Adaptive Estimation and Optimal Control in Offline Contextual MDP's

Anonymous authors
Paper under double-blind review

Abstract

Contextual MDP's are powerful tools with wide applicability in areas from biostatistics to machine learning. However, specializing them to offline datasets has been challenging due to a lack of robust, theoretically backed, methods. Our work tackles this problem by introducing a new approach towards adaptive estimation and cost optimization of contextual MDP's. This estimator, to the best of our knowledge, is the first of its kind, and is endowed with strong optimality guarantees. We achieve this by overcoming the key technical challenges evolving from the endogenous properties of contextual MDP's; such as non-stationarity, or model irregularity. Our guarantees are established under complete generality by utilizing the relatively recent and powerful statistical technique of T-estimation (Baraud, 2011). We first provide a procedure for selecting an estimator given a sample from a contextual MDP and use it derive oracle risk bounds under two distinct, but nevertheless meaningful, loss functions. We then consider the problem of determining the optimal control with the aid of the aforementioned density estimate and provide finite sample guarantees for the cost function.

1 Introduction

Contextual Markov decision processes (contextual MDPs; Hallak et al., 2015) are a core abstraction for sequential decision making with exogenous information ("context") that modulates the dynamics and costs. They underpin applications across healthcare, economics, and operations where learning a decision rule from historical logs—*offline reinforcement learning (offline RL)*—is often the only viable option. Yet, despite their practical importance, a general, assumption-light statistical theory for *estimating the transition mechanism and optimizing costs* in contextual MDPs has remained elusive: most existing approaches rely on stationarity/ergodicity, parametric modeling, or strong smoothness assumptions that are frequently violated in real data.

Goal. We develop a non-parametric framework for *offline* contextual MDPs that (i) adaptively estimates the transition density without stationarity or ergodicity assumptions, (ii) delivers *oracle risk bounds* in an empirical Hellinger metric, and (iii) transfers these guarantees to *cost minimization* and *optimal control selection* via a plug-in scheme. Our results close a gap between practice and theory by providing finite-sample guarantees under minimal conditions while remaining minimax-optimal (up to logarithmic factors) under standard smoothness regimes.

Our Approach. Let $s(x, a, g, y)$ denote the (contextual) transition density. We construct an estimator $\hat{s}$ by minimizing a penalized *pairwise comparison* functional over a rich, countable model class $\mathcal{S}$:

$$\hat{s} \in \operatorname*{argmin}_{f \in \mathcal{S}} \sup_{f' \in \mathcal{S}} \left\{ \alpha \, \mathcal{H}^2(f, f') + T(f, f') \right\} + \frac{L \, \Delta_{\mathcal{S}}(f)}{n},$$

where $\mathcal{H}^2$ is the empirical Hellinger loss, $T(\cdot, \cdot)$ compares candidates on the observed trajectory (including a density-correcting term), and $\Delta_{\mathcal{S}}$ is a data-independent complexity penalty. This is the T-estimation principle, first introduced in a seminal PTRF paper by Baraud & Birgé (2009), and rooted in modern model-selection theory, yields robust performance under dependent, non-stationary data and accommodates non-parametric model classes (e.g., piecewise-constant partitions, spline bases etc.).

Intuition. Technically, our analysis relies on (i) a martingale Bernstein inequality to control empirical deviations *without* stationarity or mixing assumptions, and (ii) the polarization identity to relate risk in squared-root density

space to Hellinger loss. These tools allow us to derive high-probability comparisons that convert into *expected* oracle inequalities after integrating the deviation parameter.

From density estimation to control: Given a user-specified, positive cost function $L(x, a, g, y)$, we study the *empirical cost*

$$\hat{C}_n(a, g) \;=\; \int L(x, a, g, y)\, \hat{s}(x, a, g, y)\, \lambda_n^{(1)}(dx, dy),$$

and show that optimizing a simple upper-confidence surrogate of $\hat{C}_n$ identifies (nearly) optimal controls from logged data. Crucially, our guarantees are *policy-agnostic*: any procedure that optimizes the plug-in cost inherits our rates, thereby decoupling statistical estimation from control optimization.

Remark 1. *Let $L(x, l, g, y) = |\varphi(g, l)^T(\psi(x) - \psi(y))|$ (Belogolovsky et al., 2021; Zhou et al., 2024), where φ and ψ are feature maps and suppose that $\tilde{\omega}_\psi$ is a joint measure between $\mu_\chi(\psi^{-1})$ with itself. Then the cost function defined as above is essentially a cost for a transport from (χ, μ_χ) to itself. In fact, it is exactly the Wasserstein-1 distance if $\tilde{\omega}_\psi$ is the optimal joint measure. Then the problem translates into selecting a control l for a given context g such that "some" risk metric is minimised (see Section 4.3 for more details).*

We now detail the two distinct challenges facing us:

1. Our first challenge is innate to non-parametric inference, and can be linked to "bandwidth selection" for kernel density estimators. For any $\sigma \in (0, 1]$, let $\mathbb{H}^\sigma(\chi)$ be the class of all Hölder smooth functions on χ (formally defined in Definition 2. See also Bergh & Löfström (1976)). Then, it is known that if $\mathscr{s}(x, l, g, \cdot) \in \mathbb{H}^{\sigma_{x,l,g}}(\chi)$, setting the bandwidth $\mathcal{O}(n^{-1/(2\sigma_{x,l,g}+1)})$ one can recover the minimax rate of estimating the density $\mathscr{s}(x, l, g, \cdot)$. However, such assumptions are impractical when there are uncountably infinite many x, l, g's.

   At this point, we remark that one cannot simply take the bandwidth to be $\mathcal{O}(n^{-1/(2\sigma+1)})$ where $\sigma = \sup_{x,l,g} \sigma_{x,l,g}$ since, without additional assumptions on the diameter of χ, Hölder spaces generally **do not** satisfy the embedding $\mathbb{H}^\alpha(\chi) \subset \mathbb{H}^\beta(\chi)$ for $\alpha < \beta$. Therefore, this problem of bandwidth selection is even more tenuous in the MDP context.

2. Our second challenge is on the collection of historical data. For regular (non-contextual) MDP's it is common practice to assume that the actions a_i depend only on the current state X_i (Sutton & Barto, 2018), and the contexts G_i are independent; an assumption which is easily violated. As an example, note that, hidden Markov models (which may correspond in real-life to the biological markers for the evolving health conditions of a single patient), are not Markovian up to any degree. However, since the contexts G_i's can themselves arise out of such an evolving time series, it is desirable to have an estimation procedure that is "robust" to non-stationarity or non-ergodicity but optimal if such extra conditions are available. We now informally state our contributions.

Contributions.

- **Non-parametric estimation for contextual MDPs.** We introduce a T-estimator for the transition density that requires only boundedness of a dominating density and *no* stationarity or ergodicity. We prove an *oracle inequality* in empirical Hellinger loss:

$$\mathrm{E}\big[\mathscr{H}^2(s, \hat{s})\big] \;\lesssim\; \inf_{f \in \mathcal{S}} \Big\{ \mathscr{H}^2(s, f) + \tfrac{\Delta_{\mathcal{S}}(f) + \dim(f) \log n}{n} \Big\}.$$

- **Minimax-rate adaptivity.** For standard smoothness classes (e.g., Hölder/Besov), our procedure attains the optimal non-parametric rate (up to poly-log order): $\mathrm{E}[\mathscr{H}^2(s, \hat{s})] \lesssim \big(\frac{\log n}{n}\big)^{\bar{\sigma}/(4+\bar{\sigma})}$.

- **Cost minimization with finite-sample guarantees.** We establish non-asymptotic bounds for the *plug-in* cost gap between the control selected by our UCB-style surrogate and the true optimal control. The bound scales as $O\Big(\|C\|_\infty \cdot \min\big\{(\frac{r_{\mathcal{S}}(f,n)}{n})^{1/2-\delta}, (\frac{r_{\mathcal{S}}(f,n)}{n})^{\delta}, (\frac{r_{\mathcal{S}}(f,n)}{n})^{1/4}\big\}\Big)$, and achieves the fastest rate at $\delta = \frac{1}{4}$.

Note for Practitioners: Our results provide a practical, assumption-light route to *offline* cost-sensitive decision making in contextual MDPs: learn a non-parametric transition model with T-estimation, plug it into the task-specific cost, and

optimize controls with UCB-style adjustments. This pipeline yields rigorous risk and optimality guarantees without relying on stationarity, ergodicity, or parametric misspecification—a favorable regime for real-world logs with shifting contexts and irregular dynamics.

The rest of the paper is organised as follows: Section 2 outlines a comprehensive discussion of relevant research works. In section 3, we formally introduce the model selection procedure and the relevant notations. In section 4, we provide our key theoretical results on estimating $\mathfrak{s}$, with the proof sketches for these, while the full proofs have been deferred till the appendix due to lack of space. Section 5 contains our theoretical guarantees on estimating the optimal cost, and finally, in Section 6, we conclude by mentioning the limitations and broader impact of our work.

2 Background and Related Research

Initial works on model selection as a form of estimation dates back to Barron et al. (1999), and for a general overview of the literature on model selection, we refer the readers to the exceptional (albeit slightly dated) monograph by Pascal Massart (Massart, 2007). T-estimators, specifically, have seen a significant amount of focus in recent years. They have been used as a general density estimation procedure in i.i.d. (Baraud & Birgé, 2009; Birgé, 2006), bivariate (Sart, 2017), Markovian (Sart, 2023) and controlled Markovian (Banerjee et al., 2025b) contexts, and seen applications in machine learning related fields like differential privacy (Sart, 2023).

In the interest of exposition, we point out that recently, T-estimators have been generalised in a series of groundbreaking papers starting with Baraud et al. (2017), with follow ups in Baraud & Birgé (2018), and Baraud & Birgé (2020) These so-called ρ-estimators estimators—based on the Bhattacharya correlation coefficient $\rho(f,g) := \int \sqrt{fg}\, d\lambda_n$—are non-parametric, yet MLE-like in their efficiency. However, theory of ρ-estimation is still in its infancy, and for the sake of brevite, we fallback on the various technical tools available to us for T-estimators.

Contextual MDP's—first formally introduced in Hallak et al. (2015)—have become a cornerstone of cost optimisation and decision making in diverse fields such as finance, economics, and healthcare (Batsis & Samothrakis, 2024; Xiao et al., 2019; Tang & Wiens, 2021; Li et al., 2022b; Javanmard & Nazerzadeh, 2019), in both online (Li et al., 2022b) and offline (Zhou et al., 2024) settings; proving to be a major tool in solving problems with real life applications (Cao et al., 2023). Much work has been done in the field of contextual MDPs with PAC/sample complexity bounds for the optimal policy (Krishnamurthy et al., 2016; Sun et al., 2019; Jiang et al., 2017). However, our work is more general; as we prove in Theorem 2, one can use any method to find the optimal control and leverage optimality guarantees endowed by our estimator—a crucial setting which till now had remained largely unexplored, formalized by the following open question.

Open question. The paper resolves the following open questions in the theory of contextual MDP's: (i) Can we provide a *complete convergence theory for contextual MDP's*—oracle risk bounds for transition function and the optimal cost without prior assumptions on the data generating process? (ii) Can we then, under further suitable assumptions, derive minimax optimality guarantees for our estimator and the associated cost function?

With that, we move on to formally define our problem.

3 Problem Formulation

We initiate this section by introducing some notation that will be used repeatedly throughout the paper. Let $\mathbb{N}$ and $\mathbb{R}$ denote the natural and real numbers, and the symbol $\lfloor\cdot\rfloor$, the floor function. All random variables in this paper will be defined with respect to a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$, where $\mathcal{F}$ is a σ-algebra and $\mathbb{F} := \{\mathcal{F}_i\}_{i\geq 0}$, with $\mathcal{F}_i \subset \mathcal{F}$, is a given filtration. Let $\{(X_i, a_i, G_i)\}$ represent a discrete-time stochastic processes adapted to $\mathbb{F}$, and taking values in $\chi \times \mathbb{I} \times \mathcal{G} \subseteq \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \times \mathbb{R}^{d_2}$. We call χ, $\mathbb{I}$, and $\mathcal{G}$ the *state*, *control*, and *context* spaces respectively. For all non-negative integers i, j, we define $\mathcal{H}_i^j := (X_j, a_j, G_j \ldots, X_i, a_i, G_i)$ and $\hbar_i^j := (x_j, l_j, g_j, \ldots, x_i, l_i, g_i)$ and note that $\hbar_i^j$ is an element of $(\chi \times \mathbb{I} \times \mathcal{G})^{j-i+1}$. The σ-field generated by $\mathcal{H}_i^j$ shall be $\mathcal{F}_i^j$. Throughout the paper, we will assume that χ, $\mathbb{I}$, and $\mathcal{G}$ are compact. $\mathbb{X} := \chi \times \mathbb{I} \times \mathcal{G} \times \chi$ shall denote the augmented state-space, and is compact by the previous assumption. When $\mathbb{X}$ is *not compact*, all of our theory still continues to hold on any restriction of $\mathfrak{s}$ on a compact subset $A \subset \chi \times \mathbb{I} \times \chi$, given by $\mathfrak{s}\mathbb{1}_A$. Observe that $\mathfrak{s}\mathbb{1}_A$ is not necessarily a conditional density, since it may not integrate upto 1.

Let $\mathbb{E}[X]$ be the expectation and $\sigma(X)$ the σ-algebra induced by X. We endow χ, $\mathbb{I}$, and $\mathcal{G}$ with integrating measures μ_χ, $\mu_\mathbb{I}$, and $\mu_\mathcal{G}$ respectively. One can assume μ's to be Lebesgue when $\chi, \mathbb{I}, \mathcal{G}$ are Euclidean or the counting measure

when they are discrete. By $\mathrm{Vol}(\mathcal{S})$ we denote the volume of the set $\mathcal{S}$ with respect to its natural measure. As an example, if $\mathcal{S} \subset \chi$, then $\mathrm{Vol}(\mathcal{S}) = \mu_\chi(\mathcal{S})$; if $\mathcal{S} \subset \mathbb{I}$, then $\mathrm{Vol}(\mathcal{S}) = \mu_{\mathbb{I}}(\mathcal{S})$, etc. $\mathcal{C}$ and c are always used to denote universal constants whose values can change from line to line. We call $m = \{k : k \subseteq \chi \times \mathbb{I} \times \chi\}$ to be a *partition* of $\chi \times \mathbb{I} \times \chi$ if $\bigcup_{k \in m} k = \chi \times \mathbb{I} \times \chi$ and $k \bigcap k' = \emptyset$ for all distinct $k, k' \in m$. Finally, to avoid trivialities, we assume throughout the paper that the number of samples, denoted by n is at least 3.

3.1 Definitions

Our objective in the paper is to select the best density (the eponymous "model") from a given class of models. To set the stage, we introduce the following definitions.

Let $\lambda_n := \sum_{i=0}^{n-1} \delta_{X_i,a_i,G_i} \otimes \mu_\chi / n$ denote the product between the point measure $n^{-1} \sum_{i=0}^{n-1} \delta_{X_i,a_i,G_i}$ and μ_χ. Formally, for any set $\mathcal{X} \subset \mathbb{X}$ such that $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3 \times \mathcal{X}_4$,

$$\lambda_n(\mathcal{X}) = \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{1}[(X_i, a_i, G_i) \in \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3] \mu_\chi(\mathcal{X}_4).$$

Observe that this is a valid measure on $\mathbb{X}$. By $\mathbb{L}_1^+(\mathbb{X}, \lambda_n)$, we denote all positive, absolutely integrable functions on $\mathbb{X}$. For $f_1, f_2 \in L_1^+(\mathbb{X}, \lambda_n)$ (not necessarily densities) on $\mathbb{X}$, we define the square of the **empirical Hellinger distance** $\mathcal{H}^2$ as

$$\mathcal{H}^2(f_1, f_2) := \frac{1}{2} \int_\chi \left(\sqrt{f_1} - \sqrt{f_2} \right)^2 d\lambda_n.$$

It follows that $\mathcal{H}$ is a nonnegative random variable adapted to $\mathcal{F}_0^n$. Our model selection procedure is as follows: For some universal constant α, let $L > 0$, and $\mathcal{S} \subset \mathbb{L}_+^2(\mathbb{X}, \lambda_n)$ be a countable (but possibly infinite) class of functions. The map $\Delta_\mathcal{S} : \mathcal{S} \to \mathbb{R}$ is said to be a penalty on $\mathcal{S}$ if $\Delta_\mathcal{S}(f) > 0 \; \forall f \in \mathcal{S}$. For convenience of notation, we define

$$\varphi(x, y) := \frac{\sqrt{y} - \sqrt{x}}{2\sqrt{x + y}}$$

For any two functions $f_1, f_2 : \chi \times \mathbb{I} \times \mathcal{G} \times \chi \to \mathbb{R}$ define $T(f_1, f_2)$ as,

$$\begin{aligned} T(f_1, f_2) := \frac{1}{n} \sum_{i=0}^{n-1} & \underbrace{\int \frac{1}{\sqrt{2}} \varphi(f_1, f_2) \delta_{X_i,a_i,G_i,X_{i+1}}}_{A} \\ & + \underbrace{\int \sqrt{\frac{f_1 + f_2}{2}} \cdot (\sqrt{f_2} - \sqrt{f_1}) \, d\lambda_n}_{B} + \underbrace{\int (f_1 - f_2) \, d\lambda_n}_{C}. \end{aligned} \tag{3.1}$$

Remark 2. *T can intuitively be thought of as a comparison of f_1 and f_2, with A comparing which one fits the $\{(X_i, a_i, G_i, X_{i+1})\}$ process better, B comparing which one fits the $\{(X_i, a_i, G_i)\}$ process better, and C penalising if f_1 or f_2 is not a proper density.*

Definition 1. *ϑ is said to be a model-selection procedure on $\mathcal{S}$ if*

$$\begin{aligned} \vartheta(f) = \sup_{f' \in \mathcal{S}} & \left[\alpha \mathcal{H}^2(f, f') + T(f, f') - L \frac{\Delta_\mathcal{S}(f')}{n} \right] \\ & + L \frac{\Delta_\mathcal{S}(f)}{n}. \end{aligned} \tag{3.2}$$

Our choice of estimator shall be the following:

$$\hat{\mathfrak{z}} := \operatorname*{argmin}_{f \in \mathcal{S}} \vartheta(f) + \frac{1}{n}. \tag{3.3}$$

Remark 3. *$\hat{\mathfrak{z}}$ depends on n and L and may not be unique. In that case, all of the choices are valid estimators.*

4 Theoretical Results

Before stating our main theorem we state the following assumptions (and discuss them below)

Assumption 1. *Any $f \in \mathcal{S}$ is finite dimensional (over the $\mathbb{L}^2$ norm) with dimension $\dim(f)$. Furthermore, $\sum_{f\in\mathcal{S}} e^{-\Delta_{\mathcal{S}}(f)} \leq 1$.*

Assumption 2. *For all $i \in \{0, \dots, n-1\}$, (X_i, a_i, G_i, X_{i+1}) admits a (possibly inhomogenous) density ϕ_i with respect to some known measure $\mu_{\mathbb{X}}$ (defined formally in Section A.3) such that $\phi(\cdot,\cdot,\cdot,\cdot) \leq \kappa$ for some constant $\kappa > 0$ and $\mu_{\mathbb{X}}(\mathbb{X}) = 1$.*

The first assumption signifies that the penalty must be large enough such that $e^{-\Delta_{\mathcal{S}}(f)}$ is summable on f; the upper bound of 1 is without losing generality. We discuss this further with an example below.

The second assumption has two parts. The requirement on the density is mild since it assumes an *upper bound* on the density, as opposed to the more prevalent (and somewhat tenuous) *lower bound* that is ubiquitous adaptive learning literature (see for instance A5, Lacour (2007)).

Now we can state our main oracle risk bound.

4.1 Oracle Risk Bound for Empirical Hellinger

The proof of the following theorem can be found in Section A.2.

Theorem 1. *Under the conditions of Assumption 1, there exists an universal constant $L_0 > 0$ such that if $L \geq L_0$, the estimator $\hat{s}$ satisfies*

$$\mathcal{C}\mathbb{E}\big[\mathcal{H}^2(s, \hat{s})\big] \leq \mathbb{E}\left[\inf_{f\in\mathcal{S}} \left\{\mathcal{H}^2(s, f) + L\frac{\Delta_{\mathcal{S}}(f)}{n}\right\}\right].$$

Classical approaches such as regularised MLE, kernel density estimation (KDE), or fitted Q-iteration typically rely on strong structural assumptions, e.g. stationarity, ergodicity, or smoothness of the transition kernel. As mentioned in Section 1, these assumptions are often violated in contextual MDPs with irregular or non-stationary dynamics, rendering such methods either inconsistent or suboptimal. To the best of our knowledge, T-estimators are the only estimators capable of accommodating both non-stationarity and irregularity, and are therefore useful despite their shortcomings and we refer the readers to various literature in adaptive estimation (Massart, 2007; Baraud & Birgé, 2009; Baraud, 2011; Sart, 2014) for more details.

Before providing a sketch of the proof, we discuss some implications of the previous theorem. Observe that the statement of the previous theorem makes no assumption on the data generating process, beyond the fact that X_i is Markovian on $(X_{i-1}, a_{i-1}, G_{i-1})$. Theorem 1 can then be interpreted as follows: $\hat{s}$ is the best estimator one can obtain for *a given sample* $(X_{i-1}, a_{i-1}, G_{i-1})$, on a *given model class* $\mathcal{S}$. There are various interesting choices for $\mathcal{S}$, and some are given in Section 4.2.

Proof sketch for Theorem 1. **Step I.** We analyze the sign of

$$T(f, \hat{s}) + L\tfrac{\Delta_{\mathcal{S}}(f)}{n} - L\tfrac{\Delta_{\mathcal{S}}(\hat{s})}{n}.$$

If $T(f, \hat{s}) + L\frac{\Delta_{\mathcal{S}}(f)}{n} - L\frac{\Delta_{\mathcal{S}}(\hat{s})}{n} \geq 0$, a direct comparison yields for some universal constants α and ε

$$\alpha\mathcal{H}^2(s, \hat{s}) \;\leq\; (1+\varepsilon)\mathcal{H}^2(s, f) + \tfrac{2L}{n}\Delta_{\mathcal{S}}(f) + 22\xi.$$

Step II. In the complementary case, using a martingale Bernstein inequality, we derive a lemma on the concentration of the risk metric (Lemma 2) which ensures that, with probability at least $1 - e^{-n\xi}$,

$$\begin{aligned}(1-\varepsilon)\mathcal{H}^2(s, f') + T(f, f') - L\tfrac{\Delta_{\mathcal{S}}(f')}{n} \\ \leq (1+\varepsilon)\mathcal{H}^2(s, f) + L\tfrac{\Delta_{\mathcal{S}}(f)}{n} + 22\xi\end{aligned}$$

for all $f' \in \mathcal{S}$. This further implies where $\nu = (1-\varepsilon)/\alpha - 1$

$$\alpha\mathcal{H}^2(f, f') \;\leq\; (2+\varepsilon+\nu^{-1})\mathcal{H}^2(f, \mathfrak{s}) + \tfrac{2L}{n}\Delta_{\mathcal{S}}(f) + 22\xi + \tfrac{1}{n}.$$

Step III. Combining the bounds from the two cases with the Hellinger triangle inequality

$$\alpha\mathcal{H}^2(\mathfrak{s}, \hat{\mathfrak{s}}) \;\leq\; 2\alpha\mathcal{H}^2(\mathfrak{s}, f) + 2\alpha\mathcal{H}^2(f, \hat{\mathfrak{s}}),$$

and applying the above control on $\mathcal{H}^2(f, \hat{\mathfrak{s}})$, we obtain with probability at least $1 - e^{-n\xi}$

$$\mathcal{C}\mathcal{H}^2(\mathfrak{s}, \hat{\mathfrak{s}}) \;\leq\; \mathcal{H}^2(\mathfrak{s}, f) + L\tfrac{\Delta_{\mathcal{S}}(f)}{n} + \xi,$$

where $\mathcal{C} = \alpha/\min\{2(2+\alpha+\varepsilon+\nu^{-1}), 44\}$. Taking complementation and integrating both sides by ξ now yields the final result. □

To state the following corollary we need one further definition concerning the squared loss function.

$$d^2(f_1, f_2) := \int_{\mathbb{X}} (\sqrt{f}_1 - \sqrt{f}_2)^2 d\mu_{\mathbb{X}}.$$

The proof of the following corollary can be found in Section A.3

Corollary 1. *Let $L \geq L_0$ for some universal constant L_0 and assume that the hypothesis in Assumptions 1, and 2 holds. Then, the selected estimator as defined in 3.3 satisfies*

$$\mathcal{C}\mathbb{E}\left[\mathcal{H}^2(\mathfrak{s}, \hat{\mathfrak{s}})\right] \leq \inf_{f\in\mathcal{S}} \left\{ d^2(\sqrt{\mathfrak{s}}, f) + \frac{\Delta_{\mathcal{S}}(f) + \dim(f)\log n}{n} \right\}$$

for some large enough constant $\mathcal{C}$ depending only on κ.

4.2 Examples

At this point, it seems reasonable to ground the abstractions of the previous section into some examples. We begin with some model classes which satisfy the Assumption 1, beginning with dyadic cuts (DeVore & Yu, 1990).

Example 1 (Piecewise Constant Estimators on Dyadic Cuts)**.** *Let $\chi = \mathbb{I} = \mathcal{G} = [0,1]$. Then, we define the space of all dyadic cuts recursively. Define $\mathcal{M}_0 := \{[0,1]^4\}$. For any $\ell \in \mathbb{N}$, let $m \in \mathcal{M}_\ell$ and $k \in m$. Thus k is an element of a partition of $[0,1]^4$, so $k \subseteq \mathbb{R}^4$. Let $k_1, k_2, \ldots, k_{2^4}$ be the 2^4 sets obtained by equally dividing k along each axis. Let*

$$S(m,k) \;:=\; m \cup \{k_1, k_2, \ldots, k_{2^4}\} \setminus k.$$

Then

$$\mathcal{M}_{\ell+1} := \left\{ \bigcup_{m\in\mathcal{M}_\ell} \bigcup_{k\in m} S(m,k) \right\} \cup \mathcal{M}_\ell.$$

The class of all piecewise constant estimators on this class of dyadic partitions is defined to be $\mathcal{S} := \bigcup_{m\in\mathcal{M}} \left\{\sum_{k\in m} a_k \mathbb{1}_k : a_k > 0\right\}$ with the corresponding penalty to be defined as $\Delta_{\mathcal{S}}(f) = |m| - (\log 3)/2$, where $|m|$ is the number of constant pieces for the piecewise function f. Obviously, $\Delta_{\mathcal{S}}(f) \geq 0$. It is now a standard result from Baraud & Birgé (2009) (see Section 3) that $\sum_{f\in\mathcal{S}} e^{-\Delta_{\mathcal{S}}(f)} \leq 1$.

It is intuitively clear from the previous construction that there is no benefit to considering piecewise constant estimators with more than n distinct bins. Consequently, our optimal estimator will lie in $\mathcal{M}_\ell$ such that $\ell = o(n)$. For a formalisation and proof of this fact, we refer the readers to Proposition 3 in Banerjee et al. (2025b), or Claim B.5 in Sart (2014).

Example 2 (Splines)**.** *Let $\varphi^{(1)}, \varphi^{(2)}, \ldots,$ be the orthonormal polynomial basis with respect to the L_2 norm, and its corresponding inner product. Let constants $c_i^{(\ell)} \in \mathbb{R}$ induce a triangular array such that for each ℓ, $\sum c_i^{(\ell)} = 1$ and consider*

$$\mathcal{M}_\ell = \left\{ \sum_{i=1}^{\ell} a_i \varphi^{(i)} : a_i \in \cup_i \{c_i^{(l)}\} \right\}$$

We now $\mathcal{S} = \bigcup_\ell \mathcal{M}_\ell$ and $\Delta_{\mathcal{S}}(f) = \ell + a$ for $a = \log(\frac{e-2}{2})$ whose specifics are given in Section B.3.

The following proposition highlights that the functions in Example 2 satisfy Assumption 1. Its proof is deferred to Section B.3

Proposition 1. *Consider the class of functions given by $\mathcal{S}$ in example 2 with corresponding penalty $\Delta_{\mathcal{S}}(f)$. Then,*

$$\dim f < \infty \text{ and } \sum_{f \in \mathcal{S}} e^{-\Delta_{\mathcal{S}}(f)} \leq 1.$$

4.3 Optimality of the Risk Bound

A natural subsequent question is now the optimality of the risk bounds derived in the previous section, which is what we dedicate this section towards. To derive the optimality bounds, we first introduce some notation.

Definition 2. *We call a function $f : A \to \mathbb{R}$ to belong to the Hölder space $\mathbb{H}_\sigma(A)$ with parameter $\sigma \in (0, 1]$ and finite norm $\|f\|_\sigma > 0$ if $|f(x) - f(y)| \leq \|f\|_\sigma \|x - y\|^\sigma \; \forall x, y \in A$.*

Recall that $\mathbb{H}^1(A)$ is the space of all Lipschitz smooth functions, and that elements of $\mathbb{H}^\sigma(A)$ are constant functions when $\sigma > 1$.

Definition 3. *Given a function $f \in L_p(\mathbb{X}), 0 < p \leq \infty$, and any integer r, we defne its modulus of smoothness of order r as*

$$\omega_r(f, t)_p := \sup_{0<|h|\leq t} \|\Delta_h^r(f, \cdot)\|_{L_p(\Omega)}, \quad t > 0,$$

where $h \in \mathbb{R}^d$ and $|h|$ is it Euclidean norm. Here, Δ_h^r, is the r-th difference operator, defined by

$$\Delta_h^r(f, x) := \sum_{k=0}^{r} (-1)^{r-k} \binom{r}{k} f(x + kh), \quad x \in \Omega \subset \mathbb{R}^d,$$

where this difference is set to zero whenever one of the points $x + kh$ is not in the support of f. It is easy to see that for any $f \in L_p(\Omega)$, we have $\omega_{\boldsymbol{r}}(f, t)_p \to 0$, Then, for any $\sigma \in (0, 1)$, Besov space $\mathbb{B}_q^\sigma(L_p(A))$ consists of all f such functions such that

$$|f|_{\mathbb{B}_q^\sigma(L_p(A))} := \begin{cases} \int_{t>0} t^{q\sigma-1} (\omega_r(f, t)_p)^q dt & 0 < q < \infty \\ \sup_{t\geq 0} t^{q\sigma-1} (\omega_r(f, t)_p)^q & q = \infty \end{cases}$$

is finite. Then, we define $\mathbb{B}^\sigma(L_p(A))$ as

$$\mathbb{B}^\sigma(L_p(A)) := \begin{cases} \mathbb{B}_p^\sigma(L_p(A)), & p \in (1, 2) \\ \mathbb{B}_\infty^\sigma(L_p(A)), & p \geq 2 \end{cases}$$

with the attached norm $\|\cdot\|_{\sigma,p}$.

Remark 4. *With $\sigma \in (0, 1)$, we restrict ourselves to isotropic Besov spaces.*

Without losing generality, let $\mathbb{X} = [0, 1]^4$. The following corollary (which follows similarly to Corollary 4.3 in Sart) establishes the optimality of $\hat{\jmath}$.

Corollary 2. *Consider the class of models in Example 1, and the Assumptions 1, and 2. Then, if $\sqrt{\jmath} \in \mathbb{H}^\sigma([0, 1]^4)$ (or $\sqrt{\jmath} \in \mathbb{B}^\sigma(L_p([0, 1]^4))$), there exists a constant $\mathcal{C}$ depending only on κ, p, and $\|\sqrt{\jmath}\|_\sigma$, (respectively $\|\sqrt{\jmath}\|_{\sigma,p}$) such that*

$$\mathcal{C}\mathbb{E}[\mathcal{H}^2(\jmath, \hat{\jmath})] \leq \left(\frac{\log n}{n}\right)^{\frac{\tilde{\sigma}}{4+\tilde{\sigma}}}$$

where $\sigma > 4(1/p - 1/2)\mathbb{1}[(1/p - 1/2) > 0]$.

5 Optimal Control Determination

One objective in obtaining the estimates of the transition probabilities is to learn the optimal control for MDP's bearing a cost for state transitions, where the optimality is considered with respect to this cost of transition between states given a control and a context, $L : \mathbb{X} \to \mathbb{R}_+$. Therefore, given a context G, we wish to solve the optimization problem

$$a^*(G) = \operatorname*{argmin}_{a \in \mathbb{I}} C_n(a, G) \tag{5.1}$$

where $C_n(a, G)$ is as defined (5.5).

Note that $a^*(G)$ can be a set of controls and hence a single optimal control might not be best and in such a scenario any control in these sets are equivalent for our purposes. In the following paragraphs, we propose a method for identifying the set of optimal controls and provide non-asymptotic guarantees for the same.

Before we proceed we make the following assumption.

Assumption 3. *This assumptions consists of two statements*

1. *We assume that the true density $\mathfrak{s} \in \mathcal{S}$ and that the cost function L is positive and continuous over all its arguments.*
2. *We assume that both $\mathfrak{s}$ and $\hat{\mathfrak{s}}$ are such that there exists $\varepsilon > 0$ with*

$$\int \mathfrak{s}\, d\lambda_n \leq (1+\varepsilon) \text{ and } \int \hat{\mathfrak{s}}\, d\lambda_n \leq (1+\varepsilon).$$

Note that this holds with $\varepsilon = 0$, when $\mathfrak{s}$, $\hat{\mathfrak{s}}$ are proper densities.

We define, for any $0 < \delta < 1/2$, the following quantity-

$$U(G, a, n) = \hat{C}_n(a,\, G) + \left(\frac{r_S(f, n)}{n}\right)^{\delta} \tag{5.2}$$

as a large sample upper confidence bound for the average cost, where $r_S(f, n) = \max(\Delta_S(f), \log n)$ (the intuition for this can be found in Corollary 1) and further define the approximate optimal control as the control which solves the optimization problem

$$\hat{a}^*(G) = \operatorname*{argmin}_{a \in \mathbb{I}} U(G, a, n). \tag{5.3}$$

We shall exhibit that under certain simple assumptions on the cost function, the upper confidence bound as defined in (5.2) preserves the same order as the actual loss over the set of controls and hence solves the problem. Also note that our analysis primarily considers $r_S(f, n) = o(n)$ and we provide an example in Example 1 where this indeed holds true, making it a very natural assumption. To that end we introduce some notation. Recall that,

$$\lambda_n := \frac{1}{n} \sum_{i=0}^{n-1} \delta_{X_i, a_i, G_i} \otimes \mu_\chi,$$

and let

$$\lambda_n^{(2)} = \int_{\chi \times \chi} \lambda_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{a_i, G_i} \text{ and } \lambda_n^{(1)} = \lambda_n / \lambda_n^{(2)}. \tag{5.4}$$

Intuitively, $\lambda_n^{(2)}$ is the empirical distribution of the context–action pairs (a_i, G_i) observed in the dataset, while $\lambda_n^{(1)}$ represents the corresponding conditional distribution of state transitions given a particular context–action pair. In other words, $\lambda_n^{(1)}$ captures how states evolve conditional on (a_i, G_i), and $\lambda_n^{(2)}$ records how often each (a_i, G_i) is observed. This decomposition allows us to express the cost functional as an empirical expectation under the observed data distribution. We next define the cost function and its estimated counterpart.

Next, as before, $L : \mathbb{X} \to \mathbb{R}_+$ denotes the cost function and $C_n(l, G)$ denotes the cost of control l for context G annealed over $\lambda_n^{(2)}$. Formally,

$$C_n(l, g) = \int L(x, l, g, y)\mathfrak{s}(x, l, g, y)\lambda_n^{(1)}(dx, dy). \tag{5.5}$$

This is the *empirical* cost of allocating control l for context G. We denote the estimated counterpart of C with $\hat{C}$, i.e.

$$\hat{C}_n(l, g) = \int L(x, l, g, y)\hat{\mathfrak{s}}(x, l, g, y)\lambda_n^{(1)}(dx, dy). \tag{5.6}$$

Theorem 2. *Let Assumptions 1, 2 and 3 hold. Then*

$$\begin{aligned}
&\mathbb{E}\left[\int \left(C(\hat{a}_n, G) - C(a, G)\right) d\lambda_n^{(2)}\right] \\
&= O\left(\|C\|_\infty \min\left(\left(\frac{r_S(f,n)}{n}\right)^{1/2-\delta}, \left(\frac{r_S(f,n)}{n}\right)^{\delta}, \right.\right. \\
&\qquad \left.\left.\left(\frac{r_S(f,n)}{n}\right)^{1/4}\right)\right)
\end{aligned}$$

Remark 5. *Note that choosing $\delta = 1/4$ gives the fastest convergence rate at $(n^{-1}\log n)^{1/4}$.*

Theorem 2 implies that upon using our scheme, on average, we select the optimal treatment most realizations, with the probability of the sub-optimal control(s) choice diminishing to 0 with increasing sample size. In the following paragraphs, we outline the key steps in the proof.

Proof sketch of Theorem 2. **Step I.** In the first step, we use the polarizing identity and Markov inequality to derive concentration bounds on the so called "bad" event that the true cost estimate is greater than its upper confidence bound as expressed in (5.2). That is we show that

$$\begin{aligned}
&\mathbb{E}\left[\int \mathbf{1}\left\{C_n(a, G) > U(G, a, n)\right\} d\lambda_n^{(2)}\right] \\
&= O((r_S(f,n)/n)^{1/2-\delta}).
\end{aligned}$$

Step II. Next we show that if two controls are such that one is optimal and the other is suboptimal and further if the optimal control is identifiable via a gap condition, then with high probability we retain the correct ordering between the upper confidence bound estimates of the control values. In essence,

$$\begin{aligned}
&\mathbb{E}\left[\int \mathbf{1}\left\{U(G, a, n) > U(G, a', n) \mid \right.\right. \\
&\qquad \left.\left. C_n(a, G) < C_n(a', G)\right\} d\lambda_n^{(2)}\right] \\
&= O\left(\left(\frac{r_S(f,n)}{n}\right)^{1/2-\delta} \frac{1}{\Delta_L}\right).
\end{aligned}$$

Step III. In the following step, we exhibit that for any pair of controls, one belonging to the true set of optimal controls 5.1 and the other belonging to the approximate set of optimal controls 5.3, then the expected L_1-norm between the two costs is diminishing with sample size. That is

$$\begin{aligned}
&\mathbb{E}\left[\int C(\hat{a}_n, G) - C(a, G)\, d\lambda^{(2)}\right] \\
&= O\left(\|C\|_\infty \left(\frac{r_S(f,n)}{n}\right)^{1/2-\delta} \frac{1}{\Delta_L}\right)
\end{aligned}$$

with Δ_L being the gap between the optimal and the suboptimal controls with respect to cost.
Step IV. Finally, following a technique similar to Bhattacharya et al. (2024), we show a gap independent upper bound for the L_1 norm between the costs by dividing the problem into settings- one where there is a gap (dependent on the sample size) and one where there is not one. That completes the proof. □

6 Conclusions

We propose a general estimator for the transition functions of contextual MDP's on continuous states spaces and exhibit finite sample oracle bounds in the randomised Hellinger and deterministic L_2 metrics. In addition, we demonstrate the effectiveness of the estimator by theoretically showing deriving cost optimality for the corresponding empirical cost. This gives a general framework towards deriving an optimal control with minimal assumptions on the data generating process; thereby acquitting contextual MDP's from its usual flaws of non-stationarity and irregularity. This method is robust and is independent of structure of the MDP involved, and should therefore be considered as an extremely effective tool for inferential purposes on MDPs and should be readily useful in optimal policy determination in offline setting-a rudimentary version of which is the optimal control determination as presented in our work.

We also introduce the problem of optimal cost determination along with the optimal control which has applications in multiple fields like healthcare, economics, etc. Again future directions in these areas lie in the inferential aspects of cost functions and optimal policy determination.

Limitations and future outlooks The main bottleneck of T-estimators developed in the paper is computation. In fact, it is known for the i.i.d. case, the objective function in eq. (3.2) can be computed in $O(ne^d)$ time for dimension d (see Section 3.2.4 Baraud & Birgé (2009) for more details). However, to the best of our knowledge, T-estimators are the only estimators capable of accommodating both non-stationarity and irregularity, and is therefore useful despite its shortcomings.

For future works, given $\mathfrak{s}$ and $\hat{\mathfrak{s}}$, we can do policy optimisation using approximate Bellman operators. This has been previously used in the offline RL setting to great effect (Li et al., 2022a;c; Banerjee et al., 2025a), but the contextual question remains open. Furthermore, the question of additional structures on the cost function, like those determined by transport maps between policies (as described in Section 1) lay unanswered, and we plan to explore this question in a future work.

References

Imon Banerjee, Harsha Honnappa, and Vinayak Rao. Off-line Estimation of Controlled Markov Chains: Minimaxity and Sample Complexity. *Operations Research*, February 2025a. ISSN 0030-364X. doi: 10.1287/opre.2023.0046. URL https://pubsonline.informs.org/doi/abs/10.1287/opre.2023.0046.

Imon Banerjee, Vinayak Rao, and Harsha Honnappa. Adaptive Estimation of the Transition Density of Controlled Markov Chains, May 2025b. URL http://arxiv.org/abs/2505.14458. arXiv:2505.14458 [math].

Y. Baraud, L. Birgé, and M. Sart. A new method for estimation and model selection:$$\rho $$-estimation. *Inventiones mathematicae*, 207(2):425–517, February 2017. ISSN 1432-1297. doi: 10.1007/s00222-016-0673-5. URL https://doi.org/10.1007/s00222-016-0673-5.

Yannick Baraud. Estimator selection with respect to Hellinger-type risks. *Probability Theory and Related Fields*, 151 (1):353–401, October 2011. ISSN 1432-2064. doi: 10.1007/s00440-010-0302-y. URL https://doi.org/10.1007/s00440-010-0302-y.

Yannick Baraud and Lucien Birgé. Estimating the intensity of a random measure by histogram type estimators. *Probability Theory and Related Fields*, 143(1):239–284, January 2009. ISSN 1432-2064. doi: 10.1007/s00440-007-0126-6. URL https://doi.org/10.1007/s00440-007-0126-6.

Yannick Baraud and Lucien Birgé. Rho-estimators revisited: General theory and applications. *The Annals of Statistics*, 46(6B):3767–3804, December 2018. ISSN 0090-5364, 2168-8966. doi: 10.1214/17-AOS1675. URL https://projecteuclid.org/journals/annals-of-statistics/volume-46/issue-6B/Rho-estimators-revisited-General-theory-and-applications/10.1214/17-AOS1675.full.

Yannick Baraud and Lucien Birgé. Robust Bayes-like estimation: Rho-Bayes estimation. *The Annals of Statistics*, 48(6):3699–3720, December 2020. ISSN 0090-5364, 2168-8966. doi: 10.1214/20-AOS1948. URL https://projecteuclid.org/journals/annals-of-statistics/volume-48/issue-6/Robust-Bayes-like-estimation-Rho-Bayes-estimation/10.1214/20-AOS1948.full. Publisher: Institute of Mathematical Statistics.

Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113(3):301–413, February 1999. ISSN 1432-2064. doi: 10.1007/s004400050210. URL https://doi.org/10.1007/s004400050210.

Alex Batsis and Spyridon Samothrakis. Contextual reinforcement learning for supply chain management. *Expert Systems with Applications*, 249:123541, September 2024. ISSN 0957-4174. doi: 10.1016/j.eswa.2024.123541. URL https://www.sciencedirect.com/science/article/pii/S0957417424004068.

Stav Belogolovsky, Philip Korsunsky, Shie Mannor, Chen Tessler, and Tom Zahavy. Inverse reinforcement learning in contextual mdps. *Machine Learning*, 110(9):2295–2334, 2021.

Jöran Bergh and Jörgen Löfström. *Interpolation Spaces: An Introduction*, volume 223 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 1976. ISBN 978-3-642-66453-3 978-3-642-66451-9. doi: 10.1007/978-3-642-66451-9. URL http://link.springer.com/10.1007/978-3-642-66451-9.

Riddhiman Bhattacharya, Thanh Nguyen, Will Wei Sun, and Mohit Tawarmalani. Active learning for fair and stable online allocations. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pp. 196–197, 2024.

Lucien Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. In *Annales de l'IHP Probabilités et statistiques*, volume 42, pp. 273–325, 2006.

Lucien Birgé and Pascal Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, September 1998. ISSN 1350-7265. URL https://projecteuclid.org/journals/bernoulli/volume-4/issue-3/Minimum-contrast-estimators-on-sieves--exponential-bounds-and-rates/bj/1174324984.full. Publisher: Bernoulli Society for Mathematical Statistics and Probability.

Junyu Cao, Esmaeil Keyvanshokooh, and Tian Liu. Safe reinforcement learning with contextual information: Theory and applications. *Available at SSRN 4583667*, 2023.

Ronald A. DeVore and Xiang Ming Yu. Degree of Adaptive Approximation. *Mathematics of Computation*, 55 (192):625–635, 1990. ISSN 0025-5718. doi: 10.2307/2008437. URL https://www.jstor.org/stable/2008437. Publisher: American Mathematical Society.

Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual Markov Decision Processes, February 2015. URL http://arxiv.org/abs/1502.02259. arXiv:1502.02259 [stat].

Adel Javanmard and Hamid Nazerzadeh. Dynamic pricing in high-dimensions. *Journal of Machine Learning Research*, 20(9):1–49, 2019.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2017.

Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Contextual-mdps for pacreinforcement learning with rich observations. *arXiv preprint arXiv:1602.02722*, 2016.

Claire Lacour. Adaptive estimation of the transition density of a Markov chain. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 43(5):571–597, September 2007. ISSN 02460203. doi: 10.1016/j.anihpb.2006.09.003. URL http://arxiv.org/abs/math/0611680. arXiv:math/0611680.

Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning. *arXiv preprint arXiv:2204.05275*, 2022a.

Yuantong Li, Chi-hua Wang, Guang Cheng, and Will Wei Sun. Dynamic matching bandit for two-sided online markets. *arXiv preprint arXiv:2205.03699*, 2022b.

Yuanzhi Li, Ruosong Wang, and Lin F Yang. Settling the horizon-dependence of sample complexity in reinforcement learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 965–976. IEEE, 2022c.

Pascal Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, Heidelberg, 2007. ISBN 978-3-540-48497-4. doi: 10.1007/978-3-540-48503-2. URL http://link.springer.com/10.1007/978-3-540-48503-2.

Mathieu Sart. Estimation of the transition density of a Markov chain. In *Annales de l'IHP Probabilités et statistiques*, volume 50, pp. 1028–1068, 2014. Issue: 3.

Mathieu Sart. Estimating the conditional density by histogram type estimators and model selection. *ESAIM: Probability and Statistics*, 21:34–55, 2017. ISSN 1292-8100, 1262-3318. doi: 10.1051/ps/2016026. URL https://www.esaim-ps.org/articles/ps/abs/2017/01/ps160026/ps160026.html. Publisher: EDP Sciences.

Mathieu Sart. Density estimation under local differential privacy and Hellinger loss. *Bernoulli*, 29(3):2318–2341, August 2023. ISSN 1350-7265. doi: 10.3150/22-BEJ1543. URL https://projecteuclid.org/journals/bernoulli/volume-29/issue-3/Density-estimation-under-local-differential-privacy-and-Hellinger-loss/10.3150/22-BEJ1543.full. Publisher: Bernoulli Society for Mathematical Statistics and Probability.

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pp. 2898–2933. PMLR, 2019.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Shengpu Tang and Jenna Wiens. Model Selection for Offline Reinforcement Learning: Practical Considerations for Healthcare Settings. In *Proceedings of the 6th Machine Learning for Healthcare Conference*, pp. 2–35. PMLR, October 2021. URL https://proceedings.mlr.press/v149/tang21a.html. ISSN: 2640-3498.

Shuai Xiao, Le Guo, Zaifan Jiang, Lei Lv, Yuanbo Chen, Jun Zhu, and Shuang Yang. Model-based Constrained MDP for Budget Allocation in Sequential Incentive Marketing. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, pp. 971–980, New York, NY, USA, November 2019. Association for Computing Machinery. ISBN 978-1-4503-6976-3. doi: 10.1145/3357384.3358031. URL https://doi.org/10.1145/3357384.3358031.

Doudou Zhou, Yufeng Zhang, Aaron Sonabend-W, Zhaoran Wang, Junwei Lu, and Tianxi Cai. Federated offline reinforcement learning. *Journal of the American Statistical Association*, 119(548):3152–3163, 2024.

A Proofs

A.1 Proof of Theorem 2

Proof. The proof of Theorem 2 is contingent on the proofs of Propositions 7, 8 and 9 which are deferred to Section B.2. Note that the proof follows via the following. For any $\Delta(n) > (\log nn^{-1})^{\delta}$ one has

$$\mathbb{E}\left[\int C(\hat{a}_n, G) - C(a, G) d\lambda^{(2)}\right]$$
$$\mathbb{E}\left[\int (C(\hat{a}_n, G) - C(a, G))\, \mathbf{1}\left\{(C(\hat{a}_n, G) - C(a, G)) > \Delta(n)\right\} d\lambda^{(2)}\right]$$
$$+ \mathbb{E}\left[\int (C(\hat{a}_n, G) - C(a, G))\, \mathbf{1}\left\{(C(\hat{a}_n, G) - C(a, G)) \leq \Delta(n)\right\} d\lambda^{(2)}\right].$$

Noting that the second term is less than $\Delta(n)$ and the first term is upper bounded by

$$C_1 \left(\frac{\log n}{n}\right)^{1/2} \frac{1}{\left(\Delta(n) - \left(\frac{\log n}{n}\right)^{\delta}\right)} + 2\, C_2 \left(\frac{\log n}{n}\right)^{1/2-\delta}$$

with constants $C_1,\, C_2$ as per Proposition 9. Therefore optimizing over $\Delta(n)$, in the domain of the constraint, one has

$$\Delta(n) = \left(\frac{\log n}{n}\right)^{\delta} + C_1^{1/2} \left(\frac{\log n}{n}\right)^{1/4}$$

as the optimal. Therefore plugging in the value of $\Delta(n)$, we get the rate and complete the proof. □

A.2 Proof of Theorem 1

This section is dedicated towards the proof of Theorem 1. Our first objective will be to prove that with probability at least $1 - e^{-n\xi}$

$$\mathcal{C}\mathcal{H}^2(\mathscr{s}, \hat{\mathscr{s}}) \leq \mathcal{H}^2(\mathscr{s}, f) + L\frac{\Delta_{\mathcal{S}}(f)}{n} + \xi. \tag{A.1}$$

To that end, we consider two cases.

Case I $\left(T(f, \hat{\mathscr{s}}) + L\frac{\Delta_{\mathcal{S}}(f)}{n} - L\frac{\Delta_{\mathcal{S}}(\hat{\mathscr{s}})}{n} \geq 0\right)$: Since $\alpha < (1 - \varepsilon)$, it follows under this case that,

$$\begin{aligned}\alpha\mathcal{H}^2(\mathscr{s}, \hat{\mathscr{s}}) &\leq (1-\varepsilon)\mathcal{H}^2(\mathscr{s}, \hat{\mathscr{s}}) + T(f, \hat{\mathscr{s}}) - L\frac{\Delta_{\mathcal{S}}(\hat{\mathscr{s}})}{n} + L\frac{\Delta_{\mathcal{S}}(f)}{n} \\ &\leq (1+\varepsilon)\mathcal{H}^2(\mathscr{s}, f) + 2L\frac{\Delta_{\mathcal{S}}(f)}{n} + 22\xi.\end{aligned} \tag{A.2}$$

Case II $\left(T(f, \hat{\mathscr{s}}) + L\frac{\Delta_{\mathcal{S}}(f)}{n} - L\frac{\Delta_{\mathcal{S}}(\hat{\mathscr{s}})}{n} \leq 0\right)$: To analyse this case we require the following lemma which is (by now) a standard fare in this literature, and is proved for completeness in Section B.1.

Lemma 2. *Set $\varepsilon = (2 + 3\sqrt{2})/8$. Under assumptions of Theorem 1, there exists a universal constant $L_0 > 0$ such that for all $L \geq L_0$ and $\xi > 0$,*

$$\forall f, f' \in \mathcal{S}, \quad (1-\varepsilon)\mathcal{H}^2(\mathscr{s}, f') + T(f, f') - L\frac{\Delta_{\mathcal{S}}(f')}{n} \leq (1+\varepsilon)\mathcal{H}^2(\mathscr{s}, f) + L\frac{\Delta_{\mathcal{S}}(f)}{n} + 22\xi$$

with probability larger than $1 - e^{-n\xi}$.

By using the above lemma, with probability larger than $1 - e^{-n\xi}$, for all $f \in \mathcal{S}$,

$$\sup_{f' \in \mathcal{S}} \left\{(1-\varepsilon)\mathcal{H}^2(\mathscr{s}, f') + T(f, f') - L\frac{\Delta_{\mathcal{S}}(f')}{n}\right\} \leq (1+\varepsilon)\mathcal{H}^2(\mathscr{s}, f) + L\frac{\Delta_{\mathcal{S}}(f)}{n} + 22\xi.$$

Therefore,

$$\begin{aligned}
\alpha\mathcal{H}^2(\mathscr{s},\hat{\mathscr{s}}) &\leq \alpha\mathcal{H}^2(\mathscr{s},\hat{\mathscr{s}}) + T(\mathscr{s},\hat{\mathscr{s}}) - L\frac{\Delta_{\mathcal{S}}(f)}{n} + L\frac{\Delta_{\mathcal{S}}(\hat{\mathscr{s}})}{n} \\
&\leq \sup_{f'\in\mathcal{S}}\left\{\alpha\mathcal{H}^2(\hat{\mathscr{s}},f') + T(\hat{\mathscr{s}},f') - L\frac{\Delta_{\mathcal{S}}(f')}{n}\right\} + L\frac{\Delta_{\mathcal{S}}(\hat{\mathscr{s}})}{n} \\
&\overset{(i)}{\leq} \vartheta(\hat{\mathscr{s}}) \\
&\overset{(ii)}{\leq} \vartheta(f) + \frac{1}{n} \\
&\leq \sup_{f'\in\mathcal{S}}\left\{\alpha\mathcal{H}^2(f,f') + T(f,f') - L\frac{\Delta_{\mathcal{S}}(f')}{n}\right\} + L\frac{\Delta_{\mathcal{S}}(f)}{n} + \frac{1}{n}
\end{aligned}$$

where (i) follows from eq. (3.2), and (ii) follows from eq. (3.3). Now observe that, with $\nu = (1-\varepsilon)/\alpha - 1 > 0$, for any pair $f, f' \in \mathcal{S}$ with probability at least $1 - e^{-n\xi}$

$$\begin{aligned}
\alpha\mathcal{H}^2(f,f') &\leq (1+\nu^{-1})\,\mathcal{H}^2(f,\mathscr{s}) + \sup_{f'\in\mathcal{S}}\left\{(1-\varepsilon)\mathcal{H}^2(\mathscr{s},f') + T(f,f') - L\frac{\Delta_{\mathcal{S}}(f')}{n}\right\} + L\frac{\Delta_{\mathcal{S}}(f)}{n} + \frac{1}{n} \\
&\overset{(i)}{\leq} (1+\nu^{-1})\,\mathcal{H}^2(f,\mathscr{s}) + \left[(1+\varepsilon)\mathcal{H}^2(\mathscr{s},f) + L\frac{\Delta_{\mathcal{S}}(f)}{n} + 22\xi\right] + L\frac{\Delta_{\mathcal{S}}(f)}{n} + \frac{1}{n} \\
&\leq (2+\varepsilon+\nu^{-1})\,\mathcal{H}^2(f,\mathscr{s}) + 2L\frac{\Delta_{\mathcal{S}}(f)}{n} + 22\xi + \frac{1}{n}.
\end{aligned}$$

with (i) follows from Lemma 2 with probability at least $1 - e^{-n\xi}$.

This leads to

$$\begin{aligned}
\alpha\mathcal{H}^2(\mathscr{s},\hat{\mathscr{s}}) &\leq 2\alpha\mathcal{H}^2(\mathscr{s},f) + 2\alpha\mathcal{H}^2(f,\hat{\mathscr{s}}) \\
&\leq 2\big(2+\alpha+\varepsilon+\nu^{-1}\big)\mathcal{H}^2(f,\mathscr{s}) + 4L\frac{\Delta_{\mathcal{S}}(f)}{n} + 44\xi + \frac{2}{n}.
\end{aligned}$$

Recall that $L > 1$ and the penalty satisfies $\Delta_{\mathcal{S}}(f) \geq 1\ \forall\ f \in \mathcal{S}$. Therefore, with $\alpha/\min\left\{2(2+\alpha+\varepsilon+\nu^{-1}), 44\right\}$ to be $\mathcal{C}$ we have, with probability at least $1 - e^{-n\xi}$, for all $f \in S$,

$$\mathcal{C}\,\mathcal{H}^2(\mathscr{s},\hat{\mathscr{s}}) \ \leq\ \mathcal{H}^2(f,\mathscr{s}) + L\frac{\Delta_{\mathcal{S}}(f)}{n} + \xi.$$

Integrating both sides by ξ completes the proof of Theorem 1

A.3 Proof of Corollary 1

Proof. Recall from the introduction that $\mathbb{X} = \chi \times \mathbb{I} \times \mathcal{G} \times \chi$ and let $\mu_{\mathbb{X}}$ be a measure on $\mathbb{X}$ formed by taking the canonical products of $\mu_\chi, \mu_{\mathbb{I}}, \mu_{\mathcal{G}}$ etc. Let $L^2(\mathbb{X}, \mu_{\mathbb{X}})$, be the space of all square–integrable functions on $\mathbb{X}$ with respect to the product measure $\mu_{\mathbb{X}}$ according to the natural L_2 metric

$$d^2(f,f') = \int_A \left(f - f'\right)^2 d\mu_{\mathbb{X}}.$$

Observe that the space of all models $\mathcal{S} \subset L^2(\mathbb{X}, \mu_{\mathbb{X}})$. We have assumed that $\mathcal{S}$ has a finite dimension, and thus admits a canonical basis expansion. Let that orthonormal basis be given by $f_1^{(b)}, f_2^{(b)}, \ldots$. Observe that, there is no restriction on $L^2(\mathbb{X}, \mu_{\mathbb{X}})$ to be finite dimensional.

Observe that the space of all models $\mathcal{S} \subset L^2(\mathbb{X}, \mu_{\mathbb{X}})$. We have assumed that $\mathcal{S}$ has a finite dimension, and thus admits a canonical basis expansion. Let that orthonormal basis be given by $f_1^{(b)}, f_2^{(b)}, \ldots$. Observe that, there is no restriction on $L^2(\mathbb{X}, \mu_{\mathbb{X}})$ to be finite dimensional.

We require the following approximation Lemma. A version of this Lemma appears in Sart (2014), who refer to Lemma 5 in Birgé (2006) for its proof, whose proof, in turn, points to Lemma 2 in Birgé & Massart (1998), which is where we also refer the reader for proof.

Lemma 3. *For any $f \in \mathcal{S}$, let the basis expansion of f be given by $\{f_1^{(b)}, f_2^{(b)}, \ldots, f_{\dim(f)}^{(b)}\}$ and let*

$$T_f = \left\{ \sum_{i=1}^{\dim f} \alpha_i f_i : \alpha_i \in \frac{2j}{\sqrt{n \dim f}}, j \in \mathbb{Z} \right\}, \quad \mathcal{M}_f = \left\{ g^2 : g \in T_f, d(g,0) \leq 2 \right\}.$$

Then, the cardinality of $\mathcal{M}_f$, can be bounded as $|\mathcal{M}_f| \leq (30n)^{\dim(f)/2}$.

Now let, $\mathcal{M} = \bigcup_g \mathcal{M}_g$, and let $\Delta_{\mathcal{M}}(f) = \inf_{g \in \mathcal{M}_f} (\Delta_{\mathcal{S}}(g) + 4 \dim g \log(n))$. Observe that, by an application of Theorem 1 using the penalty $\Delta_{\mathcal{M}}$ and the class of functions to be $\mathcal{M}$, we have :

$$\mathcal{C}\mathbb{E}\left[\mathcal{H}^2(\mathfrak{s}, \hat{\mathfrak{s}})\right] \leq \mathbb{E}\left[\inf_{f \in \mathcal{M}} \left\{ \mathcal{H}^2(\mathfrak{s}, f) + L \frac{\Delta_{\mathcal{M}}(f)}{n} \right\} \right].$$

Now, an application of Fatou's lemma for exchanging the infimum and the expectation, along with Assumption 2, we get that

$$\mathbb{E}\left[\inf_{f \in \mathcal{M}} \left\{ \mathcal{H}^2(\mathfrak{s}, f) + L \frac{\Delta_{\mathcal{M}}(f)}{n} \right\} \right] \leq \kappa \inf_{f \in \mathcal{M}} \left[d^2(\sqrt{\mathfrak{s}}, \sqrt{f}) + L \frac{\Delta_{\mathcal{M}}(f)}{n} \right] \tag{A.3}$$

At this point we observe that $d^2(\sqrt{\mathfrak{s}}, 0) = 1$ (by the virtue of $\mathfrak{s}$ being a density). Using the triangle inequality $2d^2(\sqrt{\mathfrak{s}}, 0) + 2d^2(\sqrt{f}, 0) \geq d^2(\sqrt{\mathfrak{s}}, \sqrt{f})$, it is therefore sufficient to consider f such that $d^2(\sqrt{f}, 0) \leq 2$, since otherwise one can always find a different f such that $d^2(f, \sqrt{\mathfrak{s}})$ is less than 4. Therefore, one can rewrite the right hand side of the equation in eq. (A.3) as

$$\begin{aligned} \kappa \inf_{f \in \mathcal{M}} \left[d^2(\sqrt{\mathfrak{s}}, \sqrt{f}) + L \frac{\Delta_{\mathcal{M}}(f)}{n} \right] &= \kappa \inf_{\substack{f \in \mathcal{S} \\ g \in T_f}} \left[d^2(\sqrt{\mathfrak{s}}, g) + L \frac{\Delta_{\mathcal{M}}(f)}{n} \right] \\ &\leq \kappa \inf_{f \in \mathcal{S}} \left[\inf_{g \in T_f} d^2(\sqrt{\mathfrak{s}}, g) + L \frac{\Delta_{\mathcal{S}}(g) + 4 \dim g \log(n)}{n} \right] \end{aligned}$$

Now, by construction, for all $g \in \mathcal{S}$ one can find $f \in T_f$ such that $d^2(g, f) \leq 1/n$. Therefore,

$$\inf_{g \in T_f} d^2(\sqrt{\mathfrak{s}}, g) \leq \inf_{g \in \mathcal{S}} d^2(\sqrt{\mathfrak{s}}, g) + \frac{1}{n}.$$

This completes the proof.

□

B Proofs of Auxillary Results

B.1 Proof of Lemma 2

Proof. For notational clarity, we introduce two intermediate objects, $\psi(c_1, c_2)$ and $\bar{f}$, defined by

$$\begin{aligned} \psi(c_1, c_2) &:= \frac{1}{\sqrt{2}} \frac{\sqrt{c_2} - \sqrt{c_1}}{\sqrt{c_2 + c_1}} \\ \bar{f}(x, l, y) &:= \frac{f_1(x, l, y) + f_2(x, l, y)}{2}. \end{aligned} \tag{B.1}$$

Next, for two functions f_1 and f_2, define Z_i by

$$Z_i(f_1, f_2) := \psi\left(f_1(X_i, a_i, X_{i+1}), f_2(X_i, a_i, X_{i+1})\right) - \mathbb{E}[\psi\left(f_1(X_i, a_i, X_{i+1}), f_2(X_i, a_i, X_{i+1})\right) \mid X_i, a_i]. \tag{B.2}$$

We need the following lemma, whose proof is provided in Section B.1.1:

Lemma 4.

$$\int \psi(f_1, f_2)^2 \mathscr{s}\, d\lambda_n \;\leq\; 3\left[\mathscr{H}^2\left(\mathscr{s}, f_2\right) + \mathscr{H}^2\left(\mathscr{s}, f_1\right)\right].$$

We also state the following lemma, proved by algebraic manipulations in Section B.1.2:

Lemma 5. *Recall from eq. (B.1) that* $\varphi(c_1, c_2) = (\sqrt{c_2} - \sqrt{c_1})/\sqrt{2(c_1 + c_2)}$. *Then*

$$\left(1 - \tfrac{1}{\sqrt{2}}\right)\mathscr{H}^2\left(\mathscr{s}, f_2\right) + T\left(f_1, f_2\right) \;\leq\; \left(1 + \tfrac{1}{\sqrt{2}}\right)\mathscr{H}^2\left(\mathscr{s}, f_1\right) + \frac{1}{n}\sum_{i=0}^{n-1} Z_i\left(f_1, f_2\right).$$

To proceed with the proof, adopt from Sart (2014) (see also (Massart, 2007, Chapter 2)) the following iteration of Bernstein's inequality. As before, let $\{\mathcal{F}_0^i\}_{i\geq 0}$ be a filtration and $|g_i| \leq b$ be a bounded random variable adapted to it. Then we have the following lemma.

Lemma 6. *Define the sum* $\mathscr{s}_n := \sum_{i=0}^n \left(g_i - \mathbb{E}[g_i|\mathcal{F}_0^i]\right)$ *and* $V_n := \sum_{i=0}^n \mathbb{E}[g_i^2|\mathcal{F}_0^i]$. *Then*

$$\mathbb{P}\left(\mathscr{s}_n \geq \frac{V_n}{2(\kappa - b)} + x\kappa\right) \leq \exp\left(-x\right) \tag{B.3}$$

for all $\kappa > b$, *and* $x > 0$.

Using Z_i as in eq. (B.2), set $\mathcal{Z}_n = \sum_{i=0}^{n-1} Z_i$ and

$$g_i = \psi\left(f_1(X_i, a_i, X_{i+1}),\; f_2(X_i, a_i, X_{i+1})\right).$$

Then, Lemma 6 gives us

$$\mathbb{P}\left(\mathscr{s}_n \geq \frac{V_n}{2(\kappa - b)} + x\kappa\right) \leq \exp(-x). \tag{B.4}$$

By rearrangement we reduce V_n to $n\int \psi\left(f_1, f_2\right)^2 \mathscr{s}\, d\lambda_n$. Lemma 4 then bounds $\int \psi\left(f_1, f_2\right)^2 \mathscr{s}\, d\lambda_n$ by

$$\int \psi\left(f_1, f_2\right)^2 \mathscr{s}\, d\lambda_n \;\leq\; 3\left[\mathscr{H}^2\left(\mathscr{s}, f_2\right) + \mathscr{H}^2\left(\mathscr{s}, f_1\right)\right].$$

Following eq. (B.4), we get

$$\mathbb{P}\left(\mathcal{Z}_n \geq \frac{3n\left[\mathscr{H}^2(\mathscr{s}, f_2) + \mathscr{H}^2(\mathscr{s}, f_1)\right]}{2(\kappa - b)} + x\kappa\right) \;\leq\; \exp(-x)$$

which is equivalent to

$$\mathbb{P}\left(\frac{\mathcal{Z}_n}{n} \;\geq\; \frac{3\left[\mathscr{H}^2(\mathscr{s}, f_2) + \mathscr{H}^2(\mathscr{s}, f_1)\right]}{2(\kappa - b)} + \frac{x\kappa}{n}\right) \;\leq\; \exp(-x). \tag{B.5}$$

By Lemma 5,

$$\left(1 - \tfrac{1}{\sqrt{2}}\right)\mathscr{H}^2\left(\mathscr{s}, f_2\right) + T\left(f_1, f_2\right) \;-\; \left(1 + \tfrac{1}{\sqrt{2}}\right)\mathscr{H}^2\left(\mathscr{s}, f_1\right) \;\leq\; \frac{\mathcal{Z}_n}{n}.$$

Substituting this into eq. (B.5) yields, with probability at most e^{-x},

$$\begin{aligned}&\left(1 - \tfrac{1}{\sqrt{2}}\right)\mathscr{H}^2\left(\mathscr{s}, f_2\right) + T\left(f_1, f_2\right) \;-\; \left(1 + \tfrac{1}{\sqrt{2}}\right)\mathscr{H}^2\left(\mathscr{s}, f_1\right)\\ &\quad\leq\; \frac{3\left[\mathscr{H}^2(\mathscr{s}, f_2) + \mathscr{H}^2(\mathscr{s}, f_1)\right]}{2(\kappa - b)} \;+\; \frac{x\kappa}{n}.\end{aligned}$$

Next, observe that $\psi(\cdot,\cdot) \leq 1/\sqrt{2}$. We set

$$b = 1/\sqrt{2}, \quad x = \frac{\Delta_{\mathcal{S}}(f_1) + \Delta_{\mathcal{S}}(f_2)}{\kappa} + n\zeta, \quad \kappa = \frac{2 + 11\sqrt{2}}{2\sqrt{2} - 2},$$

implying $1.5 \times (\kappa - b) = \left(1 - 1/\sqrt{2}\right)/4$. Hence, with probability at most $\exp\left(-\frac{\Delta_{\mathcal{S}}(f_1)+\Delta_{\mathcal{S}}(f_2)}{\kappa} - n\,\zeta\right)$,

$$\begin{aligned}
&\left(1 - \tfrac{1}{\sqrt{2}}\right)\mathcal{H}^2\left(\mathcal{I}, f_2\right) + T\left(f_1, f_2\right) \; - \; \left(1 + \tfrac{1}{\sqrt{2}}\right)\mathcal{H}^2\left(\mathcal{I}, f_1\right) \\
&\quad \leq \; \frac{1}{4}\left(1 - \tfrac{1}{\sqrt{2}}\right)\left[\mathcal{H}^2\left(\mathcal{I}, f_2\right) + \mathcal{H}^2\left(\mathcal{I}, f_1\right)\right] \; + \; \frac{x\kappa}{n}.
\end{aligned}$$

By rearranging terms, substituting the value of x, and bounding $\left(1 - 0.5^{0.5}\right)\mathcal{H}^2\left(\mathcal{I}, f_1\right)$ by $\left(1 + 0.5^{0.5}\right)\mathcal{H}^2\left(\mathcal{I}, f_1\right)$, we conclude

$$\frac{3}{4}\left(1 - \tfrac{1}{\sqrt{2}}\right)\mathcal{H}^2\left(\mathcal{I}, f_2\right) \; + \; T\left(f_1, f_2\right) \; \leq \; \frac{5}{4}\left(1 + \tfrac{1}{\sqrt{2}}\right)\mathcal{H}^2\left(\mathcal{I}, f_1\right) \; + \; \frac{\Delta_{\mathcal{S}}(f_1) + \Delta_{\mathcal{S}}(f_2)}{\kappa n} \; + \; \zeta.$$

Equivalently—using some trivial upper bounds such as $\kappa < 22, 1 < L$ etc., and via some rearrangements—we get

$$(1-\varepsilon)\mathcal{H}^2(\mathcal{I}, f') + T(f, f') - L\frac{\Delta_{\mathcal{S}}(f')}{n} \leq (1+\varepsilon)\mathcal{H}^2(\mathcal{I}, f) + L\frac{\Delta_{\mathcal{S}}(f)}{n} + 22\zeta.$$

Now,

$$\begin{aligned}
&\mathbb{P}\left(\bigcap_{f,f'\in\mathcal{S}}\left\{(1-\varepsilon)\mathcal{H}^2(\mathcal{I}, f') + T(f, f') - L\frac{\Delta_{\mathcal{S}}(f')}{n} \leq (1+\varepsilon)\mathcal{H}^2(\mathcal{I}, f) + L\frac{\Delta_{\mathcal{S}}(f)}{n} + 22\zeta\right\}\right) \\
&\leq \sum_{f,f'\in\mathcal{S}} \mathbb{P}\left((1-\varepsilon)\mathcal{H}^2(\mathcal{I}, f') + T(f, f') - L\frac{\Delta_{\mathcal{S}}(f')}{n} \leq (1+\varepsilon)\mathcal{H}^2(\mathcal{I}, f) + L\frac{\Delta_{\mathcal{S}}(f)}{n} + 22\zeta\right) \\
&\leq \sum_{f,f'\in\mathcal{S}} e^{\left(-\frac{\Delta_{\mathcal{S}}(f_1)+\Delta_{\mathcal{S}}(f_2)}{\kappa} - n\,\zeta\right)} \\
&\overset{(i)}{\leq} \mathcal{C}e^{-n\,\zeta}
\end{aligned}$$

where $\mathcal{C}$ is an universal constant; and (i) follows from Assumption 1, and in particularly from the fact that $\sum_f e^{-\Delta_{\mathcal{S}}(f)} \leq 1$ and $\kappa > 0$. □

B.1.1 Proof of Lemma 4

Proof. The proof will then follow by integrating both sides with respect to λ_n. It is enough to prove,

$$\left(\frac{\sqrt{f_2} - \sqrt{f_1}}{\sqrt{\bar{f}}}\right)^2 \mathcal{I} \leq 3\left[\left(\sqrt{\mathcal{I}} - \sqrt{f_2}\right)^2 + \left(\sqrt{\mathcal{I}} - \sqrt{f_1}\right)^2\right].$$

This is equivalent to proving

$$\left(\sqrt{f_2} - \sqrt{f_1}\right)^2 \mathcal{I} \leq 3\bar{f}\left[\left(\sqrt{\mathcal{I}} - \sqrt{f_2}\right)^2 + \left(\sqrt{\mathcal{I}} - \sqrt{f_1}\right)^2\right].$$

It holds by algebra that $\mathcal{I} \leq 2\left[(\sqrt{\mathcal{I}} - \sqrt{\bar{f}})^2 + \bar{f}\right]$. The left hand side can now be rewritten as

$$\begin{aligned}
\left(\sqrt{f_2} - \sqrt{f_1}\right)^2 \mathcal{I} &\leq 2\left(\sqrt{f_2} - \sqrt{f_1}\right)^2\left[(\sqrt{\mathcal{I}} - \sqrt{\bar{f}})^2 + \bar{f}\right] \\
&= 2\bar{f}\left(\sqrt{f_2} - \sqrt{f_1}\right)^2\left[\frac{(\sqrt{\mathcal{I}} - \sqrt{\bar{f}})^2}{\bar{f}} + 1\right] \\
&= 2\bar{f}\left[\frac{(\sqrt{\mathcal{I}} - \sqrt{\bar{f}})^2}{\bar{f}}\left(\sqrt{f_2} - \sqrt{f_1}\right)^2 + \left(\sqrt{f_2} - \sqrt{f_1}\right)^2\right]
\end{aligned} \tag{B.6}$$

Observe that $\left(\sqrt{f_2}-\sqrt{f_1}\right)^2/\bar{f} \leq (\sqrt{\max\{f_1,f_2\}})^2/\bar{f}$ which in turn can be upper bounded by 2. Thus,

$$\begin{aligned}\frac{(\sqrt{\mathscr{s}}-\sqrt{\bar{f}})^2}{\bar{f}}\left(\sqrt{f_2}-\sqrt{f_1}\right)^2 &\leq 2(\sqrt{\mathscr{s}}-\sqrt{\bar{f}})^2 \\ &\leq 2\frac{(\sqrt{f_2}-\sqrt{s})^2+(\sqrt{f_1}-\sqrt{s})^2}{2},\end{aligned}$$

where the second inequality follows from the convexity of the function $x \to (\sqrt{x}-\sqrt{\mathscr{s}})^2$ and Jensen's inequality. Since the fact $\left(\sqrt{f_2}-\sqrt{f_1}\right)^2 \leq 2\left[\left(\sqrt{f_2}-\sqrt{s}\right)^2+\left(\sqrt{f_1}-\sqrt{s}\right)^2\right]$ holds algebraically, we now have

$$\frac{(\sqrt{\mathscr{s}}-\sqrt{\bar{f}})^2}{\bar{f}}\left(\sqrt{f_2}-\sqrt{f_1}\right)^2+\left(\sqrt{f_2}-\sqrt{f_1}\right)^2 \leq 3\left[(\sqrt{f_2}-\sqrt{s})^2+(\sqrt{f_1}-\sqrt{s})^2\right].$$

This, when combined with eq. (B.6) completes the proof of our lemma. □

B.1.2 Proof of Lemma 5

Proof. The proof of this Lemma share similarities with the proofs of Propositions 2 and 3 in Baraud (2011) or that of Claim B3 in Sart (2014). To begin, observe that it is enough to show

$$\mathscr{H}^2(\mathscr{s},f_2)+T(f_1,f_2)-\mathscr{H}^2(\mathscr{s},f_1) \leq \frac{1}{\sqrt{2}}\left(\mathscr{H}^2(\mathscr{s},f_2)+\mathscr{H}^2(\mathscr{s},f_1)\right)+\frac{1}{n}\sum_{i=0}^{n-1}Z_i(f_1,f_2).$$

Starting from the left hand side, we substitute the expression for T from eq. (3.1), expand all squares, and cancel relevant terms. To be precise, we can write,

$$\begin{aligned}\text{L.H.S} &= \int\left(\sqrt{f_2}-\sqrt{\mathscr{s}}\right)^2 d\lambda_n - \int\left(\sqrt{f_1}-\sqrt{\mathscr{s}}\right)^2 d\lambda_n + \frac{1}{n}\sum_{i=0}^{n-1}\psi\left(f_1(X_i,a_i,X_{i+1}),f_2(X_i,a_i,X_{i+1})\right) \\ &\quad + \int\sqrt{\bar{f}}\left(\sqrt{f_2}-\sqrt{f_1}\right)d\lambda_n + \int(f_1-f_2)\,d\lambda_n. \\ &= -2\rho(f_2,\mathscr{s})+2\rho(f_1,\mathscr{s})+\frac{1}{n}\sum_{i=0}^{n-1}\psi\left(f_1(X_i,a_i,X_{i+1}),f_2(X_i,a_i,X_{i+1})\right) \\ &\quad + \int\sqrt{\bar{f}}\left(\sqrt{f_2}-\sqrt{f_1}\right)d\lambda_n \\ &= -2\rho(f_2,\mathscr{s})+2\rho(f_1,\mathscr{s})+\frac{1}{n}\sum_{i=0}^{n-1}Z_i(f_1,f_2)+\int\psi(f_1,f_2)\,\mathscr{s}\,d\lambda_n+\int\sqrt{\bar{f}}\left(\sqrt{f_2}-\sqrt{f_1}\right)d\lambda_n\end{aligned}$$

All that is now left to show is

$$-2\rho(f_2,\mathscr{s})+2\rho(f_1,\mathscr{s})+\int\psi(f_1,f_2)d\lambda_n+\int\sqrt{\bar{f}}\left(\sqrt{f_2}-\sqrt{f_1}\right)d\lambda_n$$

can be bounded above from by $0.5^{0.5}\left(\mathcal{H}^2(\mathscr{s}, f_2) + \mathcal{H}^2(\mathscr{s}, f_1)\right)$. As before, we start with the left hand side and observe that

$$
\begin{aligned}
&-2\rho(f_2, \mathscr{s}) + 2\rho(f_1, \mathscr{s}) + \int \psi(f_1, f_2)\, \mathscr{s}\, d\lambda_n + \int \sqrt{\bar{f}}\left(\sqrt{f_2} - \sqrt{f_1}\right) d\lambda_n \\
&= \int \left[-2\sqrt{f_2 \mathscr{s}} + 2\sqrt{f_1 \mathscr{s}} + \frac{\sqrt{f_2} - \sqrt{f_1}}{\sqrt{\bar{f}}}\mathscr{s} + \sqrt{\bar{f}}\left(\sqrt{f_2} - \sqrt{f_1}\right)\right] d\lambda_n \\
&= \int \left[\sqrt{\frac{f_2}{\bar{f}}}\left(\sqrt{\bar{f}} - \sqrt{\mathscr{s}}\right)^2 - \sqrt{\frac{f_1}{\bar{f}}}\left(\sqrt{\bar{f}} - \sqrt{\mathscr{s}}\right)^2\right] d\lambda_n \\
&\leq \int \sqrt{\frac{f_2}{\bar{f}}}\left(\sqrt{\bar{f}} - \sqrt{\mathscr{s}}\right)^2 d\lambda_n \\
&\leq \sqrt{2}\mathcal{H}^2(\bar{f}, \mathscr{s}).
\end{aligned}
$$

The first inequality follows trivially. The second inequality follows from the fact that $f_2/\bar{f} \leq 2$. Now, observe that the function $x \to (\sqrt{x} - \sqrt{\mathscr{s}})^2$ is convex in x when $x > 0$. Therefore, using Jensen's inequality, we can write $\sqrt{2}\mathcal{H}^2(\bar{f}, \mathscr{s}) \leq \left[\mathcal{H}^2(f_1, \mathscr{s}) + \mathcal{H}^2(f_2, \mathscr{s})\right]/\sqrt{2}$. This completes the proof. □

B.2 Proofs of Propositions 7, 8 and 9

To remind the readers about the statement of the propositions we restate them before proving each one starting with Proposition 7.

Proposition 7. *Let Assumptions 1, 2 and 3 hold. Then for any control a and context G, one has the following*

$$
\mathbb{E}\left[\int \mathbf{1}\left\{C_n(a, G) > U(G, a, n)\right\} d\lambda_n^{(2)}\right] = O((r_S(f, n)/n)^{1/2-\delta}).
$$

Proof. One starts by noting that

$$
\begin{aligned}
&\mathbb{E}\left[\int \mathbf{1}\left\{\int L(x, y, a, G)\, \mathscr{s}(x, a, G, y)\, d\lambda_n^{(1)}(dx, dy) > U(G, a, n)\right\} d\lambda_n^{(2)}\right] \\
&\leq \mathbb{E}\left[\int \mathbf{1}\left\{\left|\int L(x, y, a, G)\, \mathscr{s}(x, a, G, y)\, d\lambda_n^{(1)}(dx, dy) - \int L(x, y, a, G)\, \hat{\mathscr{s}}(x, a, G, y)\, d\lambda_n^{(1)}(dx, dy)\right| d\lambda_n^{(2)}\right.\right. \\
&\qquad \left.\left. \geq \left(\frac{r_S(f, n)}{n}\right)^{\delta}\right\}\right].
\end{aligned}
$$

Therefore, one has

$$
\begin{aligned}
&\mathbb{E}\left[\int \mathbf{1}\left\{\int L(x, y, a, G)\, \mathscr{s}(x, a, G, y)\, d\lambda_n^{(1)}(dx, dy) > U(G, a, n)\right\} d\lambda_n^{(2)}\right] \\
&\leq \left(\frac{r_S(f, n)}{n}\right)^{-\delta} \\
&\qquad \mathbb{E}\left[\int \left|\int L(x, y, a, G)\, \mathscr{s}(x, a, G, y)\, d\lambda_n^{(1)}(dx, dy) - \int L(x, y, a, G)\, \hat{\mathscr{s}}(x, a, G, y)\, d\lambda_n^{(1)}(dx, dy)\right| d\lambda_n^{(2)}\right].
\end{aligned}
$$

Noting that

$$
\begin{aligned}
&\int \left|\int L(x, y, a, G)\, \mathscr{s}(x, a, G, y)\, d\lambda_n^{(1)} - \int L(x, y, a, G)\, \hat{\mathscr{s}}(x, a, G, y)\, d\lambda_n^{(1)}\right| d\lambda_n^{(1)}\, d\lambda_n^{(2)} \\
&\leq \int\int L^2(x, y, a, G)\, \left|\mathscr{s}(x, a, G, y) - \hat{\mathscr{s}}(x, a, G, y)\right|\, d\lambda_n^{(1)}\, d\lambda_n^{(2)} \\
&\leq L_{\max} \int\int \left|\sqrt{\mathscr{s}(x, a, G, y)} + \sqrt{\hat{\mathscr{s}}(x, a, G, y)}\right| \left|\sqrt{\mathscr{s}(x, a, G, y)} - \sqrt{\hat{\mathscr{s}}(x, a, G, y)}\right| d\lambda_n^{(1)}\, d\lambda_n^{(2)}.
\end{aligned}
$$

Hence, we get

$$\begin{aligned}
&\int \left| \int L(x,y,a,G)\, \mathscr{s}(x,a,G,y)\, d\lambda_n^{(1)} - \int L(x,y,a,G)\, \hat{\mathscr{s}}(x,a,G,y)\, d\lambda_n^{(1)} \right| d\lambda_n^{(1)}\, d\lambda_n^{(2)} \\
&\leq L_{\max} \left(\int \int \left| \sqrt{\mathscr{s}(x,a,G,y)} + \sqrt{\hat{\mathscr{s}}(x,a,G,y)} \right|^2 d\lambda_n^{(1)}\, d\lambda_n^{(2)} \right)^{1/2} \\
&\quad \left(\int \int \left| \sqrt{\mathscr{s}(x,a,G,y)} - \sqrt{\hat{\mathscr{s}}(x,a,G,y)} \right|^2 d\lambda_n^{(1)}\, d\lambda_n^{(2)} \right)^{1/2} \\
&\leq L_{\max} \left(\int \int 2\left(\mathscr{s}(x,a,G,y) + \hat{\mathscr{s}}(x,a,G,y)\right) d\lambda_n^{(1)}\, d\lambda_n^{(2)} \right)^{1/2} \\
&\quad \left(\int \int \left| \sqrt{\mathscr{s}(x,a,G,y)} - \sqrt{\hat{\mathscr{s}}(x,a,G,y)} \right|^2 d\lambda_n^{(1)}\, d\lambda_n^{(2)} \right)^{1/2} \\
&\leq 2\, L_{\max} \left(1+\varepsilon\right)^{1/2} \left(\int \int \left| \sqrt{\mathscr{s}(x,a,G,y)} - \sqrt{\hat{\mathscr{s}}(x,a,G,y)} \right|^2 d\lambda_n^{(1)}\, d\lambda_n^{(2)} \right)^{1/2}.
\end{aligned}$$

Therefore, using Jensen's inequality for concave functions (here $x^{1/2}$), we have

$$\begin{aligned}
&\mathbb{E}\left[\int \left| \int L(x,y,a,G)\, \mathscr{s}(x,a,G,y)\, d\lambda_n^{(1)}(dx,dy) - \int L(x,y,a,G)\, \hat{\mathscr{s}}(x,a,G,y)\, d\lambda_n^{(1)}(dx,dy) \right|^2 d\lambda_n^{(2)} \right] \\
&\leq 2\, L_{\max} \left(1+\varepsilon\right)^{1/2} \left(\mathbb{E}\left[\int \int \left| \sqrt{\mathscr{s}(x,a,G,y)} - \sqrt{\hat{\mathscr{s}}(x,a,G,y)} \right|^2 d\lambda_n^{(1)}\, d\lambda_n^{(2)} \right] \right)^{1/2} \\
&\leq 2\, L_{\max} \left(1+\varepsilon\right)^{1/2} \left(\mathbb{E}\left[\mathscr{H}^2(\mathscr{s}, \hat{\mathscr{s}}) \right] \right)^{1/2}
\end{aligned}$$

where the last line follows as $d\lambda_n^{(2)} = d\lambda_n / d\lambda_n^{(1)}$. Therefore the proof follows immediately via Corollary 1. □

Assumption 4. *For any treatment pair of controls a, a' with $a \in a^*(G)$ and $a' \notin a^*(G)$, given any context G, one has*

$$C_n(a', G) - C_n(a, G) \geq \Delta_L$$

for some $\Delta_L > 0$.

Proposition 8. *Let Assumptions 1, 2, 3 and 4 hold. Further, assume n is large enough such that $\Delta_L > (r_S(f,n)\, n^{-1})^{\delta}$. Then for any $a \in a^*(G)$ and $a' \notin a^*(G)$, one has*

$$\mathbb{E}\left[\int \mathbf{1}\left\{ U(G,a,n) > U(G,a',n) \mid C_n(a,G) < C_n(a',G) \right\} d\lambda_n^{(2)} \right] = O\left(\left(\frac{r_S(f,n)}{n} \right)^{1/2-\delta} \frac{1}{\Delta_L} \right).$$

Proof. We start by considering the events

$$\mathcal{E}_n^{(1)} = \{ C_n(a,G) > U(G,a,n) \}$$

and

$$\mathcal{E}_n^{(2)} = \{ C_n(a',G) > U(G,a',n) \}$$

which are low probability events or error events. Therefore

$$\begin{aligned}
&\mathbb{E}\left[\int \mathbf{1}\left\{ U(G,a,n) > U(G,a',n) \mid C_n(a,G) < C_n(a',G) \right\} d\lambda_n^{(2)} \right] \\
&= \mathbb{E}\left[\int \mathbf{1}\left\{ (U(G,a,n) > U(G,a',n) \mid C_n(a,G) < C_n(a',G)) \cap \mathcal{E}_n^{(1)\,c} \cap \mathcal{E}_n^{(2)\,c} \right\} d\lambda_n^{(2)} \right] \\
&\qquad + \mathbb{E}\left[\int \mathbf{1}\left\{ \mathcal{E}_n^{(1)} \right\} d\lambda_n^{(2)} \right] + \mathbb{E}\left[\int \mathbf{1}\left\{ \mathcal{E}_n^{(2)} \right\} d\lambda_n^{(2)} \right].
\end{aligned}$$

On the set

$$\left\{(U(G,a,n) > U(G,a',n),\, C_n(a,G) < C_n(a',G)) \cap \mathcal{E}_n^{(1)\,c} \cap \mathcal{E}_n^{(2)\,c}\right\}$$

we have

$$\begin{aligned} U(G,a,n) - C(a,G) &\geq U(G,a',n) - C(a,G) \\ &\geq C(a',G) - C(a,G) \\ &\geq \Delta_L. \end{aligned}$$

Therefore

$$\left\{(U(G,a,n) > U(G,a',n),\, C_n(a,G) < C_n(a',G)) \cap \mathcal{E}_n^{(1)\,c} \cap \mathcal{E}_n^{(2)\,c}\right\} \subset \{U(G,a,n) - C(a,G) \geq \Delta_L\}.$$

Therefore

$$\begin{aligned} &\mathbb{E}\left[\int \mathbf{1}\left\{U(G,a,n) > U(G,a',n) \mid C_n(a,G) < C_n(a',G)\right\} d\lambda_n^{(2)}\right] \\ &\leq C_1 \left(\frac{r_S(f,n)}{n}\right)^{1/2} \frac{1}{\left(\Delta_L - \left(\frac{r_S(f,n)}{n}\right)^{\delta}\right)} + 2\,C_2 \left(\frac{r_S(f,n)}{n}\right)^{1/2-\delta} \\ &= O\left(\left(\frac{r_S(f,n)}{n}\right)^{1/2-\delta} \frac{1}{\Delta_L}\right). \end{aligned}$$

□

Proposition 9. *Let Assumptions 1, 2, 3 and 4 hold. For any set of controls $a \in a^*(G)$ and $\hat{a}_n \in \hat{a}^*(G)$ with context G, one has*

$$\mathbb{E}\left[\int C(\hat{a}_n, G) - C(a,G)\, d\lambda^{(2)}\right] = O\left(\|C\|_\infty \left(\frac{r_S(f,n)}{n}\right)^{1/2-\delta} \frac{1}{\Delta_L}\right)$$

Proof. One starts by noting that

$$\begin{aligned} &\mathbb{E}\left[\int (C(\hat{a}_n, G) - C(a,G))\, d\lambda_n^{(2)}\right] \\ &= \mathbb{E}\left[\int (C(\hat{a}_n, G) - C(a,G))\, \mathbf{1}\left\{\hat{a}_n \notin \hat{a}^*(G)\right\} d\lambda_n^{(2)}\right]. \end{aligned}$$

Further note that

$$(C(\hat{a}_n, G) - C(a,G)) \leq 2\,\|C\|_\infty.$$

Therefore one notes that

$$\begin{aligned} &\mathbb{E}\left[\int (C(\hat{a}_n, G) - C(a,G))\, d\lambda_n^{(2)}\right] \\ &\leq 2\|C\|_\infty\, \mathbb{E}\left[\int \mathbf{1}\left\{\hat{a}_n \notin \hat{a}^*(G)\right\} d\lambda_n^{(2)}\right] \\ &\leq 2\|C\|_\infty\, \mathbb{E}\left[\int \mathbf{1}\left\{U(G,a,n) > U(G,a',n) \mid C_n(a,G) < C_n(a',G)\right\} d\lambda_n^{(2)}\right]. \end{aligned}$$

Thus the results follow. □

B.3 Proof of Proposition 1

That $\dim f < \infty$ is obvious. We prove the summability. Observe the two facts that $|\mathcal{M}_\ell| = 2^\ell$, and for all $f \in \mathcal{M}_\ell$, $\Delta_{\mathcal{S}}(f) = e^a e^{-\ell}$. Therefore,

$$\begin{aligned}\sum_{f \in \mathcal{S}} e^a e^{-\ell} &= e^a \sum_{\ell \geq 1} \left(\frac{2}{e}\right)^\ell \\ &= e^a \frac{2}{e-2}.\end{aligned}$$

Setting $a = \log(\frac{e-2}{2})$ now completes the proof.