h_K, a_K, w)| \le L||w_{\text{opt}} - w||_2 \le L\sqrt{p}\delta.$$

Considering the interval $I_j = [w_{\text{opt},j} - \delta, w_{\text{opt},j} + \delta]$, we have that

$$\mathbb{P}\left(w_j \in I_j \mid \mathcal{D}_n\right) = \int_{I_j} p(w_j|\mathcal{D}_n) dw_j.$$

In our method, we adopt a Gaussian distribution to approximate the posterior distribution. Hence, p is Lipschitz continuous in w_j for a given mean and a given covariance matrix. Thus, we can find some constants $c_1, c_2 > 0$, such that

$$c_1\delta \le \mathbb{P}\left(w_j \in I_j \mid \mathcal{D}_n\right) \le c_2\delta.$$

This implies that

$$\mathbb{P}\left(\bigcap_{h_K,a_K}\left\{|f^{(K)}(h_K,a_K,w_{\text{opt}})-f^{(K)}(h_K,a_K,w_j)|\le L\sqrt{p}\delta\right\}\mid \mathcal{D}_n\right)\ge (c_1\delta)^p.$$

Since we randomly generate N samples from the posterior distribution of w and select the one that minimizes f, with some calculations, we have that

$$\mathbb{P}\left(|f^{(K)}(h_K,a_K,w_{\text{opt}})-f^{(K)}(h_K,a_K,w^*)|\le L\sqrt{p}\delta\mid \mathcal{D}_n\right)\ge 1-[1-(c_1\delta)^p]^N.$$

Letting $\delta=N^{-1/(2p)}$, we obtain that

$$\liminf_{N\to\infty}\mathbb{P}\left(|f^{(K)}(h_K,a_K,w_{\text{opt}})-f^{(K)}(h_K,a_K,w^*)|\le \epsilon\mid \mathcal{D}_n\right)=1 \tag{5}$$

for any ϵ. From Proposition 1, we know that

$$\mathbb{P}\left(\cap_{(h_K,a_K)}\left\{f^{(K)}(h_K,a_K,w_{\text{opt}})\le Q^{(K)}(h_K,a_K)\right\}\mid \mathcal{D}_n\right)\ge 1-\alpha.$$

Combining it with (5), we obtain that, for any ϵ,

$$\liminf_{N\to\infty}\mathbb{P}\left(\cap_{(h_K,a_K)}\left\{\widehat{f}_L^{(K)}(h_K,a_K)\le Q^{(K)}(h_K,a_K)+\epsilon\right\}\mid \mathcal{D}_n\right)\ge 1-\alpha,$$

where $\widehat{f}_L(h_K,a_K)=f^{(K)}(h_K,a_K,w^*)$. Since ϵ can be chosen arbitrarily small, we have that

$$\liminf_{N\to\infty}\mathbb{P}\left(\cap_{(h_K,a_K)}\left\{\widehat{f}_L^{(K)}(h_K,a_K)\le Q^{(K)}(h_K,a_K)\right\}\mid \mathcal{D}_n\right)\ge 1-\alpha,$$

By Corollary 7 of Wang and Blei (2019), we have

$$\liminf_{n\to\infty}\liminf_{N\to\infty}\mathbb{P}\left(\cap_{(h_K,a_K)}\left\{\widehat{f}_L^{(K)}(h_K,a_K)\le Q^{(K)}(h_K,a_K)\right\}\right)\ge 1-\alpha.$$

B.2 Proof of Theorem A1

In Algorithm 2, we employ dynamic programming and construct the pseudo-reward, $R_i^{(K)}=\max_{a_{K+1}\in\mathcal{A}}\widehat{f}_L^{(K+1)}(h_i^{(K+1)},a_{K+1})$, for each instance at the Kth stage, $K=1,2,\ldots,T-1$. Define the event

$$\Omega_K=\left\{\widehat{f}_L^{(K)}(h_K,a_K)<\mathbb{E}(R_K|H_K=h_K,A_K=a_K),\ \forall(h_K,a_K)\in\mathcal{H}\times\mathcal{A}\right\}.$$

for $K=1,2,\ldots,T$. Define the joint event as $\Omega=\bigcap_{K=1}^T\Omega_K$.

By Proposition A1, we can show that $\mathbb{P}(\Omega_K)\ge 1-\alpha/T+o(1)$ as both n and N approach infinity for any $K=1,2,\ldots,T$. Then with the Bonferroni correction, we have that $\mathbb{P}(\Omega)\ge 1-\alpha+o(1)$ when n and N approach infinity. Let $\mathbb{E}_T(\mathcal{R};s^{(1)})$ denote the average regret given the initial state $s^{(1)}$. Then we decompose the average regret into three components as follows:

$$\mathbb{E}_T(\mathcal{R};s^{(1)})=\sum_{K=1}^T\Big\{\underbrace{-\mathbb{E}_{\hat{\pi}}[\eta_K(h_K,a_K)|H_1=s^{(1)}]}_{\text{(i)}}+\underbrace{\mathbb{E}_{\pi^*}[\eta_K(h_K,a_K)|H_1=s^{(1)}]}_{\text{(ii)}}+$$
$$\underbrace{\mathbb{E}_{\pi^*}[\langle\widehat{f}_L(h_K,a_K),\pi^*(\cdot|h_K)-\hat{\pi}(\cdot|h_K)\rangle_{\mathcal{A}}|H_1=s^{(1)}]}_{\text{(iii)}}\Big\}$$

where $\eta_K(h_K,a_K)=\mathbb{E}(R_K|H_K=h_K,A_K=a_K)-\widehat{f}_L^{(K)}(h_K,a_K)$ is the model evaluation error.

We start with the last stage $K=T$ and do backward induction. We first note that, since $\hat{\pi}$ is greedy with respect to $\widehat{f}_L^{(T)}(h,a)$, the optimization error (iii) is non-positive. So it can be directly removed from the bound. Next, we consider the error term (i). Under event Ω_K, we have that

$$0\le\eta(h_T,a_T)=\mathbb{E}(R_T|H_T=h_T,A_T=a_T)-\widehat{f}_L^{(T)}(h_T,a_T)$$

Thus, we obtain that

$$-\mathbb{E}_{\hat{\pi}}[\eta(h_T, a_T)|H_1 = s^{(1)}] \leq 0$$

We repeat the same produce for $K = T-1, \ldots, 1$, and under each event Ω_K, we obtain that

$$\begin{aligned} -\mathbb{E}_{\hat{\pi}}[\eta(h_K, a_K)|H_1 = s^{(1)}] &\leq 0 \\ \mathbb{E}_{\pi^*}[\langle \widehat{f}_L(h_K, a_K), \pi^*(\cdot|h_K) - \hat{\pi}(\cdot|h_K)\rangle_{\mathcal{A}}|H_1 = s^{(1)}] &\leq 0 \end{aligned}$$

Combining the inequalities above, we have that, under the event Ω,

$$\begin{aligned} -\sum_{K=1}^{T} \mathbb{E}_{\hat{\pi}}[\eta(h_K, a_K)|H_1 = s^{(1)}] &\leq 0 \\ \sum_{K=1}^{T} \mathbb{E}_{\pi^*}[\langle \widehat{f}_L(h_K, a_K), \pi^*(\cdot|h_K) - \hat{\pi}(\cdot|h_K)\rangle_{\mathcal{A}}|H_1 = s^{(1)}] &\leq 0 \end{aligned}$$

which implies that

$$\mathbb{E}_T(\mathcal{R}; s^{(1)}) \leq \sum_{K=1}^{T} \mathbb{E}_{\pi^*}\left[\mathbb{E}(R_K|H_K = h_K, A_K = a_K) - \widehat{f}_L^{(K)}(h_K, a_K)|H_1 = s^{(1)}\right].$$

Taking the integral over the randomness of $s^{(1)}$ on both sides, as $n, N \to \infty$, with probability at least $1 - \alpha + o(1)$, we can upper bound the average regret by

$$\sum_{K=1}^{T} \mathbb{E}_{\pi^*}\left[\mathbb{E}(R_K|H_K, A_K) - \widehat{f}_L^{(K)}(H_K, A_K)\right].$$

For the specific case of BLBM, we have that

$$\begin{aligned} \mathbb{E}_T(\mathcal{R}; s^{(1)}) &\leq \sum_{K=1}^{T} \mathbb{E}_{\pi^*}\left[\mathbb{E}(R_K|H_K, A_K) - \widehat{f}_L^{(K)}(H_K, A_K)\right] \\ &\leq \sum_{K=1}^{T} \mathbb{E}_{\pi^*}\left[\mathbb{E}(R_K|H_K, A_K) - f_L^{(K)}(H_K, A_K) + f_L^{(K)}(H_K, A_K) - \widehat{f}_L^{(K)}(H_K, A_K)\right] \\ &\leq \underbrace{\sum_{K=1}^{T} \mathbb{E}_{\pi^*}\left[\mathbb{E}(R_K|H_K, A_K) - f_L^{(K)}(H_K, A_K)\right]}_{\text{(I)}} + \underbrace{\sum_{K=1}^{T} \mathbb{E}_{\pi^*}\left[f_L^{(K)}(H_K, A_K) - \widehat{f}_L^{(K)}(H_K, A_K)\right]}_{\text{(II)}} \end{aligned}$$

where (II) is 0 if we directly solve the optimization problem (4) without Monte Carlo sampling. By Corollary 7 of Wang and Blei (2019) and the Lipschitz continuous condition for $f^{(K)}$, we have that

$$\begin{aligned} \text{(I)} &\leq \zeta_{1-\alpha} T \mathbb{E}_{\pi^*}\left[\phi^T(H_K, A_K)\Lambda_K^{-1}\phi(H_K, A_K)\right] + o(n^{-1/2}) \\ &\leq \zeta_{1-\alpha} T \sqrt{\sum_{j=1}^{p} \frac{1}{cn}} + o(n^{-1/2}) \\ &\leq c'\zeta_{1-\alpha} T p^{1/2} n^{-1/2}, \end{aligned}$$

where $\zeta_{1-\alpha}$ is the $1-\alpha$ percentile of the standard normal distribution, and $\Lambda_K^{-1} = \sum_{i=1}^{n} \phi(H_{K,i}, A_{K,i})\phi^T(H_{K,i}, A_{K,i})$.

Therefore, we obtain that

$$\mathbb{E}_T(\mathcal{R}; s^{(1)}) \leq \bar{c} T p^{1/2} n^{-1/2}.$$

for some constant $\bar{c}$.

C ADDITIONAL NUMERICAL RESULTS

C.1 Data Generation for One-Stage Contextual Bandit

We first outline the details of our data generation for one-stage contextual bandit.

- Linear Signal:

$$r(s,a) = \begin{cases} 0.2s_1 + 0.25s_2 + 0.3s_3 + 0.1z & \text{if } a = 1 \\ 0.25s_1 + 0.3s_2 + 0.35s_3 + 0.1z & \text{if } a = 2 \end{cases}$$

  where $z \sim \text{Normal}(0,1)$. We draw $s \in \mathbb{R}^3$ with $s_i \sim \text{Normal}(0,1)$ for $i = 1,2,3$. For each state s, we denote $a^* = \arg\max_a \mathbb{E}[r(a,s)]$, and generate a with the probability $P(a = a^*) = 1 - \epsilon$, where $\mathbb{E}$ is taken with respect to the randomness of the reward function.

- Nonlinear Signal: We define two transformation functions,

$$\begin{aligned} f_1(s) &= [0.1\exp(4s_1) + 4/(1+\exp(-20(s_2-0.5))) + 3s_3 + 2s_4 + s_5]/2.5, \\ f_2(s) &= [0.12\exp(4s_1) + 4.8/(1+\exp(-20(s_2-0.5))) + 3.6s_3 + 2.4s_4 + 1.2s_5]/2.5, \\ r(s,a) &= \begin{cases} f_1(s) + 0.1z & \text{if } a = 1 \\ f_2(s) + 0.1z & \text{if } a = 2 \end{cases}, \end{aligned}$$

  where $z \sim \text{Normal}(0,1)$. We draw $s \in \mathcal{R}^5$ with $s_i \sim \text{Uniform}[0,1]$ for $i = 1,2,3,4,5$. We generate the actions in the same way as for the linear signal.

C.2 Data Generation for Two-Stage DTR

We next outline the details of our data generation for two-stage DTR.

- Linear Signal: We define two transformation functions,

$$\begin{aligned} f_1(s) &= 0.2s_1 + 0.25s_2 + 0.3s_3, \\ f_2(s) &= 0.25s_1 + 0.3s_2 + 0.35s_3. \end{aligned}$$

  We first randomly generate the coefficient matrix $W_1 \in \mathbb{R}^{2\times 3}$ with each entry independently drawn from $\text{Normal}(0,1)$. We then define $W_2 = W_1 + 0.05$, where the sum calculation is element-wise. We fix W_1 and W_2. For each replication, we draw the state at the first stage as $s^{(1)} = (s_1^{(1)}, s_2^{(1)})^T \in \mathbb{R}^{2\times 1}$ with $s_i^{(1)} \sim \text{Normal}(0,1)$ for $i = 1,2$. Suppose that the action of the first stage is chosen as $a^{(1)}$, then we generate the state at the second stage as

$$\begin{cases} s^{(2)} = W_1^T s^{(1)} + z & \text{if } a^{(1)} = 1 \\ s^{(2)} = W_2^T s^{(1)} + z & \text{if } a^{(1)} = 2 \end{cases},$$

  where $z = (z_1, z_2, z_3)^T \in \mathbb{R}^{3\times 1}$, and $z_i \sim \text{Normal}(0,1)$ for $i = 1,2,3$. Suppose that we get $a^{(2)}$ as the action of the second stage. We generate the reward as

$$\begin{cases} r(s,a) = f_1(s) + 0.1z' & \text{if } a^{(2)} = 1 \\ r(s,a) = f_2(s) + 0.1z' & \text{if } a^{(2)} = 2 \end{cases}, \qquad \text{where } z' \sim \text{Normal}(0,1) \text{ for } i = 1,2,3.$$

  To generate the action of the first stage, we introduce the notation,

$$\begin{aligned} g(s^{(1)}, a^{(1)}) &= \max_{a^{(2)}} \mathbb{E}_{r,s}(r(s^{(2)}, a^{(2)})) \\ a^{(1)*} &= \arg\max_{a^{(1)}} \mathbb{E}_s[g(s^{(1)}, a^{(1)})] \end{aligned}$$

  where $\mathbb{E}_{r,s}$ is taken over the randomness of the noise of the reward function and of the generation of the state $s^{(2)}$, and $\mathbb{E}_s$ is only taken over the randomness of the generation of the state $s^{(2)}$. We generate $a^{(1)}$ with the probability distribution of $P(a^{(1)} = a^{(1)*}) = 1 - \epsilon$, where ϵ is a fixed greedy parameter.

To generate the action of the second stage, we define

$$a^{(2)*} = \arg\max_{a^{(2)}} \mathbb{E}_r(r(s^{(2)}, a^{(2)})),$$

where $\mathbb{E}_r$ is taken with respect to the randomness of the reward function. We then generate $a^{(2)}$ with the probability distribution of $P(a^{(2)} = a^{(2)*}) = 1 - \epsilon$.

- Nonlinear Signal: We define two transformation functions,

$$f_1(s) = [0.1\exp(4s_1) + 4/(1 + \exp(-20(s_2 - 0.5))) + 3s_3 + 2s_4 + s_5]/2.5,$$
$$f_2(s) = [0.12\exp(4s_1) + 4.8/(1 + \exp(-20(s_2 - 0.5))) + 3.6s_3 + 2.4s_4 + 1.2s_5]/2.5.$$

We first randomly generate the coefficient matrix $W_1 \in \mathbb{R}^{2\times 5}$ with each entry independently drawn from Normal$(0, 1)$. We then define $W_2 = W_1 + 0.05$, where the sum calculation is element-wise. We fix W_1 and W_2. For each replication, we draw the state at the first stage as $s^{(1)} = (s_1^{(1)}, s_2^{(1)})^T \in \mathbb{R}^{5\times 1}$ with $s_i^{(1)} \sim$ Normal$(0, 1)$ for $i = 1, 2, 3, 4, 5$. Suppose that the action of the first stage is $a^{(1)}$, then we generate the state at the second stage as

$$\begin{cases} s^{(2)} = W_1^T s^{(1)} + z & \text{if } a^{(1)} = 1 \\ s^{(2)} = W_2^T s^{(1)} + z & \text{if } a^{(1)} = 2 \end{cases},$$

where $z = (z_1, z_2, z_3, z_4, z_5)^T \in \mathbb{R}^{3\times 1}$, and $z_i \sim$ Normal$(0, 1)$ for $i = 1, 2, 3, 4, 5$. Suppose that we get $a^{(2)}$ as the action of the second stage. We generate the reward as

$$\begin{cases} r(s, a) = f_1(s/10) + 0.1z' & \text{if } a^{(2)} = 1 \\ r(s, a) = f_2(s/10) + 0.1z' & \text{if } a^{(2)} = 2 \end{cases}, \qquad \text{where } z' \sim \text{Normal}(0, 1) \text{ for } i = 1, 2, 3, 4, 5,$$

where $s/10$ is calculated in an element-wise fashion. We generate the actions in the same way as for the linear signal.

C.3 Implementation Details and Computing Time

For BLBM, we use the RBFSampler function under its default setting in the sklearn package to generate basis with random Fourier features. For BNN, we use a two-layer neural network with 16 hidden units at each layer, and the ReLU activation function. We use SGD for optimization, with the learning rate 10^{-4}. We set the number of training epochs at 500, the batch size at 100, the number of Monte Carlo samples for gradient descent at 5, and the number of samples from the posterior distribution at 10000. We use savio_htc cluster for all the computations. For BLBM, it takes about 1.5 seconds to run for one replication on one CPU for the single-stage setting, and about 25 seconds for the two-stage setting. For BNN, it takes about 3 minutes for the single-stage setting, and about 20 minutes for the two-stage setting. We use multiple CPUs for parallelization.

C.4 Sensitivity Analysis

We conduct a sensitivity analysis by varying a number of parameters in our method, including the number of Monte Carlo samples N, the number of ensembles M for the MC gradient computation in variational inference, and the significance level α. We adopt the single-stage contextual bandit setting with a linear signal. Figure A1 reports the results. It is seen that our method is not overly sensitive to those parameters, as long as they are in a reasonable range. To the contrary, PEVI is sensitive to the choice of the parameter c, as shown in Figures 2 and A2.

C.5 Plot for Two-Stage DTR

Figure A2 reports the simulation results for the two-stage DTR setting.

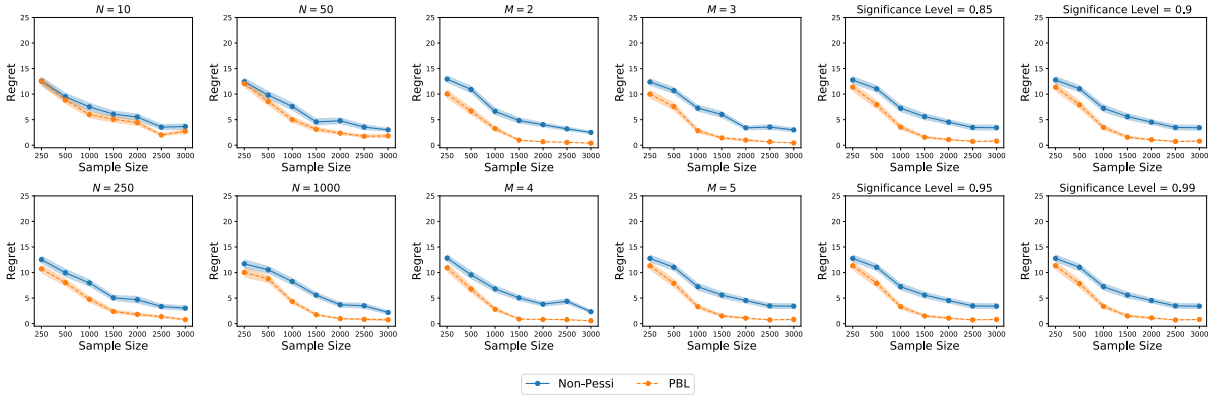
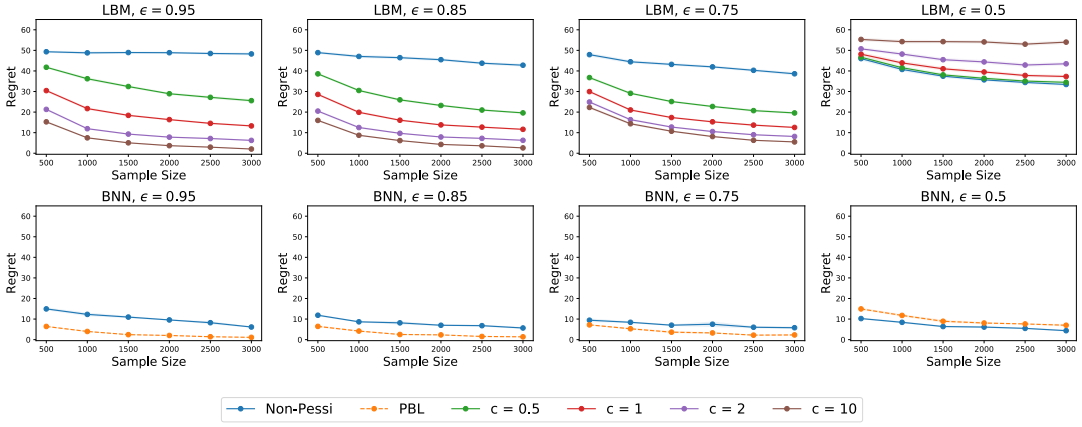
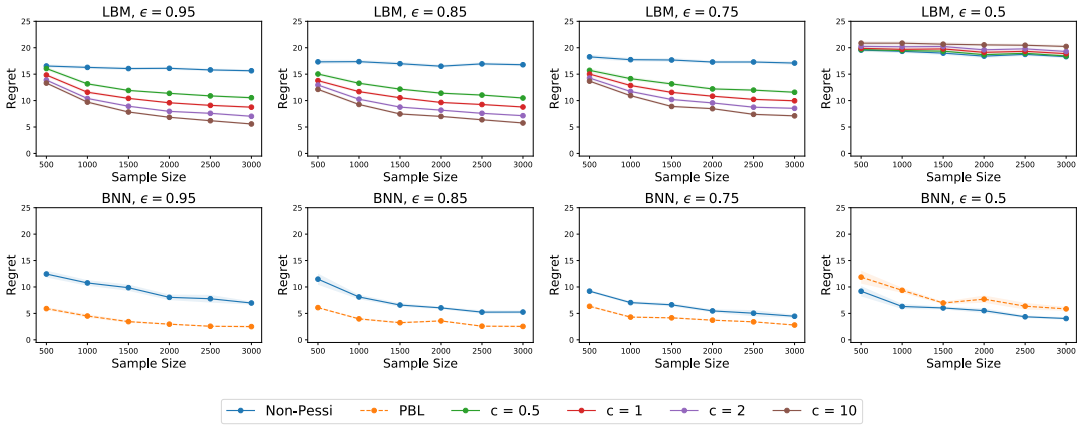


Figure A1: Sensitivity analysis for the single-stage contextual bandit simulation with a linear signal.



(a) Linear Signal



(b) Nonlinear Signal

Figure A2: The two-stage DTR simulation with linear and nonlinear signals. The methods compared include the proposed PBL method, the PEVI method of Jin et al. (2021), and the standard Q-learning method without pessimism, each of which using BLBM and BNN.