Off-policy Predictive Control with Causal Sensitivity Analysis

Myrl G. Marmarelis[1] Ali Hasan[2] Kamyar Azizzadenesheli[3] R. Michael Alvarez[1] Anima Anandkumar[1]

[1]Caltech
[2]Morgan Stanley
[3]NVIDIA

Abstract

Predictive models are often deployed for decision-making tasks for which they were not explicitly trained. When only partial observations of the relevant state are available, as in most real-world applications, there is a strong possibility of hidden confounding. Therefore, partial observability often makes the outcome of an action unidentifiable, and could render a model's predictions unreliable for action planning. We present an identification bound and propose an algorithm to account for hidden confounding during model-predictive control. To that end, we introduce a generalized causal sensitivity model for action-state dynamics. We place a constraint on the hidden confounding between trajectories of future actions and states, enabling sharp bounds on interventional outcomes. Unlike previous sensitivity models, ours accommodates hidden confounding with memory, while maintaining computational and statistical tractability. We benchmark on a wide variety of multivariate stochastic differential equations with arbitrary confounding. The results suggest that a calibrated sensitivity model helps controllers achieve higher rewards.

1 INTRODUCTION

Learning to *predict* dynamics that are partially observed may be unhelpful for *taking action* in those dynamics, especially if the hidden state confounds the relationship between action and outcome. We consider the problem of using a predictive model trained on offline trajectory data for the purpose of online control. We assume that partial observability induces hidden confounding in the offline data-generating process.

Our insights center on the *identification* of dynamics subject to intervention, like an action policy for online (closed-loop) control. We study the setting in which we only have access to a confounded predictive model that can generate samples of the dynamics, following the offline distribution of observables generated by an unknown policy acting on the hidden state. By projecting entire trajectories of possible actions into the future, a controller may find the trajectory with the best predicted outcome, and act on it. With hidden confounding, the controller needs to assess the worst-case outcome by taking into account any known constraints on the hidden confounding. This makes the action policy more conservative. If the worst case represents a truly valid instantiation of possible hidden confounding, then the controller performs as well as possible and is considered minimax optimal.

	Easy	Medium	Hard
Ours	**20.8**%	**17.9**%	**22.8**%
MSM	15.3%	13.0%	21.6%
Empirical	13.1%	15.9%	20.6%

Table 1. Results of partially identified controllers expressed as average improvement in reward over naive model predictive control (MPC) for $2^8 = 256$ i.i.d experiments in each column. Standard errors were all about 1%.

Contributions. We propose a continuity constraint on counterfactual probabilities that admits an adaptive method for partially identifying the outcomes of action trajectories. This is used as a *sensitivity model* for the hidden confounding by setting a single parameter $\Gamma \geq 0$, associated with a norm over action trajectories, that quantifies the extent of hidden confounding and can be calibrated online (Definition 4). We formally characterize sharp bounds for the partially identified outcome of an action trajectory (Lemma 2). The sharp lower bound naturally gives rise to a minimax model-predictive controller (Lemma 3). We implement such a controller by augmenting a practical algorithm that is commonly used in deep reinforcement learning (Algorithm 1). Finally, we show empirically how this algorithm yields higher rewards on average compared to alternative methods across a wide diversity of linear and nonlinear synthetic experiments (Table 1).

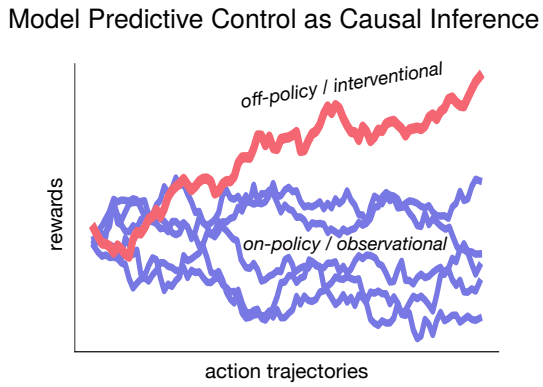

Figure 1. When planning to take actions in an online (interventional) setting, a dynamics model trained offline (on observational data) can usually be trusted more for action trajectories that remain near the reference policy. Trajectories are "off-policy" when they are generated by a new learned policy, which can be subject to hidden confounding.

2 BACKGROUND

The main question we want to answer is how to design a controller using an environment's *observable* behavior from actions taken under an unknown reference policy. Our focus is on the challenges that emerge when those actions interact with underlying state that is not observed. This is common; examples include robotics with limited sensing and text-based agents interacting with humans [Lang et al., 2024].

The field of artificial intelligence is beginning to realize the promise in deploying large (self-)supervised predictive models as *agents* to interact with the world [Acharya et al., 2025, Wang et al., 2024]. This agentic viewpoint can be cast as a predictive control problem, since the agent must learn to act from observations of the environment in order to achieve diverse goals. We wish to emphasize that without abundant feedback from the environment, and without full observability of the relevant state of the world, even the most capable predictive models could fail dramatically for planning actions to achieve a real-world goal [Saghafian, 2024]. It is a problem of *identifiability* when a model's predictions do not translate to interventions [Peters et al., 2017].

This paper considers one particular obstacle to identifiability: that of hidden confounding. Hidden confounding can easily manifest within the agentic paradigm because foundation models are seldom trained directly on the tasks that an agent would seek to accomplish. While reinforcement learning (RL) is employed to improve alignment or reasoning capabilities [Guo et al., 2025], the training process does not collect new data from the world, so the foundation model does not explore or learn from its own actions as a hypothetical agent. Interventional data are much more costly to obtain than the observational datasets that enable foundation models.

Motivating example. To motivate the problem setting, we consider an application of market impact [Guéant, 2016] of agents in the financial sector [Bai et al., 2025]. One may wish to understand how to optimally rebalance a portfolio by executing specific trades. However, the trader has not observed the full dynamics of how market participants reacted to past trades. As the trader interacts with the market, additional hidden factors may influence the evolution of the price. This may lead to the trader wanting to execute the trade according to an upper or lower bound on the expected price impact under the hidden confounding. The proposed method considers a controller that achieves this goal.

2.1 MODEL PREDICTIVE CONTROL

A *world model* that can predict the dynamics of actions and future states can readily implement agents through *model predictive control* (MPC) [Clarke et al., 1987], a widely celebrated family of algorithms for adaptive control [Fernandez-Camacho and Bordons-Alba, 1995] that project entire trajectories of states and actions into the future, select the best one, and execute the first action in that trajectory. MPC in model-based RL enables fast learning [Lale et al., 2021, 2024], as well as generalizable and multi-task agents [Hansen et al., 2024, Hu et al., 2023]. Moreover, world models trained online can be used offline to learn agents for novel tasks [Georgiev et al., 2024, Hafner et al., 2023]. A trend is emerging for offline-trained world models in realms that were traditionally suited for online RL [LeCun, 2022, Ajay et al., 2023] likely due to data accessibility and the demonstrated scalability of self-supervised learning [Chen et al., 2020].

The lack of identifiability in a partially observed system holds for MPC as well, specifically when using an offline-trained world model for novel tasks. Our goal is to provide an approach for *partially identifying* the outcomes of an agent's actions while leveraging a world model's predictions. To do so, it is necessary to assume a structural constraint on the impact of, or *sensitivity* to, hidden confounding, manifesting as a form of continuity in the action space. We propose theoretically-guaranteed conservative MPC under the worst-case scenarios admitted by the partial identification. This work is in a similar spirit, but orthogonal to "offline RL" with hidden confounding; we elaborate below.

2.2 OFFLINE REINFORCEMENT LEARNING

Offline RL refers to the class of methods for learning action policies from data collected under a reference policy that cannot be updated, and that is not from an expert—i.e., does not maximize rewards for the task of interest. Most offline RL algorithms borrow from online RL with the addition of regularization to protect against domain shift [e.g. Kumar et al., 2020]. They tend to involve learning a state-action value *Q-function* for the current action policy and iteratively

optimizing a new policy on the basis of a Bellman equation. These approaches can be efficient and robust [Panaganti et al., 2022]. Sensitivity to hidden confounding has also been incorporated through structural constraints [Bennett et al., 2024], latent variables [Pace et al., 2024], as well as adjustment through auxiliary variables [Wang et al., 2025].

Our focus on MPC diverges from those lines of work. The scope of this paper assumes access to an accurate (offline, confounded) world model, with the task of using it for online control. This regime is becoming relevant to real-world problems with the emergence of foundation models, yet also contrasts with classical control theory by allowing the dynamics—crucially, of the hidden confounders—to largely remain a black box. The constraint on the hidden confounders is meant to be adaptive to most data-generating processes.

2.3 CAUSAL INFERENCE

Our main insight is that recent theoretical tools from the intersection of causal inference and machine learning can be deployed to this context of partially identified predictive control. Causal inference is primarily concerned with the identification and estimation of causal relationships among variables. Many have studied the necessary and sufficient conditions for identifying one variable's outcome from another variable's intervention [Imbens and Rubin, 2015]. In the presence of hidden confounding, researchers have developed *sensitivity models* that impose structural constraints on the confounders, and yield tractable bounds for the *causal estimand*. Hidden confounders are distinct from latent confounders, the latter being possible to infer to some extent. In general, data cannot carry information about hidden confounders, and structural constraints can help to quantify a model's ignorance instead. Sensitivity models have a long history of improving the robustness of observational studies [Cornfield et al., 1959] and are making their way into machine-learning pipelines for the sciences [Feuerriegel et al., 2024, Haddad et al., 2023].

The push to make these methods useful in machine learning has led to more general sensitivity models: the univariate binary or discrete-intervention setting [Tan, 2006] has quickly evolved to continuous [Jesson et al., 2022, Marmarelis et al., 2023] and even multivariate [Frauen et al., 2024] interventions. Starting with Dorn and Guo [2022], progress has also been made in formally characterizing the sharpness of the bounds arising from these sensitivity models.

Considering MPC as the problem of identifying outcomes associated with *entire future trajectories* of actions, a sufficiently flexible sensitivity model should yield conservative policies in the presence of hidden confounding. Figure 1 illustrates the link between off-policy and interventions. We build on recent progress and present our analysis in the framework of *potential outcomes* introduced by Neyman [1923], which vastly simplifies notation and centers on identifiability.

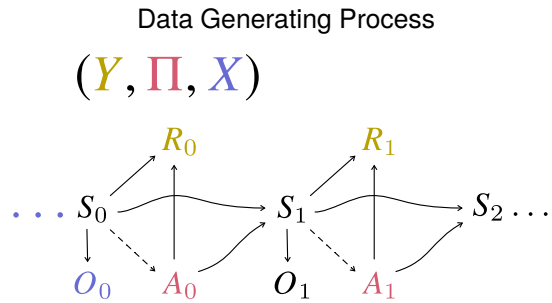

Figure 2. Graphical representation of the data-generating process. Dashed arrows correspond to the reference policy for the offline data. Dynamics are condensed into outcome Y, action trajectory Π, and observation history X; see §3.1.

3 THEORY

The main primitive underlying the data-generating process is a partially observable Markov decision process (POMDP). The POMDP consists of a sequence of state $S_t \in \mathcal{S}$, action $A_t \in \mathcal{A}$, and observation O_t random variables indexed in discrete time. O_t is the observable part of the full state S_t. There is also a reward $R_t \in \mathbb{R}_{\geq 0}$ that depends on the current state-action pair (S_t, A_t). The evolution of the POMDP is governed by a transition kernel $\tilde{T}(S_{t+1}|S_t, A_t)$ that is assumed to be unknown. All that is observed at each time step is the triplet $W_t \triangleq (O_t, A_t, R_t)$; the full state is hidden, making the process *partially observable*.

As in standard reinforcement learning, the goal of an agent is to choose actions that maximize expected future rewards with infinite horizon and discounting factor $\gamma \in (0, 1)$. Without loss of generality, denote the present context as $t = 0$. The agent acts on A_0 by picking from a set of choices $\mathcal{A}$ using the current observable state O_0 as well as any available past $(W_{-1}, W_{-2}, \cdots)$. The agent's objective is for repeated applications of its action policy to maximize $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R_t]$.

Our setting has an offline and online component. Data are collected offline under an unknown reference policy $\tilde{\pi}(A_t|S_t)$ and a predictive model is learned on the observables. We therefore assume access to samples from the conditional dynamics distributions $P_{W_0, W_1, \cdots | W_{-1}, W_{-2}, \cdots}$ that can be approximated arbitrarily well by a deep generative model. Any distribution P involving observables W_t is assumed to correspond to the offline data-generating process.

The agent must use P to act online, replacing $\tilde{\pi}$ in the data-generating process with its own policy π that aims to maximize the discounted reward in expectation. This is called *off-policy learning* because data generated from the agent's own policy are not available while learning. The domain shift between the offline and online POMDPs cannot be anticipated before the agent starts acting. In particular, because the full state S_t is unobserved, P is not guaranteed to help produce optimal actions even though it is the exact conditional

distribution of observables including actions and rewards.

3.1 HIDDEN CONFOUNDING

We simplify notation according to Figure 2 before proceeding with identification. The outcome of interest is the future discounted reward $Y \triangleq \sum_{t=0}^{\infty} \gamma^t R_t$. The agent makes plans on the basis of action trajectories taking the form $\Pi \triangleq [A_0\ A_1\ A_2\ \cdots]$ belonging to an $\mathcal{A}$-product space of finite or even infinite dimensionality, depending on the planning horizon. The agent's context is the current and past observable states, as well as actions, $X \triangleq [O_0\ \ O_{-1}\ A_{-1}\ \ O_{-2}\ A_{-2}\ \cdots]$ [Littman and Sutton, 2001].

In the MPC framework, the optimal controller is that which selects the action trajectory Π that maximizes the reward Y. This notation allows the abstraction of the dynamics in a (partially observed) Markov decision process. The optimal plan starting at a state $s \in \mathcal{S}$ is ultimately specified by

$$\pi^* \in \arg\max_{\pi \in \mathcal{T}} \mathbb{E}[Y \mid \Pi = \pi, S_0 = s]. \tag{1}$$

$\mathcal{T}$ denotes the set of all feasible action trajectories: like a power set of $\mathcal{A}$. Since S_0 is not observed, it must be inferred with all of the available information in X. However, some of the statistical variation in S_0 will probably leak through, and manifest as residual (hidden) confounding.

3.2 POTENTIAL OUTCOMES

The naive solution to action-trajectory selection would be like Equation (1) but simply using the observables instead.

$$\pi^*_{\text{naive}} \in \arg\max_{\pi \in \mathcal{T}} \mathbb{E}[Y \mid \Pi = \pi, X = x] \tag{2}$$

Clearly, if X cannot perfectly predict S_0, then these solutions might be different. A solution to Equation (2) might yield a high expected reward in the offline setting, but that is not guaranteed in the online setting in which any confounding between Π and S_0, conditioned on X, is removed. The online outcome $\mathbb{E}[Y \mid \Pi = \pi^*_{\text{naive}}, S_0 = s_0]$ is unidentifiable.

We require a simple notation for the outcomes of a potential online intervention in an instance described by the observable X. The potential-outcomes framework [Rubin, 1974, Imbens and Rubin, 2015] provides such a theory, and can flexibly handle vector-valued interventions [Marmarelis et al., 2024].

Definition 1 (Potential Outcome)**.** *For every decision-making instance, the realized outcome Y is the future reward from the offline dynamics, and the potential outcome $Y(\pi)$ associated with any action trajectory π is the future reward that would be realized online from following actions π.*

Potential outcomes and realized outcomes follow a joint distribution because each individual instance of dynamics

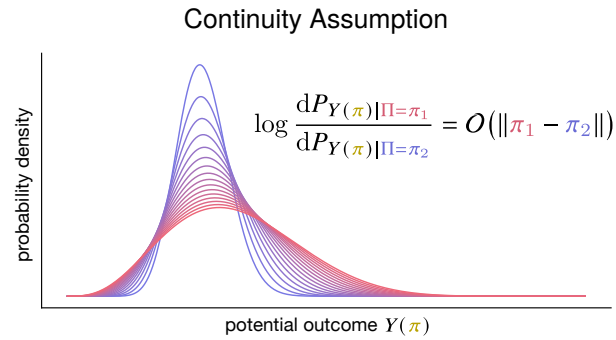

Figure 3. Continuity in the outcome counterfactual probabilities is set in terms of a norm on realized action trajectories.

is considered to have its own set of potential outcomes indexed by π. For a particular instance of state and observable (s_0, x), the *marginal* behavior of one potential outcome can be expressed as

$$\big(Y(\pi) \mid X = x\big) \sim \big(Y \mid \Pi = \pi, S_0 = s_0\big).$$

The most relevant insight is that without conditioning on the full state S_0, the offline action trajectory Π itself can reveal information about S_0 (statistically). Hence, $Y|\Pi, X$ is not predictive of $Y(\pi)$ because the underlying S_0 is not fixed across different $(\Pi = \pi)$ conditions. We formally define the counterfactual as

Definition 2 (Counterfactual)**.** *Conditional expressions of the form $Y(\pi) \mid \Pi, X$ are called counterfactual because they describe "what-if" scenarios where offline Π is observed, and we want to know the online outcome of a different π.*

Definition 3 (Causal Estimand)**.** *The quantity of interest for partial identification is $\mathbb{E}\big[Y(\pi) \mid X = x\big]$, to be evaluated at any (π, x) with support on $P_{\Pi,X}$.*

3.3 SENSITIVITY ANALYSIS

We begin our sensitivity analysis by considering the counterfactual distribution $P_{Y(\pi)|\Pi,X}$, where generally Π is called the exposure to disambiguate from π, the intervention.

Partial identification will be enabled by a mild continuity argument. The argument aligns with the reasoning that led to the classically celebrated marginal sensitivity model (MSM) for binary exposures [Tan, 2006]. The MSM constrains the Radon-Nikodym derivative of the two possible (binary) counterfactuals. In order to graduate to continuous exposure domains, Marmarelis et al. [2023] recently proposed local bounds for nearby counterfactuals. Concretely, their δMSM is derived by assuming a constraint on

$$\frac{dP_{Y(\pi)|\Pi=\alpha+\delta, X=x}}{dP_{Y(\pi)|\Pi=\alpha, X=x}} \approx 1, \tag{3}$$

for a sufficiently small value δ in the space of exposures, adapted to this paper's notation. Our analysis is inspired by a vector-valued extension of the δMSM applied to the action trajectories. By Equation (3), a form of continuity is placed on the counterfactual densities with respect to the observed trajectory Π at any feasible value α, for any potential trajectory π, and history x. In other words, it is describing how much the distribution of rewards Y for a *potential* trajectory π, denoted as $Y(\pi)$, could change with a perturbation in the *realized* action trajectory α. Any statistical dependence between $Y(\pi)$ and Π, conditioned on X, could only occur through hidden confounders that violate ignorability.

In a POMDP setting, the hidden state behaves as a hidden confounder whenever it affects the reference policy—which generates the offline realized trajectories—and the reward. It impedes identification of the reward for potential trajectories that are off-reference-policy, as in online planning. So far, the literature on causal sensitivity models has failed to provide an approach to partial identification that is *generally applicable* while also *adapting its bounds* based on how off-policy the counterfactuals in question really are.

A recently popular sensitivity model that can be used for vector valued exposures, and therefore action trajectories, is termed the CMSM [Frauen et al., 2024, Jesson et al., 2022]. While simple and surprisingly effective, the CMSM does not have a way to quantify whether some trajectories are more on-policy or off-policy than others, so it does not discriminate in its resultant bounds. On the other hand, the δMSM may provide a starting point for an adaptive sensitivity model because it considers continuity between nearby counterfactuals. We deviate from the original infinitesimal formulation of the δMSM and consider exposures and interventions in a general normed vector space of action trajectories.

First we re-frame the arguments (α, δ) by setting $\alpha = \pi$ and $\delta = \Pi - \pi$, so that $P_{Y(\pi)|\Pi=\alpha,X=x}$ becomes the identifiable quantity $P_{Y(\pi)|\Pi=\pi,X=x} = P_{Y|\Pi=\pi,X=x}$. Equation (3) transforms to

$$\frac{\mathrm{d}P_{Y(\pi)|\Pi,X=x}}{\mathrm{d}P_{Y|\Pi=\pi,X=x}} \approx 1. \tag{4}$$

This constraint is to hold almost everywhere in the joint probability space of $\big(Y(\pi), Y, \Pi\big)$, and for any (π, x) with support in $P_{\Pi,X}$. It is instructive to think of (π, x) as fixed and $\big(Y(\pi), Y, \Pi\big)$ as a triplet of random variables. The framing corresponds to the decision-making context, where for a given "state" x, we seek to evaluate possible interventions π. The existence of this Radon-Nikodym derivative can be guaranteed under the mild condition that all counterfactuals have identical support (in the outcome space $\mathcal{Y}$, shared by all potential and realized outcomes) [Kallenberg, 2002].

Suppose that a norm is defined over trajectories. If the counterfactual log-probability density functions could be assumed to be continuous in the realized trajectory Π, then the Radon-Nikodym derivative of Equation (4) could be constrained via Lipschitz continuity in $\|\Pi - \pi\|$ as illustrated in Figure 3.

Definition 4 (Sensitivity Model)**.** *Let $\Gamma \geq 1$ be the lowest constant such that*

$$\left|\log \frac{\mathrm{d}P_{Y(\pi)|\Pi,X}(Y \mid \Pi, X)}{\mathrm{d}P_{Y|\Pi,X}(Y \mid \pi, X)}\right| \leq \|\Pi - \pi\| \log \Gamma$$

almost everywhere, and for any action trajectory π. The scalar Γ is the sensitivity parameter for this model.

3.4 PARTIAL IDENTIFICATION

The proposed sensitivity model in Definition 4 ultimately places constraints on the relationships between any counterfactual density $P_{Y(\pi)|\Pi,X}(Y \mid \Pi, X)$ and its corresponding factual density $P_{Y|\Pi,X}(Y \mid \pi, X)$. These density ratios (formally Radon-Nikodym derivatives) are central to our analysis. Let us denote them as functions g_π.

$$g_\pi(Y, \Pi, X) \triangleq \frac{\mathrm{d}P_{Y(\pi)|\Pi,X}(Y \mid \Pi, X)}{\mathrm{d}P_{Y|\Pi,X}(Y \mid \pi, X)}$$

By the proposed sensitivity model, any such density ratio falls within the straightforward bounds $\left[\Gamma^{-\|\Pi-\pi\|}, \Gamma^{+\|\Pi-\pi\|}\right]$. Less straightforward is how to translate those bounds to the causal estimand given by Definition 3. We propose a change of measure. Specifically, we obtain the desired kernel by marginalizing over observational trajectories Π:

$$\begin{aligned} \tilde{g}_\pi(Y, X) &\triangleq \int_\Pi g_\pi(Y, \Pi, X)\, \mathrm{d}P_{\Pi|X}, \\ &= \frac{\mathrm{d}P_{Y(\pi)|X}(Y \mid X)}{\mathrm{d}P_{Y|\Pi=\pi,X}(Y \mid X)}. \end{aligned} \tag{5}$$

This integrated $\tilde{g}_\pi$ is the Radon-Nikodym derivative between potential and realized outcome distributions. Remarkably,

$$\mathbb{E}[Y\, \tilde{g}_\pi(Y, X) \mid \Pi = \pi, X = x] \;=\; \mathbb{E}[Y(\pi) \mid X = x]. \tag{6}$$

The kernel $\tilde{g}_\pi$ cannot be point-identified, but it does admit bounds from marginalization.

$$\mathbb{E}\big[\Gamma^{-\|\Pi-\pi\|} \;\big|\; X\big] \leq \tilde{g}_\pi(Y, X) \leq \mathbb{E}\big[\Gamma^{+\|\Pi-\pi\|} \;\big|\; X\big] \tag{7}$$

It is possible to produce sharp bounds on the expected future discounted reward, $\mathbb{E}[Y(\pi) \mid X]$, in a manner similar to other recent works on causal sensitivity analysis [Frauen et al., 2024, Oprescu et al., 2023, Dorn et al., 2024]. We seek to construct a lower-bounding function $\tilde{g}_\pi^{(-)}$ that yields

$$\mathbb{E}[Y(\pi) \mid X] \geq \mathbb{E}[Y\, \tilde{g}_\pi^{(-)}(Y, X) \mid \Pi = \pi, X], \tag{8}$$

and for which equality can hold with a feasible configuration of hidden confounders. It is important to ensure that the bounds are sharp in order to be resourceful with the sensitivity model, and for optimality results downstream.

For a putative kernel $\tilde{g}_\pi$ to be valid, it must satisfy any and all properties of a true $\mathrm{d}P_{Y(\pi)|X}/\mathrm{d}P_{Y|\Pi=\pi,X}$ that can be tested with the data (i.e., are identifiable). Crucially, we propose a necessary and sufficient condition of the form

Proposition 1 (Balancing Criterion)**.**

$$\mathbb{E}\left[\frac{\mathrm{d}P_{Y(\pi)|X}(Y \mid X)}{\mathrm{d}P_{Y|\Pi=\pi,X}(Y \mid X)} \,\middle|\, \Pi = \pi, X\right] = \int_{\mathcal{Y}} \frac{\mathrm{d}P_{Y(\pi)|X}(y \mid X)}{\mathrm{d}P_{Y|\Pi=\pi,X}(y \mid X)} \mathrm{d}P_{Y|\Pi=\pi,X}(y \mid X) \;=\; 1.$$

What remains is to construct $\tilde{g}_\pi^{(-)}$ such that it minimizes the weighted outcome expectation while obeying the sensitivity model *and* satisfying the balancing criterion. The latter is accomplished by placing a threshold on a particular quantile of $Y \mid \Pi, X$ that balances the extremes on either side of the sensitivity bounds. For high values of Y, $\tilde{g}_\pi^{(-)}$ must be as small as possible, and for low values of Y, it must be as high as possible. The line between high and low is drawn by that quantile threshold, which is denoted as $Q_\tau(\Pi, X)$.

Lemma 2 (Sharp Reward Bound)**.** *The sharp lower bound in Equation 8 can be achieved with the synthetic kernel*

$$\tilde{g}_\pi^{(-)}(Y, X) \triangleq \begin{cases} \mathbb{E}\left[\Gamma^{+\|\Pi-\pi\|} \,\middle|\, X\right] & \text{if } Y \le Q_\tau(\pi, X), \\ \mathbb{E}\left[\Gamma^{-\|\Pi-\pi\|} \,\middle|\, X\right] & \text{if } Y > Q_\tau(\pi, X), \end{cases}$$

$$\tau \triangleq \frac{\mathbb{E}\left[\Gamma^{+\|\Pi-\pi\|} \mid X\right] - 1}{\mathbb{E}\left[\Gamma^{+\|\Pi-\pi\|} \mid X\right] - \mathbb{E}\left[\Gamma^{-\|\Pi-\pi\|} \mid X\right]}.$$

4 MINIMAX CONTROL

By following the insights of Lemma 2, an MPC algorithm can be ideally conservative by choosing the action trajectory with the highest expected-reward lower bound admitted by the hidden-confounding constraints. The controller is said to be minimax-optimal if its expected discounted reward from taking actions in every time step achieves the maximum out of all controllers' total worst-case scenarios.

Recall that the outcome optimized by MPC is, in theory, $Y = \sum_{t=0}^{\infty} \gamma^t R_t$ by its selection of infinite-horizon action trajectories $\Pi = [A_0\ A_1\ A_2\ \cdots]$, using the observable state representation $X = [O_0\ O_{-1}\ A_{-1}\ O_{-2}\ A_{-2}\ \cdots]$. We take a perspective of stochastic control of an uncertain system [Bertsekas, 2025]. Uncertainty is the induced hidden confounding when the controller uses X instead of S_0.

Let $\mathcal{U}(A_0, X)$ denote the set of values an uncertainty variable U can take such that the actual instantaneous reward R_0 and state transition $X' \triangleq [O_1\ O_0\ \cdots]$ from any action $A_0 = a$ at current state representation $X = x$ is indexed by conditioning U on some value $u \in \mathcal{U}(A_0, X)$. Further, let that mapping be bijective: any such u must induce an admissible reward and state transition. In that case, the minimax controller must satisfy the Bellman equation

$$V^*(x) = \max_{a \in \mathcal{A}} \inf_{u \in \mathcal{U}(a,x)} \mathbb{E}\left[R_0 + \gamma V^*(X') \,\middle|\, A_0 = a, X = x, U = u\right]. \quad (9)$$

The sensitivity model ultimately places a constraint on how much the rewards of a controller's actions can vary from those that it predicted (by an offline world model). We assume that this is reflected in U. The value function of a stationary, deterministic policy $f : \mathcal{X} \to \mathcal{A}$ can be written as

$$V_f(x) = \inf_{u \in \mathcal{U}(f(x),x)} \mathbb{E}\left[R_0 + \gamma V_f(X') \,\middle|\, A_0 = f(x), X = x, U = u\right]. \quad (10)$$

Our theoretical MPC approach is such a policy f. It projects jointly sampled trajectories and then takes the first action of the best action trajectory from the closed set $\mathcal{T}$. By Lemma 2, our $f_{\mathrm{MPC}}(x)$ solves

$$\max_{a_0 : \pi = [a_0\ a_1\ \cdots] \in \mathcal{T}} \mathbb{E}\left[Y \tilde{g}_\pi^{(-)} \mid \Pi = \pi, X = x\right].$$

The quantity being maximized provides a lower bound on $\mathbb{E}[Y(\pi) \mid X = x]$, where $Y(\pi)$ follows the discounted-reward distribution from following the action trajectory π. Subsequent state transitions after a_0 *and* reward uncertainty are already encapsulated in $Y(\pi)$, by implication of Definition 1.

Plugging f_{MPC} into Equation 10, it can be seen that maximizing the sharp bound on the expected discounted rewards of the projected trajectories also maximizes $V_{f_{\mathrm{MPC}}}$. Details are provided in §A.2. We formalize this in the following lemma:

Lemma 3 (Minimax Control)**.** *The proposed partially identified MPC described by f_{MPC} reaches the minimax value,*

$$V_{f_{\mathrm{MPC}}}(X) = V^*(X) \quad \textit{almost everywhere.}$$

Lemma 3 illustrates the capability of the controller to achieve optimal rewards in the minimax sense.

4.1 IMPLEMENTATION

We present a concrete implementation of MPC with our partial identification strategy. State-of-the-art model-based RL algorithms [Hansen et al., 2024, Hu et al., 2023] tend to use a variant called model predictive path integral (MPPI) [Williams et al., 2015]. MPPI operates on samples of future dynamics by ranking and weighting action trajectories based on their projected rewards. We assume access to a generative model for the conditional distributions of action trajectories $P_{\Pi|X}$ and rewards $P_{Y|\Pi,X}$. In practice, the infinite horizon for discounted rewards needs to be approximated by a sufficiently long finite horizon, perhaps with a terminal value estimator if necessary.

Algorithm 1: Partially Identified MPPI (single step)

Input: dynamics models $\hat{P}_{\Pi|X}$ and $\hat{P}_{Y|\Pi,X}$,
decision-making context $x \in \mathcal{X}$,
sensitivity parameter $\Gamma \geq 1$

Output: best action $\hat{a}_0 \in \mathcal{A}$

1 Sample i.i.d action trajectories $\pi^{(1)}, \pi^{(2)}, \pi^{(3)}, \ldots$ according to $\hat{P}_{\Pi|X}(\pi \mid x)$;

2 **foreach** *search iteration* **do**

3 **foreach** *action trajectory* $\pi^{(i)}$ **do**

4 Estimate bounds for density ratio $\tilde{g}_{\pi^{(i)}}$ as $\hat{\mathbb{E}}\left[\Gamma^{\pm\|\Pi-\pi^{(i)}\|} \;\middle|\; X = x\right]$ using the π-sample;

5 Sample i.i.d reward trajectories $y^{(i,1)}, y^{(i,2)} \ldots$ according to $\hat{P}_{Y|\Pi,X}(y \mid \pi^{(i)}, x)$;

6 Estimate reward lower bound $\hat{y}^{(i)} \triangleq \hat{\mathbb{E}}\left[Y \tilde{g}^{(-)}_{\pi^{(i)}} \mid \Pi = \pi^{(i)}, X = x\right]$;

7 Update policy estimate $\hat{\pi}$ using action-reward pairs $(\pi^{(1)}, \hat{y}^{(1)}), (\pi^{(2)}, \hat{y}^{(2)}) \ldots$ as in classic MPPI;

8 Resample i.i.d action trajectories $\pi^{(1)}, \pi^{(2)} \ldots$ according to the current policy estimate $\hat{\pi}$;

9 Select and return first action $\hat{a}_0$ from policy estimate $\hat{\pi}$;

Algorithm 1 augments MPPI by lower-bounding the expected reward through Lemma 2 for each sampled action trajectory. Classic MPPI tends to use the reward sample directly. We adopt the same heuristics for ranking and weighting trajectories as in other works. During each search iteration, MPPI updates its policy estimate $\hat{\pi}$, which tends to be approximated as a multivariate Gaussian with diagonal covariance across time steps. In line 7 of Algorithm 1, the means and variances are estimated with the top-k trajectories, with weights computed through a softmax on the reward lower-bound estimates. In line 8, the action-trajectory sample is replaced with a sample of this Gaussian policy estimate, so that subsequent iterations further refine the distribution.

5 EXPERIMENTS

To fairly benchmark the core novelties of this paper, we compared variants of MPPI using the same dynamics models, and hyper-parameters tuned for the baseline algorithm that uses no partial identification. The MPPI without partial identification is referred to as "naive" as it is unaware of hidden confounding. Algorithm 1 shows MPPI augmented with our proposed sensitivity analysis. Similarly, MPPI can be augmented with the sensitivity model that has been studied in numerous recent works [Frauen et al., 2024, Kausik et al., 2024, Bennett et al., 2024], inspired by the classic MSM. In simplest terms, this sensitivity model constrains the divergence of the counterfactuals *uniformly*, rather than based on a norm in the intervention space. We refer to this baseline as "MSM". To highlight the difference in the kind of uncertainty under consideration, as well as to represent an approach from distributional RL [Bellemare et al., 2017], we present an additional baseline that takes lower conditional outcome quantiles in place of a causal sensitivity analysis. This baseline uses empirical uncertainties to emulate confounding uncertainty, so we termed it "empirical".

	Observed	Hidden	Nonlinearity
Easy	4	1	None
Medium	8	8	Sigmoid
Hard	16	16	Cubic

Table 2. Numbers of observed and hidden dimensions, as well as the type of nonlinearity, selected for the experimental settings with results in Table 1 and Figure 5.

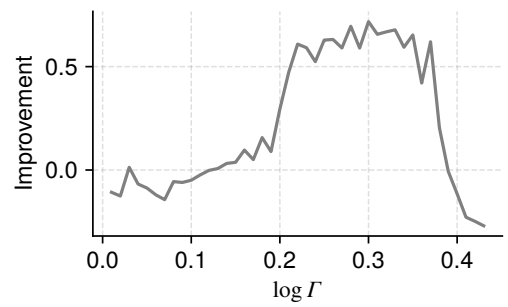

Figure 4. Reward improvement scores for the first "easy" experiment with our sensitivity analysis, as a function of $\log \Gamma$. This plot exhibits the trade-offs for increasing sensitivity. A low Γ encourages more, potentially careless action, whereas a high Γ could make the controller too conservative.

The goal is to assess the viability of these three approaches for online calibration of an offline-trained controller with hidden confounding. Each of the benchmarked methods has a single sensitivity parameter—Γ for ours and MSM, and the quantile level for the empirical baseline. We evaluated grids of these sensitivity parameters for each experiment, while ensuring that they overlapped as closely as possible in terms of relative performance. Then we identified the best-performing sensitivity value for each method and compared its total reward against that of the naive controller. These values are positive, since the possibility of no calibration is included in the search grid, setting the naive controller's total reward as a lower bound. An example calibration curve is shown in Figure 4 to illustrate how reward tends to increase, saturate, and then decrease with Γ.

For maximal generality in the simulations, we sampled multivariate stochastic differential equations (SDEs) of the Ornstein-Uhlenbeck (OU) process form [Karatzas and Shreve, 2019], with varying dimensionality and degree of nonlinearity. These processes had completely random structure, and were filtered for stability and significant confounding. We tested three distinct settings—"easy", "medium",

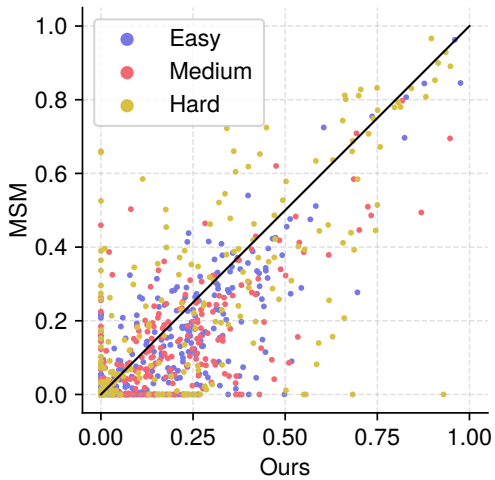

Figure 5. Comparing the pairwise results of Table 1 between our sensitivity model and the MSM baseline. We display improvements in reward over the naive controller for each of the 256 experiments. More points being to the right of the diagonal suggests that our model performed better.

and "hard"—with 256 independent experiments each. For the easy setting we trained simple linear SDE models, whereas for the medium and hard settings we trained neural models with longer windows into the past. For all settings, the controller's task was to minimize the squared value of the first dimension by controlling the second dimension. Concretely, the SDEs took the form $\mathrm{d}S_t = -h(AS_t)\mathrm{d}t + \sigma \mathrm{d}W_t$ where S_t is the full state vector, A is a mixing matrix, and $h(\cdot)$ is the optional nonlinearity given as a gradient of a convex function. R_t is given by $R_t = -\left(S_t^{[0]}\right)^2$ where the $\cdot^{[i]}$ superscript represents the ith component of S_t. The control is imposed on $S_t^{[1]}$, and the observed component is $\mathrm{d}O_t = \mathrm{d}S_t^{[0:k]}$ where k is the set of dimensions observed. The action a_t is applied to $O_t^{[1]}$ at each time step. Different realizations of processes were simulated in discrete time through the Euler-Maruyama scheme. Additional details are available in §B.

Results are mainly displayed as relative improvements in reward over the naive controller. Table 1 shows average improvements for our method compared with the baselines, across the three experimental settings described in Table 2. Improvements per experiment are plotted for our method versus the MSM in Figure 5. In aggregate, our method appears to yield a 20%+ increase in reward over the MSM.

6 DISCUSSION

The empirical results (§5) present two key findings. First, there is value to augmenting model predictive control (MPC) with a causal sensitivity analysis even in realistic settings. Second, the proposed sensitivity model (§3.3) that leverages a norm in the action-trajectory space is more helpful to MPC than more classical sensitivity models derived from the marginal sensitivity model (MSM) [Tan, 2006]. We also instantiate the partially identified MPC algorithm in the form of an augmented model predictive path integral (MPPI). MPPI is employed in several deep model-based RL algorithms that achieve the state of the art [e.g. Hansen et al., 2024].

The theoretical results (§3) motivate our sensitivity model in the context of recent developments in causal inference, and show the flexibility of the potential-outcomes notation. Our analysis (§3.4) reveals that the sharp partial identification is relatively simple, computationally tractable, and leads to minimax controllers (§4), in the sense that it finds the best policy for the worst-case scenario [Kallus and Zhou, 2021].

Notably, unlike previous work on sensitivity analysis for off-policy evaluation and learning, our approach *does not require the hidden confounders to be memoryless or static* [Kausik et al., 2024]. Instead, it allows the domain expert to select a norm that suits the action-trajectory space on which they wish to design an MPC algorithm. We look forward to further studying how to select these norms for different processes.

6.1 FUTURE WORK

A number of distinct avenues exist for extension of the current work. We consider two main threads in generative AI and online calibration.

Generative AI. The proposed methodology is general enough to suit various modalities, including text through large language models (LLMs). It appears that fine-tuning LLMs for specific tasks often reduces the diversity of their generations [Mohammadi, 2024, Kirk et al., 2024]. For this and other practical reasons, it may be more useful to use an LLM foundation model in combination with our sensitivity analysis for agents to solve tasks that are novel to the LLM. Simple algorithms in the spirit of MPC already find success in guiding LLMs [Beirami et al., 2024].

Online calibration. This paper considers the problem of partial identification and minimax control under a general class of sensitivity models. While the empirical evaluations show the utility of a calibrated sensitivity model, they do not show *how* to calibrate it online (its Γ parameter, or its choice of norm). There are numerous established solutions including bandits for online calibration. The sensitivity model's parameters are extremely low-dimensional and should therefore be easy to learn online, and much more data-efficient than wholesale online reinforcement learning.

6.2 LIMITATIONS

While we expand the existing theory on causal sensitivity analysis to make partial identification more data-adaptive,

especially in the novel action-trajectory setting, there are still fundamental limitations to the family of sensitivity models related to the MSM [Huang and Pimentel, 2025]. Our model shares the shortcomings of a pointwise hard constraint across all counterfactuals, namely that it can be untenable to use the Γ that absolutely covers all possible hidden confounders. Practically, there tends to be a Γ that is most helpful for achieving positive rewards, and this could be lower than the true Γ. A natural next step for this line of work is to turn the sensitivity model's constraint into a probabilistic statement, increasing its flexibility—especially at the tails of the conditional outcome distributions.

7 CONCLUSION

Our causal sensitivity analysis of action trajectories bridges recent developments in causal inference with off-policy learning. Model predictive control is becoming more popular for learning generalizable agents, and our contribution on dealing with partial observability is a promising step towards making them more reliable in the real world.

Acknowledgements

M. G. Marmarelis is supported in part by the Schmidt Foundation and the 2024-2025 Simoudis Discovery Prize. R. M. Alvarez and M. G. Marmarelis wish to acknowledge support from the Linde Center for Science, Society, and Policy (LCSSP). A. Anandkumar is supported in part by Bren endowed chair, ONR (MURI grant N00014-18-12624), and by the AI2050 senior fellow program at Schmidt Sciences.

References

Deepak Bhaskar Acharya, Karthigeyan Kuppan, and B Divya. Agentic ai: Autonomous intelligence for complex goals–a comprehensive survey. *IEEE Access*, 2025.

Anurag Ajay, Yilun Du, Abhi Gupta, Joshua B. Tenenbaum, Tommi S. Jaakkola, and Pulkit Agrawal. Is conditional generative modeling all you need for decision making? In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sP1fo2K9DFG.

Yahui Bai, Yuhe Gao, Runzhe Wan, Sheng Zhang, and Rui Song. A review of reinforcement learning in financial applications. *Annual Review of Statistics and Its Application*, 12(1):209–232, 2025.

Ahmad Beirami, Alekh Agarwal, Jonathan Berant, Alexander D'Amour, Jacob Eisenstein, Chirag Nagpal, and Ananda Theertha Suresh. Theoretical guarantees on the best-of-n alignment policy. *arXiv preprint arXiv:2401.01879*, 2024.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pages 449–458. PMLR, 2017.

Andrew Bennett, Nathan Kallus, Miruna Oprescu, Wen Sun, and Kaiwen Wang. Efficient and sharp off-policy evaluation in robust Markov decision processes. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=LKGuc2rY5v.

Dimitri P. Bertsekas. *Reinforcement Learning and Optimal Control*. Athena Scientific, 2nd edition, 2025.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22243–22255. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/fcbc95ccdd551da181207c0c1400c655-Paper.pdf.

David W Clarke, Coorous Mohtadi, and P Simon Tuffs. Generalized predictive control—part i. the basic algorithm. *Automatica*, 23(2):137–148, 1987.

Jerome Cornfield, William Haenszel, E Cuyler Hammond, Abraham M Lilienfeld, Michael B Shimkin, and Ernst L Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer institute*, 22(1):173–203, 1959.

Jacob Dorn and Kevin Guo. Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *Journal of the American Statistical Association*, pages 1–13, 2022.

Jacob Dorn, Kevin Guo, and Nathan Kallus. Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. *Journal of the American Statistical Association*, pages 1–12, 2024.

Eduardo Fernandez-Camacho and Carlos Bordons-Alba. *Model predictive control in the process industry*. Springer, 1995.

Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S Kohane, and Mihaela van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4):958–968, 2024.

Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. Sharp bounds for generalized causal sensitivity analysis.

Advances in Neural Information Processing Systems, 36, 2024.

Ignat Georgiev, Varun Giridhar, Nicklas Hansen, and Animesh Garg. Pwm: Policy learning with large world models. *arXiv preprint arXiv:2407.02466*, 2024.

Olivier Guéant. *The Financial Mathematics of Market Liquidity: From optimal execution to market making*. CRC Press, 2016.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Elizabeth Haddad, Myrl G Marmarelis, Talia M Nir, Aram Galstyan, Greg Ver Steeg, and Neda Jahanshad. Causal sensitivity analysis for hidden confounding: Modeling the sex-specific role of diet on the aging brain. In *International Workshop on Machine Learning in Clinical Neuroimaging*, pages 91–101. Springer, 2023.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Oxh5CstDJU.

Edward S. Hu, Richard Chang, Oleh Rybkin, and Dinesh Jayaraman. Planning goals for exploration. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6qeBuZSo7Pr.

Melody Huang and Samuel D Pimentel. Variance-based sensitivity analysis for weighting estimators results in more informative bounds. *Biometrika*, 112(1):asae040, 2025.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

Andrew Jesson, Alyson Rose Douglas, Peter Manshausen, Maëlys Solal, Nicolai Meinshausen, Philip Stier, Yarin Gal, and Uri Shalit. Scalable sensitivity and uncertainty analyses for causal-effect estimates of continuous-valued interventions. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Olav Kallenberg. *Foundations of modern probability*. Springer, 2nd edition, 2002.

Nathan Kallus and Angela Zhou. Minimax-optimal policy learning under unobserved confounding. *Management Science*, 67(5):2870–2890, 2021.

Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer, 2nd edition, 2019.

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

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=PXD3FAVHJT.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1179–1191. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/0d2b2061826a5df3221116a5085a6052-Paper.pdf.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Model learning predictive control in nonlinear dynamical systems. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 757–762. IEEE, 2021.

Sahin Lale, Peter I Renn, Kamyar Azizzadenesheli, Babak Hassibi, Morteza Gharib, and Anima Anandkumar. FALCON: Fourier adaptive learning and control for disturbance rejection under extreme turbulence. *npj Robotics*, 2(1):6, 2024.

Leon Lang, Davis Foote, Stuart Russell, Anca Dragan, Erik Jenner, and Scott Emmons. When your AIs deceive you: Challenges of partial observability in reinforcement learning from human feedback. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 93240–93299. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/a995960dd0193654d6b18eca4ac5b936-Paper-Conference.pdf.

Yann LeCun. A path towards autonomous machine intelligence (version 0.9.2). Position paper, 2022.

Michael Littman and Richard S Sutton. Predictive representations of state. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. URL https://proceedings.neurips.cc/paper_files/paper/2001/file/1e4d36177d71bbb3558e43af9577d70e-Paper.pdf.

Myrl G Marmarelis, Elizabeth Haddad, Andrew Jesson, Neda Jahanshad, Aram Galstyan, and Greg Ver Steeg. Partial identification of dose responses with hidden confounders. In *Uncertainty in Artificial Intelligence*, pages 1368–1379. PMLR, 2023.

Myrl G Marmarelis, Fred Morstatter, Aram Galstyan, and Greg Ver Steeg. Policy learning for localized interventions from observational data. In *International Conference on Artificial Intelligence and Statistics*, pages 4456–4464. PMLR, 2024.

Behnam Mohammadi. Creativity has left the chat: The price of debiasing language models. *arXiv preprint arXiv:2406.05587*, 2024.

Jerzy Neyman. On the application of probability theory to agricultural experiments. essay on principles. *Ann. Agricultural Sciences*, pages 1–51, 1923.

Miruna Oprescu, Jacob Dorn, Marah Ghoummaid, Andrew Jesson, Nathan Kallus, and Uri Shalit. B-learner: Quasi-oracle bounds on heterogeneous causal effects under hidden confounding. In *International Conference on Machine Learning*, pages 26599–26618. PMLR, 2023.

Alizée Pace, Hugo Yèche, Bernhard Schölkopf, Gunnar Ratsch, and Guy Tennenholtz. Delphic offline reinforcement learning under nonidentifiable hidden confounding. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=lUYY2qsRTI.

Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Robust reinforcement learning using offline data. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 32211–32224. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/d01bda31bbcd780774ff15b534e03c40-Paper-Conference.pdf.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Soroush Saghafian. Ambiguous dynamic treatment regimes: A reinforcement learning approach. *Management Science*, 70(9):5667–5690, 2024.

Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=ehfRiF0R3a.

Haolin Wang, Lin Liu, Jiuyong Li, Ziqi Xu, Jixue Liu, Zehong Cao, and Debo Cheng. Off-policy evaluation for multiple actions in the presence of unobserved confounders. In *THE WEB CONFERENCE 2025*, 2025. URL https://openreview.net/forum?id=LfVmZ8vaFp.

Grady Williams, Andrew Aldrich, and Evangelos Theodorou. Model predictive path integral control using covariance variable importance sampling. *arXiv preprint arXiv:1509.01149*, 2015.

Off-policy Predictive Control with Causal Sensitivity Analysis (Supplementary Material)

Myrl G. Marmarelis[1] Ali Hasan[2] Kamyar Azizzadenesheli[3] R. Michael Alvarez[1] Anima Anandkumar[1]

[1]Caltech
[2]Morgan Stanley
[3]NVIDIA

A PROOFS

A.1 LEMMA 2

Proof. We reproduce the target bound in Equation 8 below.

$$\mathbb{E}[Y(\pi) \mid X] \geq \mathbb{E}[Y\, \tilde{g}_\pi^{(-)}(Y, X) \mid \Pi = \pi, X].$$

An intermediate result is that the proposed balancing criterion (Proposition 1) is sufficient to induce sharpness.

$$\mathbb{E}\left[\frac{\mathrm{d}P_{Y(\pi)|X}(Y \mid X)}{\mathrm{d}P_{Y|\Pi=\pi,X}(Y \mid X)} \,\middle|\, \Pi = \pi, X\right] = 1$$

Following these conditions, we can use the proposed form for $\tilde{g}_\pi^{(-)}$,

$$\tilde{g}_\pi^{(-)}(Y, X) \triangleq \begin{cases} \mathbb{E}\left[\Gamma^{+\|\Pi-\pi\|} \,\middle|\, X\right] & \text{if } Y \leq Q_\tau(\pi, X), \\ \mathbb{E}\left[\Gamma^{-\|\Pi-\pi\|} \,\middle|\, X\right] & \text{if } Y > Q_\tau(\pi, X), \end{cases} \qquad \tau \triangleq \frac{\mathbb{E}\left[\Gamma^{+\|\Pi-\pi\|} \mid X\right] - 1}{\mathbb{E}\left[\Gamma^{+\|\Pi-\pi\|} \mid X\right] - \mathbb{E}\left[\Gamma^{-\|\Pi-\pi\|} \mid X\right]}.$$

In the rest of the proof, we omit X in order to relieve the notational burden. Assume that the expectations are all implicitly conditioned on X. The results are invariant to its conditioning. Additionally, the outcome space $\mathcal{Y}$, which represents cumulative non-negative rewards, must itself be assumed to be a connected subset of $\mathbb{R}_{\geq 0}$.

We note that since the construction $\tilde{g}_\pi^{(-)}(Y, X)$ is already constrained by the proposed sensitivity model as well as the putative balancing criterion, the burden of proof has shifted to the validity of the bound rather than its sharpness.

First, we prove validity. Suppose that there exists a system in which $\mathbb{E}[Y(\pi)] < \mathbb{E}[Y\, \tilde{g}_\pi^{(-)} \mid \Pi = \pi]$ for some fixed π.

By implication of Equation 6, there is an oracle $\tilde{g}_\pi$ such that

$$\mathbb{E}[Y\, \tilde{g}_\pi \mid \Pi = \pi] < \mathbb{E}[Y\, \tilde{g}_\pi^{(-)} \mid \Pi = \pi], \quad \text{and} \quad \mathbb{E}[(\tilde{g}_\pi - \tilde{g}_\pi^{(-)})Y \mid \Pi = \pi] < 0.$$

Both kernels must be balanced in the sense that $\mathbb{E}_Y[\tilde{g}_\pi - \tilde{g}_\pi^{(-)} \mid \Pi = \pi] = 0$. They are also both positive for all $y \in \mathcal{Y}$, as required for the existence of the proposed sensitivity model (Definition 4). It stands to reason that the oracle kernel $\tilde{g}_\pi$ has moved probability mass in $\tilde{g}_\pi^{(-)}$ from some points to other points (both sets with nonzero measure in $\mathcal{Y}$). In fact, for the $\tilde{g}_\pi$-weighted conditional expectation of Y to be lowered, there must be probability mass that is moved down, i.e. from a point in $\mathcal{Y}$ to a lower point in the same domain. Take any such pair of points for which that is the case:

$$y_1 > y_2 \quad \wedge \quad \tilde{g}_\pi(y_1) < \tilde{g}_\pi^{(-)}(y_1) \quad \wedge \quad \tilde{g}_\pi(y_2) > \tilde{g}_\pi^{(-)}(y_2).$$

The kernels are subject to the sensitivity bounds of Equation 7:

$$\mathbb{E}\left[\Gamma^{-\|\Pi-\pi\|}\right] \leq \tilde{g}_\pi(Y) \leq \mathbb{E}\left[\Gamma^{+\|\Pi-\pi\|}\right].$$

We must analyze where (y_1, y_2) fall around the threshold $Q_\tau(\pi)$. If they are both on one side, so either $\{Q_\tau(\pi) \geq y_1 > y_2\}$ or $\{y_1 > y_2 > Q_\tau(\pi)\}$, then the sensitivity bounds are trivially violated since $g_\pi^{(-)}$ lies at the boundary.

The remaining case is $\{y_1 > Q_\tau(\pi) \geq y_2\}$. There, it follows that

$$\tilde{g}_\pi(y_1) < \tilde{g}_\pi(y_1) = \mathbb{E}[\Gamma^{-\|\Pi-\pi\|}],$$
$$\tilde{g}_\pi(y_2) > \tilde{g}_\pi(y_2) = \mathbb{E}[\Gamma^{+\|\Pi-\pi\|}],$$

which both violate the sensitivity bounds, concluding the proof by contradiction.

Second, we prove sharpness. For any π, we seek a feasible $Y(\pi)$ such that $\mathbb{E}[Y(\pi)] = \mathbb{E}[Y\, \tilde{g}_\pi^{(-)} \mid \Pi = \pi]$. For the present purposes, it is enough to treat $Y(\pi)$ as some latent variable, as long as it satisfies the various conditions imposed by the problem. If we show that $\tilde{g}_\pi^{(-)}$ is a valid Radon-Nikodym derivative of the form

$$\tilde{g}_\pi^{(-)}(y) = \frac{\mathrm{d}P_Z(y)}{\mathrm{d}P_{Y|\Pi=\pi}(y)}$$

for some hypothetical Z, then we can have $Y(\pi) \triangleq Z$. The latent Z is completely determined by Y and π, with measure

$$P_Z(Y) = \int_Y \tilde{g}_\pi^{(-)}(y)\, \mathrm{d}P_{Y|\Pi=\pi}(y).$$

□

A.2 LEMMA 3

Proof. The main idea is that for every $Y(\pi)$, there exists a sequence $(u_0, u_1, \dots)$ giving that expected discounted reward. To do this, we will first illustrate the equivalence of the value function and the MPC problem we are solving. Next, we will define a lower bound of the value function conditioned on an uncertainty variable. Finally, we will show that the infimum of this function over the uncertainty set is equal to the value function under the proposed objective.

A point-identified MPC would solve a global optimization over future action trajectories.

$$\begin{aligned}
V(x) &= \max_{\pi=[a_0\ a_1\ \cdots]\in\mathcal{T}} \mathbb{E}\big[Y(\pi) \mid \Pi = \pi, X = x\big] \\
&= \max_{\pi=[a_0\ a_1\ \cdots]\in\mathcal{T}} \mathbb{E}\big[R_0 + \gamma R_1 + \gamma^2 R_2 + \cdots \mid \Pi = \pi, X = x\big] \\
&= \max_{\pi=[a_0\ a_1\ \cdots]\in\mathcal{T}} \mathbb{E}\big[R_0 \mid \Pi = \pi, X = x\big] \\
&\qquad\qquad + \gamma\mathbb{E}\big[R_1 + \gamma R_2 + \gamma^2 R_3 + \cdots \mid \Pi = \pi, X = x\big] \\
&= \max_{\pi=[a_0\ a_1\ \cdots]\in\mathcal{T}} \Bigg(\mathbb{E}\big[R_0 \mid A_0 = a_0, X = x\big] \\
&\qquad\qquad + \gamma\mathbb{E}\Big[\underbrace{\mathbb{E}\big[R_0 + \gamma R_1 + \gamma^2 R_2 + \cdots \mid \Pi = [a_1\ a_2\ \cdots], X = X'\big]}_{\text{(shifting one step ahead in time)}} \mid A_0 = a_0, X = x\Big]\Bigg) \\
&= \max_{a_0\in\mathcal{A}} \Bigg(\mathbb{E}\big[R_0 \mid A_0 = a_0, X = x\big] \\
&\qquad\qquad + \gamma\mathbb{E}\Bigg[\underbrace{\max_{\pi'=[a_1\ a_2\ \cdots]\in\mathcal{T}} \mathbb{E}\big[R_0 + \gamma R_1 + \gamma^2 R_2 + \cdots \mid \Pi = \pi', X = X'\big]}_{V(X')} \mid A_0 = a_0, X = x\Bigg]\Bigg) \\
&= \max_{a_0\in\mathcal{A}} \mathbb{E}\big[R_0 + \gamma V(X') \mid A_0 = a_0, X = x\big]
\end{aligned}$$

The time shift from $\mathbb{E}\big[R_1+\gamma R_2+\gamma^2 R_3+\cdots \mid \Pi = [a_0\ a_2\ \cdots], X = x\big]$ to $\mathbb{E}\big[R_0+\gamma R_1+\gamma^2 R_2+\cdots \mid \Pi = [a_1\ a_2\ \cdots], X = X'\big]$ is justified by the observable-state transition distribution $X'|A_0, X$. Recall that the rewards are structured as $\mathcal{A}\times\mathcal{X} \to \Delta(\mathbb{R}_{\geq 0})$.

The blue outer expectation in the line following the time shift is over the next observable-state transitions $X'|A_0, X$. The red inner expectation is a shifted version of the MPC objective, which we also mark as red on an earlier line. By iterated expectation, the red and blue expectations together are identical to the yellow expectation shown earlier. However, since the blue outer expectation is over a variable (X') that is invariant to future actions $a_1, a_2, \ldots$, we can safely break apart the join maximization and move the maximization over future actions inside the blue outer expectation. This gives us equivalence between the value function viewpoint and the MPC problem we are solving.

Now we define a compact uncertainty set $\mathcal{U}(A_0, X)$ conditioned on an initial action and state action transitions. By the compactness of $\mathcal{U}$, there exists a sequence $(u_0, u_1, \ldots) \in \mathcal{U}$ that minimizes the expected reward. From the definition of the sensitivity model in Definition 4, the reward is bounded below by (8). By assumption, all hidden confounding is incorporated within the uncertainty set and the infimum of the value function over all elements within the uncertainty set provides the lowest expected reward.

Under partial identification, our MPC instead maximizes a lower bound:

$$\begin{aligned}
V_{f_{\text{MPC}}}(x) &= \max_{\pi=[a_0\ a_1\ \cdots]\in\mathcal{T}} \mathbb{E}\left[Y\tilde{g}_{\pi}^{(-)} \mid \Pi = \pi, X = x\right] \\
&= \max_{\pi=[a_0\ a_1\ \cdots]\in\mathcal{T}} \inf_{(u_0,u_1,u_2,\ldots)} \mathbb{E}\left[R_0 + \gamma R_1 + \gamma^2 R_2 + \cdots \mid \Pi = \pi, X = x, U_0 = u_0, U_1 = u_1, U_2 = u_2, \ldots\right] \\
&= \max_{a_0\in\mathcal{A}} \inf_{u\in\mathcal{U}(a_0,x)} \mathbb{E}\left[R_0 + \gamma V(X') \mid A_0 = a_0, X = x\right] \\
&= V^*(x)
\end{aligned}$$

giving us the equivalence between the value functions. □

B EXPERIMENTAL DETAILS

Source code for all experiments is included in the supplementary material. Key details are listed below.

MPC instantiation. The partially identified MPPI algorithm described in Algorithm 1 was run with the following uniformly set hyperparameters: sample 512 action trajectories during search iteration, sample 64 reward trajectories for each action trajectory, and then choose the top 32 action trajectories for updating the policy estimate. We used five search iterations. Additionally, since MPPI computes weights based on a softmax on the rewards, we first normalized the rewards by dividing them by their mean across action trajectories and then set a temperature of 0.01. We used a horizon of 16 with un-discounted rewards as a practical substitute for the infinite-horizon, discounted ideal setting.

Simulation. The simulations for the easy, medium, and hard settings described in Table 2 all used a simulation time step of $\mathrm{d}t = 0.1$ and noise scale of $\sigma = 0.1$, though the learning and control was performed at unit time intervals. The mixing matrix was kept relatively sparse in order to induce a variety of dependencies among the variables in the SDE. Specifically, an entry in the mixing matrix was nonzero with probability $2/n$ where n was the dimensionality of the whole SDE—including the hidden variables. Nonzero entries were drawn from a standard normal distribution. Then, dimensions in the stochastic process were reordered such that the first dimension "received" the most influence from the other dimensions and the second dimension "gave" the most influence to the others. This arrangement made it more feasible for the second dimension to have a chance at controlling the first dimension, as the control problem was posed. Finally, through rejection sampling on candidate mixing matrices, we ensured a high degree of hidden confounding. We estimated the Pearson correlation coefficients between *a)* the hidden dimensions and the action dimension's future, and *b)* the hidden dimensions and the reward dimension's future. We required the geometric mean of *(a)* and *(b)* to be greater than 0.33, a threshold that rejected the majority of the processes. For the hard setting, we lowered that threshold to 0.20 because it was difficult to find processes that would not be rejected.

Estimation. The easy setting was learned with a linear model while the medium and hard settings relied on a neural network (multilayer perceptron) with two hidden layers of size 256 and SELU activations. The neural networks had access to the past four time points for predicting the drift term (sans noise) of the next time point.

Calibration. The grid search to choose the top sensitivity parameter across our method and the MSM & empirical baselines was tuned for efficiency and balance. Figure 6 shows the frequencies of these calibrations along respective grids.

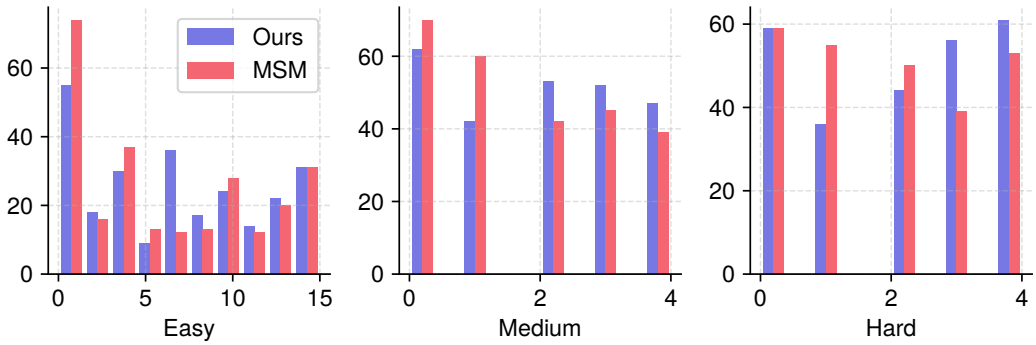

Figure 6. Histograms of the occurrence of the top reward along the calibration grids for $\log \Gamma$ that were considered for each experimental setting. Since the sensitivity parameters were generally incommensurable between our formulation and the MSM, we verified empirically that their respective grids were balanced, and with overlap in frequencies.