NON-PARAMETRIC REHEARSAL LEARNING VIA CONDITIONAL MEAN EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

In machine learning, a critical class of decision-related problems concerns preventing predicted undesirable outcomes, referred to as the *avoiding undesired future* (AUF) problem. To address this, the *rehearsal learning* framework has been proposed to model influence relations for effective decisions. However, existing rehearsal methods rely on restrictive parametric assumptions such as linear systems or additive noise, limiting their practical applicability. In this paper, we propose the first non-parametric rehearsal learning approach for AUF without assuming specific functional forms of data generation processes. Specifically, we use kernel machinery to reformulate the AUF objective into a unified representation that disentangles desirability modeling from action-induced distributional changes. To retain constraint-margin information beyond binary success labels, we present a smooth Probit surrogate with an approximation error bound. We estimate the smoothed AUF objective by scalar response regression and CME-based adjustment, with pointwise consistency at a fixed query and fixed smoothing parameter under stated spectral conditions. Such a formulation naturally accommodates nonlinear systems and non-additive noise; empirical results on synthetic and real-data-derived semi-synthetic benchmarks demonstrate effectiveness and flexibility of our approach.

1 INTRODUCTION

Machine learning (ML) has achieved great success in various real-world prediction tasks (LeCun et al., 2015). Instead of solely focusing on prediction, Zhou (2022) emphasizes another important issue, *i.e.*, if an ML model predicts that something undesired is going to occur, how to find effective actions to prevent it from happening. This problem is known as *avoiding undesired future* (AUF) problem. Consider a motivating example of bank lending illustrated in Fig. 1, where the bank seeks to proactively optimize its loan approval strategy. The objective is to maximize the probability that the target outcomes $\mathbf{Y}$ (*e.g.*, repayment rate and return on investment (ROI)) satisfy specific expectations, formalized as falling into a desired region $\mathcal{S}$ (*e.g.*, $\{\mathbf{y} \mid \text{repayment rate} \geq \alpha, \text{ROI} \geq \beta\}$). In practice, the bank typically observes user profiles $\mathbf{X}$ (*e.g.*, credit score and debt-to-income ratio) to predict $\mathbf{Y}$. If the model anticipates a high likelihood of an undesired future (*i.e.*, $\mathbf{Y} \notin \mathcal{S}$), it becomes imperative to actively alter actionable intermediate variables $\mathbf{Z}_A$ (*e.g.*, interest rate) to steer $\mathbf{Y}$ towards $\mathcal{S}$.

A main challenge in such scenario lies in the fact that historical records are collected observationally, where the values of $\mathbf{Z}_A$ could be confounded by unactionable factors (*e.g.*, market fluctuations) that affect both $\mathbf{Z}_A$ and $\mathbf{Y}$. Consequently, the data distribution under such active alteration differs significantly from the conditional distribution on the historical data. Furthermore, in high-stakes scenarios such as financial lending and healthcare, actively exploring alterations via trial-and-error is prohibitive due to the risk of financial loss in lending and ethical concerns in medical decision-making. Thus, finding optimal decision alterations to maximize the probability of $\mathbf{Y} \in \mathcal{S}$ (*i.e.*, the AUF probability) solely based on historical data becomes the core objective of the AUF problem.

To address this, one solution is by *rehearsal learning* framework (Zhou, 2022), which leverages fine-grained structural information, including both direct and bi-directed influence, to guide the selection of decision alterations. While existing rehearsal-based methods have made progress, particularly under assumptions of linear additive noise models (Qin et al., 2023; Du et al., 2025), such assumptions rarely hold in complex real-world systems, posing practical challenges. Fundamentally, these methods rely on parametric assumptions about data-generating mechanisms, often linear Gaussian models, so

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

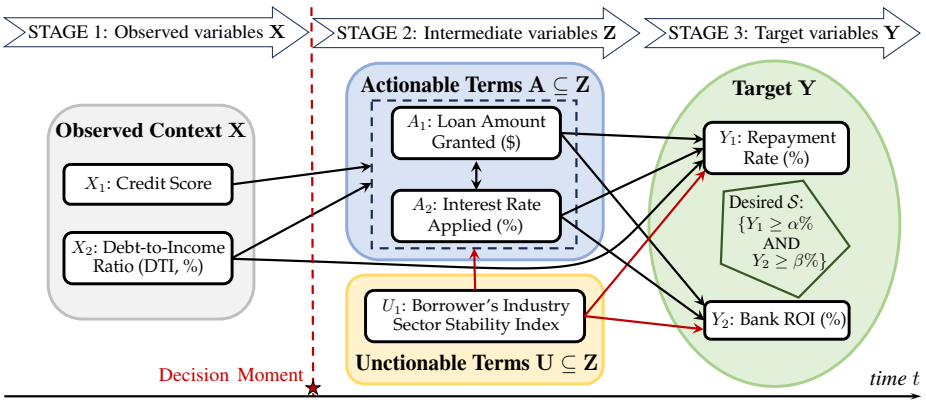

Figure 1: An example illustrating the bank lending AUF scenario.

their performance depends on how closely the assumed family matches the underlying system. When these assumptions do not match the data-generating mechanism, model misspecification can reduce the success probability of the selected alterations. While Qin et al. (2025) attempts to accommodate non-linear scenarios, it remains parametric, and resorts to a heuristic objective which can deviate from the true AUF probability as discussed in Appx. A. Consequently, this approach could not guarantee that the suggested decision aligns with the true AUF goal, often yielding suboptimal empirical results.

Considering the limitations, in this paper, we propose a non-parametric rehearsal learning framework that addresses AUF without relying on specific functional form assumptions. Our key idea is to use kernel machinery to make the AUF objective both *estimable* from observational data and *optimizable* over potential actions. Specifically, we introduce a Probit surrogate w_η for the desirability indicator $\mathbb{I}(\mathbf{Y} \in \mathcal{S})$ and bound its approximation error under a boundary-margin condition. We estimate conditional desirability by scalar kernel ridge regression (KRR) and integrate it using a conditional mean embedding (CME) of the adjustment covariates in a reproducing kernel Hilbert space (RKHS). This nested construction yields an identifiable smoothed AUF objective without requiring the response function w_η to belong to an output RKHS. For fixed smoothing, we derive explicit bias bounds and a pointwise convergence rate under fractional source and query-evaluation conditions. Finally, we employ a practical multi-start projected gradient ascent procedure to optimize the estimator and obtain decisions. For empirical evaluation, we extract a real-data-derived semi-synthetic benchmark from NHANES 2011–2018 (National Center for Health Statistics, 2018), a diabetes-related dataset, on which our method improves AUF probability from 0.402 under no-action baseline to 0.615.

Our contributions can be summarized as:

1. We develop the first non-parametric AUF framework using an existing kernel estimator. By combining response regression with CME-based adjustment, the proposed method accommodates nonlinear dependencies without specifying parametric data-generating equations.
2. We formulate a Probit-smoothed AUF objective that retains constraint-margin information beyond binary labels. Our analysis covers identifiability, Probit approximation, and estimation error. We employ multi-start projected gradient ascent to select context-dependent actions.
3. We empirically evaluate the proposed method across both synthetic and real-world-derived settings. Besides linear/nonlinear synthetic benchmarks and the standard Bermuda benchmark, we establish and introduce a semi-synthetic NHANES diabetes benchmark for the AUF problem, where our method shows strong performance against various baselines.

2 Background

In this section, we briefly review several key concepts.

2.1 Influence Structure and Feasible Alterations

We assume variables generating in a specific partial order $\mathcal{O}$ that records which variables can occur before others in the data-generating process: if $V_i \prec_{\mathcal{O}} V_j$, then V_i precedes V_j in every topological

ordering compatible with $\mathcal{O}$. In this paper, the order is treated as given side information, which can be learned from observational data (Tao et al., 2026). The variables are partitioned into context variables $\mathbf{X}$, intermediate variables $\mathbf{Z}$, and outcome variables $\mathbf{Y}$, with block-level ordering $\mathbf{X} \prec_{\mathcal{O}} \mathbf{Z} \prec_{\mathcal{O}} \mathbf{Y}$.

Context $\mathbf{X}$ is observed before decision-making, while $\mathbf{Y}$ is observed after the decision and is desired to fall in a target region $\mathcal{S}$. Intermediate variables $\mathbf{Z}$ contains both variables that can be acted upon and those cannot. The underlying data-generating process may be compatible with richer structured equation models, but our main results only require the known order with assumptions in Prop. 1.

In structural causal models (Spirtes et al., 2000), $do(V_i = v_i)$ denotes a structural operation that cuts the dependence of V_i on its original parents and replaces its structural assignment with the constant value v_i. For the AUF problem, we further need to distinguish which variables can actually be acted upon in a given real-world decision problem. We call a variable V_i actionable if its feasible alteration set Δ_{V_i} is non-empty and the decision-maker can take an action to set V_i to some value $v_i \in \Delta_{V_i}$. For such an actionable variable, we write $V_i \overset{a}{=} v_i$ to indicate a feasible alteration. For a vector of actionable variables $\mathbf{A}$, we write $\mathbf{A} \overset{a}{=} \mathbf{a}$, or equivalently use the shorthand $\mathring{\mathbf{a}}$, to make explicit that the operation is applied to actionable nodes. The notation $P(\cdot \mid \mathbf{x}, \mathring{\mathbf{a}})$ denotes the post-alteration distribution after observing context $\mathbf{x}$ and applying this feasible alteration.

2.2 Conditional mean embeddings

Let U and V be random variables in measurable domains $\mathcal{U}$ and $\mathcal{V}$, endowed with kernels $k_u : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$ and $k_v : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$, along with their associated RKHSs, denoted by $\mathcal{H}_{k_u}$ and $\mathcal{H}_{k_v}$. We assume the kernels are continuous, bounded, and positive-definite. The conditional mean embedding (CME) (Song et al., 2009; 2013) maps the conditional distribution $P(V|u)$ to an element $\mu_{V|u} \in \mathcal{H}_{k_v}$. Formally, it is defined as the expected feature map of V conditioned on u:

$$\mu_{V|u} := \mathbb{E}_{V \sim P(\cdot|u)}[k_v(v, \cdot)] = \int_{\mathcal{V}} k_v(v, \cdot)\, \mathrm{d}P(v \mid u).$$

A crucial property of CME is that it enables the evaluation of conditional expectations for any function $g \in \mathcal{H}_{k_v}$ via the inner product in $\mathcal{H}_{k_v}$ that $\mathbb{E}[g(V) \mid u] = \langle g, \mu_{V|u} \rangle_{\mathcal{H}_{k_v}}$ (reproducing property), bridging the integral operation of expectation with the algebraic operation in the Hilbert space.

3 Our approach

In this section, we formulate the AUF objective and introduce a Probit surrogate to retain constraint-margin information discarded by binary desirability labels. We estimate the resulting smoothed objective from observational data using nested KRR and apply projected gradient ascent to select feasible actions. Finally, we analyze the surrogate approximation error and the propagation of errors through the two estimation stages.

3.1 Formulation of the AUF problem

Given historical data $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}_i)\}_{i=1}^N$ (collected from observational distribution, distinct from the alterational one), we focus on the decision moment where the context $\mathbf{x}$ has just been observed, while the subsequent intermediate variables $\mathbf{Z}$ and outcomes $\mathbf{Y}$ are yet to materialize. Our objective is to determine the optimal values for the actionable subset $\mathbf{A} \subseteq \mathbf{Z}$ that maximize the probability of $\mathbf{Y}$ falling into a target desired region $\mathcal{S}$ under the feasible alteration $\mathring{\mathbf{a}}$:

$$\arg\max_{\mathbf{a}} \quad \mathbb{P}(\mathbf{Y} \in \mathcal{S} \mid \mathbf{X} = \mathbf{x}, \mathbf{A} \overset{a}{=} \mathbf{a}) \qquad \text{s.t.} \quad \mathbf{a}_{\text{left}} \preccurlyeq \mathbf{a} \preccurlyeq \mathbf{a}_{\text{right}}, \tag{1}$$

where $\preccurlyeq$ denotes element-wise inequality, and $\mathbf{a}_{\text{left}}, \mathbf{a}_{\text{right}}$ are constant vectors specifying the feasible alteration range. Unlike prior rehearsal works that rely on parametric structural equations governed by $\boldsymbol{\theta}$, our formulation is non-parametric and does not assume specific functional forms for generating process. The desired region $\mathcal{S}$ is specified as a convex polyhedron defined by l linear constraints:

$$\mathcal{S} = \left\{ \mathbf{y} \in \mathbb{R}^{d_y} \mid \mathbf{m}_k^{\mathsf{T}} \mathbf{y} \leq b_k,\ k = 1, \ldots, l \right\}. \tag{2}$$

Before constructing the surrogate, fix positive outcome scales $\mathbf{D} = \mathrm{diag}(s_1, \ldots, s_{d_y})$ and replace each nonzero coefficient pair $(\mathbf{m}_k, b_k)$ by $(\mathbf{m}_k, b_k)/\|\mathbf{D}\mathbf{m}_k\|_2$, retaining the same notation. Scale choices and the treatment of redundant constraints are specified in Appx. C.2.

3.2 SMOOTHED DESIRABILITY FUNCTION WITH IDENTIFIABILITY

To facilitate the optimization of Eq. (1), we first reformulate the AUF probability as the expectation of a desirability indicator variable. Let $\mathbb{I}(\mathbf{y} \in \mathcal{S})$ be the indicator function which takes the value 1 if $\mathbf{y} \in \mathcal{S}$ and 0 otherwise. The objective function is equivalently expressed as:

$$\mathbb{P}(\mathbf{Y} \in \mathcal{S} \mid \mathbf{x}, \mathring{\mathbf{a}}) = \mathbb{E}_{\mathbf{y} \sim P(\cdot \mid \mathbf{x}, \mathring{\mathbf{a}})}[\mathbb{I}(\mathbf{y} \in \mathcal{S})].$$

Estimating this expectation from observational data presents two challenges concerning desirability information and distributional adjustment: (i) The hard indicator assigns zero labels to all unsuccessful outcomes, discarding differences in their constraint margins. When successful samples are scarce, these binary responses may provide limited information for learning which actions favor desirable outcomes. We therefore smooth the response to retain graded information; action differentiability itself is supplied by kernel estimator introduced below. (ii) The objective depends on the alterational distribution $P(\mathbf{Y} \mid \mathbf{x}, \mathring{\mathbf{a}})$, which describes the outcome distribution after applying a feasible alteration. This distribution generally differs from observational conditional $P(\mathbf{Y} \mid \mathbf{x}, \mathbf{a})$ and cannot be directly estimated from historical data $\mathcal{D}$ due to the presence of the variables influencing both $\mathbf{A}$ and $\mathbf{Y}$.

To address the first challenge, we introduce a smooth surrogate to approximate the hard indicator. Leveraging the definition of $\mathcal{S}$ in Eq. (2), we define the smoothed desirability function $w_\eta(\mathbf{y})$ as:

$$w_\eta(\mathbf{y}) \triangleq \prod_{k=1}^{l} \Phi\left(\eta(b_k - \mathbf{m}_k^\top \mathbf{y})\right), \tag{3}$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution, and $\eta > 0$ is a scaling parameter controlling the sharpness of the approximation.

As illustrated in Fig. 2, w_η approaches the indicator away from the boundary as η increases. The construction applies to the polyhedral regions in Eq. (2), and expectation convergence holds when boundary probability is zero. For finite η, smoothing retains information discarded by the hard indicator. First, the response varies continuously with the constraint margins, assigning different desirability values to otherwise identical failure labels. Second, these graded responses may improve finite-sample learning when positive samples are scarce. By distinguishing "near-miss" outcomes from larger failures, smoothing can effectively guide the optimization toward the desired region.

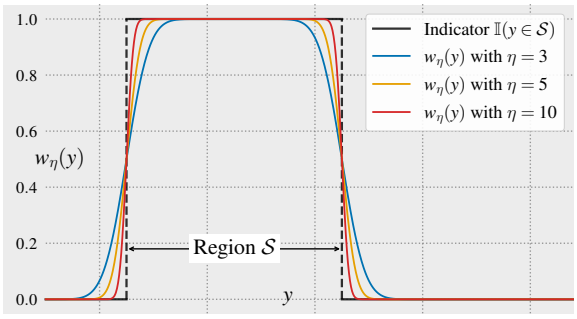

Figure 2: An example of smooth surrogate $w_\eta(\cdot)$ approximating hard indicator $\mathbb{I}(\cdot \in \mathcal{S})$ with varying scaling η.

To address the second challenge, we identify the objective through observational adjustment, independently of smoothing. We first formalize the structure of the intermediate variables $\mathbf{Z}$ below.

Definition 1 (Decomposition of $\mathbf{Z}$)**.** Given the known partial order $\mathcal{O}$, we use a disjoint partition $\mathbf{Z} = \mathbf{U} \cup \mathbf{A} \cup \mathbf{U}^{\text{post}}$ of the intermediate variables, where (i) the subset $\mathbf{A} \subseteq \mathbf{Z}$ contains variables that can be directly altered by the decision-maker; (ii) the subset $\mathbf{U}$ contains non-actionable pre-alteration variables that are unaffected by the joint alteration of $\mathbf{A}$; and (iii) the subset $\mathbf{U}^{\text{post}} := \mathbf{Z} \setminus (\mathbf{U} \cup \mathbf{A})$ contains the remaining non-actionable intermediate variables that occur after the alteration; these variables may lie downstream of $\mathbf{A}$ and are not adjusted for.

Proposition 1 (Identifiability of AUF Objective)**.** *For the decomposition in Def. 1, let $\mathbf{Y}(\mathbf{a})$ denote the outcome with alteration $\mathring{\mathbf{a}}$. Assume conditional exchangeability $\mathbf{Y}(\mathbf{a}) \perp\!\!\!\perp \mathbf{A} \mid (\mathbf{X}, \mathbf{U})$, consistency $\mathbf{Y} = \mathbf{Y}(\mathbf{A})$, and conditional overlap, with continuous-query regularity specified in Appx. C.1. Then:*

$$\mathbb{E}_{\mathbf{y} \sim P(\cdot \mid \mathbf{x}, \mathring{\mathbf{a}})}[w_\eta(\mathbf{Y})] = \mathbb{E}_{\mathbf{u} \sim P(\cdot \mid \mathbf{x})}\left[\mathbb{E}_{\mathbf{y} \sim P(\cdot \mid \mathbf{x}, \mathbf{u}, \mathbf{a})}[w_\eta(\mathbf{Y})]\right].$$

Prop. 1 identifies the objective at queries satisfying these conditions; the feasible box in Eq. (1) does not itself guarantee observational support. To streamline the notation for subsequent estimation, we formally define the smoothed AUF objective as $J_\eta(\mathbf{a}; \mathbf{x}) \triangleq \mathbb{E}_{\mathbf{y} \sim P(\cdot \mid \mathbf{x}, \mathring{\mathbf{a}})}[w_\eta(\mathbf{Y})]$ and let $\mathbf{h} \triangleq (\mathbf{x}, \mathbf{u}, \mathbf{a})$ denote the concatenated vector comprising the observed context, pre-alteration covariates, and the action variables. Accordingly, we define a kernel function $k_h(\cdot, \cdot)$ over the domain of $\mathbf{h}$.

3.3 SURROGATE AUF OBJECTIVE VIA NESTED KRR

Based on the identifiability result in Prop. 1, the AUF objective can be structured as an outer conditional expectation (over pre-alteration covariates $\mathbf{u}$) of an inner conditional expectation (over outcome $\mathbf{y}$). Hence, the evaluation naturally decomposes into a two-stage estimation problem. Let $\psi(\mathbf{h}) \triangleq \mathbb{E}_{\mathbf{y}\sim P(\cdot|\mathbf{h})}[w_\eta(\mathbf{Y})]$ denote the inner expectation given $\mathbf{h} = (\mathbf{x}, \mathbf{u}, \mathbf{a})$. With observed context $\mathbf{x}$ and specific alteration $\mathbf{a}$, $\psi(\mathbf{h})$ serves as the response variable dependent on $\mathbf{u}$, acting as the target for the outer expectation $\mathbb{E}_{\mathbf{u}\sim P(\cdot|\mathbf{x})}$. We estimate ψ by scalar KRR on the joint input space and evaluate the outer expectation using a CME of $\mathbf{U} \mid \mathbf{x}$.

Stage 1: Inner Expectation Estimation. Given observational data $\mathcal{D} = \{(\mathbf{h}_i, \mathbf{y}_i)\}_{i=1}^N$, we regress the bounded responses $w_\eta(\mathbf{y}_i) \in [0, 1]$ on $\mathbf{h}_i$:

$$\hat{\psi} = \arg\min_{f\in\mathcal{H}_{k_h}} \left\{ \frac{1}{N}\sum_{i=1}^{N} \big(f(\mathbf{h}_i) - w_\eta(\mathbf{y}_i)\big)^2 + \lambda_h \|f\|^2_{\mathcal{H}_{k_h}} \right\}. \tag{4}$$

The representer theorem gives $\hat{\psi}(\mathbf{h}) = \sum_{i=1}^N \alpha_i k_h(\mathbf{h}_i, \mathbf{h}), \boldsymbol{\alpha} = (\mathbf{K}_{hh} + N\lambda_h \mathbf{I})^{-1}\mathbf{w}_\eta$, where $\mathbf{w}_\eta = [w_\eta(\mathbf{y}_i)]_{i=1}^N$, $[\mathbf{K}_{hh}]_{ij} = k_h(\mathbf{h}_i, \mathbf{h}_j)$, $\lambda_h > 0$. Let $\mathbf{k}_h(\mathbf{h}) = [k_h(\mathbf{h}_i, \mathbf{h})]_{i=1}^N$, inner estimate is:

$$\hat{\psi}(\mathbf{h}) = \mathbf{w}_\eta^\mathsf{T}(\mathbf{K}_{hh} + N\lambda_h \mathbf{I})^{-1}\mathbf{k}_h(\mathbf{h}). \tag{5}$$

This construction requires only bounded responses; its statistical analysis concerns the regularity of ψ on the input space.

Stage 2: Outer Expectation Estimation. We now proceed to the outer expectation over the pre-alteration covariates $\mathbf{u}$. Note that the term $\boldsymbol{\alpha}^\mathsf{T} \triangleq \mathbf{w}_\eta^\mathsf{T}(\mathbf{K}_{hh} + N\lambda_h \mathbf{I})^{-1}$ in Eq. (5) is a constant row vector determined solely by the observational training data. To facilitate the analytical integration over $\mathbf{u}$, we introduce Ass. 1, a mild structural condition on the kernel that is satisfied by a lot of standard kernels, including the Gaussian RBF used in our experiments and some polynomial kernels.

Assumption 1 (Decomposable Kernels)**.** We assume that the kernel $k_h(\cdot, \cdot)$ defined on the joint space $\mathbf{h} = (\mathbf{x}, \mathbf{u}, \mathbf{a})$ admits a multiplicative decomposition:

$$k_h(\mathbf{h}_i, \mathbf{h}) = k_x(\mathbf{x}_i, \mathbf{x}) \cdot k_u(\mathbf{u}_i, \mathbf{u}) \cdot k_a(\mathbf{a}_i, \mathbf{a}).$$

With this assumption, we substitute the estimator $\hat{\psi}(\mathbf{h})$ back into the outer expectation. Since the integration is performed solely w.r.t.$\mathbf{u}$, the terms of $\mathbf{x}$ and $\mathbf{a}$ remain constant within the integral, thus $\mathbb{E}_{\mathbf{u}\sim P(\cdot|\mathbf{x})}[\hat{\psi}(\mathbf{x}, \mathbf{u}, \mathbf{a})] = \int_{\mathcal{U}} \left(\sum_{i=1}^N \alpha_i k_x(\mathbf{x}_i, \mathbf{x}) k_u(\mathbf{u}_i, \mathbf{u}) k_a(\mathbf{a}_i, \mathbf{a})\right) p(\mathbf{u} \mid \mathbf{x})\, d\mathbf{u}$ is computed as:

$$\sum_{i=1}^{N} \alpha_i \underbrace{k_x(\mathbf{x}_i, \mathbf{x})}_{\text{Context match}} \underbrace{k_a(\mathbf{a}_i, \mathbf{a})}_{\text{Action similarity}} \underbrace{\int_{\mathcal{U}} k_u(\mathbf{u}_i, \mathbf{u}) p(\mathbf{u} \mid \mathbf{x})\, d\mathbf{u}}_{\text{Pre-alteration adjustment}}. \tag{6}$$

The second KRR estimates outer CME of pre-alteration covariates as $\hat{\mu}_{\mathbf{U}|\mathbf{x}} = \sum_{j=1}^N \gamma_j(\mathbf{x}) k_u(\cdot, \mathbf{u}_j)$, with coefficients $\boldsymbol{\gamma}(\mathbf{x}) = (\mathbf{K}_{xx} + N\lambda_x \mathbf{I})^{-1}\mathbf{k}_x(\mathbf{x})$. Equivalently, this is vector-valued KRR with responses $k_u(\mathbf{u}_i, \cdot)$ and operator-valued kernel $k_x I_{\mathcal{H}_{k_u}}$ (Grünewälder et al., 2012). Leveraging this, we can explicitly expand $\int_{\mathcal{U}} k_u(\mathbf{u}_i, \mathbf{u}) p(\mathbf{u} \mid \mathbf{x})\, d\mathbf{u} = \big\langle k_u(\mathbf{u}_i, \cdot), \mu_{\mathbf{U}|\mathbf{x}} \big\rangle_{\mathcal{H}_{k_u}}$ ($\triangleq m_i(\mathbf{x})$) in Eq. (6) as:

$$\begin{aligned} \hat{m}_i(\mathbf{x}) = \big\langle k_u(\mathbf{u}_i, \cdot), \hat{\mu}_{\mathbf{U}|\mathbf{x}} \big\rangle_{\mathcal{H}_{k_u}} &= \left\langle k_u(\mathbf{u}_i, \cdot), \sum_{j=1}^{N} \gamma_j(\mathbf{x}) k_u(\cdot, \mathbf{u}_j) \right\rangle_{\mathcal{H}_{k_u}} \\ &= \sum_{j=1}^{N} \gamma_j(\mathbf{x}) \underbrace{\langle k_u(\mathbf{u}_i, \cdot), k_u(\mathbf{u}_j, \cdot)\rangle_{\mathcal{H}_{k_u}}}_{k_u(\mathbf{u}_i, \mathbf{u}_j)}. \end{aligned} \tag{7}$$

Let $\mathbf{K}_{uu} \in \mathbb{R}^{N\times N}$ denote the Gram matrix with entries $[\mathbf{K}_{uu}]_{ij} = k_u(\mathbf{u}_i, \mathbf{u}_j)$, Eq. (7) equals the i-th row of vector $\mathbf{K}_{uu}\boldsymbol{\gamma}(\mathbf{x})$. Substituting Eq. (7) into Eq. (6) yields the closed-form nested estimator of the smoothed AUF objective:

$$\begin{aligned} \hat{\mathbb{E}}_{\mathbf{u}\sim P(\cdot|\mathbf{x})}[\hat{\mathbb{E}}_{\mathbf{y}\sim P(\cdot|\mathbf{h})}[w_\eta(\mathbf{Y})]] &= \sum_{i=1}^{N} \alpha_i [\mathbf{K}_{uu}\boldsymbol{\gamma}(\mathbf{x})]_i k_x(\mathbf{x}_i, \mathbf{x}) k_a(\mathbf{a}_i, \mathbf{a}) \\ &= \left(\boldsymbol{\alpha} \odot \mathbf{k}_x(\mathbf{x}) \odot (\mathbf{K}_{uu}\boldsymbol{\gamma}(\mathbf{x}))\right)^\mathsf{T} \mathbf{k}_a(\mathbf{a}), \end{aligned}$$

Algorithm 1 Rehearsal Learning via Nested Estimator

Input: Observational data $\mathcal{D}$, context $\mathbf{x}$, region $\mathcal{S}$, feasible bounds $\mathbf{a}_{\text{left}}, \mathbf{a}_{\text{right}}$.
Hyperparameters: Kernel params $\{\sigma_h, \sigma_x, \sigma_u, \sigma_a\}$, regularization $\{\lambda_h, \lambda_x\}$, smoothing η.
// Phase 1: Offline Learning (One-time)
Compute Gram matrices $\mathbf{K}_{hh}, \mathbf{K}_{xx}, \mathbf{K}_{uu}$, and smoothed desirability vector $\mathbf{w}_\eta$ using Eq. (3).
Solve stage 1 coefs: $\boldsymbol{\alpha}^\mathsf{T} \leftarrow \mathbf{w}_\eta^\mathsf{T}(\mathbf{K}_{hh} + N\lambda_h \mathbf{I})^{-1}$, and stage 2 coefs: $\mathbf{G} \leftarrow (\mathbf{K}_{xx} + N\lambda_x \mathbf{I})^{-1}$.
// Phase 2: Context-specific Alteration (Per context $\mathbf{x}$)
Compute kernel vectors $\mathbf{k}_x(\mathbf{x})$.
Define $\boldsymbol{\gamma}(\mathbf{x}) \leftarrow \mathbf{G}\mathbf{k}_x(\mathbf{x})$, and $\boldsymbol{\omega}(\mathbf{x}) \leftarrow \boldsymbol{\alpha} \odot \mathbf{k}_x(\mathbf{x}) \odot (\mathbf{K}_{uu}\boldsymbol{\gamma}(\mathbf{x}))$.
// Phase 3: Multi-start Optimization
Set $\mathcal{S}_{\text{sols}} \leftarrow \emptyset$ and $\mathcal{A}_{\text{init}} \leftarrow$ up to Top-K indices i where $\omega_i(\mathbf{x}) > 0$.
for each $i \in \mathcal{A}_{\text{init}}$ **do**
  Initialize $\mathbf{a}^{(0)} \leftarrow \mathbf{a}_i$.
  for $t = 1$ to T **do**
    Compute gradient $\nabla_{\mathbf{a}} \hat{J}_\eta(\mathbf{a}^{(t-1)}; \mathbf{x})$ via Eq. (9).
    Update: $\mathbf{a}' \leftarrow \mathbf{a}^{(t-1)} + \nu \nabla_{\mathbf{a}} \hat{J}_\eta$ and Project: $\mathbf{a}^{(t)} \leftarrow \mathcal{P}_{\{\mathbf{a} | \mathbf{a}_{\text{left}} \preccurlyeq \mathbf{a} \preccurlyeq \mathbf{a}_{\text{right}}\}}(\mathbf{a}')$.
  end for
  Store solution: $\mathcal{S}_{\text{sols}} \leftarrow \mathcal{S}_{\text{sols}} \cup \{\mathbf{a}^{(T)}\}$.
end for
Output: $\mathbf{a}^* = \arg\max_{\mathbf{a} \in \mathcal{S}_{\text{sols}}} \hat{J}_\eta(\mathbf{a}; \mathbf{x})$.

where $\odot$ denotes Hadamard product. This is to say, let $\boldsymbol{\omega}(\mathbf{x}) = (\boldsymbol{\alpha} \odot \mathbf{k}_x(\mathbf{x}) \odot (\mathbf{K}_{uu}\boldsymbol{\gamma}(\mathbf{x})))$, the final surrogate of AUF objective simplifies to a linear function of the alteration's feature map in RKHS:

$$\hat{J}_\eta(\mathbf{a}; \mathbf{x}) = \boldsymbol{\omega}(\mathbf{x})^\mathsf{T} \mathbf{k}_a(\mathbf{a}). \tag{8}$$

Starting from the AUF adjustment formula, we derive Eq. (8) through scalar KRR and CME-based integration. Under matched kernels and ridge penalties, it coincides with the heterogeneous response estimator of Singh et al. (2024) for the same response. Applying this estimator to $\mathbb{I}(\mathbf{Y} \in \mathcal{S})$ yields the hard-indicator baseline; our smoothed response $w_\eta(\mathbf{Y})$ incorporates constraint-margin information. Whereas their analysis assumes RKHS membership, ours permits regression targets outside their RKHSs under fractional source and query-evaluation conditions. With separate hyperparameter selection under the same validation protocol, smoothing achieves higher mean AUF probabilities as shown in Sec. 4. Appx. B.1 details the correspondence and the comparison of theoretical assumptions.

3.4 Gradient-based Rehearsal Learning

Having derived the closed-form surrogate $\hat{J}_\eta(\mathbf{a}; \mathbf{x})$, our goal is to find the optimal alteration $\mathbf{a}^*$ that maximizes this objective subject to the feasible constraints:

$$\mathbf{a}^* = \arg\max_{\mathbf{a}} \hat{J}_\eta(\mathbf{a}; \mathbf{x}) \quad \text{s.t.} \quad \mathbf{a}_{\text{left}} \preccurlyeq \mathbf{a} \preccurlyeq \mathbf{a}_{\text{right}}.$$

We specify all kernels involved in our approach (*i.e.*, k_x, k_u, k_a) as Radial Basis Function (RBF) kernels. For a generic variable $\mathbf{v} \in \{\mathbf{x}, \mathbf{u}, \mathbf{a}\}$, the corresponding kernel is defined as:

$$k_v(\mathbf{v}, \mathbf{v}') = \exp\left(-\frac{\|\mathbf{v} - \mathbf{v}'\|_2^2}{2\sigma_v^2}\right),$$

where σ_v denotes the bandwidth parameter specific to each variable type. This specification naturally satisfies the decomposable assumption in Ass. 1 and provides the necessary smoothness for the subsequent gradient derivation. Under this specification, the objective function becomes a weighted sum of Gaussian bumps centered at historical actions $\mathbf{a}_i$:

$$\hat{J}_\eta(\mathbf{a}; \mathbf{x}) = \sum_{i=1}^{N} \omega_i(\mathbf{x}) \exp\left(-\frac{\|\mathbf{a}_i - \mathbf{a}\|_2^2}{2\sigma_a^2}\right).$$

Recall that the weight $\omega_i(\mathbf{x}) = [\boldsymbol{\alpha} \odot \mathbf{k}_x(\mathbf{x}) \odot (\mathbf{K}_{uu}\boldsymbol{\gamma}(\mathbf{x}))]_i$ encapsulates the confounding-adjusted desirability of the i-th historical sample given the current context $\mathbf{x}$. The coefficients $\omega_i(\mathbf{x})$ can be

negative, and the unconstrained KRR estimator $\hat{J}_\eta(\mathbf{a};\mathbf{x})$ need not lie in $0, 1$. We maximize this fitted objective and evaluate the selected alteration by its empirical success frequency.

For fixed data, the coefficients are independent of the query action, so gradients follow from the action kernel. By applying the chain rule, the gradient is given by:

$$\nabla_{\mathbf{a}}\hat{J}_\eta(\mathbf{a};\mathbf{x}) = \sum_{i=1}^{N} \omega_i(\mathbf{x})\nabla_{\mathbf{a}} k_a(\mathbf{a}_i,\mathbf{a}) = \frac{1}{\sigma_a^2}\sum_{i=1}^{N}\left[\omega_i(\mathbf{x})k_a(\mathbf{a}_i,\mathbf{a})\right](\mathbf{a}_i-\mathbf{a}). \tag{9}$$

This gradient interpretation is intuitive: it acts as a force vector that pulls the candidate $\mathbf{a}$ towards historical actions $\mathbf{a}_i$ that have high desirability weights $\omega_i(\mathbf{x})$, while being repelled by those with negative weights. Meanwhile, objective $\hat{J}_\eta(\mathbf{a};\mathbf{x})$ presents a potentially non-convex landscape since structured as a sum of Gaussians with mixed-sign weights $\omega_i(\mathbf{x})$, thus random initialization strategy risks entrapment in poor local optima. We initialize projected gradient ascent at up to K historical actions with the largest positive weights, using K uniformly sampled feasible actions when none are available. Specifically, we construct the candidate set $\mathcal{A}_{\text{init}}$ by selecting historical $\mathbf{a}_i$ whose corresponding weight values $\omega_i(\mathbf{x})$ are among top-K (up to total number of) positive entries in $\boldsymbol{\omega}(\mathbf{x})$:

$$\mathcal{A}_{\text{init}} = \{\mathbf{a}_i \in \mathcal{D} \mid \omega_i(\mathbf{x}) > 0 \text{ and is among top-}K\}.$$

Alg. 1 outlines the full decision-making procedure. It starts with an offline stage where kernel Gram matrices and corresponding KRR coefficients are computed from historical data. Given a newly observed context $\mathbf{x}$, the algorithm then computes the context-dependent mixing weights $\boldsymbol{\omega}(\mathbf{x})$, which are subsequently used to initialize a multi-start projected gradient ascent procedure that iteratively performs gradient ascent and projects the updates onto the feasible set, returning the best candidate alteration $\mathbf{a}^*$. With offline quantities cached, the dense implementation requires $\mathcal{O}(N^3)$ fitting time, $\mathcal{O}(N^2)$ storage, and $\mathcal{O}(N^2 + KTNd_a)$ time per context, where K, T, and d_a denote the number of restarts, iterations per restart, and action dimension, respectively.

3.5 Theoretical analysis

We first establish the approximation bound for the smooth desirability function as follows.

Theorem 1. *Let $\mathcal{S}$ be a nonempty bounded convex polyhedron defined by Eq. (2), with $\mathbf{m}_k \neq \mathbf{0}$ and Probit surrogate w_η from Eq. (3). If the conditional density $p(\mathbf{y} \mid \mathbf{x}, \mathring{\mathbf{a}})$ is bounded, then, as $\eta \to \infty$,*

$$\left|\mathbb{P}(\mathbf{Y} \in \mathcal{S} \mid \mathbf{x}, \mathring{\mathbf{a}}) - \mathbb{E}[w_\eta(\mathbf{Y}) \mid \mathbf{x}, \mathring{\mathbf{a}}]\right| = \mathcal{O}\left(\eta^{-1}\right).$$

Thm. 1 quantifies the approximation bias introduced by smoothed desirability response. Experiments show higher mean AUF probabilities with smoothing (Tab. 1). Building on scalar and vector-valued regression theory (Fischer & Steinwart, 2020; Li et al., 2024), we analyze two-stage estimation under fractional source and query-evaluation conditions. This gives pointwise consistency at fixed smoothing for targets that may lie outside their RKHSs, complementing the RKHS-based uniform guarantees of Singh et al. (2024). The following guarantee is pointwise at fixed smoothing; Appx. B.1 compares assumptions and provides a concrete AUF example satisfying our conditions with $\psi_\eta \notin \mathcal{H}_h$.

Theorem 2. *Fix $\eta > 0$, the kernels, and pairs $(\mathbf{x}, \mathbf{a})$. Under the regularity conditions in Appx. C.3, let $\rho, q > 0$ be the associated exponents, with $\rho + q < 1/2$. For deterministic $0 < \lambda_h, \lambda_x \leq 1$ with $N\lambda_h^2 \geq c_0 > 0$, the nested estimator satisfies:*

$$|\hat{J}_\eta(\mathbf{a};\mathbf{x}) - J_\eta(\mathbf{a};\mathbf{x})| = \mathcal{O}_p\left(\lambda_h^\rho + \lambda_h^{-q}\lambda_x^\rho + \frac{1}{\sqrt{N}\lambda_h} + \frac{1}{\sqrt{N}\lambda_x\lambda_h^q}\right).$$

In Thm. 2, ρ characterizes bias decay and q the growth of the inner population RKHS norm as regularization vanishes. The regularization schedules derived in Appx. C.3 yield $|\hat{J}_\eta(\mathbf{a};\mathbf{x}) - J_\eta(\mathbf{a};\mathbf{x})| = \mathcal{O}_p\left(N^{-\frac{\rho^2}{2(1+\rho)(\rho+q)}}\right)$, explicitly characterizing the learning speed of our nested estimator. This bound gives pointwise consistency of the smoothed objective at fixed context, action and η. For example, if $\rho = 1/3$ and $q = 1/12$, choosing $\lambda_h \asymp N^{-3/10}$ and $\lambda_x \asymp N^{-3/8}$ gives $\mathcal{O}_p(N^{-1/10})$. Cor. 1 further bounds AUF regret by twice the uniform smoothing and estimation errors plus global optimization error. Decision consistency further requires these errors to vanish; while the proposed multi-start gradient ascent does not guarantee global optimization accuracy as the target could be non-convex.

4 EXPERIMENTS

In this section, we present empirical results validating the effectiveness of our proposed method. Detailed data generation procedures and benchmark construction details are provided in Appx. D.

4.1 EXPERIMENTAL SETUP

We evaluate our method against rehearsal learning baselines, including QWZ23 (Qin et al., 2023), MICNS (Du et al., 2024), CARE (Du et al., 2025), and Grad-Rh (Qin et al., 2025). It is worth noting that CARE is theoretically optimal for linear additive Gaussian data, and Grad-Rh is a heuristic approach compatible with non-linear scenarios. All experiments were conducted on an NVIDIA Tesla A100 GPU and two Intel Xeon Platinum 8358 CPUs.

All methods use exactly the same observational training set, containing $1{,}000$ samples. For our method, we use 80% of this training set for model fitting and the remaining 20% for hyperparameter selection, and then fix the selected hyperparameters and refit our estimator on the entire training set, ensuring that no additional training data are used. All methods are evaluated on with the same seeds and $1{,}000$ Monte Carlo draws per seed. We report the mean AUF probability across these seeds. Selection criteria and implementation details are provided in Appx. D.1 and the accompanying code. We also evaluate a hard-indicator variant that replaces the Probit response in both validation-stage fitting and final estimation. Each variant selects its own hyperparameters using the same validation criteria, candidate ranges for bandwidths and regularization, with all other settings unchanged.

4.2 DATASETS

Linear settings. We utilize the Bermuda dataset (Courtney et al., 2017; Andersson & Bates, 2018) that records environmental variables in the Bermuda area, with the objective of maintaining a high net coral ecosystem calcification (NEC) level. The data-generating graph structure is known (Courtney et al., 2017), and the parameters are derived by fitting linear models. The dimensions of $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{Y}$ are 3, 7, and 1, respectively, and the desired region is defined as $\mathcal{S} = \{\text{NEC} \in [0.5, 2]\}$. Additionally, we employ the synthetic dataset consistent with Du et al. (2025), where the dimensions of $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{Y}$ are 2, 4, and 2, respectively, and the desired region $\mathcal{S}$ for $\mathbf{Y}$ is a rectangle area.

Nonlinear settings. We first extract a real-data-derived semi-synthetic diabetes benchmark from NHANES 2011–2018 (National Center for Health Statistics, 2018), using continuous actionable lifestyle / metabolic variables and two glycemic outcomes, HbA1c and fasting plasma glucose (FPG). Its desired region requires both outcomes to fall within normal glycemic ranges, $\mathcal{S}_{\text{NHANES}} = \{(\text{HbA1c}, \text{FPG}) \mid 0 \leq \text{HbA1c} \leq 5.7,\ 0 \leq \text{FPG} \leq 100 \text{ mg/dL}\}$, where HbA1c reflects longer-term blood-glucose exposure and FPG measures fasting blood glucose; together they characterize clinically meaningful glycemic status. We also implement the motivating loan approval example illustrated in Fig. 1. To facilitate visualization, we focus on altering solely on the interest rate (A_2). The objective is to ensure a high repayment rate (Y_1) while maintaining a favorable return on ROI (Y_2), thus the desired region is defined as $\mathcal{S} = \{(Y_1, Y_2) \mid 0.6 \leq Y_1 \leq 1.0, 0.3 \leq Y_2\}$. Furthermore, to evaluate performance under complex conditions, we implement several synthetic datasets featuring non-linear mechanisms (including square, trigonometric, and Sigmoid transformations) and non-Gaussian randomness (including uniform, beta, and exponential distributions). Appx. D.6 provides the construction details of the NHANES benchmark.

4.3 RESULTS

Experimental results are presented in Tab. 1. Baseline results denote the natural AUF probability, *i.e.*, $\mathbb{P}(\mathbf{Y} \in \mathcal{S} \mid \mathbf{x})$, observed without any alterations. In the completed linear additive Gaussian settings, our method outperforms all rehearsal learning baselines except CARE. This performance gap is expected, as CARE is theoretically proven to be optimal under these specific conditions. However, it is crucial to highlight that CARE (and most other baselines, excluding Grad-Rh) know the true parametric family in this case, in contrast, our approach is non-parametric thus does not use prior knowledge of the underlying functional forms.

In the nonlinear lending and synthetic settings, our method achieves the highest mean AUF probability among the methods reported in Tab. 1. On NHANES, all adapted action methods improve over the

Table 1: AUF probability $\mathbb{P}(\mathbf{Y} \in \mathcal{S} \mid \mathbf{x}, \mathring{\mathbf{a}})$ (mean $\pm$ standard deviation over seeds, with 1,000 Monte Carlo draws per seed). Bold marks the highest mean.

Dataset	No action	QWZ23	MICNS	CARE	Grad-Rh	Hard indicator	Ours
Bermuda	0.222 $\pm$ 0.032	0.676 $\pm$ 0.027	0.668 $\pm$ 0.079	**0.706 $\pm$ 0.065**	0.674 $\pm$ 0.030	0.667 $\pm$ 0.032	0.679 $\pm$ 0.025
Lin-Syn1	0.166 $\pm$ 0.160	0.938 $\pm$ 0.032	0.912 $\pm$ 0.038	**0.956 $\pm$ 0.019**	0.934 $\pm$ 0.022	0.881 $\pm$ 0.018	0.944 $\pm$ 0.007
NHANES	0.402 $\pm$ 0.257	0.562 $\pm$ 0.217	0.536 $\pm$ 0.195	0.526 $\pm$ 0.211	0.426 $\pm$ 0.203	0.533 $\pm$ 0.211	**0.615 $\pm$ 0.142**
BankExp	0.576 $\pm$ 0.136	0.563 $\pm$ 0.019	0.576 $\pm$ 0.014	0.640 $\pm$ 0.054	0.698 $\pm$ 0.055	0.769 $\pm$ 0.066	**0.807 $\pm$ 0.063**
Non-Syn1	0.064 $\pm$ 0.014	0.130 $\pm$ 0.083	0.114 $\pm$ 0.091	0.078 $\pm$ 0.058	0.291 $\pm$ 0.107	0.402 $\pm$ 0.032	**0.415 $\pm$ 0.040**
Non-Syn2	0.020 $\pm$ 0.040	0.030 $\pm$ 0.040	0.266 $\pm$ 0.284	0.010 $\pm$ 0.020	0.222 $\pm$ 0.210	0.263 $\pm$ 0.285	**0.293 $\pm$ 0.296**

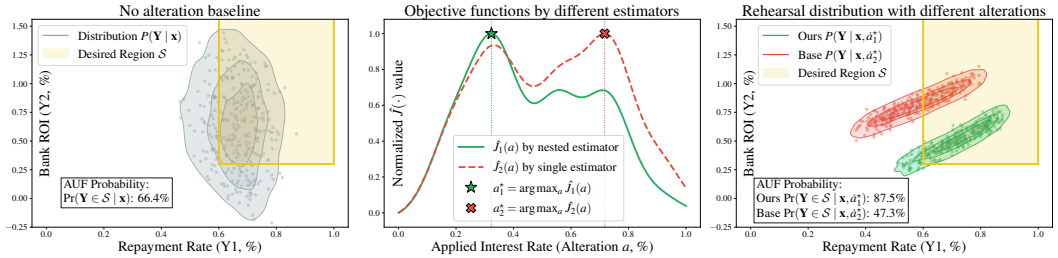

Figure 3: Illustrative example of loan interest rate optimization as shown in Fig. 1. The left panel shows the distribution conditioned on context $\mathbf{x}$ without active alterations. The middle panel shows normalized estimated objectives and the actions selected by each estimator. The right panel illustrates the rehearsal distributions under the alterations optimized by the two estimators.

no-action baseline on average under the fitted NHANES generator, and our non-parametric adapter obtains the highest mean AUF probability. This result highlights the potential of rehearsal learning for real-world medical decision problems. Using the same validation protocol for both variants, smoothing yields higher mean AUF probabilities on all six benchmarks. To illustrate the necessity of optimizing the alterational objective $\mathbb{E}_{\mathbf{y}\sim P(\cdot|\mathbf{x},\mathring{\mathbf{a}})}[w_\eta(\mathbf{Y})]$ (identifiable as in Prop. 1), rather than directly maximizing $\mathbb{E}_{\mathbf{y}\sim P(\cdot|\mathbf{x},\mathbf{a})}[w_\eta(\mathbf{Y})]$ (pure conditional), we present a comparative analysis on the lending task in Fig. 1. Initially, distribution $P(\mathbf{Y} \mid \mathbf{x})$ shows a baseline probability of 66.4% within the desired region $\mathcal{S}$ without active alterations. Driven by the goal to actively perform feasible alterations to increase the probability of outcomes falling into $\mathcal{S}$, we compare optimal alterations suggested by maximizing the two estimators. Maintaining consistent hyperparameters (*e.g.*, kernel bandwidth), the landscape estimated by single estimator ($\hat{\mathbb{E}}_{\mathbf{y}\sim P(\cdot|\mathbf{x},\mathbf{a})}[w_\eta(\mathbf{Y})]$) is misled by the unactionable industry stability U that influences both $\mathbf{A}$ and $\mathbf{Y}$, suggesting a suboptimal action a_2^*. In this example, the nested and purely conditional estimators select a_1^* and a_2^*, respectively; success probabilities are evaluated under post-alteration distributions. Specifically, the resulting rehearsal distributions confirm that the misled alteration yields a detrimental effect (AUF probability 47.3%), performing even worse than the no-alteration baseline; while our method successfully shifts outcomes into $\mathcal{S}$ with an AUF probability of 87.5%, illustrating the importance of confounder adjustment for AUF action selection for optimizing $\mathbb{E}_{\mathbf{y}\sim P(\cdot|\mathbf{x},\mathring{\mathbf{a}})}[w_\eta(\mathbf{Y})]$.

5 Conclusion

In this paper, we presented the first non-parametric rehearsal learning approach for the AUF problem, addressing the key limitations of existing rehearsal-based methods that rely on restrictive parametric assumptions. By avoiding specific functional forms of the data generation process, our approach extends rehearsal learning to nonlinear and non-additive settings. Methodologically, we introduced a smooth Probit approximation of the desirability indicator and estimated the resulting AUF objective using nested KRR. This estimator enables direct and effective learning of the AUF objective from observational data, grounded in the theory of CMEs. Theoretically, we established identifiability, bounded Probit approximation, and proved pointwise consistency for the smoothed objective at fixed η under stated spectral conditions. Empirically, we evaluated the method on synthetic and real-data-derived semi-synthetic benchmarks and added a NHANES diabetes benchmark to stress-test continuous actionable variables and two-dimensional clinical success regions.

AI USE STATEMENT

In this work, we used generative AI tools to polish the language and improve the readability of the manuscript, and to assist with literature retrieval and the identification of relevant prior work. We reviewed all AI-assisted revisions and checked the identified references against the original sources for accuracy and relevance. We take responsibility for the final content of this work, including text, claims, and references.

REFERENCES

Virginia Aglietti, Xiaoyu Lu, Andrei Paleyes, and Javier González. Causal Bayesian optimization. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pp. 3155–3164, 2020.

American Diabetes Association Professional Practice Committee. 2. diagnosis and classification of diabetes: Standards of care in diabetes–2024. *Diabetes Care*, 47(Supplement 1):S20–S42, 2024. doi: 10.2337/dc24-S002.

Andreas Andersson and Nicholas Bates. In situ measurements used for coral and reef-scale calcification structural equation modeling including environmental and chemical measurements, and coral calcification rates in bermuda from 2010 to 2012 (BEACON project), 2018. http://lod.bco-dmo.org/id/dataset/720788.

Lynton Ardizzone, Carsten Lüth, Jakob Kruse, Carsten Rother, and Ullrich Köthe. Guided image generation with conditional invertible neural networks. *arXiv preprint arXiv:1907.02392*, 2019.

Elias Bareinboim, Andrew Forney, and Judea Pearl. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems 28*, pp. 1342–1350, 2015.

David A Bruns-Smith. Model-free and model-based policy evaluation when causality is uncertain. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 1116–1126, 2021.

Ruichu Cai, Jie Qiao, Kun Zhang, Zhenjie Zhang, and Zhifeng Hao. Causal discovery from discrete data using hidden compact representation. In *Advances in Neural Information Processing Systems 31*, pp. 2666–2674, 2018.

Travis A Courtney, Mario Lebrato, Nicholas R Bates, Andrew Collins, Samantha J De Putron, Rebecca Garley, Rod Johnson, Juan-Carlos Molinero, Timothy J Noyes, Christopher L Sabine, et al. Environmental controls on modern scleractinian coral and reef-scale calcification. *Science Advances*, 3(11):e1701356, 2017.

Wen-Bo Du, Tian Qin, Tian-Zuo Wang, and Zhi-Hua Zhou. Avoiding undesired future with minimal cost in non-stationary environments. In *Advances in Neural Information Processing Systems 37*, pp. 135741–135769, 2024.

Wen-Bo Du, Hao-Yi Lei, Lue Tao, Tian-Zuo Wang, and Zhi-Hua Zhou. Enabling optimal decisions in rehearsal learning under CARE condition. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 14536–14561, 2025.

Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21(205):1–38, 2020. URL https://www.jmlr.org/papers/v21/19-734.html.

Steffen Grünewälder, Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, and Massimilano Pontil. Conditional mean embeddings as regressors. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 1803–1810, 2012.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1861–1870, 2018.

YangBo He and Zhi Geng. Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 9(84):2523–2547, 2008.

Nathan Kallus and Angela Zhou. Confounding-robust policy evaluation in infinite-horizon reinforcement learning. In *Advances in Neural Information Processing Systems 33*, pp. 22293–22304, 2020.

Chinmaya Kausik, Yangyi Lu, Kevin Tan, Maggie Makar, Yixin Wang, and Ambuj Tewari. Offline policy evaluation and optimization under confounding. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, pp. 1459–1467, 2024.

Murat Kocaoglu, Alex Dimakis, and Sriram Vishwanath. Cost-optimal learning of causal graphs. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1875–1884, 2017.

Finnian Lattimore, Tor Lattimore, and Mark D. Reid. Causal bandits: Learning good interventions via causal inference. In *Advances in Neural Information Processing Systems 29*, pp. 1181–1189, 2016.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Sanghack Lee and Elias Bareinboim. Structural causal bandits: Where to intervene? In *Advances in Neural Information Processing Systems 31*, pp. 2573–2583, 2018.

Sanghack Lee and Elias Bareinboim. Characterizing optimal mixed policies: Where to intervene and what to observe. In *Advances in Neural Information Processing Systems 33*, pp. 8565–8576, 2020.

Jiachen Li, Edwin Zhang, Ming Yin, Qinxun Bai, Yu-Xiang Wang, and William Yang Wang. Offline reinforcement learning with closed-form policy improvement operators. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 20485–20528, 2023.

Zhu Li, Dimitri Meunier, Mattes Mollenhauer, and Arthur Gretton. Towards optimal Sobolev norm rates for the vector-valued regularized least-squares algorithm. *Journal of Machine Learning Research*, 25(181):1–51, 2024. URL https://www.jmlr.org/papers/v25/23-1663.html.

Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Causal bandits with unknown graph structure. In *Advances in Neural Information Processing Systems 34*, pp. 24817–24828, 2021.

Krikamol Muandet, Motonobu Kanagawa, Sorawit Saengkyongam, and Sanparith Marukatat. Counterfactual mean embeddings. *Journal of Machine Learning Research*, 22(162):1–71, 2021.

National Center for Health Statistics. National health and nutrition examination survey data, 2011–2018, 2018. https://www.cdc.gov/nchs/nhanes/.

Min Woo Park and Sanghack Lee. On transportability for structural causal bandits. *CoRR*, abs/2511.17953, 2025.

Min Woo Park, Andy Arditi, Elias Bareinboim, and Sanghack Lee. Structural causal bandits under markov equivalence. In *Advances in Neural Information Processing Systems 38*, pp. 109621––109669, 2025.

Judea Pearl. *Causality: Models, Reasoning and Inference, 2nd Edition*. Cambridge University Press, 2009.

Vitchyr H. Pong, Ashvin Nair, Laura Smith, Catherine Huang, and Sergey Levine. Offline meta-reinforcement learning with online self-supervision. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 17811–17829, 2022.

Dan Qiao and Yu-Xiang Wang. Offline reinforcement learning with differential privacy. In *Advances in Neural Information Processing Systems 36*, pp. 61395–61436, 2023.

Tian Qin, Tian-Zuo Wang, and Zhi-Hua Zhou. Rehearsal learning for avoiding undesired future. In *Advances in Neural Information Processing Systems 36*, pp. 80517–80542, 2023.

Tian Qin, Tian-Zuo Wang, and Zhi-Hua Zhou. Gradient-based nonlinear rehearsal learning with multivariate alterations. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, pp. 26859–26867, 2025.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pp. 1177–1184, 2007.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.

Rajat Sen, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Identifying best interventions through online importance sampling. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3057–3066, 2017.

Kirankumar Shiragur, Jiaqi Zhang, and Caroline Uhler. Causal discovery with fewer conditional independence tests. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 45060–45078, 2024.

Rahul Singh, Liyuan Xu, and Arthur Gretton. Kernel methods for causal functions: dose, heterogeneous and incremental response curves. *Biometrika*, 111(2):497–516, 2024. doi: 10.1093/biomet/asad042.

Rahul Singh, Liyuan Xu, and Arthur Gretton. Sequential kernel embedding for mediated and time-varying dose response curves. *Bernoulli*, 31(4):3013–3033, 2025. doi: 10.3150/24-BEJ1836.

Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th International Conference on Machine Learning*, pp. 961–968, 2009.

Le Song, Kenji Fukumizu, and Arthur Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.

Yuda Song, Yifei Zhou, Ayush Sekhari, J. Andrew Bagnell, Akshay Krishnamurthy, and Wen Sun. Hybrid RL: Using both offline and online data can make RL efficient. In *Proceedings of the 11th International Conference on Learning Representations*, 2023.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2000.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction, 2nd Edition*. MIT Press, 2018.

Lue Tao, Tian-Zuo Wang, Yuan Jiang, and Zhi-Hua Zhou. Avoiding undesired future with sequential decisions. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence*, pp. 6245–6253, 2025.

Yu-Xuan Tao, Tian-Zuo Wang, and Zhi-Hua Zhou. Order-based rehearsal learning. *arXiv preprint arXiv:2605.04955*, 2026.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018.

Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*, pp. 682–688, 2000.

Christina Winkler, Daniel Worrall, Emiel Hoogeboom, and Max Welling. Learning likelihoods with conditional normalizing flows. *CoRR*, abs/1912.00042, 2019.

Houssam Zenati, Bariscan Bozkurt, and Arthur Gretton. Doubly-robust estimation of counterfactual policy mean embeddings. In *Advances in Neural Information Processing Systems 38*, pp. 57044—-57100, 2025.

Junzhe Zhang and Elias Bareinboim. Near-optimal reinforcement learning in dynamic treatment regimes. In *Advances in Neural Information Processing Systems 32*, pp. 13401–13411, 2019.

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference in Uncertainty in Artificial Intelligence*, pp. 647–655, 2009.

Kun Zhang, Biwei Huang, Jiji Zhang, Clark Glymour, and Bernhard Schölkopf. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 1347–1353, 2017.

Zhi-Hua Zhou. Rehearsal: Learning from prediction to decision. *Frontiers of Computer Science*, 16 (4):1–3, 2022.

A Discussion

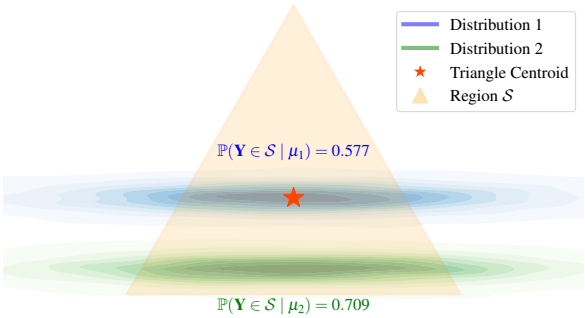

Figure 4: A counter-example showing the misalignment between distance minimization and true AUF objective.

Prior rehearsal learning methodologies have been predominantly confined to the assumption that the underlying generation mechanisms follow linear additive noise models. Qin et al. (2025) use conditional normalizing flows for nonlinear generation mechanisms and select alterations through distance-based surrogate objectives. To circumvent the intractability of directly optimizing the AUF probability, they aggregate losses based on post-alteration samples' distances to the Chebyshev center of $\mathcal{S}$. For illustration, consider the simpler mean-distance heuristic $\|\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{x},\mathring{\mathbf{a}}} - \mathbf{c}\|_2$, where $\mathbf{c}$ denotes the centroid of region $\mathcal{S}$. This objective can differ from the AUF objective (maximizing $\mathbb{P}(\mathbf{Y} \in \mathcal{S} \mid \mathbf{x}, \mathring{\mathbf{a}})$). As illustrated in Fig. 4 ($\mathcal{S}$ is an equilateral triangle, a common polytope), the strategy minimizing this mean-distance heuristic (blue) perfectly aligns the distribution mean with the centroid ($\|\boldsymbol{\mu}_1 - \mathbf{c}\| = 0$) yet only yields a success probability of 0.577. Conversely, an alternative strategy (green) does not align the distribution mean with the centroid ($\|\boldsymbol{\mu}_2 - \mathbf{c}\| > 0$) but achieves a higher probability of 0.709. This demonstrates that the mean-distance surrogate can result in suboptimal decisions, diverging from the goal of AUF.

We instead estimate the Probit-smoothed AUF objective by scalar KRR on $w_\eta(\mathbf{Y})$ followed by CME-based adjustment over $\mathbf{U} \mid \mathbf{X}$. Thm. 1 bounds the smoothing error for bounded polyhedral regions under bounded conditional density, while Thm. 2 gives pointwise consistency at fixed η under the stated spectral conditions. Response sharpness, kernel bandwidths, and ridge penalties jointly determine the fitted objective. The separately tuned response comparisons in Appx. D.5 select these parameters through observational validation against a common binary success target. This criterion compares response families on a common prediction scale, while subsequent generator evaluation measures the success of their selected alterations.

B Related Work

Rehearsal learning. Rehearsal learning exploits influence relations among variables to support decision-making that optimizes the AUF probability (Zhou, 2022). Existing rehearsal approaches adopt parametric formulations, relying on assumptions like linear additive systems to derive probabilistic constraints (Qin et al., 2023; Du et al., 2024; Tao et al., 2025) or tractable objectives (Du et al., 2025). Qin et al. (2025) extend the parametric paradigm to nonlinear settings using conditional normalizing flows (Winkler et al., 2019; Ardizzone et al., 2019) to estimate generation mechanisms, but replace AUF objective with heuristic surrogates. The distance objective can select a different alteration from AUF probability maximization; Appx. A gives a concrete example. Our approach estimates a smoothed desirability expectation with an explicit approximation bound.

Reinforcement Learning. RL has demonstrated strong empirical performance across various decision-making problems (Sutton & Barto, 2018). Online algorithms (Schulman et al., 2017; Haarnoja et al., 2018) use repeated environment interaction, while offline and hybrid approaches exploit previously collected data (Song et al., 2023; Pong et al., 2022; Li et al., 2023; Qiao & Wang, 2023). Offline policy evaluation must also account for variables influencing both observed actions and subsequent outcomes (Kallus & Zhou, 2020; Bruns-Smith, 2021; Kausik et al., 2024). The AUF task considered here selects a single context-dependent alteration from observational data and evaluates a specified outcome event.

Influence structure and response estimation. A substantial body of work has investigated the use of structural models for decision-making, with most approaches rooted in the framework of structural causal models (SCMs) (Spirtes et al., 2000). Several studies focus on system identification (He & Geng, 2008; Zhang & Hyvärinen, 2009; Kocaoglu et al., 2017; Zhang et al., 2017; Cai et al., 2018; Shiragur et al., 2024). However, these methods typically emphasize recovering influence structure,

which supplies information for downstream estimation and alteration selection. Additionally, kernel mean embeddings have been used to estimate response curves and counterfactual distributions (Singh et al., 2024; Muandet et al., 2021; Zenati et al., 2025). Singh et al. (2024) derive closed-form KRR estimators for dose, heterogeneous, and incremental response curves. Under matched kernels and regularization, their heterogeneous response estimator coincides with Eq. (8) when the response is replaced by $w_\eta(\mathbf{Y})$. Our study focuses on AUF decision-making, combining desirability approximation with feasible, context-dependent alteration selection to maximize the probability of reaching a specified outcome region. In parallel, bandit methods exploit influence graphs to identify reward-maximizing actions (Bareinboim et al., 2015; Lattimore et al., 2016; Sen et al., 2017; Lee & Bareinboim, 2018; Zhang & Bareinboim, 2019; Lee & Bareinboim, 2020; Lu et al., 2021; Park et al., 2025; Park & Lee, 2025). Our formulation conditions alteration selection on the observed context and defines desirability through a specified region of the outcome space.

B.1 Relationship to kernel response estimation

Our goal is to select feasible, context-dependent actions that maximize the AUF probability. Under matched kernels and ridge penalties, our hard-indicator estimator coincides with the heterogeneous response estimator of Singh et al. (2024) applied to $\mathbb{I}(\mathbf{Y} \in \mathcal{S})$. For the same bounded response and identification setup, Singh et al. (2024) impose RKHS source conditions with $c, c_2 \in (1, 2]$ and spectral-decay conditions, obtaining uniform consistency. Ass. 2 instead combines fractional source conditions with explicit query-evaluation bounds and compatible conditional versions, yielding pointwise consistency at fixed smoothing. This extends the covered source regime without asserting that the complete assumption set is uniformly weaker.

A concrete AUF example. Generate observations using independent exogenous noises:

$$X \sim \text{Bernoulli}(1/2), \qquad U \mid X = x \sim \text{Bernoulli}(1/4 + x/2), \qquad A \mid x, u \sim N(x + u, 1).$$

Let $p_{xu}(a) = \frac{1}{4} + \frac{1}{2}e^{-5(a-x-u)^2/4}$, $B \mid x, u, a \sim \text{Bernoulli}(p_{xu}(a))$, and set $Y = 2(1 - B) + \varepsilon$, where $\varepsilon \sim \text{Unif}[-1/4, 1/4]$ is independent. For $\mathcal{S} = [-1/2, 1/2]$ and feasible actions $[-1, 3]$, success is equivalent to $B = 1$. Here U affects both action assignment and outcomes. Use equality kernels for X, U and $k_a(a, a') = e^{-(a-a')^2/2}$. For every fixed $\eta > 0$,

$$\psi_\eta(x, u, a) = C_\eta + D_\eta e^{-5(a-x-u)^2/4}, \qquad C_\eta, D_\eta > 0.$$

This target is outside $\mathcal{H}_h$: its action sections have a nonzero limit at infinity, whereas Gaussian RKHS functions vanish there. Even after removing the constant, the Gaussian Fourier criterion excludes the remaining bump from $\mathcal{H}_a$.

Nevertheless, Ass. 2 holds with $\alpha = 1/4$ and $\beta = 3/4$. Within each group, the Gaussian–Hermite eigenvalues are $\tau_j = (1 - r)r^j$, where $r = (3 - \sqrt{5})/2$. Target coefficients satisfy $b_{2m+1} = 0$ and:

$$|b_{2m}|^2 \lesssim \frac{r^{2m} + s^{2m}}{\sqrt{m+1}}, \qquad s = \frac{6 - \sqrt{5}}{6 + \sqrt{5}}.$$

Since $s^2/r^{3/2} < 1$, the fractional source sum $\sum_j |b_j|^2/\tau_j^{3/4}$ is finite. Mehler's formula gives the query bound uniformly on the compact action interval; the finitely many group weights only change constants. The outer conditions hold because X, U are finite-valued, and continuity ensures compatibility of the source expansions with the stated conditional versions.

For AUF, we introduce a Probit desirability response that retains constraint-margin information beyond binary success labels. With hyperparameters selected separately under the same validation protocol, smoothing achieves higher mean AUF probabilities on all six benchmarks (Tab. 1). Appx. D.5 provides additional controlled comparisons across sample sizes and tuning protocols. These results support the practical value of smoothing for AUF decision-making.

C Proofs

In this section, we provide proof for claims in the main text.

C.1 Proof of Proposition 1

We specify the conditions underlying Prop. 1 and prove the adjustment identity. Fix $\eta > 0$. The alteration $\mathring{\mathbf{a}}$ jointly sets $\mathbf{A} = \mathbf{a}$ while leaving the remaining structural mechanisms unchanged. In particular, $(\mathbf{X}, \mathbf{U})$ are unaffected by the alteration.

Identification conditions. The following conditions formalize conditional exchangeability, consistency, and overlap (Pearl, 2009; Singh et al., 2024).

(i) **Conditional exchangeability.** For each candidate action $\mathbf{t}$, $\mathbf{Y}(\mathbf{t}) \perp\!\!\!\perp \mathbf{A} \mid (\mathbf{X}, \mathbf{U})$. For continuous actions, this condition holds throughout the neighborhood of query considered below.

(ii) **Consistency.** Potential outcomes refer to the same structural mechanisms as the observed outcome, with only the joint action assignment replaced. Thus $\mathbf{Y} = \mathbf{Y}(\mathbf{A})$ almost surely.

(iii) **Conditional overlap.** For $P(\mathbf{U} \mid \mathbf{x})$-almost every $\mathbf{u}$, a discrete action query satisfies

$$P(\mathbf{A} = \mathbf{a} \mid \mathbf{X} = \mathbf{x}, \mathbf{U} = \mathbf{u}) > 0.$$

For a continuous action query, the conditional action law has a Lebesgue density that is positive almost everywhere in an open neighborhood of $\mathbf{a}$. This neighborhood may depend on $(\mathbf{x}, \mathbf{u})$.

Continuous-query regularity. Write

$$q_\eta(\mathbf{x}, \mathbf{u}, \mathbf{t}) = \mathbb{E}[w_\eta(\mathbf{Y}(\mathbf{t})) \mid \mathbf{X} = \mathbf{x}, \mathbf{U} = \mathbf{u}],$$
$$r_\eta(\mathbf{x}, \mathbf{u}, \mathbf{t}) = \mathbb{E}[w_\eta(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}, \mathbf{U} = \mathbf{u}, \mathbf{A} = \mathbf{t}].$$

For continuous actions, we use the following sufficient regularity conditions on the overlap neighborhood: conditional on $(\mathbf{X}, \mathbf{U}) = (\mathbf{x}, \mathbf{u})$, the potential-response process $\mathbf{t} \mapsto w_\eta(\mathbf{Y}(\mathbf{t}))$ has almost surely continuous sample paths, and $r_\eta(\mathbf{x}, \mathbf{u}, \cdot)$ admits a continuous version. Because $0 \le w_\eta \le 1$, bounded convergence also implies continuity of $q_\eta(\mathbf{x}, \mathbf{u}, \cdot)$. We use these versions for pointwise evaluation. The conditional-context statement initially holds for $P_{\mathbf{X}}$-almost every $\mathbf{x}$. At a prespecified continuously valued context, we additionally require counterfactual mean and adjusted observational mean in the proposition to equal their continuous extensions from a supported neighborhood of contexts satisfying the preceding conditions. The regularity requirements concern generating process and conditional responses; smoothness of fitted kernel estimator alone does not establish them.

Proof of Prop. 1. First consider a discrete action with positive conditional mass. Conditional exchangeability and consistency yield:

$$\begin{aligned} q_\eta(\mathbf{x}, \mathbf{u}, \mathbf{a}) &= \mathbb{E}[w_\eta(\mathbf{Y}(\mathbf{a})) \mid \mathbf{X} = \mathbf{x}, \mathbf{U} = \mathbf{u}, \mathbf{A} = \mathbf{a}] \\ &= \mathbb{E}[w_\eta(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}, \mathbf{U} = \mathbf{u}, \mathbf{A} = \mathbf{a}] = r_\eta(\mathbf{x}, \mathbf{u}, \mathbf{a}). \end{aligned}$$

For continuous actions, fix an admissible $(\mathbf{x}, \mathbf{u})$. Let h be a bounded measurable function supported in the overlap neighborhood, and let $\mathbf{A}_n$ be countably valued functions of $\mathbf{A}$ converging to $\mathbf{A}$ within that neighborhood. Consistency, conditional exchangeability, and bounded convergence give:

$$\begin{aligned} \mathbb{E}[h(\mathbf{A}) w_\eta(\mathbf{Y}) \mid \mathbf{x}, \mathbf{u}] &= \lim_{n \to \infty} \mathbb{E}[h(\mathbf{A}) w_\eta(\mathbf{Y}(\mathbf{A}_n)) \mid \mathbf{x}, \mathbf{u}] \\ &= \lim_{n \to \infty} \mathbb{E}[h(\mathbf{A}) q_\eta(\mathbf{x}, \mathbf{u}, \mathbf{A}_n) \mid \mathbf{x}, \mathbf{u}] \\ &= \mathbb{E}[h(\mathbf{A}) q_\eta(\mathbf{x}, \mathbf{u}, \mathbf{A}) \mid \mathbf{x}, \mathbf{u}]. \end{aligned}$$

Since this identity holds for every such h, the definition of conditional expectation implies that $r_\eta(\mathbf{x}, \mathbf{u}, \mathbf{t}) = q_\eta(\mathbf{x}, \mathbf{u}, \mathbf{t})$ holds for conditional-action-almost every $\mathbf{t}$ in the neighborhood. Positive conditional density makes the equality hold Lebesgue-almost everywhere there. Continuity of both functions therefore extends it to the specified query $\mathbf{a}$.

Since $(\mathbf{X}, \mathbf{U})$ are unaffected by the alteration, iterated expectation now gives:

$$\begin{aligned} \mathbb{E}[w_\eta(\mathbf{Y}) \mid \mathbf{x}, \mathring{\mathbf{a}}] = \mathbb{E}[w_\eta(\mathbf{Y}(\mathbf{a})) \mid \mathbf{X} = \mathbf{x}] &= \int q_\eta(\mathbf{x}, \mathbf{u}, \mathbf{a}) \, dP(\mathbf{u} \mid \mathbf{x}) \\ &= \int r_\eta(\mathbf{x}, \mathbf{u}, \mathbf{a}) \, dP(\mathbf{u} \mid \mathbf{x}) = \mathbb{E}_{\mathbf{u} \sim P(\cdot \mid \mathbf{x})} \left[\mathbb{E}[w_\eta(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}, \mathbf{U} = \mathbf{u}, \mathbf{A} = \mathbf{a}] \right]. \end{aligned}$$

The stipulated continuous extensions preserve this identity at a prespecified continuous context, completing the proof. Note that post-alteration variables $\mathbf{U}^{\text{post}}$ are marginalized within the conditional outcome response. They are not included in the adjustment set because they may be affected by $\mathbf{A}$ and mediate its effect on $\mathbf{Y}$. □

Scope. The proposition establishes identification only at queries satisfying the stated conditions. For identification throughout a feasible action box, these conditions must hold at every action in that box. Physical feasibility alone does not establish conditional overlap, and a kernel prediction outside observational support is an extrapolation rather than an identified causal value. Even within support, weak overlap can impair finite-sample estimation; estimation guarantees require the additional regularity conditions stated in the subsequent analysis. When covariates and actions interleave, including covariates affected by earlier actions, identification instead requires sequential g-formula under consistency, sequential exchangeability, and positivity (Singh et al., 2025). Sequential kernel embeddings provide a natural extension to that setting, which is beyond present single-stage analysis.

C.2 Proof of Theorem 1

Constraint normalization. Let $(\bar{\mathbf{m}}_k, \bar{b}_k)$ denote supplied coefficients and $\mathbf{D} = \mathrm{diag}(s_1, \ldots, s_{d_y})$, with $s_j > 0$, specify fixed reference scales for the outcomes. We use

$$\mathbf{m}_k = \frac{\bar{\mathbf{m}}_k}{\|\mathbf{D}\bar{\mathbf{m}}_k\|_2}, \qquad b_k = \frac{\bar{b}_k}{\|\mathbf{D}\bar{\mathbf{m}}_k\|_2}.$$

Thus $b_k - \mathbf{m}_k^{\mathsf{T}}\mathbf{y}$ is the signed constraint margin in the scaled outcome coordinates $\mathbf{D}^{-1}\mathbf{y}$. Replacing $(\bar{\mathbf{m}}_k, \bar{b}_k)$ by $(c\bar{\mathbf{m}}_k, c\bar{b}_k)$, $c > 0$, leaves this margin and w_η unchanged. Under a change of units $\mathbf{y}' = \mathbf{C}\mathbf{y}$, where $\mathbf{C}$ is positive diagonal, the corresponding transformations $\bar{\mathbf{m}}'_k = \mathbf{C}^{-T}\bar{\mathbf{m}}_k$ and $\mathbf{D}' = \mathbf{C}\mathbf{D}$ preserve the normalized margins. The reference scales are part of the smoothing specification: changing them while keeping the outcome units fixed can change the finite-η objective.

For full-dimensional target regions, we remove duplicate and redundant inequalities and use the irredundant facet representation; l denotes the number of retained facets. Its normalized coefficients are unique up to permutation, so the surrogate is invariant to equivalent halfspace representations for fixed $\mathbf{D}$. Normalization alone would not provide invariance to duplicated constraints. For lower-dimensional regions, we retain a specified normalized representation and claim invariance to positive row rescaling, but not to arbitrary changes of representation.

The scales and retained constraints are fixed as η varies. Thm. 1 applies to these normalized coefficients. Its constant may depend on the chosen scales, the region, and the conditional density bound, but not on arbitrary positive rescaling of the supplied inequalities.

Lemma 1 (Vershynin, 2018, Proposition 2.1.2)**.** *Let $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-u^2}{2}}\, du$ denote the CDF of the standard normal distribution. For any $t > 0$, the following inequality holds:*

$$\Phi(-t) \le \frac{1}{\sqrt{2\pi}t} e^{\frac{-t^2}{2}}.$$

Lemma 2. *Let $\{u_k\}_{k=1}^{l}$ and $\{v_k\}_{k=1}^{l}$ be two sequences of real numbers such that $u_k, v_k \in [0, 1]$ for all $k = 1, \ldots, l$. Then, the absolute difference of their products is bounded by the sum of their absolute pointwise differences, as shown in:*

$$\left| \prod_{k=1}^{l} u_k - \prod_{k=1}^{l} v_k \right| \le \sum_{k=1}^{l} |u_k - v_k|.$$

Proof of Lemma 2. The result follows from a standard telescoping sum argument:

$$\left| \prod_{k=1}^{l} u_k - \prod_{k=1}^{l} v_k \right| = \left| \sum_{k=1}^{l} \left(\prod_{i=1}^{k-1} v_i \right) (u_k - v_k) \left(\prod_{j=k+1}^{l} u_j \right) \right| \le \sum_{k=1}^{l} |u_k - v_k|.$$

The inequality holds as $|u_i|, |v_i| \le 1$, thus the magnitude of coefficients is bounded by 1. □

Theorem 1. *Let $\mathcal{S}$ be a nonempty bounded convex polyhedron defined by Eq. (2), with $\mathbf{m}_k \neq \mathbf{0}$ and Probit surrogate w_η from Eq. (3). If the conditional density $p(\mathbf{y} \mid \mathbf{x}, \mathring{\mathbf{a}})$ is bounded, then, as $\eta \to \infty$,*

$$|\mathbb{P}(\mathbf{Y} \in \mathcal{S} \mid \mathbf{x}, \mathring{\mathbf{a}}) - \mathbb{E}[w_\eta(\mathbf{Y}) \mid \mathbf{x}, \mathring{\mathbf{a}}]| = \mathcal{O}\big(\eta^{-1}\big)\,.$$

Proof of Thm. 1. Fix the query and write $p(\mathbf{y} \mid \mathbf{x}, \mathring{\mathbf{a}}) \le M$ almost everywhere. Use the normalized constraint margins $s_k(\mathbf{y}) = b_k - \mathbf{m}_k^\top \mathbf{y}$, and define:

$$\Delta(\mathbf{y}) = |\mathbb{I}(\mathbf{y} \in \mathcal{S}) - w_\eta(\mathbf{y})|, \qquad \mathcal{S}_1 = \{\mathbf{y} : s_k(\mathbf{y}) \ge -1 \text{ for all } k\}.$$

The fixed relaxation $\mathcal{S}_1$ is bounded. Indeed, an unbounded sequence in $\mathcal{S}_1$, normalized and restricted to a convergent subsequence, would yield a unit vector $\mathbf{v}$ satisfying $\mathbf{m}_k^\top \mathbf{v} \le 0$ for all k. For any $\mathbf{y}_0 \in \mathcal{S}$, the ray $\mathbf{y}_0 + t\mathbf{v}$, $t \ge 0$, would then lie in $\mathcal{S}$, contradicting its boundedness. Choose a ball B_R containing $\mathcal{S}_1$. Outside $\mathcal{S}_1$, some margin is less than -1, so $\Delta(\mathbf{y}) = w_\eta(\mathbf{y}) \le \Phi(-\eta)$.

For every $\mathbf{y}$, Lemma 2 gives

$$\Delta(\mathbf{y}) \le \sum_{k=1}^{l} |\mathbb{I}(s_k(\mathbf{y}) \ge 0) - \Phi(\eta s_k(\mathbf{y}))| = \sum_{k=1}^{l} \Phi(-\eta |s_k(\mathbf{y})|).$$

Let $V_{d_y-1}(R)$ denote the volume of a $(d_y - 1)$-dimensional ball of radius R, with $V_0(R) = 1$. Integrating along the direction $\mathbf{m}_k / \|\mathbf{m}_k\|_2$, every orthogonal section of B_R has volume at most $V_{d_y-1}(R)$. Consequently,

$$\begin{aligned}\int_{B_R} \Phi(-\eta |s_k(\mathbf{y})|)\, d\mathbf{y} &\le \frac{V_{d_y-1}(R)}{\|\mathbf{m}_k\|_2} \int_{\mathbb{R}} \Phi(-\eta|t|)\, dt \\ &= \frac{V_{d_y-1}(R)}{\|\mathbf{m}_k\|_2} \frac{\sqrt{2/\pi}}{\eta},\end{aligned}$$

where $\int_0^\infty \Phi(-t)\, dt = 1/\sqrt{2\pi}$ follows by interchanging the Gaussian integrals. Combining the two regions yields

$$\begin{aligned}|\mathbb{P}(\mathbf{Y} \in \mathcal{S} \mid \mathbf{x}, \mathring{\mathbf{a}}) - \mathbb{E}[w_\eta(\mathbf{Y}) \mid \mathbf{x}, \mathring{\mathbf{a}}]| &\le M \int_{\mathcal{S}_1} \Delta(\mathbf{y})\, d\mathbf{y} + \Phi(-\eta) \\ &\le \frac{M V_{d_y-1}(R)\sqrt{2/\pi}}{\eta} \sum_{k=1}^{l} \frac{1}{\|\mathbf{m}_k\|_2} + \Phi(-\eta).\end{aligned} \tag{10}$$

Lemma 1 bounds the remaining tail, giving $\mathcal{O}(\eta^{-1})$ and hence the stated rate. □

Discussion. Thm. 1 is stated for bounded closed convex polyhedra, whose boundedness allows a rate guarantee under bounded joint conditional density. The same $\mathcal{O}(\eta^{-1})$ rate extends straightforwardly to unbounded convex polyhedra if each projection $\mathbf{m}_k^\top \mathbf{Y}$, conditional on $(\mathbf{x}, \mathring{\mathbf{a}})$, has a bounded density near b_k. Indeed, the preceding product bound reduces the error to one-dimensional Gaussian-tail integrals near each threshold, plus exponentially small tails away from it. Absolute continuity alone ensures expectation convergence for either type of polyhedron, but gives no universal rate. These approximation statements require no RKHS membership of w_η. Rehearsal learning primarily addresses real-world decision problems, where desired outcomes are typically specified within bounded ranges reflecting physical limits or practical requirements. The bounded-region assumption therefore accommodates a broad range of practically relevant target specifications.

C.3 Proof of Theorem 2

We analyze scalar KRR for $\psi(\mathbf{h}) = \mathbb{E}[w_\eta(\mathbf{Y}) \mid \mathbf{H} = \mathbf{h}]$ and vector-valued KRR for $m(\mathbf{x}) = \mathbb{E}[k_u(\mathbf{U}, \cdot) \mid \mathbf{X} = \mathbf{x}]$. Write $\mathcal{H}_v = \mathcal{H}_{k_v}$ and $\mathcal{G}_x = \mathcal{H}_x \otimes \mathcal{H}_u$, the vector-valued RKHS with kernel $k_x I_{\mathcal{H}_u}$. For $v \in \{h, x\}$, let

$$(L_v f)(z) = \int k_v(z, z') f(z')\, dP_V(z')$$

be the integral operator on $L_2(P_V)$, and let (τ_{vj}, e_{vj}) denote its positive eigenpairs, with the e_{vj} orthonormal in $L_2(P_V)$. At query points we use the canonical versions

$$e_{vj}(z) = \tau_{vj}^{-1} \int k_v(z, z') e_{vj}(z')\, dP_V(z'). \tag{11}$$

The following fractional source conditions allow the regression targets to lie outside their respective RKHSs (Fischer & Steinwart, 2020; Li et al., 2024).

Assumption 2 (Spectral regularity at the query)**.** Fix $\eta > 0$, the kernels, and an identified query $(\mathbf{x}, \mathbf{a})$.

(i) The observations $(\mathbf{h}_i, \mathbf{y}_i)$ are i.i.d. The kernels are measurable, their RKHSs are separable, and $k_v(z, z) \le 1$ for $v \in \{x, u, a\}$. The kernel $k_h = k_x k_u k_a$ is defined on the Cartesian input space as in Ass. 1.

(ii) For some $0 < \alpha < \beta < 1$ and finite R_h, R_x, the full L_2 source expansions are

$$\begin{aligned} \psi &= \sum_j \tau_{hj}^{\beta/2} c_j e_{hj}, \qquad \sum_j c_j^2 \le R_h^2, \\ m &= \sum_j \tau_{xj}^{\beta/2} d_j e_{xj}, \quad \sum_j \|d_j\|_{\mathcal{H}_u}^2 \le R_x^2, \end{aligned} \tag{12}$$

where $c_j \in \mathbb{R}$ and $d_j \in \mathcal{H}_u$; there is no component in the nullspace of the corresponding integral operator.

(iii) For finite D_x, D_h, the query-evaluation bounds are

$$\sum_j \tau_{xj}^{\alpha} e_{xj}(\mathbf{x})^2 \le D_x^2, \qquad \int \sum_j \tau_{hj}^{\alpha} e_{hj}(\mathbf{x}, \mathbf{u}, \mathbf{a})^2 \, dP(\mathbf{u} \mid \mathbf{x}) \le D_h^2. \tag{13}$$

With the versions in Eq. (11), the source series in Eq. (12) equal the identified conditional targets $m(\mathbf{x})$ and $\psi(\mathbf{x}, \mathbf{u}, \mathbf{a})$ for $P(\mathbf{u} \mid \mathbf{x})$-almost every $\mathbf{u}$.

Part (iii) specifies the representatives needed for fixed-query evaluation; an L_2 source condition alone does not identify values on a slice of observational probability zero. These conditions require neither $w_\eta \in \mathcal{H}_{k_y}$ nor $k_x(\mathbf{x}, \cdot) \in \mathrm{Ran}(C_{XX})$. They are additional distribution-dependent assumptions, rather than consequences of Gaussian kernels or smoothing. All constants below may depend on η, the kernels, and the fixed query, but not on N or the regularization parameters. Set $\rho = (\beta - \alpha)/2$ and $q = (1 - \beta)/2$. Then $\rho, q > 0$ and $\rho + q = (1 - \alpha)/2 < 1/2$; conversely, $\beta = 1 - 2q$ and $\alpha = 1 - 2(\rho + q)$. These exponents describe the assumed regularity of the targets and query evaluation.

Define the population ridge solutions

$$\begin{aligned} \psi_{\lambda_h} &= \arg\min_{f \in \mathcal{H}_h} \left\{ \mathbb{E}[(w_\eta(\mathbf{Y}) - f(\mathbf{H}))^2] + \lambda_h \|f\|_{\mathcal{H}_h}^2 \right\}, \\ m_{\lambda_x} &= \arg\min_{g \in \mathcal{G}_x} \left\{ \mathbb{E}[\|k_u(\mathbf{U}, \cdot) - g(\mathbf{X})\|_{\mathcal{H}_u}^2] + \lambda_x \|g\|_{\mathcal{G}_x}^2 \right\}. \end{aligned} \tag{14}$$

Their empirical counterparts are $\hat{\psi}$ from Eq. (4) and $\hat{m}(\mathbf{x}) = \hat{\mu}_{\mathbf{U}|\mathbf{x}}$ from Stage 2.

Lemma 3 (Population bias and ridge norm)**.** *Under Ass. 2, for* $0 < \lambda_h, \lambda_x \le 1$,

$$\begin{aligned} b_h(\lambda_h) &\triangleq \int |\psi_{\lambda_h}(\mathbf{x}, \mathbf{u}, \mathbf{a}) - \psi(\mathbf{x}, \mathbf{u}, \mathbf{a})| \, dP(\mathbf{u} \mid \mathbf{x}) \le D_h R_h \lambda_h^{\rho}, \\ b_x(\lambda_x) &\triangleq \|m_{\lambda_x}(\mathbf{x}) - m(\mathbf{x})\|_{\mathcal{H}_u} \le D_x R_x \lambda_x^{\rho}, \quad \|\psi_{\lambda_h}\|_{\mathcal{H}_h} \le R_h \lambda_h^{-q}. \end{aligned} \tag{15}$$

Proof. For either regression, write $F = \sum_j \tau_j^{\beta/2} a_j e_j$, with coefficients in $\mathbb{R}$ or $\mathcal{H}_u$ and $\sum_j \|a_j\|^2 \le R^2$. The population ridge normal equation gives

$$F_\lambda = \sum_j \frac{\tau_j}{\tau_j + \lambda} \tau_j^{\beta/2} a_j e_j, \qquad F_\lambda - F = -\sum_j \frac{\lambda \tau_j^{\beta/2}}{\tau_j + \lambda} a_j e_j. \tag{16}$$

Indeed, the functions $\sqrt{\tau_j} e_j$ are orthonormal in the observable component of the input RKHS, and the ridge solution has no component orthogonal to this span. Thus the first series defines the RKHS solution with the canonical versions in Eq. (11). For $0 < r < 1$,

$$\frac{\lambda t^r}{t + \lambda} = \lambda^r \frac{(t/\lambda)^r}{1 + t/\lambda} \le \lambda^r, \qquad t > 0. \tag{17}$$

Apply Cauchy–Schwarz in Eq. (16), splitting $\tau_j^{\beta/2} = \tau_j^{\rho}\tau_j^{\alpha/2}$, to obtain

$$\|F_\lambda(z) - F(z)\| \le \left(\sum_j \left(\frac{\lambda\tau_j^\rho}{\tau_j+\lambda}\right)^2 \|a_j\|^2\right)^{1/2} \left(\sum_j \tau_j^\alpha e_j(z)^2\right)^{1/2}$$
$$\le R\lambda^\rho \left(\sum_j \tau_j^\alpha e_j(z)^2\right)^{1/2}.$$

The evaluation bounds ensure convergence of the source and bias series at the required points. Using Eq. (13) directly proves the outer bound. For the inner bound, Cauchy–Schwarz under the probability measure $P(\mathbf{u} \mid \mathbf{x})$ gives

$$b_h(\lambda_h) \le R_h\lambda_h^\rho \left(\int \sum_j \tau_{hj}^\alpha e_{hj}(\mathbf{x},\mathbf{u},\mathbf{a})^2\, dP(\mathbf{u}\mid\mathbf{x})\right)^{1/2} \le D_h R_h \lambda_h^\rho.$$

Finally, the RKHS expansion yields

$$\|\psi_{\lambda_h}\|_{\mathcal{H}_h}^2 = \sum_j \frac{\tau_{hj}^{\beta+1}}{(\tau_{hj}+\lambda_h)^2} c_j^2 \le \lambda_h^{\beta-1} \sum_j c_j^2 \le R_h^2 \lambda_h^{-2q},$$

where $t^{\beta+1}/(1+t)^2 \le 1$ for $t > 0$ and $0 < \beta < 1$. □

Lemma 4 (Sampling error for Hilbert-valued ridge regression)**.** *Let $(V_i, Z_i)_{i=1}^N$ be i.i.d., with Z_i in a separable Hilbert space $\mathcal{B}$, $\|Z_i\|_{\mathcal{B}} \le 1$, and $k(V,V) \le 1$. Let $\mathcal{G} = \mathcal{H}_k \otimes \mathcal{B}$, and let $F_\lambda, \hat{F}_\lambda$ minimize the population and empirical squared loss plus $\lambda\|F\|_{\mathcal{G}}^2$, respectively. For deterministic $\lambda > 0$ and $0 < \epsilon < 1$,*

$$\mathbb{P}\left(\|\hat{F}_\lambda - F_\lambda\|_{\mathcal{G}} \le \frac{1}{\lambda\sqrt{N\epsilon}}\right) \ge 1 - \epsilon. \tag{18}$$

Proof. Identify $F(v) = T\phi(v)$ with a Hilbert–Schmidt operator $T : \mathcal{H}_k \to \mathcal{B}$, where $\phi(v) = k(v,\cdot)$ and $\|T\|_{\mathrm{HS}} = \|F\|_{\mathcal{G}}$. Write $C = \mathbb{E}[\phi(V)\otimes\phi(V)]$ and $\hat{C} = N^{-1}\sum_i \phi(V_i)\otimes\phi(V_i)$. Comparison of the population ridge objective with the zero function gives

$$\mathbb{E}\|Z - T_\lambda\phi(V)\|_{\mathcal{B}}^2 + \lambda\|T_\lambda\|_{\mathrm{HS}}^2 \le \mathbb{E}\|Z\|_{\mathcal{B}}^2 \le 1.$$

Define the residual scores

$$S_i = (Z_i - T_\lambda\phi(V_i)) \otimes \phi(V_i).$$

The population normal equation implies $\mathbb{E}S_i = \lambda T_\lambda$, while bounded kernel diagonals give

$$\mathbb{E}\|S_i\|_{\mathrm{HS}}^2 \le \mathbb{E}\|Z - T_\lambda\phi(V)\|_{\mathcal{B}}^2 \le 1.$$

Subtracting the population and empirical normal equations yields the exact identity

$$\hat{T}_\lambda - T_\lambda = \left(\frac{1}{N}\sum_i S_i - \mathbb{E}S_i\right)(\hat{C} + \lambda I)^{-1}.$$

Independence across observations, centering, and $\|(\hat{C}+\lambda I)^{-1}\|_{\mathrm{op}} \le \lambda^{-1}$ imply

$$\mathbb{E}\left\|\frac{1}{N}\sum_i (S_i - \mathbb{E}S_i)\right\|_{\mathrm{HS}}^2 \le \frac{1}{N}, \qquad \mathbb{P}\left(\|\hat{T}_\lambda - T_\lambda\|_{\mathrm{HS}} > \frac{1}{\lambda\sqrt{N\epsilon}}\right) \le \epsilon.$$

The norm identity proves Eq. (18). The argument also applies to deterministic sequences $\lambda = \lambda_N$. □

Theorem 2. *Fix $\eta > 0$, the kernels, and pairs $(\mathbf{x},\mathbf{a})$. Under the regularity conditions in Appx. C.3, let $\rho, q > 0$ be the associated exponents, with $\rho + q < 1/2$. For deterministic $0 < \lambda_h, \lambda_x \le 1$ with $N\lambda_h^2 \ge c_0 > 0$, the nested estimator satisfies:*

$$|\hat{J}_\eta(\mathbf{a};\mathbf{x}) - J_\eta(\mathbf{a};\mathbf{x})| = \mathcal{O}_p\left(\lambda_h^\rho + \lambda_h^{-q}\lambda_x^\rho + \frac{1}{\sqrt{N}\lambda_h} + \frac{1}{\sqrt{N}\lambda_x\lambda_h^q}\right).$$

Proof of Thm. 2. Fix $0 < \epsilon < 1/2$ and write $t_v = (\lambda_v\sqrt{N\epsilon})^{-1}$ for $v \in \{h, x\}$. Apply Lemma 4 to $(V, Z) = (\mathbf{H}, w_\eta(\mathbf{Y}))$ and $(\mathbf{X}, k_u(\mathbf{U}, \cdot))$. A union bound gives, with probability at least $1 - 2\epsilon$,

$$\|\hat{\psi} - \psi_{\lambda_h}\|_{\mathcal{H}_h} \le t_h, \qquad \|\hat{m} - m_{\lambda_x}\|_{\mathcal{G}_x} \le t_x. \tag{19}$$

No independence between the two stages is required. Normalized kernel diagonals and Lemma 3 then imply

$$\begin{aligned} \int |\hat{\psi}(\mathbf{x}, \mathbf{u}, \mathbf{a}) - \psi(\mathbf{x}, \mathbf{u}, \mathbf{a})|\, dP(\mathbf{u} \mid \mathbf{x}) &\le D_h R_h \lambda_h^\rho + t_h, \\ \|\hat{m}(\mathbf{x}) - m(\mathbf{x})\|_{\mathcal{H}_u} &\le D_x R_x \lambda_x^\rho + t_x, \\ \|\hat{\psi}\|_{\mathcal{H}_h} &\le R_h \lambda_h^{-q} + t_h. \end{aligned} \tag{20}$$

By the product-kernel structure, the section $\hat{\psi}_{\mathbf{x},\mathbf{a}} = \hat{\psi}(\mathbf{x}, \cdot, \mathbf{a})$ belongs to $\mathcal{H}_u$ and

$$\|\hat{\psi}_{\mathbf{x},\mathbf{a}}\|_{\mathcal{H}_u} \le \sqrt{k_x(\mathbf{x}, \mathbf{x}) k_a(\mathbf{a}, \mathbf{a})}\|\hat{\psi}\|_{\mathcal{H}_h} \le \|\hat{\psi}\|_{\mathcal{H}_h}.$$

The empirical formula in Eq. (8) equals $\hat{J}_\eta(\mathbf{a}; \mathbf{x}) = \langle \hat{\psi}_{\mathbf{x},\mathbf{a}}, \hat{m}(\mathbf{x}) \rangle_{\mathcal{H}_u}$. For every realized dataset,

$$\begin{aligned} \Delta_{\text{est}} &\triangleq |\hat{J}_\eta(\mathbf{a}; \mathbf{x}) - J_\eta(\mathbf{a}; \mathbf{x})| \\ &\le \int |\hat{\psi}(\mathbf{x}, \mathbf{u}, \mathbf{a}) - \psi(\mathbf{x}, \mathbf{u}, \mathbf{a})|\, dP(\mathbf{u} \mid \mathbf{x}) + \|\hat{\psi}_{\mathbf{x},\mathbf{a}}\|_{\mathcal{H}_u} \|\hat{m}(\mathbf{x}) - m(\mathbf{x})\|_{\mathcal{H}_u}. \end{aligned}$$

Combining these inequalities proves the explicit bound, with probability at least $1 - 2\epsilon$,

$$\Delta_{\text{est}} \le D_h R_h \lambda_h^\rho + t_h + \big(R_h \lambda_h^{-q} + t_h\big)\big(D_x R_x \lambda_x^\rho + t_x\big). \tag{21}$$

Since $N\lambda_h^2 \ge c_0$ and $\lambda_h \le 1$, we have $t_h \le (c_0\epsilon)^{-1/2} \le (c_0\epsilon)^{-1/2}\lambda_h^{-q}$. Thus, for each fixed ϵ, the right-hand side of Eq. (21) is bounded by a constant depending on ϵ times

$$\lambda_h^\rho + \lambda_h^{-q}\lambda_x^\rho + \frac{1}{\sqrt{N}\lambda_h} + \frac{1}{\sqrt{N}\lambda_x\lambda_h^q}.$$

As ϵ can be chosen arbitrarily small, this proves the stated $\mathcal{O}_p$ bound. □

Choice of regularization and convergence rate. Write $\lambda_h \asymp N^{-a}$ and $\lambda_x \asymp N^{-b}$. The four decay exponents in Thm. 2 are

$$a\rho, \qquad b\rho - aq, \qquad \tfrac{1}{2} - a, \qquad \tfrac{1}{2} - b - aq.$$

Balancing the first, second, and fourth terms gives

$$b = \frac{1}{2(1+\rho)}, \qquad a = \frac{\rho}{2(1+\rho)(\rho+q)}, \qquad \nu = \frac{\rho^2}{2(1+\rho)(\rho+q)}.$$

Indeed, $a\rho = b\rho - aq = \tfrac{1}{2} - b - aq = \nu$, $\quad \tfrac{1}{2} - a - \nu = \frac{q}{2(\rho+q)} > 0$. Moreover, $0 < a < b < 1/2$, so $N\lambda_h^2 \to \infty$. All four terms are therefore $\mathcal{O}(N^{-\nu})$, proving the claimed rate. For $\rho = 1/3$ and $q = 1/12$, the schedules are $\lambda_h \asymp N^{-3/10}$ and $\lambda_x \asymp N^{-3/8}$. The four terms then have orders $N^{-1/10}$, $N^{-1/10}$, $N^{-1/5}$, and $N^{-1/10}$, respectively. These schedules give pointwise consistency of the smoothed objective at a fixed query and fixed η under Ass. 2.

Corollary 1 (Decision consistency)**.** *Fix a context* $\mathbf{x}$*, the kernels, and a nonempty compact feasible action set* $\mathcal{A}$ *on which the identification conditions of Prop. 1 hold for every* $\mathbf{a} \in \mathcal{A}$*. Let* $J(\mathbf{a}; \mathbf{x}) = \mathbb{P}(\mathbf{Y} \in \mathcal{S} \mid \mathbf{x}, \mathring{\mathbf{a}})$ *be continuous on* $\mathcal{A}$*, and write* $\mathcal{A}^\star = \arg\max_{\mathbf{a}\in\mathcal{A}} J(\mathbf{a}; \mathbf{x})$*. For a deterministic sequence* $\eta_N \to \infty$*, assume the following conditions hold.*

(i) The region $\mathcal{S}$ *satisfies the geometric conditions of Thm. 1. There is a finite* M *such that* $p(\mathbf{y} \mid \mathbf{x}, \mathring{\mathbf{a}}) \le M$ *for every* $\mathbf{a} \in \mathcal{A}$ *and almost every* $\mathbf{y}$*.*

(ii) Ass. 1 and 2 hold at $\eta = \eta_N$ *with fixed* α, β, D_x, R_x *and inner source radius* $R_h(\eta_N)$*. The target identities hold for every* $\mathbf{a} \in \mathcal{A}$*, and the inner query bound is uniform:*

$$\sup_{\mathbf{a}\in\mathcal{A}} \int \sum_j \tau_{hj}^\alpha e_{hj}(\mathbf{x}, \mathbf{u}, \mathbf{a})^2\, dP(\mathbf{u} \mid \mathbf{x}) \le D_h^2 < \infty,$$

with D_h *independent of* N*.*

Use the regularization schedules above, with $\nu = \rho^2/[2(1+\rho)(\rho+q)]$. Let $\hat{\mathbf{a}}_N \in \mathcal{A}$ be a measurable decision satisfying the global optimization error bound

$$\sup_{\mathbf{a}\in\mathcal{A}} \hat{J}_{\eta_N}(\mathbf{a};\mathbf{x}) - \hat{J}_{\eta_N}(\hat{\mathbf{a}}_N;\mathbf{x}) \le \epsilon_{\mathrm{opt},N}.$$

If $(1+R_h(\eta_N))N^{-\nu} \to 0$ and $\epsilon_{\mathrm{opt},N} = o_p(1)$, then, for any $\mathbf{a}^\star \in \mathcal{A}^\star$,

$$J(\mathbf{a}^\star;\mathbf{x}) - J(\hat{\mathbf{a}}_N;\mathbf{x}) \xrightarrow{p} 0, \qquad \mathrm{dist}(\hat{\mathbf{a}}_N, \mathcal{A}^\star) \xrightarrow{p} 0.$$

In particular, a unique maximizer $\mathbf{a}^\star$ implies convergence of the decisions themselves, $\hat{\mathbf{a}}_N \xrightarrow{p} \mathbf{a}^\star$.

Proof. Define the uniform smoothing and estimation errors over the feasible action set by

$$B_N = \sup_{\mathbf{a}\in\mathcal{A}} |J(\mathbf{a};\mathbf{x}) - J_{\eta_N}(\mathbf{a};\mathbf{x})|,$$
$$E_N = \sup_{\mathbf{a}\in\mathcal{A}} |J_{\eta_N}(\mathbf{a};\mathbf{x}) - \hat{J}_{\eta_N}(\mathbf{a};\mathbf{x})|.$$

The uniform density bound and Eq. (10) give $B_N \le C_{\mathcal{S}}/\eta_N$, with $C_{\mathcal{S}}$ independent of N and $\mathbf{a}$. For estimation, fix $0 < \epsilon < 1/2$ and write $t_v = (\lambda_v\sqrt{N\epsilon})^{-1}$ and $R_N = R_h(\eta_N)$. The labels $w_{\eta_N}(\mathbf{Y})$ remain in $[0,1]$, so Lemma 4 yields the event in Eq. (19) for every N. On this event, the uniform query bound makes Eq. (21) hold simultaneously for all $\mathbf{a} \in \mathcal{A}$. Thus, with probability at least $1-2\epsilon$,

$$E_N \le D_h R_N \lambda_h^\rho + t_h + \big(R_N\lambda_h^{-q} + t_h\big)\big(D_x R_x \lambda_x^\rho + t_x\big).$$

Substituting the regularization schedules gives the uniform estimation rate, retaining its dependence on the smoothing parameter:

$$E_N = \mathcal{O}_p\big((1+R_h(\eta_N))N^{-\nu}\big).$$

For any $\mathbf{a}^\star \in \mathcal{A}^\star$, adding and subtracting the estimated objective at the two actions yields

$$\begin{aligned} 0 \le J(\mathbf{a}^\star;\mathbf{x}) - J(\hat{\mathbf{a}}_N;\mathbf{x}) &\le 2B_N + 2E_N + \hat{J}_{\eta_N}(\mathbf{a}^\star;\mathbf{x}) - \hat{J}_{\eta_N}(\hat{\mathbf{a}}_N;\mathbf{x}) \\ &\le 2B_N + 2E_N + \epsilon_{\mathrm{opt},N}. \end{aligned} \tag{22}$$

The three terms vanish in probability under the stated conditions, proving convergence of the achieved objective value. For example, if $R_h(\eta) = \mathcal{O}(1+\eta^\kappa)$ for some $\kappa \ge 0$, choosing $\eta_N = N^{\nu/(\kappa+1)}$ gives

$$B_N + E_N = \mathcal{O}_p\Big(N^{-\nu/(\kappa+1)}\Big).$$

To obtain convergence of the decisions, fix $\delta > 0$ and set $\mathcal{A}_\delta = \{\mathbf{a} \in \mathcal{A} : \mathrm{dist}(\mathbf{a}, \mathcal{A}^\star) \ge \delta\}$. If $\mathcal{A}_\delta$ is nonempty, compactness and continuity imply a strictly positive gap

$$\gamma_\delta = J(\mathbf{a}^\star;\mathbf{x}) - \max_{\mathbf{a}\in\mathcal{A}_\delta} J(\mathbf{a};\mathbf{x}) > 0.$$

Consequently, objective-value convergence bounds the probability of selecting an action in $\mathcal{A}_\delta$:

$$\mathbb{P}(\mathrm{dist}(\hat{\mathbf{a}}_N, \mathcal{A}^\star) \ge \delta) \le \mathbb{P}(J(\mathbf{a}^\star;\mathbf{x}) - J(\hat{\mathbf{a}}_N;\mathbf{x}) \ge \gamma_\delta) \to 0.$$

If $\mathcal{A}_\delta$ is empty, the event on the left is empty as well. This proves convergence to the optimal set and, when it is a singleton, to the unique optimal action. □

D Experimental details and additional results

In this section, we provide details for the experiments in Sec. 4. The implemented variable roles and our estimator settings are given below, followed by the generating mechanisms.

Table 2: Variable roles in the six implemented alteration tasks. NHANES context abbreviations denote family history of diabetes and income ratio.

Task	$\mathbf{X}$	$\mathbf{U}$	Jointly assigned $\mathbf{A}$	$\mathbf{U}^{\text{post}}$
Bermuda	Light, Temp, Sal	$\varnothing$	DIC, TA, Ω_A, Chla, Nut	pHsw, CO_2
Lin-Syn1	X_1, X_2	U_2	A_1, A_2	U_1
NHANES	Age, Sex, Race, FamilyHx, Education, Income	$\varnothing$	BMI, CalIntake, CarbIntake, FiberIntake, Sedentary_min	SBP, DBP, TotalCholesterol, HDL
BankExp	X_1, X_2	U_1	A_2	$\varnothing$
Non-Syn1	X	U	A_1, A_2	$\varnothing$
Non-Syn2	X_1, X_2	U_1	A_1, A_2	U_2

D.1 Implemented variable roles and estimator settings

Table 2 lists the variables used by our estimator in each task. Stage 1 uses $(\mathbf{X}, \mathbf{U}, \mathbf{A})$, while Stage 2 uses $(\mathbf{X}, \mathbf{U})$ to integrate the pre-alteration variables. The variables in $\mathbf{U}^{\text{post}}$ are marginalized within the outcome regression. During evaluation, the generator fixes the query $\mathbf{X} = \mathbf{x}$ and jointly assigns the listed $\mathbf{A}$; every other variable follows its original conditional mechanism.

In Lin-Syn1, U_1 is generated after the assigned A_1 and is excluded from both estimation stages. BankExp contains no A_1 node in its implemented generator. In Non-Syn2, U_2 is generated from (X_1, U_1), and A_1 does not enter the outcome equation. For Bermuda and NHANES, $\mathbf{U} = \varnothing$ and the adjustment vector is exactly $\mathbf{1}$.

Hyperparameter selection. For each benchmark, we split the observational training set 80:20 into fitting and validation subsets, stratifying by success labels when possible. Bandwidth parameters, λ_h, and η jointly minimize validation mean squared error against the observed binary success indicator. For nonempty $\mathbf{U}$, λ_x minimizes the validation kernel feature error under a fixed reference kernel. Preprocessing and distance-based bandwidth statistics use only the fitting subset during selection. After selection, we combine the two subsets and refit one estimator on all observations, recomputing these statistics while keeping the selected hyperparameters fixed. The selection routine receives neither query contexts nor post-alteration evaluation outcomes. Candidate grids, bandwidth conventions, optimization settings, seeds, test inputs, and reproduction commands are provided in the accompanying code. Bermuda's generating coefficients are fitted after standardizing the original data; its alteration bounds and NEC target are in these standardized units.

Outcome scales. The implemented constraint margins use $\mathbf{D} = \mathbf{I}$ in the reported outcome units for Bermuda and the synthetic tasks, and $\mathbf{D} = \text{diag}(1, 10)$ for NHANES when HbA1c is measured in percentage points and FPG in mg/dL. These fixed response scales are distinct from the standardization of kernel inputs.

Initialization and projection. The synthetic and Bermuda implementations use all historical alterations with $\omega_i > 10^{-9}$ as starts. When this set is empty, they draw uniformly distributed feasible starts using the dataset-specific count in the accompanying code. Projection follows each gradient update, and iteration stops when consecutive alterations differ by less than 10^{-6}. NHANES instead selects up to 20 highest-weight historical alterations whose weights exceed the median, then fills the remaining starts uniformly from the action box. It projects before the first step and after every update, and caps the gradient norm at 10.

Final fitting and evaluation. For the main benchmark results in Tab. 1, we use the same observational datasets for all approaches, containing $1,000$ observational samples for each benchmark. Hyperparameters are selected using an $80 : 20$ fitting/validation split, after which the estimator is refitted on the complete observational dataset. For each seed, the AUF probability is estimated using $1,000$ Monte Carlo draws from the true post-alteration distribution.

D.2 Linear data

Bermuda data. The Bermuda data is an environmental dataset that records a collection of marine and biogeochemical variables measured in the Bermuda region (Courtney et al., 2017). The variables involved in the generation process include:

- Light: Light levels at the ocean bottom;

- Temp: Temperature at the ocean bottom;
- Sal: Sea surface salinity;
- DIC: Dissolved inorganic carbon in seawater;
- TA: Total alkalinity of seawater;
- Ω_A: Aragonite saturation state of seawater;
- Chla: Chlorophyll-*a* concentration at the sea surface;
- Nut: The first principal component (PC1) of NH_4, NO_2 + NO_3, and SiO_4;
- pHsw: Seawater pH;
- CO_2: Partial pressure of CO_2 in seawater (P_{CO_2});
- NEC: Net ecosystem calcification.

The influence graph over these variables is illustrated in Fig. 5. Following Aglietti et al. (2020); Qin et al. (2023), the actionable variables that can be altered by the decision-maker are DIC, TA, Ω_A, Chla, and Nut, which can be altered into values with constraint $[-1.0, 1.0]$. The desired region is $\mathcal{S} = \{\text{NEC} \in [0.5, 2]\}$, following the specifications in Sec. 4.

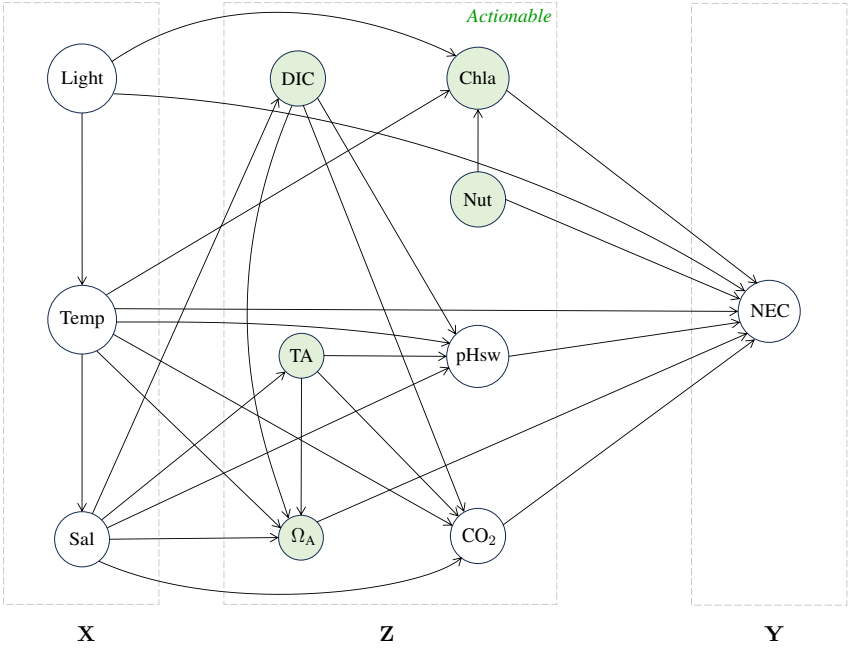

Figure 5: The graph structure of Bermuda data.

Linear synthetic data. The influence graph over the variables in our linear synthetic dataset is illustrated in Fig. 6. The desired region of $\mathbf{Y}$ is specified as $\{(Y_1, Y_2) \mid 0.0 \leq Y_1 \leq 2.0,\ 0.0 \leq Y_2 \leq 2.0\}$, and the actionable variables are A_1 and A_2 with $A_1, A_2 \in [-3.0, 3.0]$. The generation mechanisms underlying the natural observational process are given by:

$$
\begin{cases}
X_1 := \mathcal{N}(0, 0.1); \\
X_2 := \mathcal{N}(0, 0.1); \\
U_2 := 10.0 \times X_2 + \mathcal{N}(0, 0.1); \\
A_1 := 10.0 \times X_1 + \mathcal{N}(0, 0.1); \\
U_1 := 0.5 \times A_1 + 1.3 \times U_2 + \mathcal{N}(0, 0.1); \\
A_2 := 2.0 \times A_1 + 0.4 \times U_2 + \mathcal{N}(0, 0.1); \\
Y_1 := -1.0 \times A_1 + 0.9 \times A_2 + \mathcal{N}(0, 0.1); \\
Y_2 := 1.6 \times A_1 - 0.9 \times A_2 + \mathcal{N}(0, 0.1).
\end{cases}
$$

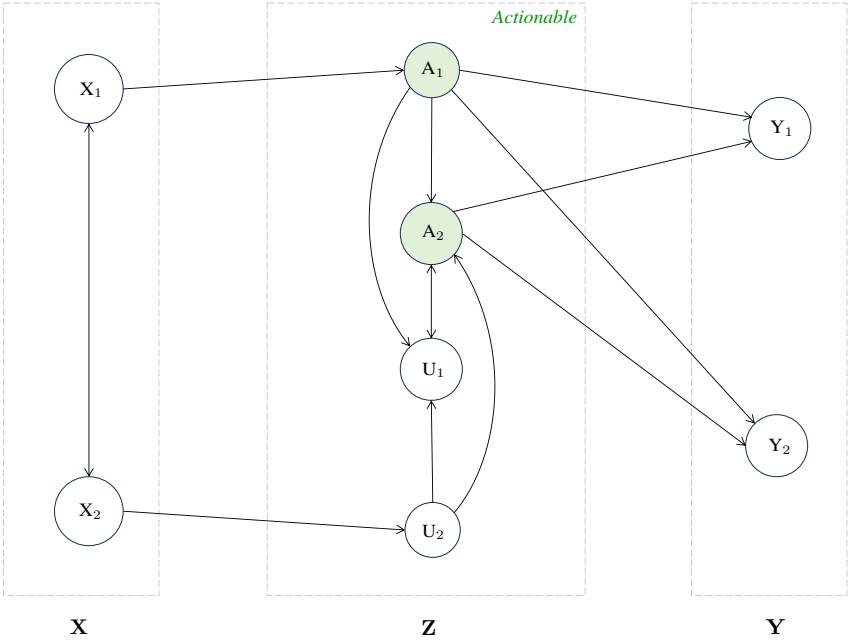

Figure 6: The graph structure of linear synthetic data.

D.3 NONLINEAR DATA

Bank data. The influence graph over the variables in our bank dataset is illustrated in Fig. 1. The implemented BankExp task assigns only A_2; A_1 appears in the motivating example but is absent from this generator. The desired region of $\mathbf{Y}$ is $\mathcal{S} = \{(Y_1, Y_2) \mid 0.6 \leq Y_1 \leq 1.0, 0.3 \leq Y_2\}$, following the specifications in Sec. 4. The generating mechanisms used in our BankExp implementation are:

- X_1: Credit Score ($X_1 := \mathcal{U}(0, 1)$);
- X_2: Debt-to-Income Ratio ($X_2 := \mathcal{U}(0, 1)$);
- U_1: Borrower's Industry Sector Stability Index ($U_1 \sim \text{Beta}(2, 2)$);
- A_1: Loan Amount Granted (Excluded in implementation);
- A_2: Interest Rate Applied ($A_2 := U_1 + 0.5X_1 + 0.5X_2 - 0.5 + \mathcal{N}(0, 0.2^2)$); feasible alterations satisfy $a_2 \in [0, 1]$;
- Y_1: Repayment Rate ($Y_1 := \text{Sigmoid}(2.0 \times U_1^{1.1} - 1.5 \times A_2 + 0.2 \times X_2 + 0.4) + \mathcal{N}(0, 0.005^2)$);
- Y_2: Bank ROI (Return on Investment) ($Y_2 := 0.8A_2 + 0.5U_1 + \mathcal{N}(0, 0.05^2)$).

Nonlinear synthetic data 1. The influence graph over the variables in our nonlinear synthetic dataset 1 can be described as: $X \to A_1, X \to Y; U \to A_1, U \to A_2, U \to Y; A_1 \to Y; A_2 \to Y$. Desired region of $\mathbf{Y}$ is specified as $\{Y \mid 1.5 \leq Y \leq 2.0\}$, and actionable variables are A_1 and A_2 with $A_1, A_2 \in [-1.0, 1.0]$. Generation mechanisms underlying observational process are given by:

$$
\begin{cases}
X := \mathcal{U}(-1, 1); \\
U := \mathrm{Exp}(1); \\
A_1 := 0.8X + 0.2U + \mathcal{N}(0, 0.5^2); \\
A_2 := 0.5\sin(U) + \mathcal{N}(0, 0.5^2); \\
Y := 1.5 - (A_1 - X)^2 - (A_2 - \ln(U+1))^2 + 0.2\sin(A_1 A_2) + \mathcal{N}(0, 0.01^2).
\end{cases}
$$

Nonlinear synthetic data 2. The influence graph over the variables in our nonlinear synthetic dataset 2 is illustrated in Fig. 7. The desired region of $\mathbf{Y}$ is specified as $\{Y_1 \mid 0.9 \leq Y_1 \leq 1.5\}$, and the actionable variables are A_1 and A_2 with $A_1, A_2 \in [-1.0, 1.0]$. The generation mechanisms underlying the natural observational process are given by:

$$
\begin{cases}
X_1 := \mathcal{U}(0, 1); \\
X_2 := \mathcal{U}(0, 1); \\
A_1 := 1.0X_1 - 1.0X_2 + 0.2X_1^2 - 0.2X_2^2 + 0.1X_1X_2 + \mathcal{N}(0, 0.1^2); \\
U_1 := -1.0X_1 + 2.0X_2 - 0.2X_1^2 + 0.4X_2^2 + 0.1X_1X_2 + \mathcal{N}(0, 0.1^2); \\
U_2 := -1.0X_1 + 4.0U_1 - 0.2X_1^2 + 0.8U_1^2 + 0.1X_1U_1 + \mathcal{N}(0, 0.1^2); \\
A_2 := -1.0X_1 - 0.5X_2 + 0.3U_1 - 0.2X_1^2 - 0.1X_2^2 + 0.06U_1^2 + 0.1(X_1X_2 + X_1U_1 \\
\qquad + X_2U_1) + \mathcal{N}(0, 0.1^2); \\
Y_1 := 0.5 + \frac{1}{60}\Big(0.2(X_1 + X_1^2) - 5.0(U_1 + U_1^2) - 1.0(U_2 + U_2^2) + 5.0(A_2 + A_2^2)\Big) \\
\qquad + 0.5 \sum_{p_i, p_j \in \mathbf{Pa}(Y_1), i<j} p_i p_j + \exp(0.5A_2)\varepsilon_Y.
\end{cases}
$$

Here $\varepsilon_Y \sim \mathcal{N}(0, 0.1^2)$ is independent of all upstream disturbances. The multiplicative term gives non-additive outcome noise, with $\mathrm{Var}(Y_1 \mid X_1, U_1, U_2, A_2) = 0.01\exp(A_2)$.

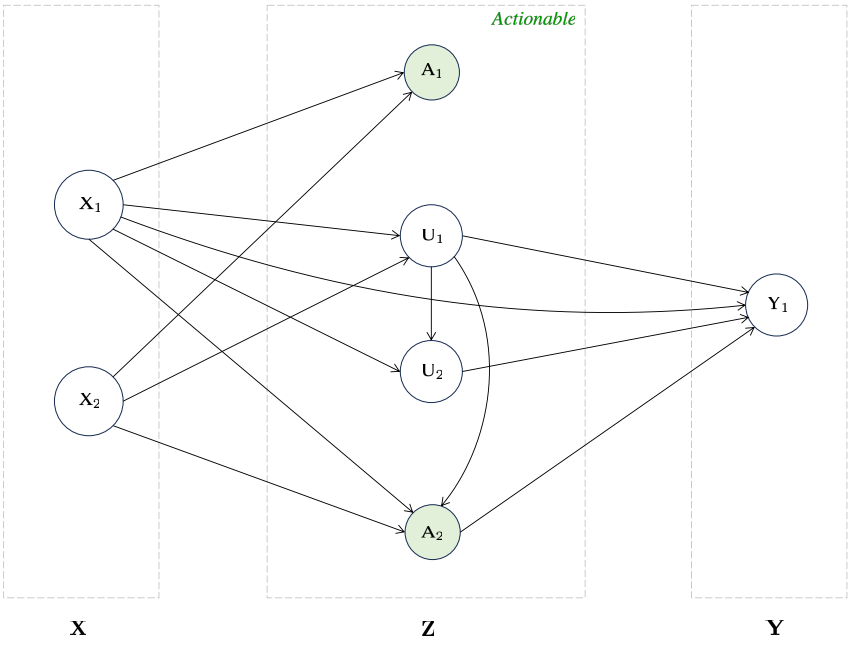

Figure 7: The graph structure of nonlinear synthetic data 2.

D.4 SCALABILITY AND SENSITIVITY EXPERIMENTS

We report additional simple experiments on Non-Syn1 to evaluate computational scalability and hyperparameter sensitivity. Unless otherwise stated, AUF probabilities in this section are averaged over 5 random seeds.

Kernel approximation for scalability. The exact nested KRR estimator requires solving Gram-matrix systems in the offline fitting stage. The reported timings cover offline fitting only; online action selection additionally requires context-weight computation and multi-start optimization. To test whether standard kernel approximations can reduce the offline cost, we compare exact KRR with Nyström (Williams & Seeger, 2000) and random Fourier feature (RFF) (Rahimi & Recht, 2007) approximations in Tab. 3.

Table 3: Scalability results on Non-Syn1. Time is the offline fitting time in seconds.

Method	N	Rank/Dim.	Time (s)	AUF Prob. (↑)
Exact	1k	–	0.131	0.424
Nyström	1k	100	0.022	0.424
Nyström	1k	200	0.035	0.424
RFF	1k	500	0.087	0.360
Exact	5k	–	2.118	0.453
Nyström	5k	100	0.292	0.453

Nyström preserves the AUF probability of exact KRR in these runs while reducing offline time by roughly $4\times$–$7\times$. RFF also reduces runtime but shows a clearer speed-performance trade-off in this setting. These results indicate that the proposed estimator can directly benefit from standard low-rank kernel approximations when larger observational datasets are available.

Sensitivity to regularization. We separately fix $\eta = 10$ and vary the inner and outer KRR parameters λ_h and λ_x around their default values. As shown in Tab. 4, the method exhibits a broad plateau around default scale: many combinations between $0.2\times \sim 2\times$ achieve nearly identical AUF probabilities. Performance mainly degrades when both regularization levels are substantially larger.

Table 4: Sensitivity to regularization on Non-Syn1. Entries are mean AUF probabilities with $\eta = 10$.

λ_h/λ_x	$0.2\times$	$0.5\times$	$1\times$	$2\times$	$5\times$
$0.2\times$	0.456	0.456	0.456	0.456	0.448
$0.5\times$	0.456	0.456	0.456	0.448	0.448
$1\times$	0.456	0.456	0.456	0.448	0.376
$2\times$	0.456	0.448	0.448	0.376	0.352
$5\times$	0.448	0.448	0.382	0.352	0.358

D.5 CONTROLLED COMPARISONS OF SMOOTH AND EXACT RESPONSES

We compare two responses within the same kernel estimator: the finite-η Probit response and the exact indicator of the desired event. Within each repetition, both receive identical observational data, query, feasible alteration range, and optimization budget. Each selected alteration is evaluated using 2000 shared Monte Carlo disturbances, independently of parameter selection. Each comparison contains 40 independent training-set/query repetitions.

Observation-only parameter selection. Both response families evaluate 48 joint candidates using the same three observational validation folds, for 144 inner regression fits per family. Bandwidth multipliers are drawn from $\{0.25, 0.5, 1, 2, 4\}$ and ridge penalties from $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. The finite family assigns eight candidates to each $\eta \in \{3, 5, 10, 20, 50, 100\}$; the exact family evaluates the same 48 kernel-parameter rows. Thus smoothing selection is included within the common candidate budget.

For candidate c, let $\hat{\psi}_c^{(-f(i))}$ denote the inner model fitted without observation i's fold. All responses are scored against the same binary target:

$$\widehat{R}(c) = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\psi}_c^{(-f(i))}(\mathbf{h}_i) - \mathbb{I}(\mathbf{y}_i \in \mathcal{S})\right)^2.$$

The minimum score selects the candidate, with ties resolved by its fixed index. Predictions remain untruncated, and no simulated alteration outcomes enter this criterion. Means, population standard deviations, and median positive pairwise distances are fitted within each training fold; outcomes retain their original units. After selection, preprocessing and the estimator are refitted on the full observational sample. For nonempty $\mathbf{U}$, six outer penalties are compared using a fixed reference RBF kernel on $\mathbf{U}$, with fold-training median bandwidth. For held-out $(\mathbf{x}_j, \mathbf{u}_j)$, the loss is

$$k_u^0(\mathbf{u}_j, \mathbf{u}_j) - 2\boldsymbol{\gamma}_j^\mathsf{T}\mathbf{k}_u^0(\mathbf{u}_j) + \boldsymbol{\gamma}_j^\mathsf{T}\mathbf{K}_{uu}^0\boldsymbol{\gamma}_j, \qquad \boldsymbol{\gamma}_j = (\mathbf{K}_{xx} + n_f\lambda_x\mathbf{I})^{-1}\mathbf{k}_x(\mathbf{x}_j),$$

where n_f is the fold-training size and k_x uses the selected context bandwidth. The same reference kernel scores every outer candidate, independently of its response and fitted U bandwidth.

Table 5: Smooth-versus-exact response comparisons. Success probabilities and differences are percentage points. Each row contains 40 training-set repetitions and 2000 evaluation draws per query.

Condition	Smooth	Exact	Paired difference
Synthetic 1, separate selection	17.178	16.066	1.111
Synthetic 1, shared parameters	16.725	4.805	11.920
Synthetic 2, separate selection	19.390	15.996	3.394
Synthetic 2, shared parameters	21.628	18.370	3.258
NHANES, separate selection	59.636	55.716	3.920
BankExp, separate response environment	82.646	80.399	2.248

This result show the gain of the smooth approach.

D.6 NHANES BENCHMARK

We build one real-data-derived semi-synthetic AUF benchmark from the U.S. National Health and Nutrition Examination Survey (NHANES) 2011–2018 cycles (National Center for Health Statistics, 2018). The benchmark is designed as a contextual diabetes-control task: after observing immutable demographic and socioeconomic context variables, a method chooses feasible values for actionable lifestyle/metabolic variables, and the target outcomes are two glycemic markers, glycated hemoglobin (HbA1c) and fasting plasma glucose (FPG). A structural generator fitted to the cross-sectional measurements supplies the benchmark post-alteration distributions.

Data source and cohort. The raw NHANES modules are merged by participant identifier (`SEQN`) across the 2011–2012, 2013–2014, 2015–2016, and 2017–2018 cycles. We harmonize cross-cycle aliases such as `BPXOSY1`/`BPXSY1` and `BPXODI1`/`BPXDI1`, remove NHANES sentinel missing codes, keep adults with both outcomes observed, and apply complete-case selection on the benchmark variables. The resulting benchmark has 7,465 complete-case participants. The generator fits these records without NHANES survey weights, strata, or primary sampling units.

Table 6: NHANES benchmark summary. The validation column reports the mean absolute error between real and generated correlation matrices.

Benchmark	Complete cases	Actionable variables $\mathbf{Z}_a$	Post-alteration variables $\mathbf{U}^{\text{post}}$	Outcomes $\mathbf{Y}$	Corr. MAE
NHANES	7,465	BMI, CalIntake, CarbIntake, FiberIntake, Sedentary_min	SBP, DBP, TotalCholesterol, HDL	HbA1c, FPG	0.0483

Actionability and influence relations. The benchmark specifies a temporal and physiological ordering: immutable context variables $\mathbf{X}$ precede intermediate variables $\mathbf{Z}$, and glycemic outcomes $\mathbf{Y}$ are downstream targets. The immutable context variables are Age, Sex, Race, FamilyHx_Diabetes, Education, and Income_Ratio. Expert actionability assumptions mark BMI, daily calorie intake,

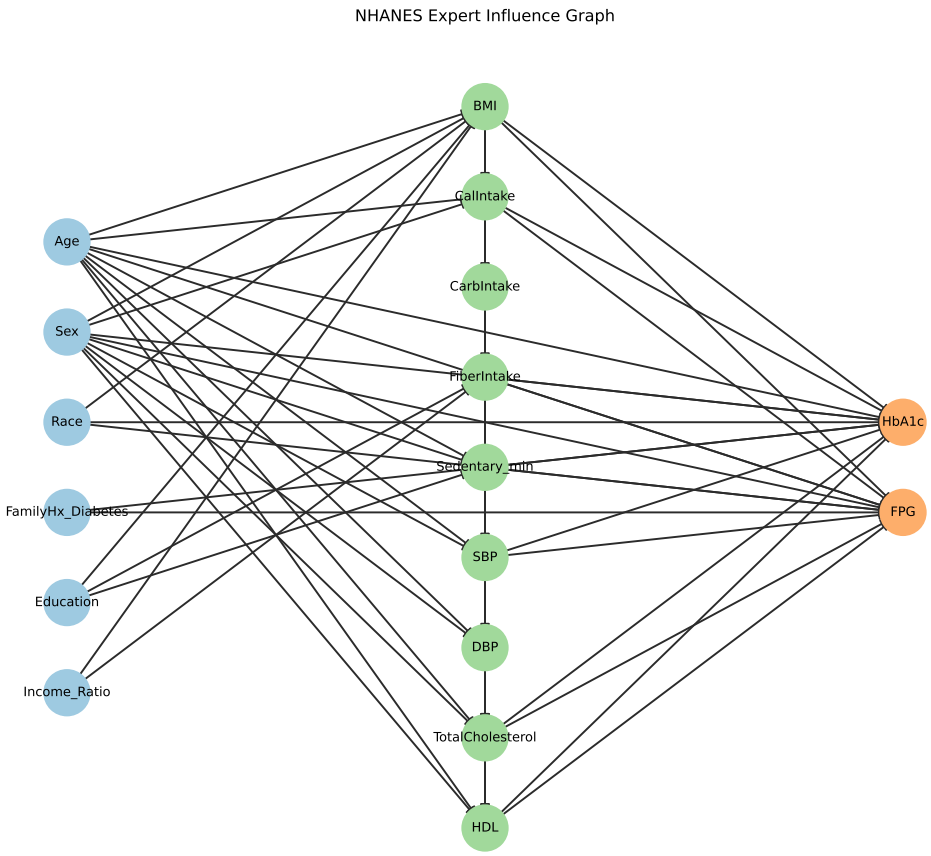

Figure 8: NHANES influence graph specified by human experts for the benchmark. Blue nodes denote context variables; green nodes denote intermediate variables, where BMI, CalIntake, CarbIntake, FiberIntake, and Sedentary_min are actionable and SBP, DBP, TotalCholesterol, and HDL are non-actionable; orange nodes denote glycemia outcomes.

carbohydrate intake, fiber intake, and sedentary minutes as actionable. Blood pressure and lipid measurements are treated as non-actionable because the benchmark models them as physiological consequences rather than direct decision levers. The benchmark DAG encodes following influence relations. Context variables influence BMI, diet, sedentary behavior, blood pressure, lipids, and both outcomes. BMI influences diet, blood pressure, lipids, HbA1c, and FPG. Calorie intake influences carbohydrate/fiber intake, lipids, HbA1c, and FPG. Fiber intake and sedentary time influence both glycemic outcomes, and the physiological intermediates SBP, total cholesterol, HDL influence both outcomes.

Typed structural generator. For each node, the generator fits a conditional model suited to the variable type. The parent sets for these node-wise models follow the expert-specified influence graph in Fig. 8. Discrete variables use empirical categorical sampling for root nodes and gradient-boosted classifiers for non-root nodes. Continuous variables use gradient-boosted regressors with residual bootstrap; nonnegative variables are modeled on a log-transformed scale and clipped back to valid support. Each alteration jointly assigns BMI, calorie intake, carbohydrate intake, fiber intake, and sedentary minutes within their empirical $[1\%, 99\%]$ quantile intervals. The generator then samples the four physiological intermediates and both outcomes from their fitted conditional mechanisms. Thus $\mathbf{U} = \varnothing$ and $\mathbf{U}^{\text{post}} = \{\text{SBP}, \text{DBP}, \text{TotalCholesterol}, \text{HDL}\}$; the outcome regression uses $(\mathbf{X}, \mathbf{A})$ and the adjustment vector is $\mathbf{1}$. The action box specifies direct assignments in this generator, including an idealized assignment of BMI.

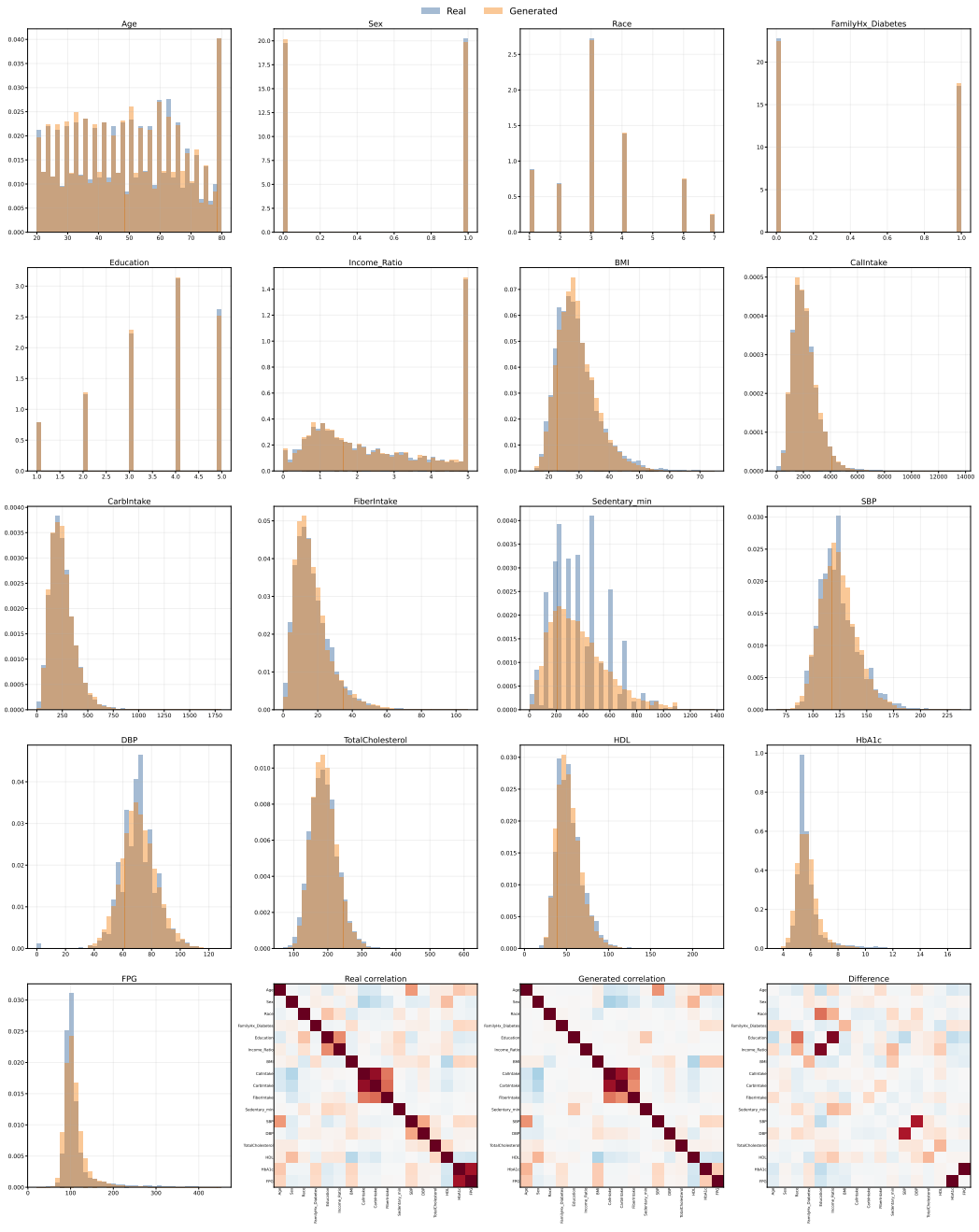

Figure 9: NHANES generator validation plots comparing real complete-case samples against generated observational samples. In the histogram panels, blue histograms denote real samples and orange histograms denote generated samples. The three heatmaps show the real correlation matrix, the generated correlation matrix, and their difference, respectively.

Success region. We use one clinically interpretable two-dimensional glycemic target:

$$\mathcal{S}_{\text{NHANES}} = \{(\text{HbA1c}, \text{FPG}) \mid \text{HbA1c} \leq 5.7\%,\ \text{FPG} \leq 100\ \text{mg/dL}\}.$$

These thresholds correspond to common normoglycemia screening boundaries (American Diabetes Association Professional Practice Committee, 2024). Both outcomes have nonnegative support under the generator. With HbA1c expressed in percentage points and FPG in mg/dL, we fix reference

scales of one percentage point and $10\ \mathrm{mg/dL}$, respectively. The normalized response is:

$$w_\eta(\mathbf{y}) = \Phi\big(\eta(5.7 - y_{\mathrm{HbA1c}})\big)\Phi\big(\eta(100 - y_{\mathrm{FPG}})/10\big).$$

Input columns are standardized using training-sample means and population standard deviations; action bounds use the same transformation. Selected alterations are returned to raw units for evaluation, which counts outcomes satisfying both inequalities.

Selection and evaluation. Our NHANES estimator uses one observational training set of 1000 samples. Hyperparameter selection uses an 80:20 split of this set, followed by refitting on the combined subsets as described in Appx. D.1. The independent response comparisons in Appx. D.5 use separate repetitions and a common observational validation rule.

Validation against observed statistics. The fitted generator is validated by comparing generated samples against the processed complete-case data. The mean absolute difference between real and generated correlation matrices is 0.0483. Figure 9 compares marginal distributions and pairwise correlations between the observed complete-case data and generated samples.