PRACTICAL PERFORMATIVE POLICY LEARNING WITH STRATEGIC AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper studies performative policy learning, where agents adjust their features in response to a deployed policy to improve their potential outcomes, thereby inducing endogenous distribution shifts. While there is growing interest in strategic environments—including strategic classification Hardt et al. (2016) and performative prediction Perdomo et al. (2020)—existing approaches often rely on restrictive parametric assumptions, such as micro-level utility models in strategic classification or macro-level distribution maps in performative prediction, which limit scalability. In this paper, we relax parametric assumptions on both micro-level agent behavior and macro-level data distribution. We identify a practical information structure in performative distribution shifts and introduce the evaluation vector—defined as the policy values at all feasible points a given agent can reach—as an effective mediator on the causal path from the deployed model to the shifted data. We propose a gradient-based policy optimization algorithm that uses a differentiable classifier as a surrogate for the high-dimensional distribution map. Our methodology exploits batch feedback, achieving superior sample efficiency over bandit feedback that summarizes responses into scalars. We provide theoretical convergence guarantees and demonstrate the method's efficacy in challenging high-dimensional settings.

1 INTRODUCTION

The interaction between machine learning models and society has become increasingly important because human behavior can induce distribution shifts. The canonical learning paradigm assumes a static data distribution and often fails in strategic environments. A growing body of work studies shifts caused by agents' strategic behavior. Unlike classical exogenous shifts, these arise endogenously from responses to deployed algorithms in pursuit of better outcomes.

One line of work is strategic classification, initiated by Brückner et al. (2012); Hardt et al. (2016). It typically models agent behavior as parametric utility maximization: an agent trades off the benefit of being classified into a desirable class under a manipulated feature vector against the cost of manipulation. Thus, strategic classification characterizes micro-level distribution shift through each agent's best response, given their original features, model knowledge, and manipulation cost. Recent work relaxes these assumptions Ghalme et al. (2021); Lechner et al. (2023), but practical model optimization still relies on parametric micro-level utility models Levanon & Rosenfeld (2021).

Performative prediction Perdomo et al. (2020) is a more general formulation that relaxes stringent assumptions about agents' utilities. From a causal perspective, it treats the predictive model as a complex intervention on the data distribution, yielding the distribution map $\mathcal{D}(\theta)$ from model parameters to the shifted distribution. Proposed methods learn the performatively optimal θ and this map through bandit feedback Jagadeesan et al. (2022) or gradient-based optimization under parametric specifications such as $\mathcal{N}(\mu(\theta), \sigma^2)$ Miller et al. (2021); Izzo et al. (2021). In practice, both θ and the covariate x can be high-dimensional, challenging both approaches.

Our objective is to make performative learning practical under mild structural assumptions. First, because θ is often high-dimensional, learning $\mathcal{D}(\theta)$ requires many deployments with sufficiently varied parameters, which is inefficient. Second, the deployed function f_θ, rather than its parameterization, is what agents observe and respond to. For a neural network, θ has little meaning apart from the architecture, making a parameter-to-distribution map opaque and difficult to learn. This functional

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.

view is also reflected in the micro-level mechanism in Eq. (4). We therefore model $\mathcal{D}(f_\theta)$ and exploit behavioral structure to reduce its effective dimension.

To simplify $\mathcal{D}(f_\theta)$, we draw on bounded rationality Conlisk (1996) as a behavioral lens. Bounded rationality suggests that agents may process only a limited subset of the information available to them due to cognitive and evaluation costs. In our context, agents interact with a potentially complex deployed prediction function f_θ through feasible alternatives they can practically consider, rather than by optimizing over the full feature space. We therefore model manipulation as selecting the highest-utility option from a finite consideration set. This finite-choice abstraction is consistent with classical ideas of satisficing Simon (1955) and consideration-set formation under evaluation costs Hauser & Wernerfelt (1990). More broadly, behavioral structure can be crucial for making complex strategic systems tractable Mazumdar et al. (2025).

This perspective is well suited to performative learning, where a principal interacts with many agents, such as loan applicants or platform users. Unlike the fully rational agents often assumed to solve complex nonconvex optimization problems in strategic classification, a boundedly rational agent with feature x need not process the entire function f_θ. Its relevant information can instead be summarized by finitely many evaluations, e.g., $(f_\theta(x_i))_{i=1}^m$, over a consideration set $(x_i)_{i=1}^m$.

We focus on policy learning with binary treatment, a decision-making task more general than the binary classification setting commonly studied in strategic classification. Incentive-aware machine learning often concerns decisions rather than predictions, such as loan approval or college admission—both canonical examples in related work. Throughout the paper, f_θ denotes a prediction function and π_θ a policy function; either can serve as an intervention on the data distribution.

To support gradient-based optimization for performative learning, we model the decision outcome directly, bypassing the agent's internal decision process (e.g., utility maximization). Concretely, we approximate the true performative distribution $\mathcal{D}(\pi_\theta)$ by $p(u' \mid v, (\pi_\theta(u,v))_{u\in\mathcal{U}})$, where u and u' denote the pre- and post-manipulation features, and v is the fixed context feature. Here, $(\pi_\theta(u,v))_{u\in\mathcal{U}}$, referred to as the **evaluation vector**, is the core component of our methodology: it records the policy values at all possible points an agent can achieve. Beyond addressing tractability issues, this distributional model moves beyond the setting of deterministic manipulation and naturally extends to nuanced micro-level scenarios, such as private types Levanon & Rosenfeld (2022) and random or noisy responses Jagadeesan et al. (2021).

In addition, we establish a convergence guarantee for our gradient-based policy optimization algorithm through an analysis in a smooth reproducing kernel Hilbert space (RKHS). Specifically, we assume that the response probabilities and their learned estimators are realizable in this space, so that function-approximation error controls the error in the response derivative used by our policy gradient. This route avoids the parametric distribution maps and finite-difference gradient estimation used in prior performative-learning algorithms Izzo et al. (2021), which can be fragile in high-dimensional settings.

1.1 Related Work

Strategic Classification. Strategic classification studies classifiers robust to manipulation. Recent work relaxes the original framework Brückner et al. (2012); Hardt et al. (2016) by allowing bounded manipulation, partial information, and stochastic responses Björkegren et al. (2020); Ghalme et al. (2021); Jagadeesan et al. (2021); Lechner et al. (2023); Shao et al. (2023). A causal line studies how classifiers induce genuine improvement, not superficial gaming Kleinberg & Raghavan (2020); Shavit et al. (2020); Bechavod et al. (2021); Chen et al. (2023); Podimata (2025). Recent work also studies behaviorally realistic departures from fully rational best responses Lv et al. (2026). Closest to us, Miller et al. (2020); Horowitz & Rosenfeld (2023) use the causal/non-causal feature distinction to characterize feature and outcome shifts. Our work shares this structural perspective, but targets scalable policy optimization without parametric utility models.

Performative Prediction. Perdomo et al. (2020) introduce the distribution map $\mathcal{D}(\theta)$ and performative risk $\mathbb{E}_{z\sim\mathcal{D}(\theta)}\ell(z;\theta)$, distinguishing performative stability (fixed points) from performative optimality (risk minimization). Stability has been studied Mendler-Dünner et al. (2020); Drusvyatskiy & Xiao (2022); Mofakhami et al. (2023), but stable points can be suboptimal Miller et al. (2021), motivating direct optimization. Izzo et al. (2021) approximate performative gradients under parametric distribution maps; see also Lin & Zrnic (2024); Yan & Cao (2023). Such methods are limited

by the data scarcity of bandit feedback Jagadeesan et al. (2022). Recent complexity results show that finding stable solutions can become intractable beyond weak-feedback regimes Anagnostides et al. (2026), further motivating exploitable response structure. We instead use micro-level behavior structure. Following work that views prediction as intervention and studies outcome performativity Mendler-Dünner et al. (2022); Kim & Perdomo (2023), we reduce the dimensionality of the functional intervention π_θ. Discrete agent responses provide tractability Bracale et al. (2025).

Policy Learning with Strategic Agents. Policy learning seeks high-value decisions, often via counterfactual prediction. In static settings, optimal policies reduce to conditional average treatment effect (CATE)-based cutoff rules Manski (2004); Athey & Wager (2021). Munro (2024) extends policy learning to strategic settings where the policy π itself acts as an intervention, showing that optimal policies need not retain this cutoff form. This motivates our move beyond CATE-only estimation. However, Munro's functional analysis lacks complexity reduction and scales only to one-dimensional features. Tsirtsis et al. (2024) also study strategic policy learning with discrete features, but without a practical learning algorithm. Sahoo & Wager (2025) study policy learning among competing agents using a related two-part policy-gradient decomposition. We instead study a single principal interacting with strategic agents.

2 Problem Setup

2.1 Multi-Period Setting

We use Neyman–Rubin potential outcome notation Rubin (1974) and model learning as a multi-period interaction between a principal and agents. For each period $t = 1, 2, \ldots, T$, the event flow is:

1. The principal releases policy $\pi = \pi_{\theta_t}$.
2. n new agents arrive, manipulate features, and generate post-manipulation features $X_i^\pi = x_i$ and potential outcomes $Y_i(1), Y_i(0)$, for $i = 1, \ldots, n$.
3. Agents report the post-manipulation features.
4. The principal observes x_i and assigns binary treatment $Z_i \sim \text{Bernoulli}(\pi_{\theta_t}(x_i))$.
5. Agents observe Z_i, and the principal observes outcomes $Y_i(Z_i)$.
6. The principal updates the policy parameter θ_t.

We assume the principal releases policy π. Even otherwise, deployed policies are often stable enough for external agents (e.g., credit consultancies) to reverse-engineer them from historical data Ghalme et al. (2021). Transparency can be deliberate when policies aim to incentivize improvement, such as skill development in college admissions. Finally, we simplify notation by treating potential outcomes $Y_i(1)$ and $Y_i(0)$ as implicitly dependent on manipulated feature X_i^π, i.e., $Y_i(1; X_i^\pi)$ and $Y_i(0; X_i^\pi)$.

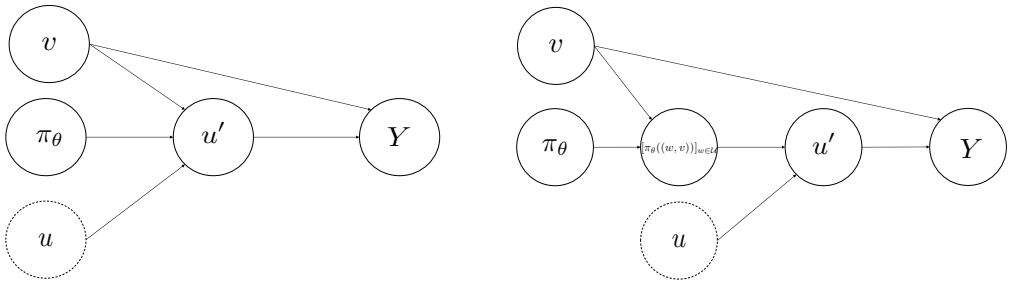
In each period, new agents arrive with features sampled independently from the same fixed distribution. Informed of π_θ, agents exert effort to modify specific features, submitting the final x_i. We focus on scenarios where this manipulation causally influences potential outcomes. Consequently, a well-designed policy can incentivize feature changes to improve outcomes Y. The principal then allocates treatment Z_i based on $\pi_{\theta_t}(x_i)$ and observes the outcomes to update θ_t.

For example, in loan applications, the bank (principal) evaluates applicants (agents). The treatment is a loan offer: $Y(0) = 0$, while $Y(1)$ reflects expected return net of default risk. This risk depends on manipulable features such as asset liquidity; agents may liquidate high-risk stocks to improve repayment capacity. A good policy incentivizes such improvement, whereas a poor one may encourage riskier allocations and increase default probability.

2.2 Gradient-based Performative Policy Learning

We now formulate the principal's learning problem. Let X_i^π denote the post-manipulation covariate vector observed after agent i responds to policy π, and let x be one of its realized values. Following Munro (2024), we define the conditional average treatment effect (CATE) by

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i^\pi = x]. \tag{1}$$



(a) Default case

(b) With proposed mediator

Figure 1: **(Left)** The original graph involving latent manipulable feature u (dashed node), fixed feature v, policy π_θ, post-manipulation feature u', and outcome Y. Here, the policy function π_θ acts as a complex functional intervention. **(Right)** Under our feature partition, the policy's influence is fully mediated by the evaluation vector $[\pi_\theta((w, v))]_{w\in\mathcal{U}}$, simplifying the causal mechanism.

Thus $\tau(x)$ is indexed by the realized post-manipulation covariate, not the agent's pre-manipulation state. We assume policy affects only the distribution of X_i^π, not the conditional potential-outcome contrast at fixed x; performativity thus creates covariate shift without concept shift.

We denote the post-manipulation density by $p(x; \pi_\theta)$, referred to as the *performative distribution*, which captures the policy's influence on the data. The policy value is $V(\pi_\theta) := \mathbb{E}_{x\sim p(\cdot;\pi_\theta)}[\pi_\theta(x)\tau(x)]$. The principal seeks an optimal policy π^* to maximize performative value. We adopt a supervised learning approach that uses interaction data as batch feedback, avoiding the limitations of bandit feedback. In online-learning terms, we target best-arm identification rather than cumulative regret minimization. The performative policy gradient is

$$\nabla_\theta V(\pi_\theta) = \mathbb{E}_{x\sim p(\cdot;\pi_\theta)}[\nabla_\theta \pi_\theta(x)\tau(x) + \pi_\theta(x)\tau(x)\nabla_\theta \log p(x; \pi_\theta)]. \tag{2}$$

The gradient decomposes into policy and density components, matching the performative gradient Izzo et al. (2021). Although derived from REINFORCE Williams (1992), the main challenge is formulating the **density gradient** $\nabla_\theta \log p(x; \pi_\theta)$. Prior work simplifies this by parameterizing the distribution and identifying π_θ with θ, yielding $\nabla_\theta p_\phi(x; \theta)$. This scales poorly with high-dimensional x and θ.

3 Main Methodology

3.1 Partitioning the Feature Space

We partition the feature space $\mathcal{X}$ into manipulable features $u \in \mathcal{U}$ and fixed features $v \in \mathcal{V}$. We treat $\mathcal{U}$ as discrete, representing primitive types or discretized features, consistent with bounded rationality.

Assumption 3.1. Each feature $x \in \mathcal{X}$ decomposes into a discrete manipulable component $u \in \mathcal{U}$ and a fixed component $v \in \mathcal{V}$, where $\mathcal{U}$ is finite.

Importantly, this is only a **modeling assumption**: the set $\mathcal{U}$ is the principal's finite consideration alphabet, chosen before policy learning and held fixed throughout optimization. Beyond making function-valued interventions technically tractable, discretization implements bounded rationality as a counterpart to the fully rational agents typically assumed in strategic classification. We argue that agents should not be assumed capable of searching the full feature space and solving a complex non-convex optimization problem, such as Eq. (4), when the policy π is nonlinear. Instead, they compare finitely many feasible alternatives, so the model does not require them to recover a global optimum. Indeed, even in our simulations, finding the global optimum is infeasible when π is parameterized as a neural network. Appendix C gives continuous-feature value bounds, reduces piecewise-constant policies to finite comparisons, and characterizes when refining the agents' grid no longer changes their selected policy cells. Section 5.3 tests this resolution gap empirically.

3.2 Encapsulating Impact via Evaluation Vectors

Leveraging this partition, we summarize the policy's influence on agent i by the *evaluation vector* $[\pi_\theta((u, v_i))]_{u\in\mathcal{U}}$, which evaluates π_θ over all $u \in \mathcal{U}$ with v_i fixed. Figure 1 shows that this vector mediates the path from π_θ to the reported signal $X_i^\pi = (u_i', v_i)$.

Assumption 3.2 (Causal Mechanism)**.** Conditional on the original manipulable feature u_0 and fixed feature v, the deployed policy affects the post-manipulation feature only through its evaluation vector:

$$u' = u'(u_0, v, \pi_\theta(\cdot, v)) \,. \tag{3}$$

Equivalently, policies that agree on $\{\pi_\theta(u, v) : u \in \mathcal{U}\}$ induce the same response distribution for covariates (u_0, v); policy values at contexts $v' \neq v$ are irrelevant.

This mechanism restricts an agent's view of the policy to its achievable alternatives and excludes cases such as cross-agent interference. It includes the classic strategic-classification response

$$u' = \operatorname{argmax}_u \ \pi_\theta(u, v) - c\|u - u_0\|, \tag{4}$$

where c is the manipulation-cost coefficient. We use only the policy-information structure encoded by this mechanism, without imposing its parametric utility.

Henceforth, u denotes the reported manipulable feature; when needed, u' and u_0 distinguish the reported and original features. Combining Assumptions 3.1 and 3.2 yields the following factorization.

Proposition 3.3. *The performative distribution satisfies*

$$p(x; \pi_\theta) = p(u \mid v, \zeta(v, \pi_\theta))p(v), \tag{5}$$

and consequently

$$\nabla_\theta \log p(x; \pi_\theta) = \nabla_\zeta \log p(u \mid v, \zeta)\, \nabla_\theta \zeta(v, \pi_\theta), \tag{6}$$

where $\zeta(v, \pi_\theta) = [\pi_\theta((w, v))]_{w\in\mathcal{U}}$.

The proof is in Appendix E.1. The transformation reduces the policy intervention from the high-dimensional parameter θ to the low-dimensional vector ζ. The observed v remains an ordinary covariate, while its policy-invariant marginal $p(v)$ need not be modeled.

End-to-end distribution-map approaches instead model $p(u \mid v, \theta)$. This is difficult because θ is high-dimensional, its relation to behavior is opaque, and only T distinct parameter vectors are deployed. Replacing θ by the function evaluations ζ respects the true intervention and turns the Tn agent responses into supervised feedback rather than T scalar bandit observations.

We introduce a differentiable classifier h_γ, termed the *behavior model*, to approximate $p(u \mid v, \zeta(v, \pi_\theta))$ without specifying agents' decision processes. We fit it by cross-entropy on labels u_i and inputs $(v_i, \zeta(v_i, \pi_{\theta_t}))$, using either a neural network or a Gaussian process classifier.

3.3 Strategic Policy Gradient

In addition to the behavior model h_γ, we estimate the CATE $\hat{\tau}(x)$ using meta-learners Künzel et al. (2019). Substituting these into Eq. (2) yields the gradient estimator:

$$\hat{g}_t = \frac{1}{n}\sum_{i=1}^{n}[\nabla_{\theta_t}\pi_{\theta_t}(x_i)\hat{\tau}(x_i) + \pi_{\theta_t}(x_i)\hat{\tau}(x_i)\nabla_{\theta_t}\log q(u_i \mid v_i, \zeta(v_i, \pi_{\theta_t}))]. \tag{7}$$

The resulting algorithm, *Strategic Policy Gradient*, is detailed in Algorithm 1. By using the behavior model to capture agents' reported features, it enables gradient-based optimization without parametric assumptions on microfoundations (e.g., utility maximization) or macro-level distributions.

We begin with a warm-up stage to pre-train the behavior model h_γ, CATE estimator $\hat{\tau}(x)$, and policy π_θ. In this phase, we perform repeated gradient ascent with the vanilla gradient, temporarily ignoring the performative component. This yields nT_0 samples across T_0 distinct θ_t values, creating sufficient variation in the evaluation vector ζ to learn h_γ before estimating performative responses. We then switch to the full gradient estimator in Eq. (7), using the learned behavior model.

We next highlight two implementation details. First, the behavior model requires careful discretization. The granularity $|\mathcal{U}|$, also the number of classification classes, trades off resolution against per-epoch sample efficiency. We further study this trade-off in Appendix D.3 and the experiments.

Algorithm 1 Strategic Policy Gradient

Input: Time horizon T, warm-up rounds T_0, learning rate η_1, η_2
// Warm-up stage, update θ_t using vanilla gradient
for $t = 1$ **to** T_0 **do**
  Deploy policy π_{θ_t}
  $\mathcal{D}_t \leftarrow \{(x_i, z_i, Y_i(z_i), [\pi_{\theta_t}((u, v_i))]_{u\in\mathcal{U}})\}_{i=1}^n$
  Update CATE estimator $\hat{\tau}$
  $\theta_{t+1} \leftarrow \theta_t + \frac{\eta_1}{n}\sum_{i=1}^n \nabla_\theta \pi_{\theta_t}(x_i)\hat{\tau}(x_i)$
end for
// Merge data collected in warm-up stage
$\mathcal{D}_{\text{warm}} \leftarrow \{\mathcal{D}_t\}_{t=1}^{T_0}$
Train h_γ on $\mathcal{D}_{\text{warm}}$
// Update θ_t with full gradient
for $t = T_0 + 1$ **to** T **do**
  Deploy policy π_{θ_t}
  $\theta_{t+1} \leftarrow \theta_t + \eta_2 \hat{g}_t$
end for

Second, optimization requires care. Although momentum is often useful for non-convex objectives, first-order momentum can reduce stability here: as the deployed policy continually shifts the data distribution, past gradients may become stale and potentially misleading. We therefore recommend Adagrad Duchi et al. (2011), which normalizes updates using accumulated squared gradients without a directional first moment and is empirically more stable.

4 Theoretical Analysis

4.1 Overview

Convergence to performatively stable points through repeated retraining is well understood Perdomo et al. (2020), but convergence to performative optimality is more challenging. The key step is to control the difference between the true and estimated performative gradients through the behavior-model and CATE estimation errors. Unlike finite-difference methods for parametric distribution maps Izzo et al. (2021), our differentiable behavior model scales to high-dimensional policies.

For fixed (u, v), write $p_{u,v}(\zeta) = p(u \mid v, \zeta)$ and $q_{t,u,v}(\zeta) = q_t(u \mid v, \zeta)$, where $\Omega = [0, 1]^{|\mathcal{U}|}$ is the evaluation-vector domain. The lemma below converts RKHS approximation error into gradient error.

Lemma 4.1. *Let f, g lie in an RKHS $\mathcal{H}$ with kernel K. Suppose that K has continuous mixed second derivatives on $\Omega \times \Omega$ and that each derivative representer $\partial_{\zeta_i} K(\zeta, \cdot)$ belongs to $\mathcal{H}$. Then*

$$\|\nabla_\zeta f(\zeta) - \nabla_\zeta g(\zeta)\| \le \|f - g\|_{\mathcal{H}} \left\| (\|\partial_{\zeta_i} K(\zeta, \cdot)\|_{\mathcal{H}})_{i=1}^{|\mathcal{U}|} \right\|. \tag{8}$$

The result follows from the derivative reproducing property; see Appendix E.2. For standard smooth kernels (e.g., Gaussian) on compact Ω, the derivative representers are uniformly bounded.

Assumption 4.2 (RKHS regularity)**.** For all (u, v) and post-warm-up epochs t, $p_{u,v}$ and $q_{t,u,v}$ lie in a common RKHS $\mathcal{H}$ as functions of ζ, uniformly in v, and $q_t(\cdot \mid v, \zeta)$ is a valid conditional distribution.

4.2 Theoretical convergence guarantee

Assumption 4.3 (Technical assumptions)**.** 1. (**RKHS boundedness**.) Uniformly over $\zeta \in \Omega$, $\|K(\zeta, \cdot)\|_{\mathcal{H}} \le R$ and $\max_i \|\partial_{\zeta_i} K(\zeta, \cdot)\|_{\mathcal{H}} \le G_K$.

2. (**Performative regularity**.) There is $\iota > 0$ such that, for every u, v, t, ζ, $p_{u,v}(\zeta), q_{t,u,v}(\zeta) \ge \iota$ and $\|\nabla_\zeta p_{u,v}(\zeta)\| \le G_p$.

3. (**Bounded policy gradient**.) $\|\nabla_\theta \pi_\theta(x)\| \le G_\pi$.

4. (**Bounded CATE**.) $|\tau(x)|, |\hat{\tau}(x)| \le B_\tau$; the estimate may be clipped to this range.

5. (**CATE estimation**.) $\mathbb{E}_{X\sim\mathcal{D}_0}[|\tau(X) - \hat{\tau}(X)|] \leq \epsilon_\tau$, where $\mathcal{D}_0$ is the warm-up distribution.

6. (**Policy-value regularity**.) $V(\pi_\theta)$ is l-smooth in θ.

We use Euclidean norms and write $m = |\mathcal{U}|$. The same lower bound holds for the warm-up conditionals, enabling importance weighting from $\mathcal{D}_0$ to current distributions.

Theorem 4.4 (Convergence of strategic policy gradient)**.** *Fix $\delta_1, \delta_2 \in (0, 1)$. Suppose that, with probability at least $1 - \delta_2$, the behavior-model approximation event satisfies*

$$\sup_{1\leq t\leq T} \sup_{u\in\mathcal{U},\, v\in\mathcal{V}} \|q_{t,u,v} - p_{u,v}\|_{\mathcal{H}} \leq \varepsilon_{\mathcal{H}},$$

and that the empirical gradient noise is simultaneously bounded with probability at least $1 - \delta_1$ by $B_n = \kappa\sqrt{\log(2T/\delta_1)/n}$ at every post-warm-up epoch. For a constant step size $0 < \eta \leq 1/l$,

$$\min_{1\leq t\leq T} \|\nabla_\theta V(\pi_{\theta_t})\|^2 \leq \frac{4B_\tau}{\eta T} + (G_E + B_n)^2 \tag{9}$$

with probability at least $1 - \delta_1 - \delta_2$, where

$$G_E = \frac{\epsilon_\tau G_\pi}{\iota} + \frac{\epsilon_\tau G_p G_\pi \sqrt{m}}{\iota^2} + \varepsilon_{\mathcal{H}} B_\tau G_\pi \sqrt{m} \left(\frac{R G_p}{\iota^2} + \frac{G_K \sqrt{m}}{\iota} \right). \tag{10}$$

The proof is in Appendix E.3. The theorem gives convergence to a neighborhood of a critical point whose radius is controlled by CATE error, behavior-model error, and finite-batch noise. The bound vanishes jointly as $T \to \infty$, $n \to \infty$, and $(\epsilon_\tau, \varepsilon_{\mathcal{H}}) \to 0$; fixed estimation error correctly leaves a nonzero error floor. Under strong concavity, it also controls distance to the optimum.

5 Experiment with Synthetic Data

5.1 Basic Setting

Experiments begin on synthetic data with $|\mathcal{U}| = 5$ and $\dim \mathcal{V} = 20$. Fixed features v are drawn from a Gaussian, while u is generated via $\mathbb{P}(u \mid v) = \mathrm{Softmax}(Wv)$ with a predefined W (details in Appendix A.1). We primarily employ the classic best-response mechanism:

$$u' = \mathrm{argmax}_u\, \pi_\theta(u, v) - c|u - u_0|. \tag{11}$$

We report results for $c = 0.1$; other values are in Appendix A.4. We also evaluate two classical noisy decision mechanisms, with Gaussian and Gumbel noise; see Appendix A.6.

We consider three baselines, detailed in Appendix A.3: ***Cutoff Rule***, $\pi(x) = \mathbb{I}(\tau(x) > 0)$; ***Vanilla Policy Gradient***, adapted from repeated gradient descent Perdomo et al. (2020), which ignores distribution shifts and updates via $\mathbb{E}_{x\sim p(\cdot)}[\nabla_\theta \pi_\theta(x)\tau(x)]$; and ***End2end Policy Gradient***, which models the performative distribution $p(u \mid v, \pi_\theta)$ directly as $p(u \mid v, \theta)$. All gradient-based methods share the CATE estimates and settings; only the post-warm-up gradient differs (after epoch 30).

We use a multilayer perceptron (MLP) as the behavior model; Appendix A.7 shows that Gaussian process classifiers also work. Any moderately expressive input-differentiable model suffices.

5.2 Results and Discussion

The policy value curves with $c = 0.1$ are shown in Figure 2a. The distributions of the manipulable feature, both before and after incentivization by each policy, are displayed in Figure 2b. Detailed results can be found in Appendix A.4. The error bars show $\pm$ one standard error. All methods were evaluated over 10 independent random seeds.

First, our method converges stably to high value. The cutoff is performatively stable—a fixed point of vanilla policy gradient—whereas End2end Policy Gradient remains trapped nearby. Thus performative stability can be far from optimality. Because larger u implies higher CATE, Figure 2b and Table 2 show that our method shifts the most agents and yields the largest mean increase in u.

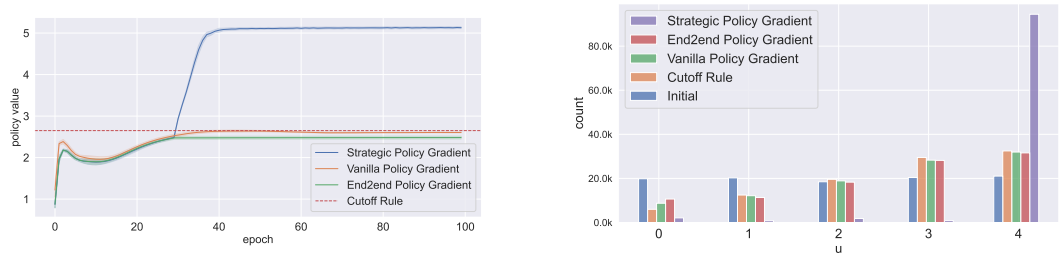


Figure 2: Deterministic best responses ($c = 0.1$). (a) After a shared 30-epoch warm-up, Strategic Policy Gradient surpasses the stable cutoff. (b) It produces the largest shift of u toward high values.

Figure 3 tests response noise, behavior-model architecture, and gradient informativeness. Under both Gaussian perturbations and multinomial-logit responses, Strategic Policy Gradient (SPG) breaks through the stable cutoff and converges to a higher plateau. Replacing the multilayer perceptron (MLP) classifier with a Gaussian process (GP) classifier gives nearly identical policy-value trajectories. The correction-gradient panel explains End2end's failure: its norm stays near zero after warm-up, whereas SPG receives a strong transient signal and then decays toward zero as the policy stabilizes. Full gradient, response, and GP diagnostics appear in Appendices A.5, A.6, and A.7.

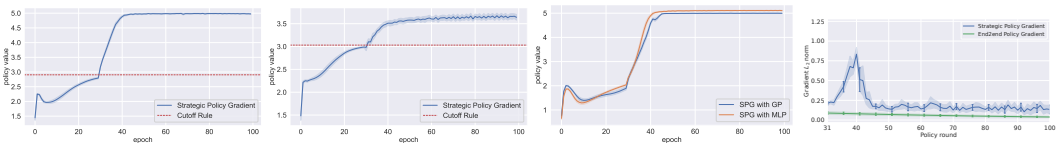


Figure 3: Synthetic robustness, left to right: Gaussian perturbations, multinomial-logit responses, GP versus MLP behavior models, and correction-gradient norms. SPG clears the stable cutoff under both noisy mechanisms and is robust to classifier family. End2end's correction gradient stays near zero, whereas SPG exhibits an informative post-warm-up signal that decays as the policy stabilizes.

5.3 Examining Coarse Discretization

We examine coarse discretization of the manipulable feature u, separating its true cardinality M (agents' view) from the principal's $|\mathcal{U}|$ bins. We test $M \in \{15, 50\}$ and $|\mathcal{U}| \in \{2, 3, 4, 5, 10\}$. With a 90-epoch warm-up for $|\mathcal{U}| = 10$, Figures 4a–4b show that 10 bins closely approach the optimum, while increasing M from 15 to 50 has negligible effect. Proposition C.5 explains this saturation: after the agents' grid resolves each principal cell, only near-tie types can switch cells. Proposition C.6 then separates principal-resolution, agent-grid, estimation, and optimization errors.

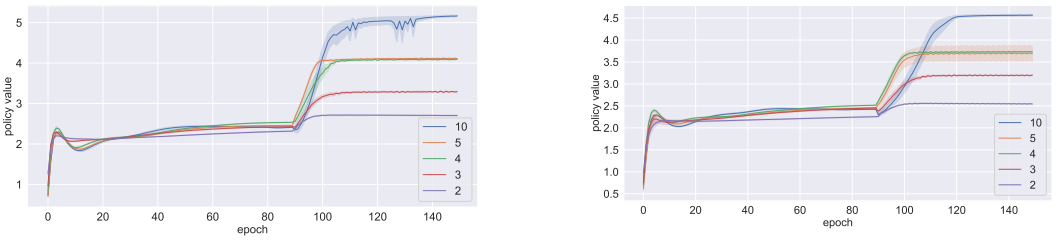


Figure 4: Policy value under coarse discretization. Panels use true cardinalities (a) $M = 15$ and (b) $M = 50$; each curve denotes a principal resolution $|\mathcal{U}|$.

6 Experiment with Semi-Synthetic Data

6.1 Basic Setting

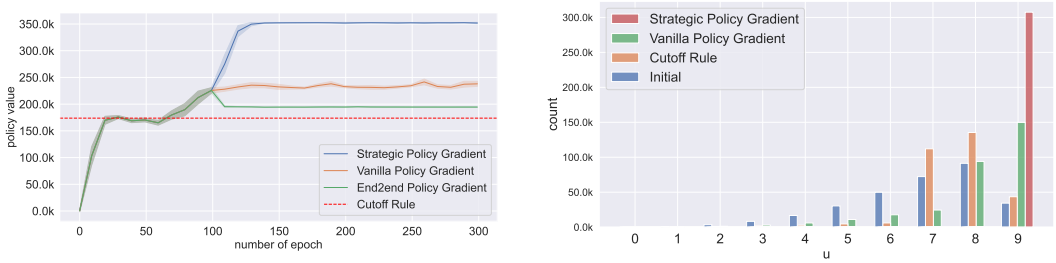
We conduct semi-synthetic experiments on a loan-default dataset with 307,508 records. Applicant attributes and loan terms serve as fixed features v. We construct a manipulable external credit score u from v and define Y as expected revenue, inducing a complex CATE. To mimic sequential decision-making, we split the data into 300 batches, discretize u into 10 levels, and use a more expressive policy network; Appendix B.1 gives details.

We also evaluate *End2end Policy Gradient*. Its behavior model takes the full 25,741-dimensional policy vector together with the semi-synthetic response history. As deployments add new response batches, its input dimension and training set are both large, so fitting scales poorly with data volume and $\dim(\theta)$; our evaluation vector remains 10-dimensional (Appendix A.3).

6.2 Results and Discussion

We evaluate cost coefficients $c \in \{0.1, 0.15, 0.2\}$, which are higher than in the synthetic setup. Figure 5 shows $c = 0.1$, with full tabulations—including End2end—in Tables 4–6. Our method converges rapidly to the highest value. End2end peaks during warm-up but settles near 1.95×10^5, below vanilla and far below SPG (Table 4); the same ordering persists for $c = 0.15$ and $c = 0.2$. Unlike the synthetic case, vanilla policy gradient exceeds the cutoff, so the static optimum is not performatively stable here. Since higher u reduces default risk, Figure 5b and Table 4 show that vanilla improves u more than the cutoff, whereas SPG maximizes it for every agent. At $c = 0.1$ and $c = 0.15$, End2end moves more agents than vanilla but achieves lower final value. Feature movement alone is therefore insufficient; policy value also depends on allocation quality after manipulation.

Appendix Figures 11–13 report the complete trajectories at all costs. The semi-synthetic gradient diagnostics in Figure 8 show that End2end again learns little correction signal, whereas SPG exhibits a pronounced post-warm-up correction that decays as the policy converges. Thus the value gap reflects learned performative information rather than merely a larger total update.



(a) (b)

Figure 5: Semi-synthetic experiments at $c = 0.1$. (a) Policy-value curves for all four methods. (b) Distributions of u for cutoff, vanilla, and strategic policies.

7 Conclusion

We study performative policy learning without parametric models of either micro-level decisions or macro-level distribution shifts. Our limited-manipulation framework treats the policy function $\mathcal{D}(\pi_\theta)$, rather than only its parameter vector $\mathcal{D}(\theta)$, as the intervention and exploits batch feedback for sample efficiency. The resulting behavior model connects strategic and performative perspectives, and synthetic and semi-synthetic experiments validate the approach. Gradient diagnostics attribute these gains to a learned performative correction rather than a larger vanilla update.

High-dimensional performative learning remains under-explored. Appendix C gives continuous-to-discrete value bounds and an oracle inequality separating principal-resolution, agent-grid, estimation, and optimization errors. These results motivate adaptive discretization and continuous-action extensions, together with online behavior diagnosis as responses evolve.

Reproducibility statement

The proposed estimator and optimization procedure are specified in Section 3.3 and Algorithm 1. Appendices A and B document the data-generating processes, preprocessing, model architectures, hyperparameters, baselines, random-seed protocol, and additional experiments. The assumptions and complete proofs of the theoretical results appear in Appendix E.

Ethics statement

This work studies policies that may be deployed in high-stakes allocation settings. The experiments are synthetic or semi-synthetic and involve no live deployment or human-subject intervention. Real-world use would require domain-specific review of privacy, fairness across subgroups, legal constraints, and the incentives created by the policy, together with ongoing monitoring for harmful gaming or distribution shifts. The method is intended to support outcome-improving responses rather than deceptive feature manipulation.

Artificial intelligence use statement

In this work, generative artificial intelligence (AI) tools were used to check and locally formalize author-developed mathematical derivations and proofs, including the discretization analysis; suggest corrections to proof exposition; and polish grammar and readability. They were not used to formulate the research questions, originate the core methodology, design or implement experiments, generate or clean data, write experimental code, or interpret results. All AI-assisted suggestions were independently reviewed, and every mathematical change was re-derived and verified by the authors. The authors take full responsibility for the final text, claims, proofs, and artifacts.

References

Ioannis Anagnostides, Rohan Chauhan, Ioannis Panageas, Tuomas Sandholm, and Jingming Yan. On the computational complexity of performative prediction. *arXiv preprint arXiv:2601.20180*, 2026. URL https://arxiv.org/abs/2601.20180.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021. doi: 10.3982/ECTA15732.

Yahav Bechavod, Katrina Ligett, Steven Wu, and Juba Ziani. Gaming helps! learning from strategic interactions in natural dynamics. In *International Conference on Artificial Intelligence and Statistics*, pp. 1234–1242. PMLR, 2021.

Daniel Björkegren, Joshua E Blumenstock, and Samsun Knight. Manipulation-proof machine learning. *arXiv preprint arXiv:2004.03865*, 2020. URL https://arxiv.org/abs/2004.03865.

Daniele Bracale, Subha Maity, Yuekai Sun, and Moulinath Banerjee. Learning the distribution map in reverse causal performative prediction. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 973–981. PMLR, 2025. URL https://proceedings.mlr.press/v258/bracale25b.html.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Michael Brückner, Christian Kanzow, and Tobias Scheffer. Static prediction games for adversarial learning problems. *The Journal of Machine Learning Research*, 13(1):2617–2654, 2012.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Yatong Chen, Jialu Wang, and Yang Liu. Learning to incentivize improvements from strategic agents. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=W98AEKQ38Y.

John Conlisk. Why bounded rationality? *Journal of economic literature*, 34(2):669–700, 1996.

Dmitriy Drusvyatskiy and Lin Xiao. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*, 48(2):954–998, 2022. doi: 10.1287/moor.2022.1287.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011. URL http://jmlr.org/papers/v12/duchi11a.html.

Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 385–394, 2005.

Ganesh Ghalme, Vineet Nair, Itay Eilat, Inbal Talgam-Cohen, and Nir Rosenfeld. Strategic classification in the dark. In *International Conference on Machine Learning*, pp. 3672–3681. PMLR, 2021.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pp. 111–122, 2016.

John R Hauser and Birger Wernerfelt. An evaluation cost model of consideration sets. *Journal of consumer research*, 16(4):393–408, 1990.

Guy Horowitz and Nir Rosenfeld. Causal strategic classification: A tale of two shifts. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13233–13253. PMLR, 2023. URL https://proceedings.mlr.press/v202/horowitz23a.html.

Zachary Izzo, Lexing Ying, and James Zou. How to learn when data reacts to your model: performative gradient descent. In *International Conference on Machine Learning*, pp. 4641–4650. PMLR, 2021.

Meena Jagadeesan, Celestine Mendler-Dünner, and Moritz Hardt. Alternative microfoundations for strategic classification. In *International Conference on Machine Learning*, pp. 4687–4697. PMLR, 2021.

Meena Jagadeesan, Tijana Zrnic, and Celestine Mendler-Dünner. Regret minimization with performative feedback. In *International Conference on Machine Learning*, pp. 9760–9785. PMLR, 2022.

Michael P Kim and Juan C Perdomo. Making decisions under outcome performativity. In *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*, volume 251 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 79:1–79:15. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2023. doi: 10.4230/LIPIcs.ITCS.2023.79.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation (TEAC)*, 8(4):1–23, 2020.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

Tosca Lechner, Ruth Urner, and Shai Ben-David. Strategic classification with unknown user manipulations. In *International Conference on Machine Learning*, pp. 18714–18732. PMLR, 2023.

Sagi Levanon and Nir Rosenfeld. Strategic classification made practical. In *International Conference on Machine Learning*, pp. 6243–6253. PMLR, 2021.

Sagi Levanon and Nir Rosenfeld. Generalized strategic classification and the case of aligned incentives. In *International Conference on Machine Learning*, pp. 12593–12618. PMLR, 2022.

Licong Lin and Tijana Zrnic. Plug-in performative optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 30546–30565. PMLR, 2024. URL https://proceedings.mlr.press/v235/lin24ab.html.

Xinpeng Lv, Yunxin Mao, Renzhe Xu, Chunyuan Zheng, Yikai Chen, Haoxuan Li, Yang Shi, Jinxuan Yang, Zhouchen Lin, Yuanlong Chen, Yuanxing Zhang, Shaowu Yang, Wenjing Yang, and Haotian Wang. Beyond rational illusion: Behaviorally realistic strategic classification. *arXiv preprint arXiv:2605.19674*, 2026. URL https://arxiv.org/abs/2605.19674.

Charles F Manski. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4): 1221–1246, 2004.

Eric Mazumdar, Kishan Panaganti, and Laixi Shi. Tractable multi-agent reinforcement learning through behavioral economics. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=stUKwWBuBm.

Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. *Advances in Neural Information Processing Systems*, 33:4929–4939, 2020.

Celestine Mendler-Dünner, Frances Ding, and Yixin Wang. Anticipating performativity by predicting from predictions. *Advances in Neural Information Processing Systems*, 35:31171–31185, 2022. doi: 10.52202/068431-2260.

John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*, pp. 6917–6926. PMLR, 2020.

John P Miller, Juan C Perdomo, and Tijana Zrnic. Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning*, pp. 7710–7720. PMLR, 2021.

Mehrnaz Mofakhami, Ioannis Mitliagkas, and Gauthier Gidel. Performative prediction with neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 11079–11093. PMLR, 2023.

Evan Munro. Treatment allocation with strategic agents. *Management Science*, 71(1):123–145, 2024. doi: 10.1287/mnsc.2022.01629.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pp. 7599–7609. PMLR, 2020.

Chara Podimata. Incentive-aware machine learning; robustness, fairness, improvement & causality. *ACM SIGecom Exchanges*, 22(2):82–101, 2025. URL https://www.sigecom.org/exchanges/volume_22/2/PODIMATA.pdf.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350.

Roshni Sahoo and Stefan Wager. Policy learning with competing agents. *Operations Research*, 73 (5):2515–2529, 2025. doi: 10.1287/opre.2022.0687.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Han Shao, Avrim Blum, and Omar Montasser. Strategic classification under unknown personalized manipulation. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1151.

Yonadav Shavit, Benjamin Edelman, and Brian Axelrod. Causal strategic linear regression. In *International Conference on Machine Learning*, pp. 8676–8686. PMLR, 2020.

Herbert A Simon. A behavioral model of rational choice. *The quarterly journal of economics*, pp. 99–118, 1955.

Stratis Tsirtsis, Behzad Tabibian, Moein Khajehnejad, Adish Singla, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. Optimal decision making under strategic behavior. *Management Science*, 70 (12):8506–8519, 2024. doi: 10.1287/mnsc.2021.02567.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992. doi: 10.1007/BF00992696.

Wenjing Yan and Xuanyu Cao. Zero-regret performative prediction under inequality constraints. *Advances in Neural Information Processing Systems*, 36:1298–1308, 2023. doi: 10.52202/075280-0063.

Ding-Xuan Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of computational and Applied Mathematics*, 220(1-2):456–463, 2008.

A Experiment Details for Synthetic Data

A.1 Detailed Data Generation Mechanism

We generate potential outcomes that depend linearly on (u, v), assigning larger positive weights to larger u. Thus, a good policy incentivizes agents to increase u and improve outcomes. Specifically, we use the linear model

$$Y_i(z_i) = (z_i + 1)\langle x_i, \beta\rangle + \epsilon_i \tag{12}$$

where ϵ_i is i.i.d. standard Gaussian noise.

We draw v from a standard multivariate Gaussian distribution, and generate the discrete feature u according to

$$\mathbb{P}(u \mid v) = \mathrm{Softmax}(Wv) \tag{13}$$

The weight vector β and matrix W are sampled once from Gaussian distributions using a fixed random seed and remain unchanged throughout the simulation study.

A.2 Hyperparameter Setting

This subsection gives the synthetic-experiment hyperparameters for data generation and policy learning.

1. Dimension of features: $\dim \mathcal{V} = 20$, $|\mathcal{U}| = 5$.
2. Dimension of policy parameter θ: 1,351.
3. Batch size: $n = 3{,}000$.
4. Validation-set size: $n_{\mathrm{eval}} = 10{,}000$. The policy-evaluation samples used to generate the figures remain fixed throughout each run.
5. Time horizon: $T = 100$.
6. Warm-up duration: $T_0 = 30$.
7. Weight vector β in the CATE: for $u \in \{0, 1, 2, 3, 4\}$, the weights are linearly spaced over $[-5, 5]$; the weight for each coordinate of v is drawn from a standard Gaussian distribution.
8. Weight matrix W for generating $u \mid v$: each entry is drawn from $\mathcal{N}(1, 1)$, and the matrix is fixed across all experiments.
9. Policy network π_θ: 2-layer MLP with rectified linear unit (ReLU) activation and Sigmoid output. Hidden layer size: $2(\dim \mathcal{V} + |\mathcal{U}|) = 50$.
10. Behavior network h_γ: 3-layer MLP with ReLU activation. Hidden layer size: $2(\dim \mathcal{V} + |\mathcal{U}|) = 50$.
11. Optimizer of π_θ: Adam Kingma & Ba (2015) with learning rate 0.05 for warm-up stage, and Adagrad Duchi et al. (2011) with learning rate 0.05 after warm-up stage.
12. Optimizer of h_γ: stochastic gradient descent (SGD) with learning rate 0.1.

The dimension of θ exceeds 1,000, well beyond the typically sub-10-dimensional settings considered in prior work.

A.3 Baselines

In this subsection, we explain the three baselines in detail.

The *Cutoff Rule* is optimal without strategic behavior and is implemented using the **oracle** CATE in both the synthetic and semi-synthetic experiments.

The *Vanilla Policy Gradient* applies the idea of repeated gradient descent Perdomo et al. (2020), though gradient ascent is used here to maximize policy value. This baseline represents the method that is performativity-aware but does not model performativity explicitly.

The *End2end Policy Gradient* follows the same training pattern as Strategic Policy Gradient, including warm-up and early stopping for the behavior model. It differs only in how it models $p(u \mid v, \pi_\theta)$. The

behavior model takes (θ, v) as input (concat$[\theta; v]$) to predict post-manipulation u, rather than taking $(\zeta(v, \pi_\theta), v)$. This baseline models performativity through the distribution map $\mathcal{D}(\theta)$ and tests the benefit of treating the prediction function, rather than its parameters, as the intervention.

For all gradient-based methods, we estimate the CATE $\tau(x)$ using linear regression in the synthetic experiment and XGBoost Chen & Guestrin (2016) in the semi-synthetic experiment, then plug the estimate into the gradient.

We do not include zeroth-order optimization Flaxman et al. (2005); Miller et al. (2021). Our central setting has a high-dimensional parameter θ, as in modern neural networks. Zeroth-order methods use bandit feedback and are sample-inefficient because they compress each batch into a single scalar, especially for a potentially non-convex objective Miller et al. (2021); Jagadeesan et al. (2021).

In the semi-synthetic experiment, End2end uses a width-128 MLP on (θ, v) under the same warm-up and post-warm-up protocol. Its input expands from 1,351 to 25,741 policy parameters, while behavior-model fitting also grows with the response history. Thus its computational cost scales sharply in both data size and $\dim(\theta)$, whereas our mediator remains the 10-dimensional evaluation vector.

A.4 Experiment Results and Strategic Behavior

Because the manipulation-cost coefficient c changes the response pattern, Tables 1–3 report results for three values, $c \in \{0.05, 0.1, 0.15\}$.

Each table reports the best policy value over the trajectory and the value at the final epoch. We also report the proportion of agents whose u improves and the signed mean change in u, averaged over all agents. The unreported manipulation rate is nearly identical to the improvement rate for every method, indicating little harmful manipulation. Hence, the mean movement among agents who change can be recovered by dividing Move by %Change.

The manipulation induced by our method weakens as c increases, as reflected by lower improvement rates and mean movement. However, the mean movement among agents who do improve remains nearly unchanged, indicating that agents with improvement potential remain well incentivized.

We also observe that the strategic behavior induced by the other three baselines is insensitive to the cost coefficient here. Since the cutoff policy, CATE> 0 is the performative stable point of these two methods, we can explain this phenomenon through the mechanism of the cutoff policy. Under this policy, only agents whose CATE lies just below zero improve u; agents with positive CATE or CATE far below zero have no incentive to do so. We call the agents with CATE just below the threshold as potential improvers, and the size of this group will decrease only for a much larger cost coefficient. To illustrate it, we provide the following policy value curve for cutoff policy under different manipulation cost coefficients c. A good policy must not only induce improvement but also allocate treatment

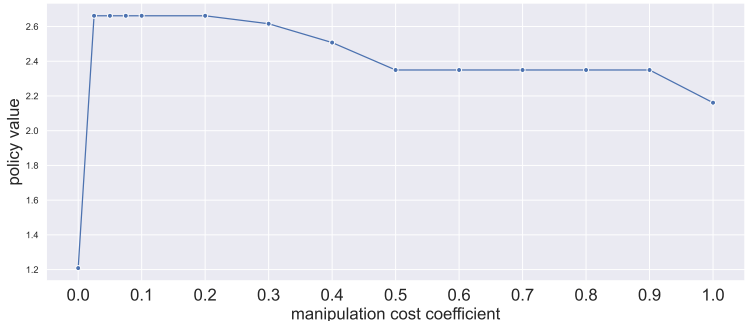


Figure 6: Policy value of the cutoff policy under different manipulation-cost coefficients c.

appropriately given the estimated CATE. For example, an applicant with a poor fixed credit profile should not receive high treatment propensity solely because improvement is possible. Assigning propensities $(0.7, 0.8)$ or $(0.1, 0.2)$ to two u levels can induce the same improvement but produce very different policy values. Thus, the behavioral indices only complement, rather than replace, policy value.

Table 1: Synthetic-data results with $c = 0.05$

METHOD	POLICY VALUE (BEST)	POLICY VALUE (FINAL EPOCH)	%CHANGE	MOVE
CUTOFF	2.92 ± 0.00	2.92 ± 0.00	37.50	0.74
VANILLA	2.89 ± 0.02	2.86 ± 0.02	30.24	0.59
END2END	2.62 ± 0.02	2.55 ± 0.07	20.73	0.41
STRATEGIC	5.22 ± 0.03	5.19 ± 0.04	76.54	1.92

Table 2: Synthetic-data results with $c = 0.1$

METHOD	POLICY VALUE (BEST)	POLICY VALUE (FINAL EPOCH)	%CHANGE	MOVE
CUTOFF	2.92 ± 0.00	2.92 ± 0.00	37.50	0.74
VANILLA	2.96 ± 0.02	2.89 ± 0.01	30.82	0.59
END2END	2.76 ± 0.06	2.76 ± 0.06	24.29	0.48
STRATEGIC	5.26 ± 0.03	5.23 ± 0.03	74.38	1.87

Table 3: Synthetic-data results with $c = 0.15$

METHOD	POLICY VALUE (BEST)	POLICY VALUE (FINAL EPOCH)	%CHANGE	MOVE
CUTOFF	2.92 ± 0.00	2.92 ± 0.00	37.50	0.74
VANILLA	2.99 ± 0.02	2.92 ± 0.01	31.12	0.59
END2END	2.86 ± 0.06	2.86 ± 0.06	26.20	0.51
STRATEGIC	5.25 ± 0.06	5.21 ± 0.07	70.88	1.80

A.5 BEHAVIOR-MODEL TRAINING AND GRADIENT DIAGNOSTICS

This subsection provides representative behavior-model training traces and aggregate gradient diagnostics. Figure 7 compares Strategic Policy Gradient with End2end Policy Gradient; the methods differ only in whether the response model conditions on the evaluation vector or the raw parameter vector θ.

The End2end behavior model trains poorly, consistent with the weak relationship between individual parameter coordinates and agent responses. The rightmost panel of Figure 3 reports the synthetic performative correction-gradient norm over 10 seeds with 90% confidence intervals. End2end remains near zero throughout, whereas SPG develops a pronounced signal after warm-up and then decays toward a small steady level as the policy converges. Figure 8 complements that panel with total-gradient norms in both settings and the semi-synthetic correction gradient. In the larger semi-synthetic experiment, SPG again exhibits a substantial early post-warm-up signal before the gradients settle near zero; End2end never develops a comparable correction signal.

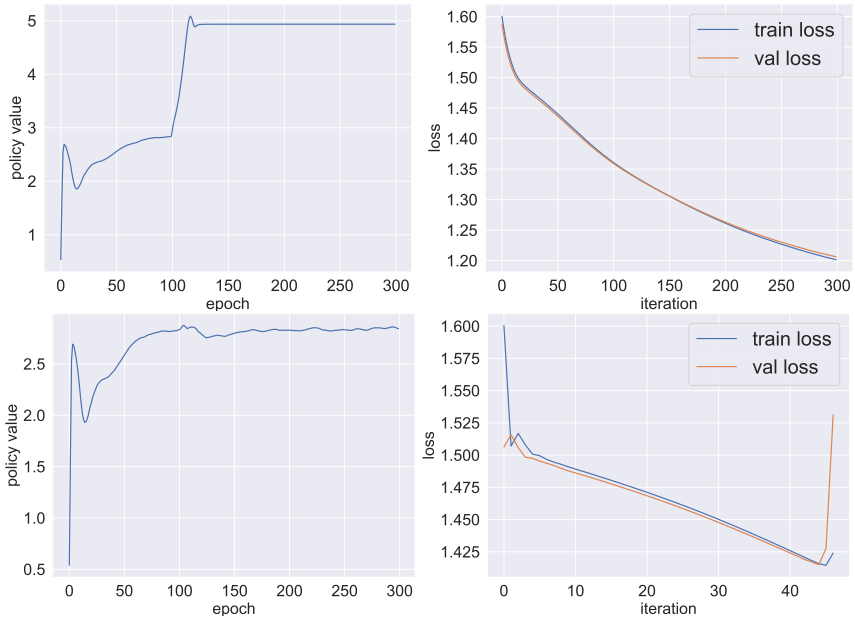
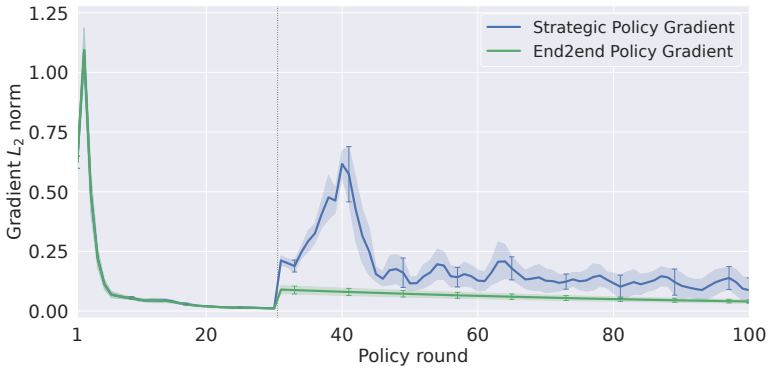
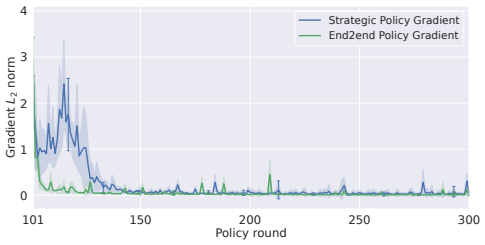


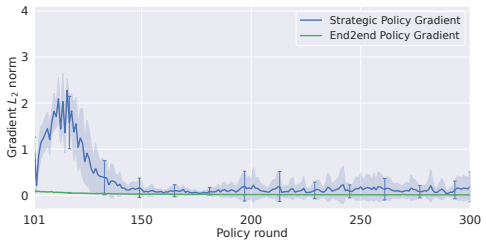
Figure 7: Training traces for *Strategic Policy Gradient* (top: policy value on the left and behavior-model loss on the right) and *End2end Policy Gradient* (bottom).



(a) Synthetic total-gradient norm.



(b) Semi-synthetic total-gradient norm.



(c) Semi-synthetic correction-gradient norm.

Figure 8: Additional gradient diagnostics over 10 seeds; shaded regions and error bars show 90% confidence intervals. Together with Figure 3, these panels separate the total update from the learned performative correction.

A.6 Decision processes with Gaussian and Gumbel noise

We evaluate two noisy mechanisms. First, we add Gaussian noise to perceived propensity, inspired by the parameter-noise setting in Jagadeesan et al. (2021):

$$u' = \text{argmax}_u \, \pi_\theta(u, v) + \epsilon_\pi - c|u - u_0|, \tag{14}$$

where $\epsilon_\pi \sim \mathcal{N}(0, c^2)$. Second, we consider the celebrated Multinomial Logit model. We add a multiplier of 5 to prevent u' from approaching a uniform distribution.

$$u' \sim \text{Softmax}\left(5(\pi_\theta(u, v) - c|u - u_0|)\right) \tag{15}$$

The first two panels of Figure 3 compare our method against the cutoff policy under noisy utility (14) and Softmax (15) mechanisms. *Strategic Policy Gradient* remains effective, demonstrating robustness to flexible micro-mechanisms without requiring a parametric decision-process model.

A.7 Gaussian Process (GP) Classifier as Behavior Model

While we primarily use MLPs to learn $p(u \mid v, \zeta(v, \pi_\theta))$, we also examine GP classifiers for theoretical consistency, since GP kernels induce RKHSs. The third panel of Figure 3 compares MLP and GP policy values under best-response manipulation ($c = 0.1$). Despite using different objectives—cross-entropy (CE) for the MLP and the negative evidence lower bound (nELBO) for the GP—the two behavior models produce nearly identical policy trajectories. Figure 9 shows that both training objectives decrease smoothly after warm-up.

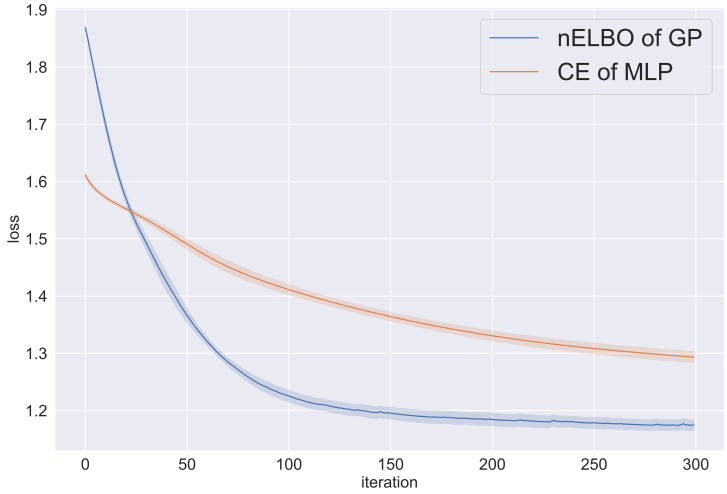


Figure 9: Behavior-model training objectives for the MLP and GP classifiers in the synthetic best-response setting. CE and nELBO both decrease smoothly, complementing the policy-value comparison in Figure 3.

B Experiment Details for Semi-Synthetic Data

B.1 Details of Semi-Synthetic Experiment

B.1.1 Data and Preprocessing

The experiment utilizes the Loan Defaulter dataset[1], containing 307,508 loan records. We select 56 features with low missing rates to form the fixed feature vector v. Although some features are binary flags, the resulting space $\mathcal{V}$ is significantly more complex than the Gaussian distributions used in the synthetic setting.

For preprocessing, we apply label encoding to categorical features and normalize numerical features via mean and standard deviation. Due to the severe class imbalance (normal cases far outnumber defaults), we apply the Synthetic Minority Oversampling Technique (SMOTE) Chawla et al. (2002) to ensure the fitted probability models are effective functions of the input v.

B.1.2 Outcome and Feature Generation

To generate organic potential outcomes and manipulable features, we train two Random Forest Breiman (2001) models on the binary default label: a "strong" model (ensemble of 200 trees, F1-score 0.99) and an "ordinary" model (ensemble of 40 trees with restricted depth, F1-score 0.90). Figure 10 illustrates their predicted probability distributions.

We treat the non-default probability predicted by the ordinary model as the initial manipulable feature, u_0, representing an external credit score (e.g., a liquidity assessment). Agents can manipulate u (e.g., by liquidating assets) to improve this score, thereby genuinely improving their repayment ability. Consequently, the true probability of default is modeled as a weighted combination:

$$g(u, v) = \lambda f_{\text{strong}}(v) + (1 - \lambda)u \tag{16}$$

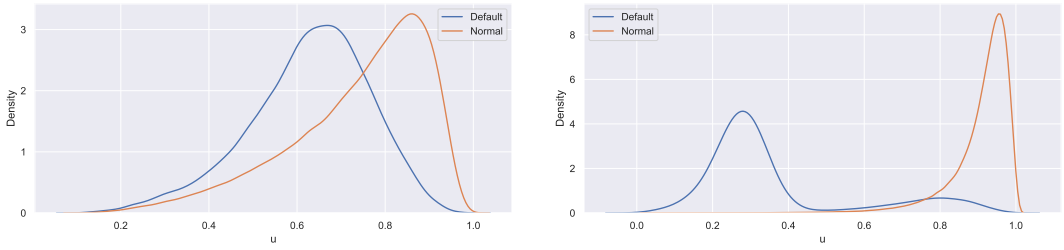
where $\lambda \in [0, 1]$ balances the fixed and manipulable components, and $u_0 = f_{\text{ordinary}}(v)$. We set $\lambda = 0.4$ to allow for strong strategic incentives. The domain is discretized into 10 subintervals, $\mathcal{U} = \{0, \ldots, 9\}$. Agents manipulate u via the best-response mechanism:

$$u' = \text{argmax}_u \, \pi_\theta(u, v) - c|u - u_0| \tag{17}$$

The principal's decision $z_i \in \{0, 1\}$ corresponds to loan rejection or approval. We set $Y_i(0) = 0$. The expected revenue for approval ($z_i = 1$) is defined as:

$$Y_i(1) = (g(u_i', v) - \beta(1 - g(u_i', v))) \times \text{amt}_i \tag{18}$$

where amt_i is the loan amount, introducing heterogeneity. The multiplier β penalizes defaults; we set $\beta = 4$, meaning the loss from a default is four times the gain from a repayment. This parameter choice yields a baseline policy value of $-30,570$ for "treat all" versus 0 for "control all."



(a) Ordinary Model (b) Strong Model

Figure 10: Distribution of predicted non-default probabilities generated by ordinary (left) and strong (right) random forest models.

[1] https://www.kaggle.com/datasets/gauravduttakiit/loan-defaulter

B.2 Features and Causal Parameters

1. Manipulable feature u: We discretize the predicted probability of no default generated by the ordinary random forest and let this discretized variable be the manipulable feature $u \in \{0, 1, \ldots, 9\}$:

$$u = \lfloor 10 \times \text{probability of no default} \rfloor$$

2. Fixed features v: In addition to the manipulable feature u, the remaining 55 variables in the dataset are treated as fixed features v.
3. Treatment z: the binary variable $z \in \{0, 1\}$ indicates whether the principal approves the loan request, with $z = 1$ for approval and $z = 0$ for denial.

B.3 Hyperparameter Setting

This subsection gives the semi-synthetic experiment hyperparameters:

1. Dimension of features: $\dim \mathcal{V} = 55$, $|\mathcal{U}| = 10$.
2. Dimension of policy parameter θ: 25,741.
3. Number of total rounds: $T = 300$.
4. Number of warm-up rounds: $T_0 = 100$.
5. Policy network π_θ: 3-layer MLP with ReLU activation and Sigmoid output. Hidden layer size: $2(\dim \mathcal{V} + |\mathcal{U}|) = 130$.
6. Behavior network h_γ: 3-layer MLP with ReLU activation and Softmax output. Hidden layer size: $2(\dim \mathcal{V} + |\mathcal{U}|) = 130$.
7. End2end behavior network: MLP on $\mathrm{concat}[\theta; v]$ with hidden width 128, trained by full-batch SGD with learning rate 0.01 and validation-based early stopping (patience 3, at most 300 epochs).
8. Optimizer of π_θ: Adam with learning rate 0.01 for warm-up stage, and Adagrad with learning rate 0.01 after warm-up stage.
9. Optimizer of h_γ: SGD with learning rate 0.01.

Because the semi-synthetic objective is substantially more nonlinear than the synthetic one, we use Adam during warm-up and Adagrad thereafter. Consistent with the main text, large first-moment momentum can destabilize both our method and the vanilla gradient; we therefore use lower momentum coefficients than the default $(0.9, 0.999)$.

The behavior models use random $80\%/20\%$ training/validation splits, with validation loss for early stopping. Policy evaluation is used only for reporting and never affects updates. The original baselines are evaluated on the full dataset at every epoch; because raw-θ End2end is substantially more expensive, it is evaluated at rounds $1, 10, 20, \ldots, 300$.

B.4 Experiment Results

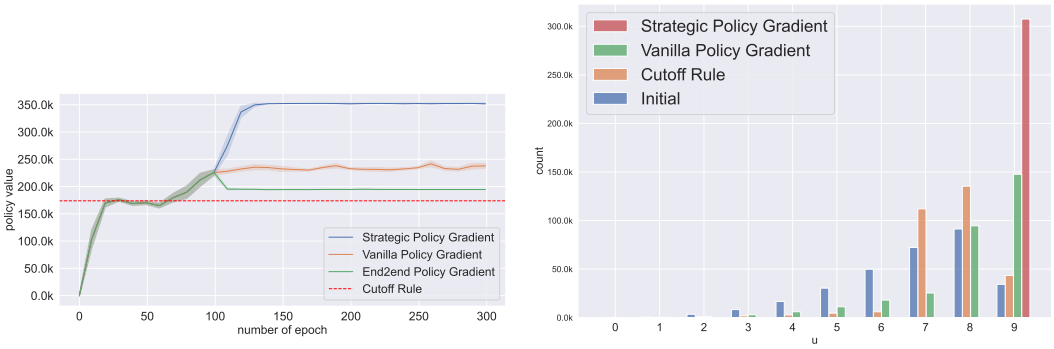
Figures 11–13 show the latest trajectories for all four methods at each cost, including End2end Policy Gradient. Tables 4–6 report the best and final policy values, the proportion of agents whose u improves, and the signed mean change in u. For End2end, "Best" is the maximum over 31 recorded evaluations (rounds $1, 10, 20, \ldots, 300$), while the other methods retain their original all-300-round best values. End2end means and sample standard deviations use 10 seeds; its final standard deviations round to 0.00 at two decimals but are nonzero. Table 4 complements the main-text Figure 5.

Across all three costs, End2end attains its best value during warm-up and then settles near 1.95×10^5. It remains below vanilla at the final epoch and substantially below Strategic Policy Gradient, consistent with the weak raw-θ correction signal and its poor scaling in both sample size and policy dimension.

In the semi-synthetic experiment, we consider $c \in \{0.1, 0.15, 0.2\}$. Because Strategic Policy Gradient already induces all agents to raise u to its highest level at lower costs, we do not repeat experiments for $c < 0.1$. Oscillation increases at $c = 0.2$ and total policy value declines, plausibly because

the behavior model h_γ becomes harder to learn when more agents leave u unchanged, making the gradient estimate $\nabla_\theta q(u \mid v, \pi_\theta)$ noisier. For still larger costs, such as $c = 0.3$, our method approaches the vanilla gradient, as noted in the main text.

Finally, we discuss CATE estimation. The CATE structure is much more complex than in the synthetic case. As a diagnostic, we estimate CATE from data collected under a constant allocation policy, e.g., $\pi(x) = 0.9$, and then apply the estimated-CATE cutoff rule. Its average policy value is about 1.29 (details omitted), as expected. This result illustrates the feedback between policy learning and the data it generates: the policy determines the data used for its next update. From a pure-exploration perspective, a good update process should pursue policy value while adaptively collecting the data needed for future updates; the constant-allocation policy is a counterexample. Although the gradient-based methods do not encode this principle explicitly, they may implement it implicitly, and their exploration dynamics merit further study.



(a) (b)

Figure 11: Semi-synthetic results for $c = 0.1$: policy value for all four methods with 90% confidence bands (left), and the distribution of u for cutoff, vanilla, and strategic policies (right).



(a) (b)

Figure 12: Semi-synthetic results for $c = 0.15$: policy value for all four methods with 90% confidence bands (left), and the distribution of u for cutoff, vanilla, and strategic policies (right).



(a) (b)

Figure 13: Semi-synthetic results for $c = 0.2$: policy value for all four methods with 90% confidence bands (left), and the distribution of u for cutoff, vanilla, and strategic policies (right).

† End2end best is the best observed value on its 31-point reporting grid; the other rows use all 300 rounds.

Table 4: Semi-synthetic results with $c = 0.1$

METHOD	$\times 10^5$ POLICY VALUE (BEST)	$\times 10^5$ POLICY VALUE (FINAL EPOCH)	%CHANGE	MOVE
CUTOFF	1.74 ± 0.00	1.74 ± 0.00	34.41	0.79
VANILLA	2.61 ± 0.05	2.35 ± 0.04	50.17	1.19
END2END†	2.27 ± 0.11	1.95 ± 0.00	59.14	1.30
STRATEGIC	3.52 ± 0.00	3.51 ± 0.02	88.83	2.18

Table 5: Semi-synthetic results with $c = 0.15$

METHOD	$\times 10^5$ POLICY VALUE (BEST)	$\times 10^5$ POLICY VALUE (FINAL EPOCH)	%CHANGE	MOVE
CUTOFF	1.74 ± 0.00	1.74 ± 0.00	34.07	0.76
VANILLA	2.50 ± 0.04	2.28 ± 0.08	51.64	1.20
END2END†	2.41 ± 0.06	1.95 ± 0.00	58.86	1.28
STRATEGIC	3.50 ± 0.03	3.48 ± 0.08	87.40	2.08

Table 6: Semi-synthetic results with $c = 0.2$

METHOD	$\times 10^5$ POLICY VALUE (BEST)	$\times 10^5$ POLICY VALUE (FINAL EPOCH)	%CHANGE	MOVE
CUTOFF	1.71 ± 0.00	1.71 ± 0.00	31.17	0.61
VANILLA	2.53 ± 0.08	2.26 ± 0.01	61.79	1.25
END2END†	2.26 ± 0.25	1.96 ± 0.00	55.01	1.07
STRATEGIC	3.13 ± 0.11	3.08 ± 0.09	66.97	1.26

C Practicality and Approximation Theory of Discretization

C.1 Why a Finite Manipulable Alphabet Is Operationally Meaningful

Assumption 3.1 can be read in two complementary ways. First, some manipulable variables are intrinsically discrete (credentials, product tiers, or report categories). Second, a continuous feature can be represented through a finite consideration alphabet chosen by the principal. The second interpretation is not merely computational: boundedly rational agents are more plausibly modeled as comparing a finite menu than as globally optimizing a non-convex neural policy over a continuum.

The finite alphabet also resolves two statistical bottlenecks. If $u \in [0,1]$ were continuous, the behavior model would receive the function $\pi_\theta(\cdot, v)$ as an input and output a conditional density over reports. Functional inputs are normally represented through evaluations on a grid, while continuous conditional-density derivatives are substantially harder to estimate than class probabilities. Discretization therefore converts both sides into a supervised multiclass problem: the input is an evaluation vector and the output is a probability mass function.

The key question is whether this finite representation approximates a continuous manipulation problem. We first give three structural answers: a value bound under stable responses, a best-response bound for deterministic agents, and an exact finite-candidate reduction for piecewise-constant policies. We then combine discretization, estimation, and optimization errors in a single oracle inequality.

C.2 Approximation Guarantees for Continuous Manipulable Features

Let the underlying manipulable domain be a compact set $\mathcal{A} \subset \mathbb{R}^{d_u}$. A quantizer $Q_h : \mathcal{A} \to \mathcal{U}_h$ has mesh radius h if

$$\sup_{u \in \mathcal{A}} \|u - Q_h(u)\| \le h.$$

For a continuous policy π, define its quantized counterpart by $\pi_h(u, v) = \pi(Q_h(u), v)$. Let $\mathcal{D}_\pi$ denote the post-manipulation distribution induced by deploying π, and let W_1 be the 1-Wasserstein distance on the post-manipulation covariates.

Proposition C.1 (Value approximation under response stability)**.** *Let $\Pi_h = \{\pi_h : \pi \in \Pi\}$. Suppose every $\pi \in \Pi$ is L_π-Lipschitz in u, $|\tau(x)| \le B_\tau$, and $g_\pi(x) = \pi(x)\tau(x)$ is L_g-Lipschitz uniformly over Π. Assume the response map is Hölder-stable on $\Pi \cup \Pi_h$: for some $\beta \in (0,1]$ and $L_D < \infty$,*

$$W_1(\mathcal{D}_\pi, \mathcal{D}_{\widetilde{\pi}}) \le L_D \|\pi - \widetilde{\pi}\|_\infty^\beta.$$

Then, for every $\pi \in \Pi$,

$$|V(\pi_h) - V(\pi)| \le B_\tau L_\pi h + L_g L_D (L_\pi h)^\beta =: \varepsilon_h. \tag{19}$$

Consequently,

$$\sup_{\pi \in \Pi} V(\pi) - \sup_{\rho \in \Pi_h} V(\rho) \le \varepsilon_h.$$

Thus a stable stochastic response ($\beta = 1$) yields an $O(h)$ policy-value loss. Deterministic best responses can be less stable because the argmax changes discontinuously near ties. The next result gives a verifiable sufficient condition and also handles the case where agents themselves optimize only over the grid.

Proposition C.2 (Deterministic best-response approximation)**.** *For agent type $s = (u_0, v)$, let $F_\pi(u; s)$ be the utility from reporting u, and let $r_\pi(s)$ be its unique maximizer over $\mathcal{A}$. Assume the uniform quadratic-growth condition*

$$F_\pi(r_\pi(s); s) - F_\pi(u; s) \ge \frac{\mu}{2} \|u - r_\pi(s)\|^2 \tag{20}$$

for some $\mu > 0$.

(i) *If*

$$\sup_{u,s} |F_\pi(u; s) - F_{\widetilde{\pi}}(u; s)| \le L_F \|\pi - \widetilde{\pi}\|_\infty,$$

then

$$\|r_\pi(s) - r_{\widetilde{\pi}}(s)\| \le 2\sqrt{\frac{L_F}{\mu}}\, \|\pi - \widetilde{\pi}\|_\infty^{1/2}. \tag{21}$$

Coupling agents with the same type s therefore gives

$$W_1(\mathcal{D}_\pi, \mathcal{D}_{\widetilde{\pi}}) \le 2\sqrt{L_F/\mu}\,\|\pi - \widetilde{\pi}\|_\infty^{1/2},$$

so deterministic responses satisfy Proposition C.1 with $\beta = 1/2$ whenever quadratic growth holds uniformly.

(ii) *Suppose $\mathcal{U}_h$ is an h-net of $\mathcal{A}$ and $F_\pi(\cdot; s)$ is L_u-Lipschitz. For a grid best response $r_{\pi,h}(s) \in \arg\max_{u\in\mathcal{U}_h} F_\pi(u; s)$,*

$$0 \le F_\pi(r_\pi(s); s) - F_\pi(r_{\pi,h}(s); s) \le L_u h, \tag{22}$$

$$\|r_{\pi,h}(s) - r_\pi(s)\| \le \sqrt{\frac{2L_u h}{\mu}}. \tag{23}$$

If $g_\pi(u, v)$ is L_g-Lipschitz in u, averaging over agent types gives a policy-value discrepancy of at most

$$L_g\sqrt{\frac{2L_u h}{\mu}}. \tag{24}$$

The quadratic-growth condition is substantive rather than cosmetic. Without a margin or response-stability condition, an $O(h)$ utility regret need not imply vanishing policy-value error: a nearly flat utility surface can move the selected action by $O(1)$ under an arbitrarily fine grid. This distinction explains why coarse-discretization experiments should evaluate downstream policy value, not only the agent's utility approximation.

C.3 Piecewise-Constant Policies Reduce Manipulation to Finite Comparisons

We next show that the principal can make a finite comparison set endogenous. For simplicity, let $u \in [0, 1]$, fix v, and write u_0 for the original feature. Let $U(u_0, u, v, \pi)$ denote the agent's utility. The standard strategic-classification example is

$$U(u_0, u, v, \pi_\theta) = \pi_\theta(u, v) - c\|u - u_0\|_2. \tag{25}$$

Assumption C.3 (Regular manipulation)**.** For an agent with original feature u_0, fixed feature v, and deployed policy π:

1. If $U(u_0, u_1, v, \pi) > U(u_0, u_2, v, \pi)$, then $u_1 \succ u_2$.
2. If $\pi(u_1, v) = \pi(u_2, v)$ and $|u_0 - u_1| < |u_0 - u_2|$, then $u_1 \succ u_2$.

Given knots $0 < a_1 < \cdots < a_k < 1$, define $I_0 = [0, a_1)$, $I_j = [a_j, a_{j+1})$ for $1 \le j < k$, and $I_k = [a_k, 1]$. A piecewise-constant policy has $\pi(u, v) = \pi_j(v)$ for $u \in I_j$.

Proposition C.4 (Finite-candidate reduction)**.** *Under Assumption C.3, any attained continuous best response distinct from u_0 is a knot or a boundary point; if a preferred plateau excludes its right endpoint, a maximizing sequence approaches the adjacent knot from the left.*

More concretely, let $G_\Delta = \{0, \Delta, 2\Delta, \ldots, 1\}$ with $1/\Delta \in \mathbb{N}$, let $J_\Delta = \{j : I_j \cap G_\Delta \ne \varnothing\}$, and define

$$P_j^\Delta(u_0) = \arg\min_{u\in I_j\cap G_\Delta} |u - u_0|, \qquad j \in J_\Delta.$$

Every best response on G_Δ belongs to $\bigcup_{j\in J_\Delta} P_j^\Delta(u_0)$. Each $P_j^\Delta(u_0)$ contains at most two points, so the agent need compare at most $2(k+1)$ candidates, or at most $k+1$ under deterministic distance-tie breaking. As $\Delta \downarrow 0$, candidates for cells not containing u_0 approach knots or boundaries, while the home-cell candidate converges to the no-change report u_0.

Proposition C.4 is stronger than merely asserting that a fine grid approximates a continuum: a plateau policy collapses the continuous optimization problem to one local projection per plateau. The evaluation vector therefore describes exactly the policy information needed for this finite comparison, even when the physical feature itself is continuous. The next result explains when refining the underlying action grid stops changing those comparisons.

Proposition C.5 (Resolution saturation under cell margins)**.** *Consider the piecewise-constant policy above and distance-based utility*

$$U_s(u;\pi) = b_j(s) - c(s)|u - u_0|, \qquad u \in I_j,$$

where $s = (u_0, v)$ and $0 \le c(s) \le \bar{c}$. Define the limiting score of cell j by

$$S_j(s) = b_j(s) - c(s)\, d(u_0, \overline{I_j}),$$

interpreting an excluded endpoint by its one-sided limit. Let $j^*(s)$ be the unique best cell and let

$$\Gamma(s) = S_{j^*(s)}(s) - \max_{j \neq j^*(s)} S_j(s)$$

be its cell margin. For all sufficiently fine G_Δ, if $\Gamma(s) > 2\bar{c}\Delta$, then the grid best response selects the same cell $j^*(s)$ as the limiting continuous comparison.

Suppose further that the downstream contribution $R_s(u) = \pi(u, v)\tau(u, v)$ satisfies $|R_s(u)| \le B_R$ and is L_R-Lipschitz within each cell. If $r(s)$ is the limiting distance-minimizing response in the best cell and $r_\Delta(s)$ is the grid best response, then

$$|\mathbb{E}R_s(r_\Delta(s)) - \mathbb{E}R_s(r(s))| \le L_R\Delta + 2B_R\,\mathbb{P}(\Gamma(s) \le 2\bar{c}\Delta)\,. \tag{26}$$

Consequently, the margin condition $\mathbb{P}(\Gamma(s) \le t) \le C_\Gamma t^\kappa$ yields an $O(\Delta^{\min\{1,\kappa\}})$ value error. Under a uniform positive margin, the selected cell is eventually exact; if R_s is also constant within each cell, the aggregate value is eventually unchanged by further grid refinement.

This proposition separates two effects of increasing the agents' true cardinality. Away from policy-cell ties, the selected cell stabilizes after a finite resolution. Further refinement only changes the within-cell report by at most one grid step. Hence a visible dependence on M should persist mainly when many agents lie close to indifference boundaries or when outcomes vary sharply within a principal cell.

C.4 Choosing the Resolution

Finer discretization reduces approximation bias but increases the number of behavior classes and the amount of data needed to estimate their probability derivatives. The next result separates the errors due to the principal's policy alphabet, the agents' action grid, behavior learning, and numerical optimization.

Proposition C.6 (Joint continuous-to-learned value guarantee)**.** *Let V denote value under continuous agent responses, let V_M denote value when agents optimize over an M-point action grid, and let $\widehat{V}_{N,M}$ be a learned criterion based on N response observations. For a principal class Π_m, suppose*

$$\sup_{\pi \in \Pi} V(\pi) - \sup_{\rho \in \Pi_m} V(\rho) \le \varepsilon_m, \tag{27}$$

$$\sup_{\rho \in \Pi_m} |V_M(\rho) - V(\rho)| \le \delta_M, \tag{28}$$

$$\sup_{\rho \in \Pi_m} |\widehat{V}_{N,M}(\rho) - V_M(\rho)| \le e_{N,m}. \tag{29}$$

If the returned policy $\widehat{\rho}_{m,M} \in \Pi_m$ satisfies

$$\widehat{V}_{N,M}(\widehat{\rho}_{m,M}) \ge \sup_{\rho \in \Pi_m} \widehat{V}_{N,M}(\rho) - \xi,$$

then

$$\sup_{\pi \in \Pi} V(\pi) - V_M(\widehat{\rho}_{m,M}) \le \varepsilon_m + \delta_M + 2e_{N,m} + \xi. \tag{30}$$

Thus principal discretization, agent-grid approximation, statistical estimation, and optimization contribute additively rather than being conflated in a single discretization term.

Corollary C.7 (Resolution choice under multiclass estimation)**.** *Suppose a quasi-uniform m-point principal discretization of a d_u-dimensional action space has mesh $h_m \le C_Q m^{-1/d_u}$ and that, with the notation of Proposition C.6,*

$$\varepsilon_m \le C_p m^{-\alpha_p/d_u}, \qquad \delta_M \le C_a M^{-\alpha_a/d_u}, \qquad e_{N,m} \le C_e\sqrt{m/N}.$$

Then

$$\sup_{\pi\in\Pi} V(\pi) - V_M(\widehat{\rho}_{m,M}) \le C_p m^{-\alpha_p/d_u} + C_a M^{-\alpha_a/d_u} + 2C_e\sqrt{m/N} + \xi. \tag{31}$$

Balancing the principal approximation and estimation terms gives

$$m_N \asymp N^{d_u/(2\alpha_p+d_u)}, \qquad \varepsilon_N \asymp N^{-\alpha_p/(2\alpha_p+d_u)}, \tag{32}$$

up to the separate agent-grid and optimization terms. Here $\alpha_p = 1$ corresponds to a Lipschitz-stable response, Proposition C.2 gives exponent $1/2$ under quadratic-growth deterministic responses, and Proposition C.5 gives $\alpha_a = \min\{1, \kappa\}$ under a cell-margin exponent κ.

The distinction between m and M is essential. The principal chooses $m = |\mathcal{U}|$, which controls both policy approximation and the dimension of the multiclass behavior model. The physical action grid M enters only through δ_M. For the piecewise-constant policies used in our coarse experiment, Proposition C.5 sharpens the generic power law: after the M-grid resolves the principal cells, only near-tie types and within-cell outcome variation remain. Hence increasing M from 15 to 50 can have negligible effect even while changing m remains consequential, as observed in Figure 4.

Proposition C.6 is an oracle decomposition, not an additional claim that Algorithm 1 globally optimizes a non-concave objective. Theorem 4.4 controls stationarity of the learned policy; under strong concavity or a gradient-dominance condition, that stationarity bound can be converted into the optimization term ξ. Without such structure, ξ remains explicit and records non-convex optimization error.

D Additional Discussion

D.1 Discussion on Assumption 3.2

Assumption 3.2 permits the response to depend directly on the fixed feature v; accordingly, the behavior model conditions on both v and the evaluation vector. The restriction is only that, conditional on (u_0, v), all dependence on the deployed policy is mediated by $\zeta(v, \pi_\theta)$. It can fail if agents react to policy values at other contexts $v' \neq v$, to architectural or parameterization details not reflected in predictions, or to aggregate behavior through cross-agent or equilibrium effects. In such cases, two deployments with the same (v, ζ) may induce different response distributions, so $p(u \mid v, \zeta)$ is not identified without a richer mediator. Designed policy perturbations or additional state variables may address these settings, but they lie beyond the scope of this paper.

D.2 Discussion on Technical Assumptions for Theorem 4.4

The RKHS bounds in Assumption 4.3 are satisfied by standard smooth kernels on the compact evaluation-vector domain; the constants R and G_K remain explicit in Theorem 4.4. The positivity condition is required by the score term $\nabla \log q$ and also provides overlap between the warm-up and post-deployment distributions. In implementation, the estimated probabilities can be stabilized by softmax smoothing or clipping, whereas the lower bound on the true response probabilities is a substantive assumption that can be supported by response noise or deliberate exploration.

The CATE condition is stated under the warm-up distribution. Positivity lets us transfer this error to a later deployment through the likelihood ratio $p_t(u \mid v)/p_0(u \mid v)$, as shown in Eq. (33). Finally, the convergence result assumes smoothness, not global concavity, and therefore guarantees an approximate critical point. Strong concavity or another gradient-dominance condition would additionally translate the gradient bound into distance or value suboptimality.

D.3 Discussion of $|\mathcal{U}|$

The relevant design variable is not the finest physically possible action grid, but the principal's effective alphabet. Increasing $|\mathcal{U}|$ reduces the approximation term in Proposition C.1, yet enlarges the multiclass output and the evaluation vector that the behavior model must differentiate. Multiclass generalization bounds depend on class complexity, such as the Natarajan dimension Shalev-Shwartz & Ben-David (2014); in our setting, rare bins also reduce the local sample size available for estimating response derivatives.

Corollary C.7 formalizes this bias–estimation tension. Its $\sqrt{|\mathcal{U}|/N}$ term is a stylized rate rather than a universal neural-network guarantee; replacing it by another valid complexity rate leaves the balancing argument unchanged. Proposition C.5 separately explains why the agents' underlying grid can stop mattering: once principal cells are resolved, only near-tie types switch cells. The two-resolution experiment in Figure 4 is consistent with both predictions.

This perspective also suggests adaptive discretization. The principal can begin with a coarse alphabet and split only cells that have sufficient occupancy and substantial within-cell policy or CATE variation. Such refinement preserves data support in low-information regions while allocating additional classes where the value bound is likely dominated by approximation bias.

E Proofs

E.1 Proof of Proposition 3.3

Proof. Because the deployed policy changes only the manipulable response, the marginal distribution of v is policy invariant. Assumption 3.2 further states that, conditional on v, the policy affects the reported u only through $\zeta(v, \pi_\theta)$. Hence

$$p(x; \pi_\theta) = p(u, v; \pi_\theta) = p(u \mid v, \zeta(v, \pi_\theta))p(v).$$

Taking logarithms, differentiating with respect to θ, and applying the chain rule gives

$$\nabla_\theta \log p(x; \pi_\theta) = \nabla_\zeta \log p(u \mid v, \zeta)\, \nabla_\theta \zeta(v, \pi_\theta),$$

because $p(v)$ does not depend on θ. □

E.2 Proof of Lemma 4.1

Proof. Under the stated kernel regularity, the derivative reproducing property (Zhou, 2008) gives, for each coordinate i,

$$\partial_{\zeta_i} f(\zeta) = \langle f, \partial_{\zeta_i} K(\zeta, \cdot)\rangle_{\mathcal{H}}\,.$$

Applying this identity to $f - g$ and using Cauchy–Schwarz coordinatewise,

$$\begin{aligned}\|\nabla_\zeta f(\zeta) - \nabla_\zeta g(\zeta)\|^2 &= \sum_{i=1}^{|\mathcal{U}|} \left|\langle f - g, \partial_{\zeta_i} K(\zeta, \cdot)\rangle_{\mathcal{H}}\right|^2 \\ &\le \|f - g\|_{\mathcal{H}}^2 \sum_{i=1}^{|\mathcal{U}|} \|\partial_{\zeta_i} K(\zeta, \cdot)\|_{\mathcal{H}}^2.\end{aligned}$$

Taking square roots proves the claim. □

E.3 Proof of Theorem 4.4

Proof. Let $m = |\mathcal{U}|$ and define

$$\bar{g}_t = \mathbb{E}[\hat{g}_t \mid \theta_t], \qquad E_t = \bar{g}_t - \nabla_\theta V(\pi_{\theta_t}).$$

We first bound the deterministic gradient-estimation error E_t on the behavior-model event in Theorem 4.4.

Because the marginal distribution of v is policy invariant and the warm-up conditional mass satisfies $p_0(u \mid v) \ge \iota$,

$$\mathbb{E}_{X\sim\mathcal{D}_t}|\hat{\tau}(X) - \tau(X)| = \mathbb{E}_{X\sim\mathcal{D}_0}\left[\frac{p_t(u \mid v)}{p_0(u \mid v)}|\hat{\tau}(X) - \tau(X)|\right] \le \frac{\epsilon_\tau}{\iota}. \tag{33}$$

For each (u, v), the reproducing property and Lemma 4.1 imply, uniformly in t and ζ,

$$|q_{t,u,v}(\zeta) - p_{u,v}(\zeta)| \le R\varepsilon_{\mathcal{H}}, \tag{34}$$

$$\|\nabla_\zeta q_{t,u,v}(\zeta) - \nabla_\zeta p_{u,v}(\zeta)\| \le G_K\sqrt{m}\,\varepsilon_{\mathcal{H}}. \tag{35}$$

Using the lower probability bound and $\|\nabla_\zeta p_{u,v}\| \le G_p$ gives

$$\|\nabla_\zeta \log q_{t,u,v} - \nabla_\zeta \log p_{u,v}\| \le \varepsilon_{\mathcal{H}} \left(\frac{RG_p}{\iota^2} + \frac{G_K\sqrt{m}}{\iota} \right). \tag{36}$$

Moreover, the evaluation-vector Jacobian has operator norm at most its Frobenius norm:

$$\|\nabla_\theta \zeta(v, \pi_\theta)\|_{\mathrm{op}} \le \|\nabla_\theta \zeta(v, \pi_\theta)\|_{\mathrm{F}} \le G_\pi \sqrt{m}. \tag{37}$$

Combining (36) and (37),

$$\|\nabla_\theta \log q_{t,u,v} - \nabla_\theta \log p_{u,v}\| \le \varepsilon_{\mathcal{H}} G_\pi \sqrt{m} \left(\frac{RG_p}{\iota^2} + \frac{G_K\sqrt{m}}{\iota} \right). \tag{38}$$

We also have

$$\|\nabla_\theta \log p_{u,v}\| \le \frac{G_p G_\pi \sqrt{m}}{\iota}. \tag{39}$$

Write the difference between the conditional mean estimator and the true gradient as

$$\begin{aligned} E_t = {} & \mathbb{E}_{\mathcal{D}_t}\big[\nabla_\theta \pi_\theta(X)(\hat{\tau}(X) - \tau(X))\big] \\ & + \mathbb{E}_{\mathcal{D}_t}\big[\pi_\theta(X)(\hat{\tau}(X) - \tau(X))\nabla_\theta \log p_{u,v}\big] \\ & + \mathbb{E}_{\mathcal{D}_t}\big[\pi_\theta(X)\hat{\tau}(X)(\nabla_\theta \log q_{t,u,v} - \nabla_\theta \log p_{u,v})\big]. \end{aligned}$$

Applying (33), (38), and (39), together with $0 \le \pi_\theta \le 1$, gives

$$\|E_t\| \le \frac{\epsilon_\tau G_\pi}{\iota} + \frac{\epsilon_\tau G_p G_\pi \sqrt{m}}{\iota^2} + \varepsilon_{\mathcal{H}} B_\tau G_\pi \sqrt{m} \left(\frac{RG_p}{\iota^2} + \frac{G_K\sqrt{m}}{\iota} \right) = G_E. \tag{40}$$

Let $\Delta_t = \hat{g}_t - \bar{g}_t$. On the joint sampling event in the theorem, $\|\Delta_t\| \le B_n$ for all t, so the total inexact-gradient error $e_t = E_t + \Delta_t$ satisfies $\|e_t\| \le G_E + B_n$. The update is $\theta_{t+1} = \theta_t + \eta(\nabla V_t + e_t)$, where $\nabla V_t = \nabla_\theta V(\pi_{\theta_t})$. By l-smoothness,

$$\begin{aligned} V_{t+1} &\ge V_t + \eta\langle \nabla V_t, \nabla V_t + e_t\rangle - \frac{l\eta^2}{2}\|\nabla V_t + e_t\|^2 \\ &= V_t + \frac{\eta}{2}\|\nabla V_t\|^2 + \frac{\eta}{2}(1 - l\eta)\|\nabla V_t + e_t\|^2 - \frac{\eta}{2}\|e_t\|^2 \\ &\ge V_t + \frac{\eta}{2}\|\nabla V_t\|^2 - \frac{\eta}{2}(G_E + B_n)^2, \end{aligned} \tag{41}$$

where the identity $2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a - b\|^2$ is used with $b = \nabla V_t + e_t$, and the last line uses $\eta \le 1/l$. Summing (41) over $t = 1, \ldots, T$ yields

$$\frac{\eta}{2}\sum_{t=1}^{T} \|\nabla V_t\|^2 \le V_{T+1} - V_1 + \frac{\eta T}{2}(G_E + B_n)^2.$$

Since $0 \le \pi_\theta \le 1$ and $|\tau| \le B_\tau$, $|V_t| \le B_\tau$, so $V_{T+1} - V_1 \le 2B_\tau$. Dividing by $\eta T/2$ and applying the minimum–average inequality gives (9). A union bound over the behavior-model and sampling events gives probability at least $1 - \delta_1 - \delta_2$. □

Remark on behavior-model rates. Theorem 4.4 keeps $\varepsilon_{\mathcal{H}}$ explicit. Membership of $p_{u,v}$ in an RKHS alone does not imply the previously claimed $O(n^{-1/2})$ RKHS-norm error for a regularized estimator: for $\lambda > 0$, one must also account for the approximation term between $p_{u,v}$ and the population regularized minimizer. Under an appropriate source condition and a probability-simplex-constrained RKHS estimator, standard regularized empirical-risk bounds can make both this approximation term and the sampling term vanish. The explicit formulation above separates that statistical estimation question from the inexact-gradient convergence argument.

E.4 Proof of Proposition C.1

Proof. The quantizer and Lipschitzness of π imply $\|\pi_h - \pi\|_\infty \le L_\pi h$. Add and subtract the value of the original reward function under the discretized response distribution:

$$\begin{aligned} V(\pi_h) - V(\pi) = \mathbb{E}_{\mathcal{D}_{\pi_h}}[(\pi_h(X) - \pi(X))\tau(X)] \\ + \mathbb{E}_{\mathcal{D}_{\pi_h}}[g_\pi(X)] - \mathbb{E}_{\mathcal{D}_\pi}[g_\pi(X)]. \end{aligned}$$

The first term has absolute value at most $B_\tau L_\pi h$. By the Kantorovich–Rubinstein inequality and response stability, the second is at most $L_g W_1(\mathcal{D}_{\pi_h}, \mathcal{D}_\pi) \le L_g L_D (L_\pi h)^\beta$. This proves (19).

For the optimal-value statement, take any $\pi \in \Pi$. Since its quantization π_h belongs to Π_h, $V(\pi) \le V(\pi_h) + \varepsilon_h \le \sup_{\rho \in \Pi_h} V(\rho) + \varepsilon_h$. Taking the supremum over $\pi \in \Pi$ proves the result. □

E.5 Proof of Proposition C.2

Proof. For part (i), write $r = r_\pi(s)$, $\widetilde{r} = r_{\widetilde{\pi}}(s)$, and $\delta = \|\pi - \widetilde{\pi}\|_\infty$. Optimality of $\widetilde{r}$ under $F_{\widetilde{\pi}}$ gives

$$\begin{aligned} F_\pi(r; s) - F_\pi(\widetilde{r}; s) &\le |F_\pi(r; s) - F_{\widetilde{\pi}}(r; s)| \\ &\quad + F_{\widetilde{\pi}}(r; s) - F_{\widetilde{\pi}}(\widetilde{r}; s) \\ &\quad + |F_{\widetilde{\pi}}(\widetilde{r}; s) - F_\pi(\widetilde{r}; s)| \\ &\le 2L_F\delta. \end{aligned}$$

Combining this with quadratic growth for F_π yields $(\mu/2)\|r - \widetilde{r}\|^2 \le 2L_F\delta$, which is (21).

For part (ii), choose $\bar{u}_h \in \mathcal{U}_h$ with $\|\bar{u}_h - r_\pi(s)\| \le h$. Since $r_{\pi,h}(s)$ maximizes utility on the grid,

$$F_\pi(r_{\pi,h}(s); s) \ge F_\pi(\bar{u}_h; s) \ge F_\pi(r_\pi(s); s) - L_u h.$$

This proves (22); combining it with (20) proves (23). Finally, Lipschitzness of g_π gives the pointwise reward difference $L_g\|r_{\pi,h}(s) - r_\pi(s)\|$, and averaging proves (24). □

E.6 Proof of Proposition C.4

Proof. Let $u^\star \ne u_0$ be an attained continuous optimum and suppose that it lies strictly inside a plateau I_j. If $u_0 \in I_j$, staying at u_0 yields the same policy value and a strictly smaller movement cost, contradicting optimality. If $u_0 < a_j$, the included left endpoint a_j has the same policy value and is closer to u_0. If $u_0 \ge a_{j+1}$, moving within I_j toward a_{j+1} preserves the policy value and strictly reduces distance. Thus a changed attained optimum cannot lie in a plateau interior. The remaining attained optima are knots or the boundaries 0 and 1; for an excluded right endpoint, a maximizing sequence may approach the adjacent knot from the left.

Now restrict reports to G_Δ. If a best response $u^\star \in I_j \cap G_\Delta$ does not belong to $P_j^\Delta(u_0)$, then there exists $w \in P_j^\Delta(u_0)$ with $|w - u_0| < |u^\star - u_0|$. Because w and $u^\star$ lie on the same policy plateau, Assumption C.3 implies $w \succ u^\star$, a contradiction. Hence every best response lies in the stated union over nonempty cells. In one dimension a point has at most two nearest grid points in an interval, proving the cardinality bound. For cells not containing u_0, these projections converge to adjacent knots or boundaries; the projection in the home cell converges to u_0. □

E.7 Proof of Proposition C.5

Proof. For sufficiently small Δ, every cell contains a grid point. Let

$$S_{j,\Delta}(s) = \max_{u \in I_j \cap G_\Delta} U_s(u; \pi).$$

The grid point nearest to the projection of u_0 onto $\overline{I_j}$ lies within Δ of that projection (using a one-sided point for an excluded endpoint). Therefore

$$|S_{j,\Delta}(s) - S_j(s)| \le c(s)\Delta \le \bar{c}\Delta \tag{42}$$

for every cell. If $j \ne j^*(s)$, then

$$\begin{aligned} S_{j^*,\Delta}(s) - S_{j,\Delta}(s) &\ge S_{j^*}(s) - S_j(s) - 2\bar{c}\Delta \\ &\ge \Gamma(s) - 2\bar{c}\Delta. \end{aligned}$$

Thus $\Gamma(s) > 2\bar{c}\Delta$ preserves the maximizing cell.

On this event, both responses are distance-minimizing representatives of the same cell and differ by at most Δ, so their downstream contributions differ by at most $L_R\Delta$. On the complementary near-tie event, boundedness gives the crude bound $2B_R$. Taking expectations proves (26). Substituting $\mathbb{P}(\Gamma(s) \le t) \le C_\Gamma t^\kappa$ gives

$$L_R\Delta + 2B_R C_\Gamma (2\bar{c}\Delta)^\kappa = O(\Delta^{\min\{1,\kappa\}}).$$

A uniform margin eliminates the near-tie event for all sufficiently small Δ; if R_s is cell-constant, the remaining within-cell term is zero. □

E.8 Proof of Proposition C.6

Proof. Let $V^* = \sup_{\pi\in\Pi} V(\pi)$. Add and subtract the best continuous-response and grid-response values over Π_m:

$$\begin{aligned} V^* - V_M(\widehat{\rho}_{m,M}) = V^* &- \sup_{\rho\in\Pi_m} V(\rho) \\ &+ \sup_{\rho\in\Pi_m} V(\rho) - \sup_{\rho\in\Pi_m} V_M(\rho) \\ &+ \sup_{\rho\in\Pi_m} V_M(\rho) - V_M(\widehat{\rho}_{m,M}). \end{aligned}$$

The first term is at most ε_m, and the second is at most δ_M. For the third, uniform criterion error and ξ-optimality give

$$\begin{aligned} \sup_{\rho\in\Pi_m} V_M(\rho) - V_M(\widehat{\rho}_{m,M}) &\le \sup_{\rho\in\Pi_m} \widehat{V}_{N,M}(\rho) - \widehat{V}_{N,M}(\widehat{\rho}_{m,M}) + 2e_{N,m} \\ &\le \xi + 2e_{N,m}. \end{aligned}$$

Summing the three bounds proves (30). □

E.9 Proof of Corollary C.7

Proof. Substituting the assumed rates into Proposition C.6 gives (31). Let $r = \alpha_p/d_u$. Up to constants, the m-dependent terms are $m^{-r} + m^{1/2}N^{-1/2}$. Balancing them gives $m^{r+1/2} \asymp N^{1/2}$, hence $m \asymp N^{1/(2r+1)} = N^{d_u/(2\alpha_p+d_u)}$. Substitution into either term yields $N^{-r/(2r+1)} = N^{-\alpha_p/(2\alpha_p+d_u)}$; the M-grid and optimization terms remain additive. □