
Linear Strategic Classification with Endogenous Improvements

Siddharth Shrivastava¹
Naresh Manwani²

¹IIT Hyderabad

Mahvith Akshintala¹
Sujit Gujar²

B Vamsha Vardhan Reddy²
Ganesh Ghalme¹

²IIT Hyderabad

Abstract

Classical models of strategic classification treat agents’ responses to a deployed classifier as cosmetic: features change, but true labels do not. We study an improvement-aware variant in which strategic responses can change outcomes: agents choose post-deployment features strategically at a cost, and labels are then drawn from a fixed conditional outcome law. We formalize this problem for linear classifiers under a single-index qualification model and linear-decomposable costs. Under these assumptions, agents’ best responses reduce to a closed-form shift of the decision boundary, and the resulting improvement-aware classification problem is learnable with linear sample complexity. As post-deployment labels are unavailable a priori, we begin by establishing a generalization bound with oracle access to the outcome law. One of our main contributions is a practical plug-in algorithm that replaces the oracle with an estimate, together with a bound on its improvement-aware error. We also show that the optimal classifier for standard strategic classification is a better proxy for the improvement-aware objective than the non-strategic Bayes-optimal classifier. Experiments on semi-synthetic and synthetic data support these findings and the plausibility of our assumptions.

of receiving a favorable decision. Classical strategic classification treats these responses as manipulation: the feature vector changes, but the underlying label or qualification remains fixed.

This immutability assumption is often too restrictive. In many applications, the same actions that change observable features can also improve the underlying outcome of interest. Coursework may improve academic readiness, professional training may increase job productivity, and financial discipline may improve repayment probability. Thus, a classifier can both induce strategic feature changes and incentivize genuine improvement. We call this setting *improvement-aware* strategic classification.

In this paper, we study improvement-aware strategic classification with *endogenous improvement* for binary label prediction tasks. In particular, we assume that the original population is governed by a fixed conditional outcome law $\eta(x) = \mathbb{P}(Y = 1 \mid X = x)$. The deployed classifier affects agents by changing their incentives over feasible feature modifications. Thus, the classifier induces a change in the marginal distribution of features, but not the structural relation between features and outcomes. In this sense, improvement is endogenous; it is not externally imposed by the analyst but arises because agents move to regions of the feature space that, under the original data-generating process, are associated with better outcomes.

Our framework makes two assumptions: 1) the class-probability function η follows a single-index qualification model, $\eta(x) = g(w^{\star\top}x)$, for a non-decreasing link $g : \mathbb{R} \rightarrow [0, 1]$ and non-negative weights $w^{\star} \in \mathbb{R}_{\geq 0}^d$, and 2) the manipulation cost is linear-decomposable, i.e., the total cost is a weighted sum of feature-wise improvement costs. Assumption 1 guarantees that non-strategic Bayes optimal classifier is linear, and subsumes the monotonicity of η in features. Monotonicity of the outcome in improvement-relevant features is a standard assumption in the improvement-aware strategic classification literature (Kleinberg and Raghavan, 2020; Miller et al., 2020). The single-index qualification model also encompasses many class-probability mod-

1 INTRODUCTION

Strategic classification (Hardt et al., 2016) studies prediction settings in which agents respond to a deployed classifier by modifying observable features at a cost. Such behavior arises in credit scoring, admissions (Hardt et al., 2016; Dong et al., 2018), and hiring (Kleinberg and Raghavan, 2020), applicants may adjust credit utilization, obtain certifications, or enroll in test-preparation programs to improve their chances

els commonly used in the literature, including linear and sigmoidal models (Foster et al., 2018; Audibert and Tsybakov, 2007; Yara and Terada, 2026). We will discuss these assumptions in detail in Section 2.

Paper Organization. Section 2 introduces the model and assumptions. Section 3 gives preliminary structural results. Section 4 studies learnability under oracle access to η , and Section 5 gives an oracle-free plug-in algorithm with generalization guarantees. Section 6 reports experiments on synthetic and semi-synthetic data, and Section 7 proposes a decision rule for setting where η is difficult to learn.

1.1 Related work

Hardt et al. (2016) formalized strategic classification as a Stackelberg game in which agents manipulate observable features at a cost. Follow-up work studies the social cost of manipulation (Milli et al., 2019; Liu et al., 2022; Hossain et al., 2025), information asymmetry (Tang et al., 2021; Ghalme et al., 2021; Shao et al., 2023), sequential settings (Cohen et al., 2023; Shekhtman and Dean, 2024) and differentiable surrogates for end-to-end training (Levanon and Rosenfeld, 2021; Eilat et al., 2022). Classical strategic classification treats manipulation as feature movement that leaves the label unchanged, an assumption we relax.

Another line lets agents genuinely improve rather than game, closely related to performative prediction (Perdomo et al., 2020; Mendler-Dünner et al., 2020). Work here designs classifiers that incentivize effort over gaming (Kleinberg and Raghavan, 2020; Ahmadi et al., 2022; Levanon and Rosenfeld, 2022), studies fairness under improvement (Efthymiou et al., 2025; Alhanouti and Naghizadeh, 2025), grounds the improvement/gaming distinction in causal feature structure (Miller et al., 2020; Horowitz and Rosenfeld, 2023; Góis et al., 2026), and analyzes dynamics, persistence, and welfare (Miller et al., 2020; Horowitz and Rosenfeld, 2023; Góis et al., 2026). Chen et al. (2023) and Xie et al. (2024) (1-D thresholds with active experimentation) study incentivizing improvement under deterministic qualification change.

Learning-theoretic work on strategic classification establishes PAC-learnability and strategic VC-dimension results (Sundaram et al., 2023; Lechner and Urner, 2022), incentive-aware sample complexity bounds (Zhang and Conitzer, 2021), online regret bounds (Ahmadi et al., 2023), and learnability gaps between strategic and standard PAC learning (Cohen et al., 2024). Closest to ours, Attias et al. (2025) prove PAC bounds under exogenously specified improvement sets. Sharma and Sun (2026) characterize proper learnability with improvements, give positive results for Euclidean-ball

improvement sets, and extend to bounded label noise and online learning. In both, improvement sets are fixed independently of the classifier, post-improvement labels follow a fixed target concept (up to label noise in Sharma and Sun (2026)), and ties are broken in the worst case.

Prior work thus treats qualification change as deterministic or classifier-independent (Chen et al., 2023; Xie et al., 2024; Attias et al., 2025; Sharma and Sun, 2026). It also separates improvement from manipulation through exogenous structure, such as causal graphs, improvable/non-improvable feature labels, or classifier-independent improvement sets, which demands domain knowledge often unavailable in practice. We instead model post-response qualification as η evaluated at the adapted features, that is the observational conditional $\mathbb{P}(Y = 1 \mid X)$ remains valid under change in agent features. Under this modeling choice, manipulation and probabilistic improvement coexist in a single model without a prespecified causal/non-causal feature partition, and the manipulation region depends on the deployed classifier. This yields a PAC learning guarantee under η , with sample complexity linear in the dimension d (Corollary 4.4), and generalization guarantee (Theorem 5.3), with no direct analogue in Attias et al. (2025); Sharma and Sun (2026).

2 SETTINGS

Let $\mathcal{X} \subseteq \mathbb{R}^d$ denote the feature space, and $\mathcal{D}$ be a distribution over $\mathcal{X}$. Also let $\mathcal{X}_0 := \text{supp}(\mathcal{D}) \subseteq \mathcal{X}$ denote the support of the original feature distribution. Agents may report feature vectors in a feasible report set $\mathcal{A} \subseteq \mathcal{X}$, with $\mathcal{X}_0 \subseteq \mathcal{A}$. For simplicity we take $\mathcal{A}$ to be common across agents; the model can also allow agent-dependent feasible sets $\mathcal{A}(x) \subseteq \mathcal{X}$. We define the class-probability function as $\eta(x) = \mathbb{P}(y = 1 \mid x)$, where the label for x is drawn as $y \mid x \sim \text{Bernoulli}(\eta(x))$. The training dataset D consists of unaltered points $(x_i, y_i)_{i=1}^n$, where the x_i ’s are independently sampled from $\mathcal{D}$ and y_i are independent realizations of $\eta(x_i)$. For brevity, we write y_x to denote the label y conditioned on x , throughout the paper. Further, we define $\text{err}(\cdot)$ as the standard expected 0–1 loss, $\text{err}(f) = \mathbb{E}_{x \sim \mathcal{D}}[\mathbb{1}(f(x) \neq y_x)]$.

Definition 2.1 (Agent Best Response). For a classifier f , an agent with original feature vector $x \in \mathcal{X}_0$ reports

$$x^f \in \arg \max_{z \in \mathcal{A}} \{\beta f(z) - c(x, z)\},$$

$\beta > 0$ is the benefit from positive classification and $c(x, x')$ is the cost of changing features from x to x' .

Strategic reporting induces a mismatch between the training distribution and the distribution observed at deployment. A strategically robust classifier accounts

for this best-response-induced distribution shift and adjusts its decision rule accordingly. The goal in standard strategic classification setting is to minimize the strategic error given below.

Definition 2.2 (Strategic Error (Hardt et al., 2016)). The strategic error of a classifier $f \in \mathcal{F}$ is defined as

$$\text{err}_s(f) = \mathbb{E}_{x \sim \mathcal{D}} [\mathbb{1}[f(x^f) \neq y_x]]. \quad (1)$$

This definition implicitly assumes labels are immutable. We adapt the above objective to reflect the resulting change in qualification/labels in improvement-aware settings as follows.

Definition 2.3 (Strategic Improvement Error). The strategic improvement error of a classifier $f \in \mathcal{F}$ is defined as

$$\text{err}_{\text{Imp}}(f) = \mathbb{E}_{x \sim \mathcal{D}} [\mathbb{1}[f(x^f) \neq y_{x^f}]]. \quad (2)$$

We consider endogenous improvement in qualification. This modeling choice captures strategic behavior as inducing true qualification changes. In particular, the probabilistic gain in qualification is a function of distance from classifier boundary, in contrast to models that either partition agents exogenously into gaming versus improving types (Horowitz and Rosenfeld, 2023; Chen et al., 2023) or treat improvement as deterministic (Attias et al., 2025). To characterize when and how feature improvements translate to qualification improvements, we assume the following condition on the model.

Assumption 1 (Single-index qualification model). There exists a continuous non-decreasing function $g : \mathbb{R} \rightarrow [0, 1]$ and a vector $w^* \in \mathbb{R}_{\geq 0}^d \setminus \{0\}$ such that $\eta(x) = g(w^{*\top} x)$ for all $x \in \mathcal{X}$. Moreover, the Bayes threshold $b^* := \inf\{t \in \mathbb{R} : g(t) \geq 1/2\}$ is finite.

The single-index qualification model ensures that the true data-generating process respects a natural alignment between feature improvements and qualification improvements, enabling us to design classifiers that leverage this structure. This class of models covers several standard probabilistic classification models, including logistic, probit, and other monotone-link generalized linear models; the sigmoid link is a common example (Foster et al., 2018; Yara and Terada, 2026). Furthermore, Assumption 1 is particularly justified in domains where features represent positive attributes, skill investments, or substantive credentials rather than mere signals. We validate this assumption on real-world datasets in Appendix H. Next we observe that Assumption 1 guarantees that the optimal classifier is linear.

Observation 2.4 (Existence of a linear Non-strategic Bayes optimal classifier). *Under 0-1 loss in the Single-index qualification model (Assumption 1), there exists*

a Non-strategic Bayes optimal classifier of the form $f^*(x) = \mathbb{1}\{w^{*\top} x \geq b^*\}$ for some threshold $b^* \in \mathbb{R}$. In particular, if $b^* := \inf\{t \in \mathbb{R} : g(t) \geq 1/2\}$, then the classifier above is Bayes-optimal.

This observation (proof in Appendix B) allows us to restrict the space of classifiers to the set of linear classifiers with non-negative weights and hence, throughout the paper, we will assume that

$$\mathcal{F} := \{x \mapsto \{0, 1\} : f_{w,b}(x) = \mathbb{1}(w^\top x \geq b)\},$$

where $w_i \geq 0$ for all $i \in [d]$, $w \neq 0$, and $b \in \mathbb{R}$, writing f and $f_{w,b}$ interchangeably.

Definition 2.5 (Linear-decomposable cost). A cost function $c : \mathcal{X}_0 \times \mathcal{A} \rightarrow \mathbb{R}_+$ is decomposable if there exist $\alpha_i > 0$ for $i \in [d]$ such that

$$c(x, z) = \sum_{i=1}^d \alpha_i (z_i - x_i)_+.$$

Here, $(z_i - x_i)_+ = \max(0, z_i - x_i)$.

The cost model is a weighted one-sided ℓ_1 action cost, where the parameters α_i denotes the per-unit cost or difficulty of improving a given feature i . It is closely related to the linear/separable manipulation costs used in strategic classification (Hardt et al., 2016; Ahmadi et al., 2021) and to coordinate-wise action costs used in algorithmic recourse (Ustun et al., 2019). This structure captures settings where each feature can be improved independently at some per-unit cost, *e.g.*, hours spent on test preparation, tuition for certification courses, or effort invested in building work experience.

Assumption 2. Cost function is linear-decomposable.

First, we prove certain structural results for the improvement-aware strategic classification.

3 PRELIMINARIES

All proofs for this section are deferred to Appendix C. Throughout, we assume feasibility: for every $x \in \mathcal{X}_0$ and $i \in [d]$, $x + te_i \in \mathcal{A}$ for all $t \in [0, \beta/\alpha_i]$. Our first result characterizes the agents' best response to a linear classifier and shows that, with non-negative weights, agents who respond do so by improving their feature values, *i.e.*, the best-response feature vector is coordinate-wise no smaller than the original one.

Lemma 3.1. *For $f_{w,b} \in \mathcal{F}$, let $S(w) := \beta \max_{j \in [d]} w_j / \alpha_j$ and let j^* be the smallest index in $\arg \max_{j \in [d]} w_j / \alpha_j$. Then for every $x \in \mathcal{X}_0$, the selected best response is*

$$x^f = \begin{cases} x + \frac{b - w^\top x}{w_{j^*}} e_{j^*}, & \text{if } b - S(w) \leq w^\top x < b \\ x, & \text{otherwise.} \end{cases}$$

Lemma 3.1 shows that, under $f_{w,b}$ exactly the agents with $w^\top x \in [b - S(w), b)$ move, and they land on the decision boundary; since $w \neq 0$, $w_{j^*} > 0$ and x^f is well defined. This is the linear-decomposable analogue of the Euclidean-cost best response used by Levanon and Rosenfeld (2022); here, agents move along the most cost-effective coordinate.

Remark 3.2. Lemma 3.1 implies $x^f \geq x$ coordinate-wise: agents never decrease a feature. Since η is non-decreasing in each coordinate (Assumption 1), $\eta(x^f) \geq \eta(x)$, i.e., best responses can only raise an agent's qualification, which is a central theme of the paper.

Theorem 3.3. *For $f_{w,b} \in \mathcal{F}$, let $f_s(x) = \mathbf{1}\{w^\top x \geq b_s\}$ with $b_s := b + \beta \max_{j \in [d]} w_j / \alpha_j$. Then $f_s(x^{f_s}) = f_{w,b}(x)$ for every $x \in \mathcal{X}_0$. Consequently, $\mathbf{err}_s(f_s) = \mathbf{err}(f_{w,b})$.*

Each unit of cost raises the score $w^\top x$ by at most $S(w)/\beta$, attained by moving along e_{j^*} . Since positive classification is worth β , an agent can profitably raise its score by at most $S(w)$. Under f_s (the strategic counterpart of $f_{w,b}$), agents with $w^\top x \geq b$ can therefore reach $b_s = b + S(w)$, while agents with $w^\top x < b$ cannot, so $f_s(x^{f_s}) = f_{w,b}(x)$.

Corollary 3.4. *Let $f_{w^*,b^*} \in \arg \min_{f \in \mathcal{F}} \mathbf{err}(f)$, i.e., it is a non-strategic Bayes-optimal classifier (Observation 2.4). Then, its shift $f_s^* := f_{w^*,b_s^*}$ is strategic optimal among linear classifiers in $\mathcal{F}$.*

In one dimension, with η strictly increasing, let $\theta^*, \theta_s, \theta_{\text{Imp}}$ be the thresholds minimizing $\mathbf{err}, \mathbf{err}_s, \mathbf{err}_{\text{Imp}}$, respectively. Then $\theta^* \leq \theta_{\text{Imp}} \leq \theta_s$. (Lemma C.1)

We proved important structural properties of an optimal improvement-aware classifier. However, learning an optimal improvement-aware classifier requires post-deployment labels, which depend on agents' responses x^f (hence on f itself) and are unobserved before deployment. In Section 4, we assume oracle access to these labels, as in a controlled deployment; learning remains nontrivial since $\mathcal{D}$ is unknown and both x^f and its label depend on the classifier f , and we establish PAC Learnability. In Section 5, we replace the oracle with a plug-in estimate $\hat{\eta}$ and derive generalization bounds depending on its estimation error.

4 PAC LEARNABILITY WITH ORACLE ACCESS

We establish PAC learnability of improvement-aware strategic classification under oracle access by reducing it to uniform convergence over an induced loss class. Throughout, we fix the deterministic tie-breaking rule of Section 3, so that the best response x^f is uniquely determined for every $x \in \mathcal{X}_0$ and $f \in \mathcal{F}$. In this section, we assume oracle access to the true conditional law $\eta(\cdot)$,

which allows the learner to construct, for any $f \in \mathcal{F}$, the improvement-aware labels of Definition 4.1.

Definition 4.1 (Oracle labels). Let $(x_i, y_i)_{i=1}^n$ be i.i.d., with $x_i \sim \mathcal{D}$ supported on $\mathcal{X}_0$ and $y_i \mid x_i \sim \text{Bernoulli}(\eta(x_i))$, and let $u_1, \dots, u_n \sim \text{Unif}[0, 1]$ be i.i.d., independent of $(x_i, y_i)_{i=1}^n$, drawn once and shared across all $f \in \mathcal{F}$. For $f \in \mathcal{F}$, define

$$y_i^f := \mathbf{1}\{y_i = 1 \vee u_i \leq p^f(x_i)\},$$

where $p^f(x) := \frac{\eta(x^f) - \eta(x)}{1 - \eta(x)} \in [0, 1]$ is the true improvement probability (Appendix I).

Since u_i is independent of (x_i, y_i) , $\mathbb{P}(y_i^f = 1 \mid x_i) = \eta(x_i) + (1 - \eta(x_i))p^f(x_i) = \eta(x_i^f)$; that is for every fixed f , the posterior distribution of y_i^f given x_i is $\text{Bernoulli}(\eta(x_i^f))$. Hence the empirical improvement-aware error

$$\widehat{\mathbf{err}}_{\text{Imp}}(f) := \frac{1}{n} \sum_{i=1}^n \mathbf{1}[f(x_i^f) \neq y_i^f]$$

satisfies $\mathbb{E}[\widehat{\mathbf{err}}_{\text{Imp}}(f)] = \mathbf{err}_{\text{Imp}}(f)$ for any fixed $f \in \mathcal{F}$.

Theorem 4.2. *Let the sample and oracle labels be as in Definition 4.1, suppose the induced oracle-loss class $\mathcal{H} := \{h_f : f \in \mathcal{F}\}$, where*

$$h_f(x, y, u) := \mathbf{1}[f(x^f) \neq \mathbf{1}\{y = 1 \vee u \leq p^f(x)\}]$$

has finite VC dimension $\tilde{d} \geq 1$. Then for every $\delta \in (0, 1)$ and $n \geq \tilde{d}$, with probability at least $1 - \delta$ over sample $(x_i, y_i, u_i)_{i=1}^n$,

$$\sup_{f \in \mathcal{F}} |\mathbf{err}_{\text{Imp}}(f) - \widehat{\mathbf{err}}_{\text{Imp}}(f)| \leq \varepsilon_n(\delta),$$

$$\text{where } \varepsilon_n(\delta) = \frac{2}{\sqrt{n}} + 4\sqrt{\frac{2\tilde{d}\log(en/\tilde{d})}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

The main difficulty is that the post-response label is random and its law depends on f . Independently re-sampling labels across classifiers introduces classifier-specific randomness, which can prevent uniform convergence over infinite classes. We therefore couple the labels across classifiers through a single auxiliary variable u_i per point (Definition 4.1), so that the loss of every f is a fixed function h_f of the same augmented point $z_i = (x_i, y_i, u_i)$, and uniform convergence can be controlled through the VC dimension of $\mathcal{H}$. The coupling preserves each classifier's conditional post-response label distribution, $y_i^f \mid x_i \sim \text{Bernoulli}(\eta(x_i^f))$, so $\mathbb{E}[h_f(z)] = \mathbf{err}_{\text{Imp}}(f)$, while $\widehat{\mathbf{err}}_{\text{Imp}}(f) = \frac{1}{n} \sum_{i=1}^n h_f(z_i)$ by definition. The problem thus reduces to uniform convergence for the binary class $\mathcal{H}$ over i.i.d. sample: symmetrization together with Massart's lemma and the Saucer's lemma bounds the expected deviation by the first two terms

of $\varepsilon_n(\delta)$, and McDiarmid's inequality yields the stated high-probability term. The full proof is in Appendix D. The following lemma yields $\tilde{d} = \mathcal{O}(d)$ in our setting.

Lemma 4.3. *Let $\mathcal{X} \subseteq \mathbb{R}^d$ and let $\mathcal{F}$ be the class of linear classifiers of Section 2. Under Assumptions 1 and 2, the induced oracle-loss class $\mathcal{H}$ of Theorem 4.2 has VC dimension $\tilde{d} < 11(d+1) = \mathcal{O}(d)$.*

Write $\theta = (w, b)$. Under Assumption 2 and the tie-breaking rule, an agent who moves does so along the single coordinate $j^* = \min \arg \max_j w_j / \alpha_j$, so the parameter space splits into d regions $\Theta_j = \{\theta : j^* = j\}$ that do not depend on the input. On each Θ_j , Section 3 gives the best response in closed form: whether the agent moves, and the post-response decision $f(x^f)$, are determined by the signs of two affine functions of θ . For the oracle label, monotonicity of the link g (Assumption 1) rewrites the event $\{u \leq p^f(x)\}$ as the threshold condition $w^{*\top} x^f \geq \kappa(x, \tau)$, where κ depends only on the input through the generalized inverse of g . Substituting the closed form x^f turn this into the sign of a third affine function of θ , while the degenerate cases (no move, $\eta(x) = 1$, or infinite κ) reduce to events that do not depend on θ . Hence, on each region, $h_f(x, y, u)$ is a fixed Boolean function of the signs of three affine functions of $\theta \in \mathbb{R}^{d+1}$ with coefficients depending only on (x, u) , and a sign-pattern counting bound (Lemma D.3) gives $\tilde{d} < 11(d+1)$. Notably, the bound does not depend on the form of g ; only its monotonicity and continuity are used. The full proof is in Appendix D.

Corollary 4.4 (Agnostic PAC learnability). *Under the conditions of Theorem 4.2 and Lemma 4.3, let $\hat{f} \in \arg \min_{f \in \mathcal{F}} \widehat{\text{err}}_{\text{Imp}}(f)$. Then for every $\delta \in (0, 1)$ and $n \geq 11(d+1)$, with probability at least $1 - \delta$,*

$$\text{err}_{\text{Imp}}(\hat{f}) \leq \inf_{f \in \mathcal{F}} \text{err}_{\text{Imp}}(f) + 2\varepsilon_n(\delta),$$

$$\text{where } 2\varepsilon_n(\delta) = \mathcal{O}\left(\sqrt{\frac{d \log(n/d)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right).$$

Remark 4.5. Oracle access to η specifies how labels respond to strategic movement but not the population distribution $\mathcal{D}$; since $\text{err}_{\text{Imp}}(f)$ couples f with the induced response x^f , its minimizer over $\mathcal{F}$ depends upon $\mathcal{D}$ and must be learned from samples. In Section 5, when η is replaced by an estimator $\hat{\eta}$, the resulting plug-in analysis incurs an additional approximation term of order $\sup_{z \in \mathcal{A}} |\hat{\eta}(z) - \eta(z)|$.

5 ORACLE FREE LEARNING WITH ESTIMATED η

We now propose an improvement-aware strategic learning algorithm STRAT-IMP-AWARE that targets the

Algorithm 1 STRAT-IMP-AWARE

```

1: Input: Training Set  $\mathcal{D}_{tr}$ , Validation Set  $\mathcal{D}_{val}$ , Epochs  $T$ , Learning rate  $\gamma$ , Utility  $\beta$ , Cost-coeff.  $\vec{\alpha}$ , Differentiable Estimate  $\hat{\eta}$ , Guard  $\delta_g$ 
2: Initialize  $w \in \mathbb{R}_{\geq 0}^d \setminus \{0\}$ ,  $b \in \mathbb{R}$ ;  $\mathcal{L}_{best} \leftarrow \infty$ 
3: for  $epoch = 1 \dots T$  do
4:   for batch  $(X, Y) \in \mathcal{D}_{tr}$  do
5:      $\hat{\pi} \leftarrow \text{SIM-IMP-PROB}(X, w, b, \beta, \vec{\alpha}, \hat{\eta}, \delta_g)$ 
6:      $\mathcal{L}_{batch} \leftarrow \text{STRAT-IMP-LOSS}(X, Y, \hat{\pi}, w, b, \beta, \vec{\alpha})$ 
7:      $w \leftarrow (w - \gamma \nabla_w \mathcal{L}_{batch})_+$ 
8:      $b \leftarrow b - \gamma \nabla_b \mathcal{L}_{batch}$ 
9:   end for
10:   $\mathcal{L}_{val} \leftarrow 0$ ;  $m \leftarrow 0$ 
11:  for batch  $(X, Y) \in \mathcal{D}_{val}$  do
12:     $\hat{\pi} \leftarrow \text{SIM-IMP-PROB}(X, w, b, \beta, \vec{\alpha}, \hat{\eta}, \delta_g)$ 
13:     $\mathcal{L}_{val} \leftarrow \mathcal{L}_{val} + \text{STRAT-IMP-LOSS}(X, Y, \hat{\pi}, w, b, \beta, \vec{\alpha})$ 
14:     $m \leftarrow m + 1$ 
15:  end for
16:   $\mathcal{L}_{val} \leftarrow \mathcal{L}_{val} / m$ 
17:  if  $\mathcal{L}_{val} < \mathcal{L}_{best}$  then
18:     $\mathcal{L}_{best} \leftarrow \mathcal{L}_{val}$ ,  $w_{best} \leftarrow w$ ,  $b_{best} \leftarrow b$ 
19:  end if
20: end for
21: return Best Model  $(w_{best}, b_{best})$ 
```

Algorithm 2 SIM-IMP-PROB

```

1: Input: Feature  $x$ , Model  $(w, b)$ , Utility  $\beta$ , Cost-coeff.  $\vec{\alpha}$ , Differentiable Estimate  $\hat{\eta}$ , Guard  $\delta_g$ 
2:  $j^* \leftarrow \min \arg \max_{j \in [d]} w_j / \alpha_j$ 
3:  $S(w) \leftarrow \beta w_{j^*} / \alpha_{j^*}$ 
4:  $x^f \leftarrow x$ 
5: if  $0 < b - w^\top x \leq S(w)$  then
6:    $x^f \leftarrow x + \frac{b - w^\top x}{w_{j^*}} e_{j^*}$ 
7: end if
8: return  $\hat{\pi}_{\text{Imp}} \leftarrow \text{clamp}_{[0,1]} \left( \frac{\hat{\eta}(x^f) - \hat{\eta}(x)}{\max\{1 - \hat{\eta}(x), \delta_g\}} \right)$ 
```

improvement-aware classification error under endogenous feature manipulation and label improvement, without the need of oracle access. The algorithm extends Levanon and Rosenfeld (2022) in three ways: (i) decomposable costs replace L_2 costs, yielding a closed-form strategic shift $S(w) = \beta \cdot \max_{j \in [d]} w_j / \alpha_j$; (ii) label improvement is modeled through each agent's improvement probability under the deployed model, computed from a plug-in estimate $\hat{\eta}$; and (iii) training differentiates through both the strategic shift and these estimated improvement probabilities.

Algorithm 3 STRAT-IMP-LOSS

```

1: Input: Point  $(x, y)$ , Imp-probability  $\hat{\pi}_{\text{Imp}}$ , Model  $(w, b)$ , Utility  $\beta$ , Cost-coeff.  $\vec{\alpha}$ 
2:  $q \leftarrow y + (1 - y) \hat{\pi}_{\text{Imp}}$ 
3:  $S(w) \leftarrow \beta \max_{j \in [d]} \left( \frac{w_j}{\alpha_j} \right)$ 
4:  $\text{score} \leftarrow w^\top x - b + S(w)$ 
5: return  $\ell \leftarrow q(1 - \text{score})_+ + (1 - q)(1 + \text{score})_+$ 
```

Improvement probability. Given a sample $\{x_i, y_i\}_{i=1}^n$ from the pre-deployment distribution and the current model (w, b) , SIM-IMP-PROB (Algorithm 2) first computes the best response x^f , which under non-negative weights and decomposable costs has a closed form (Section 3): if $0 < b - w^\top x \leq S(w)$, the agent moves along the most cost-effective coordinate j^* exactly onto the decision boundary; otherwise (already positive, or too far to move profitably) $x^f = x$. Since η is unknown, SIM-IMP-PROB uses a plug-in estimate $\hat{\eta}$ to compute and return the improvement probability $\hat{\pi}_{\text{Imp}} = \text{clamp}_{[0,1]}((\hat{\eta}(x^f) - \hat{\eta}(x)) / \max\{1 - \hat{\eta}(x), \delta_g\})$. STRAT-IMP-LOSS (Algorithm 3) then forms $q = y + (1 - y)\hat{\pi}_{\text{Imp}}$, the probability that post-response label is positive: a point with $y = 1$ remains positive, and a point with $y = 0$ becomes positive with probability $\hat{\pi}_{\text{Imp}}$. Equivalently, $q = \mathbb{P}(y_{\text{Imp}} = 1 \mid x, y)$ for the improvement-aware label $y_{\text{Imp}} = \mathbf{1}(y = 1 \vee u \leq \hat{\pi}_{\text{Imp}})$ with $u \sim \text{Unif}[0, 1]$. With η in place of $\hat{\eta}$, $\mathbb{E}[q \mid x] = \eta(x^f)$ under Assumption 1 whenever $1 - \eta(x) \geq \delta_g$, so that $y_{\text{Imp}} \mid x \sim \text{Bernoulli}(\eta(x^f))$ (Appendix I).

Loss. STRAT-IMP-LOSS applies a hinge surrogate to the shifted score $M(x) = w^\top x - b + S(w)$; since $f(x^f) = \mathbf{1}(w^\top x \geq b - S(w))$, the shift accounts for strategic responses without computing x^f inside the loss. Rather than sampling y_{Imp} , we take the expectation over u in closed form:

$$\mathbb{E}_u[(1 - \tilde{y}M(x))_+] = q(1 - M(x))_+ + (1 - q)(1 + M(x))_+ \quad (3)$$

where $\tilde{y} = 2y_{\text{Imp}} - 1$. This is the conditional expectation over u of the sampled-label empirical risk analyzed in Section E: it has the same population risk, deterministic function of (w, b) and removes the sampling noise in u .

Training. At each update, q is recomputed from (w, b) , and gradients propagate through the differentiable estimator $\hat{\eta}$. For a manipulating point,

$$x^f = x + \frac{b - w^\top x}{w_{j^*}} e_{j^*}, \quad \frac{\partial x_{j^*}^f}{\partial b} = \frac{1}{w_{j^*}}, \quad \nabla_w x_{j^*}^f = -\frac{x^f}{w_{j^*}}.$$

The coordinate j^* and manipulation status are held fixed during differentiation, yielding exact gradients wherever the corresponding branch is differentiable. On the movement branch, the bound $|\partial \ell / \partial q| \leq 2\beta w_{j^*} / \alpha_{j^*}$ cancels the inverse factor in the response derivatives; elsewhere, $q = y$ locally. The loss is continuous across manipulation boundaries but may jump across the tie hyperplanes $\{w_j / \alpha_j = w_k / \alpha_k\}$, where the selected response coordinate changes. Algorithm 1 uses projected (sub)gradient steps on $\mathbb{R}_{\geq 0}^d \times \mathbb{R}$, as required by Theorem 3.3, and returns the model with the lowest validation loss across epochs, which also guards against

degradation from crossing the tie hyperplanes. Algorithms 2 and 3 are applied row-wise, with per-point losses averaged within each batch and validation loss averaged across batches. Validation improvement probabilities are recomputed under the current model, and $(\cdot)_+$ is applied entrywise. More details are given in Appendix H.

Lemma 5.1 (Gradient bound, Informal). *Under mild boundedness conditions, the gradients of the loss of Algorithm 3 are bounded by a constant that depends neither on the classifier’s parameters nor on the training data, so they do not blow up as the weights shrink.*

The formal statement, its assumptions, and the proof are given in Appendix E. Lemma 5.1 keeps the iterates bounded: since per-point (sub)gradients are bounded by constants G_w, G_b that do not depend on the training sample, and the projection onto $\mathbb{R}_{\geq 0}^d$ is non-expansive, after N updates every iterate satisfies $\|w\|_2 \leq \|w^{(0)}\|_2 + \gamma N G_w$ and $|b| \leq |b^{(0)}| + \gamma N G_b$. Hence the returned model lies in $\mathcal{W}_{k,B} := \{(w, b) : w \in \mathbb{R}_{\geq 0}^d, \|w\|_2 \leq k, |b| \leq B\}$ with $k := \|w^{(0)}\|_2 + \gamma N G_w$ and $B := |b^{(0)}| + \gamma N G_b$, both fixed before training; our generalization analysis is uniform over this class.

We next derive a generalization bound for STRAT-IMP-AWARE when the outcome law η is unknown. Whereas Theorem 4.2 assumes oracle access to η , Algorithm 1 replaces η by an independently trained estimator $\hat{\eta}$: it uses $\hat{\eta}$ to compute each agent’s improvement probability $\hat{\pi}_{\text{Imp}}$ under the deployed model (Algorithm 2), and minimizes the plug-in strategic improvement loss, a hinge surrogate that weights each point by the resulting probability of a positive post-response label (Algorithm 3). The bound separates the resulting error into a generalization term and a plug-in estimation term. Throughout, $\hat{\eta} = g(\hat{w}^\top \cdot)$ with $\hat{w} \in \mathbb{R}_{\geq 0}^d$ is estimated independently of the training sample, of size n , and the guard satisfies $\delta_g \leq 1 - \sup_{z \in \mathcal{A}} \hat{\eta}(z)$. Let $r := \sup_{z \in \mathcal{A}} \|z\|_2 < \infty$, $\rho_{\text{Imp}} := r + \frac{\beta}{\min_{j \in [d]} \alpha_j}$, $C_{k,B} := k\rho_{\text{Imp}} + B$, $\varepsilon_\eta := \sup_{z \in \mathcal{A}} |\hat{\eta}(z) - \eta(z)|$.

Lemma 5.2. *Let $\mathfrak{d} := \text{Pdim}(\Pi_{k,B}) + \text{Pdim}(\mathcal{M}_{k,B})$, where Pdim denotes pseudo-dimension Anthony and Bartlett (2009) and $\Pi_{k,B}, \mathcal{M}_{k,B}$ be the classes of maps $x \mapsto \hat{\pi}_{\text{Imp}}(x; w, b)$ and $x \mapsto M_{w,b}(x)$ over $\mathcal{W}_{k,B}$. Then $\mathfrak{d} < 12(d + 1)$, and the loss class of Algorithm 3 over $\mathcal{W}_{k,B}$ has the covering numbers of a class of pseudo-dimension $\mathfrak{d}$, up to absolute constants.*

Detailed proof of Lemma 5.2 is given in Appendix E.

Theorem 5.3. *Let $(\tilde{w}, \tilde{b}) \in \mathcal{W}_{k,B}$ be the model returned by STRAT-IMP-AWARE. Under Assumptions 1 and 2, for every $\delta \in (0, 1)$, with probability at least*

$1 - \delta$,

$$\begin{aligned} \text{err}_{\text{Imp}}(f_{\tilde{w}, \tilde{b}}) &\leq \widehat{L}_{\text{hinge}}^{\widehat{\mathfrak{d}}}(\tilde{w}, \tilde{b}) + c(1 + C_{k,B})\sqrt{\mathfrak{d}/n} \\ &\quad + (1 + C_{k,B})\sqrt{\frac{\log(1/\delta)}{2n}} + 4C_{k,B}\varepsilon_{\eta}, \end{aligned}$$

where $\mathfrak{d}$ is as in Lemma 5.2 and $c > 0$ is an absolute constant.

Proof overview of Theorem 5.3. The proof combines three bounds, each uniform over $\mathcal{W}_{k,B}$.

1. By Theorem 3.3, $f_{w,b}(x^f) = \mathbf{1}(M_{w,b}(x) \geq 0)$ with $|M_{w,b}(x)| \leq C_{k,B}$, so err_{Imp} is bounded by a hinge risk weighted by $\eta(x^f)$.
2. The weights $q_{w,b}$ used by Algorithm 3 have conditional mean within $2\varepsilon_{\eta}$ of $\eta(x^f)$ (Lemma E.1 in Appendix E), so, replacing $\eta(x^f)$ by conditional mean of $q_{w,b}$ in true loss L_{hinge}^{η} costs at most $4C_{k,B}\varepsilon_{\eta}$ and $\widehat{L}^{\widehat{\mathfrak{d}}}$ is an unbiased estimate of the resulting risk.
3. Since the loss is Lipschitz in $\widehat{\pi}_{\text{Imp}}$ and $M_{w,b}$, Lemma 5.2 controls the covering numbers of the loss class, and Dudley’s entropy integral and McDiarmid’s inequality give the two $\sqrt{\cdot/n}$ terms. $\square$

Detailed proof of Theorem 5.3 is given in Appendix E. *Remark 5.4.* Beyond the empirical risk $\widehat{L}_{\text{hinge}}^{\widehat{\mathfrak{d}}}$, the bound consists of two terms of order $\mathcal{O}(n^{-1/2})$, with $\mathfrak{d} = \mathcal{O}(d)$, and the estimation term $4C_{k,B}\varepsilon_{\eta}$. The estimator enters the guarantee only through ε_{η} , and not through the loss used to obtain it: any independently estimated $\widehat{\eta}$ with $\varepsilon_{\eta} = \mathcal{O}(m^{-1/2})$ yields the same conclusion. We give explicit theoretical upper bound under additional conditions in Appendix F for completeness.

6 EXPERIMENTS

We now validate our theoretical results and assumptions, and evaluate performance of our proposed STRAT-IMP-AWARE on both synthetic and semi-synthetic data. Our code is available [here].

Experimental Setting. We evaluate our algorithm on four datasets Adult (Becker and Kohavi, 1996), HELOC (FICO, 2018), Law School (Wightman, 1998) and ACS Income (Ding et al., 2021) and on Synthetic dataset. We partition each dataset in a 45/45/5/5 ratio: the first 45% is used to estimate η and the rest as training, validation and test sets. We perform two major experiments on these datasets. Additional details on experimental setup with ablation studies are given in the Appendix H. Dataset details are summarized in Table 4 in Appendix H.2. The class-probability function $\eta(x)$ is modeled using Logistic Regression, following the Assumption 1. On all the datasets, we consider the features for which $\eta(x)$ is monotonically increasing, more

details can be found in Appendix H. We use a single-layer neural network with a d -dimensional input and a biased linear output and optimize the model parameters using a STRAT-IMP-LOSS objective. The model is trained using stochastic gradient descent (SGD).

We implement two benchmark algorithms namely SERM (Levanon and Rosenfeld, 2022) and Attias et al. (2025). The classifier is trained using an objective that anticipates agent manipulation. During training, we adjust the loss by adding a strategic gain term, which represents the maximum score change within their budget β . For more details, refer Appendix H.2.

Comparison with Attias et al. (2025): The comparison is necessarily model-mismatched because the two frameworks model improvement differently. Attias et al. (2025) use classifier-independent improvement sets, whereas our model induces improvement endogenously through strategic best responses. We therefore use their method as a representative improvement-aware baseline rather than as a strictly model-matched comparison.

Robustness to Training Set Size. Figure 1 evaluates the improvement-aware strategic error on test data across the datasets as a function of the number of training samples and shows that the improvement-aware linear classifier f_{Imp}^* consistently achieves lowest improvement error regardless of training set size. Furthermore, the strategic linear classifier f_s^* generally outperforms the standard optimal linear classifier f^* , while achieving comparable performance on the Law School dataset. This clear hierarchy empirically demonstrates that f_s^* serves as a better proxy for improvement-aware settings than the naive f^* , even when training data is limited.

Improvement Rates. We evaluate the agent’s label improvement across the four classifiers: SVM, SERM (Levanon and Rosenfeld, 2022), STRAT-IMP-AWARE and Attias et al. (2025). For each deployed classifier, we count an agent as manipulating if its selected best response differs from its original feature. Among manipulating agents, we count an agent as improved if the sampled post-response label is positive after manipulation when the original label was negative. The improvement rate reported in Table 1 is the fraction of negative labeled manipulating agents who improve. Total counts of manipulating and improved agents are quantified using the test data.

7 OUTCOME LAW UNCERTAINTY

Access to η determines the classifier for deployment under improvement aware setting. For instance, we show in Section 4 that if the learner has *oracle access* to

Dataset	Metric	SVM	SERM	STRAT-IMP-AWARE	Attias et al. (2025)
Adult	Manip. agents	1168 [†]	197 [†]	147	183
	Improved agents	153 [†]	128 [†]	104	73
	Imp Error (%) ↓	62.40 [†]	15.50 [†]	15.80	17.90
HELOC	Manip. agents	141	44	92	66
	Improved agents	35	17	47	25
	Imp Error (%) ↓	38.90	32.50	22.0	25.70
Law School	Manip. agents	62 [†]	13 [†]	69	60
	Improved agents	31 [†]	9 [†]	41	31
	Imp Error (%) ↓	5.28 [†]	30.07 [†]	4.71	5.48
ACS Income	Manip. agents	4091	794	2640	135
	Improved agents	1330	725	2310	85
	Imp Error (%) ↓	40.60	19.50	6.4	36.1
Synthetic	Manip. agents	412	38	221	293
	Improved agents	148	35	192	238
	Imp Error (%) ↓	36.60	21.10	6.30	8.10

Table 1: STRAT-IMP-AWARE achieves the lowest strategic improvement error and induces more responses and label improvements than SERM on four of five datasets. Relative to SVM, it induces fewer responses on four of five datasets and achieves a higher improvement rate among responding agents on all five. This suggests that improvement-aware training encourages feature changes that enhance qualifications. [†] class-weighted results addressing severe class imbalance.

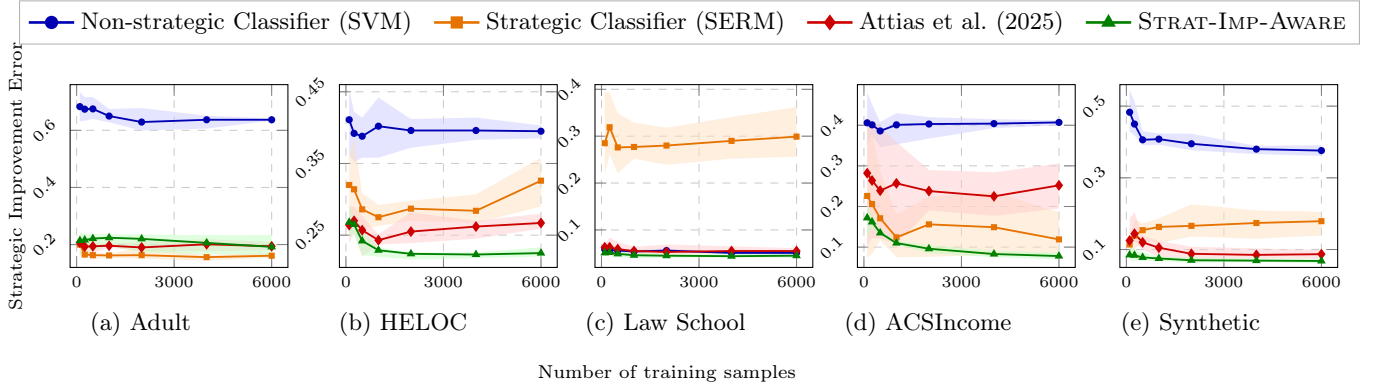


Figure 1: Strategic improvement error ($\widehat{\text{err}}_{\text{Imp}}$) vs. number of training samples. Lines show the mean over 5 runs; shaded regions show the min-max range.

η the optimal improvement aware classifier is learnable. However, oracle access may not be a practical assumption in many real world settings, except where multiple controlled trials are feasible, since post-deployment labels are function of the deployed classifier. Another interesting idea we explored in this paper, is to simulate an η labels from the input data to evaluate the classifiers under consideration. Since the outcome law η governing pre- and post deployment label generation is fixed, we first learn $\hat{\eta}$, an estimate of η and use it to learn optimal improvement aware classifier with an additive error in η estimate. Learning η comes with additional complication that it may not be always possible. We discuss this aspect more in Appendix F and give sufficient conditions under which we can learn η .

Our next result (proof in Appendix G) provides a

recommendation in cases where these conditions are difficult to verify or fail to hold. In particular, we compare two classifiers namely, Bayes optimal classifier, which accounts for neither gaming nor improvement (naive viewpoint), and the optimal strategic classifier which incorporates only strategic gaming (pessimistic viewpoint). We show that when post-deployment labels generation process is difficult to learn it is always safe to choose the optimal strategic classifier.

Theorem 7.1. *Under Assumptions 1 and 2, the feasibility and tie-breaking conditions from Section 3, the strategic-optimal classifier f_s^* and the Bayes-optimal classifier f^* satisfy*

$$\text{err}_{\text{Imp}}(f_s^*) \leq \text{err}_{\text{Imp}}(f^*).$$

8 CONCLUSION

When improvements are endogenous, strategic classification is no longer a purely defensive problem: agents may respond to the deployed classifier by improving their qualifications, not only by gaming it. Our results show that this richer setting remains statistically tractable: under a single-index qualification model and linear-decomposable costs, best responses act as a closed-form shift of the decision boundary, and improvement-aware learning is PAC learnable under oracle access to η , with sample complexity linear in the dimension. Experiments on synthetic and semi-synthetic data support the plausibility of our assumptions and the effectiveness of STRAT-IMP-AWARE when η is estimated. Our guarantees rest on two choices: a single-index structure, which lets $\hat{\eta}$ extrapolate to best-response points beyond the support of the pre-deployment data, and independent agent responses under a fixed conditional outcome law. Future work includes non-decomposable costs e.g., ℓ_2 costs, heterogeneous agents, repeated interactions and causal models that separate genuine improvement from gaming.

AI use statement

In this work, we used generative AI tools in three limited capacities: (i) identifying grammatical and typographical errors in the text, (ii) assisting with the derivation of proofs, and (iii) debugging and writing code. All LLM-assisted proofs were verified line by line by the authors, and all LLM-generated code was reviewed and tested for correctness. Generative AI tools were not used to formulate research questions, hypotheses or claims, and the ideas and contributions presented here are our own. We conducted a manual literature survey to ensure that prior work is properly attributed. We take responsibility for the final content of this work, including all text, claims, proofs, and code, whether or not generative AI tools were used in producing them.

References

- Saba Ahmadi, Hedyeh Beyhaghi, Avrim Blum, and Keziah Naggita. The strategic perceptron. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 6–25, 2021.
- Saba Ahmadi, Hedyeh Beyhaghi, Avrim Blum, and Keziah Naggita. On classification of strategic agents who can both game and improve. *arXiv preprint arXiv:2203.00124*, 2022.
- Saba Ahmadi, Avrim Blum, and Kunhe Yang. Fundamental bounds on online strategic classification. In *Proceedings of the 24th ACM Conference on Economics and Computation*, pages 22–58, 2023.
- Sura Alhanouti and Parinaz Naghizadeh. Anticipating gaming to incentivize improvement: Guiding agents in (fair) strategic classification. *arXiv preprint arXiv:2505.05594*, 2025.
- Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.
- Idan Attias, Avrim Blum, Keziah Naggita, Donya Saleh, Dravyansh Sharma, and Matthew Walter. Pac learning with improvements. *arXiv preprint arXiv:2503.03184*, 2025.
- Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.
- Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. URL <https://uci.edu>. DOI: <https://doi.org>.
- Yatong Chen, Jialu Wang, and Yang Liu. Learning to incentivize improvements from strategic agents. *Transactions on Machine Learning Research*, 2023.
- Lee Cohen, Saeed Sharifi-Malvajerdi, Kevin Stangl, Ali Vakilian, and Juba Ziani. Sequential strategic screening. In *International Conference on Machine Learning*, pages 6279–6295. PMLR, 2023.
- Lee Cohen, Yishay Mansour, Shay Moran, and Han Shao. Learnability gaps of strategic classification. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1223–1259. PMLR, 2024.
- Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.
- Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70, 2018.
- Valia Efthymiou, Ekaterina Fedorova, and Chara Podimata. Desirable effort fairness and optimality trade-offs in strategic learning. *arXiv preprint arXiv:2510.19098*, 2025.
- Itay Eilat, Ben Finkelshtein, Chaim Baskin, and Nir Rosenfeld. Strategic classification with graph neural networks. *arXiv preprint arXiv:2205.15765*, 2022.
- FICO. Explainable machine learning challenge dataset. <https://community.fico.com/s/explainable-machine-learning-challenge>, 2018. Accessed: 2026-10-07.
- Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. Uniform convergence of gradients for non-convex learning and optimization. *Advances in neural information processing systems*, 31, 2018.

- Ganesh Ghalme, Vineet Nair, Itay Eilat, Inbal Talgam-Cohen, and Nir Rosenfeld. Strategic classification in the dark. In *International Conference on Machine Learning*, pages 3672–3681. PMLR, 2021.
- António Góis, Sophia Gunluk, Nir Rosenfeld, Nidhi Hegde, Simon Lacoste-Julien, and Dhanya Sridhar. The role of causal features in strategic classification for robustness and alignment. *arXiv preprint arXiv:2605.27163*, 2026.
- Paul Goldberg and Mark Jerrum. Bounding the vapnik-chervonenkis dimension of concept classes parameterized by real numbers. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 361–369, 1993.
- Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016.
- Guy Horowitz and Nir Rosenfeld. Causal strategic classification: A tale of two shifts. In *International Conference on Machine Learning*, pages 13233–13253. PMLR, 2023.
- Safwan Hossain, Evi Micha, Yiling Chen, and Ariel Procaccia. Strategic classification with externalities. In *International Conference on Learning Representations*, volume 2025, pages 28338–28361, 2025.
- 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.
- Tosca Lechner and Ruth Urner. Learning losses for strategic classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7337–7344, 2022.
- Sagi Levanon and Nir Rosenfeld. Strategic classification made practical. In *International Conference on Machine Learning*, pages 6243–6253. PMLR, 2021.
- Sagi Levanon and Nir Rosenfeld. Generalized strategic classification and the case of aligned incentives. In *International Conference on Machine Learning*, pages 12593–12618. PMLR, 2022.
- Lydia T Liu, Nikhil Garg, and Christian Borgs. Strategic ranking. In *International Conference on Artificial Intelligence and Statistics*, pages 2489–2518. PMLR, 2022.
- 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.
- John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*, pages 6917–6926. PMLR, 2020.
- Smitha Milli, John Miller, Anca D Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 230–239, 2019.
- Mehryar Mohri and Andres Munoz Medina. Learning theory and algorithms for revenue optimization in second price auctions with reserve. In *International conference on machine learning*, pages 262–270. PMLR, 2014.
- Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.
- Han Shao, Avrim Blum, and Omar Montasser. Strategic classification under unknown personalized manipulation. *Advances in Neural Information Processing Systems*, 36:26452–26484, 2023.
- Dravyansh Sharma and Alec Sun. Conservative classifiers do consistently well with improving agents: characterizing statistical and online learning. *Advances in Neural Information Processing Systems*, 38:101320–101346, 2026.
- Eliot Shekhtman and Sarah Dean. Strategic usage in a multi-learner setting. In *International Conference on Artificial Intelligence and Statistics*, pages 2665–2673. PMLR, 2024.
- Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. *Advances in neural information processing systems*, 23, 2010.
- Ravi Sundaram, Anil Vullikanti, Haifeng Xu, and Fan Yao. Pac-learning for strategic classification. *Journal of Machine Learning Research*, 24(192):1–38, 2023.
- Wei Tang, Chien-Ju Ho, and Yang Liu. Linear models are robust optimal under strategic behavior. In *International Conference on Artificial Intelligence and Statistics*, pages 2584–2592. PMLR, 2021.
- Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 10–19, 2019.
- Linda F. Wightman. Lsac national longitudinal bar passage study. Lsac research report series, Law School Admission Council, Newtown, PA, 1998. URL <https://github.com/damtharvey/law-school-dataset>.
- Tian Xie, Zhiqun Zuo, Mohammad Mahdi Khalili, and Xueru Zhang. Learning under imitative strategic behavior with unforeseeable outcomes. *arXiv preprint arXiv:2405.01797*, 2024.

Atsutomo Yara and Yoshikazu Terada. Nonparametric logistic regression with deep learning. *Bernoulli*, 32 (2):952–977, 2026.

Hanrui Zhang and Vincent Conitzer. Incentive-aware pac learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5797–5804, 2021.

CHECKLIST

1. For all models and algorithms presented, check if you include:
 - (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
 - (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
 - (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
2. For any theoretical claim, check if you include:
 - (a) Statements of the full set of assumptions of all theoretical results. [Yes]
 - (b) Complete proofs of all theoretical results. [Yes]
 - (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
 - (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
 - (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
 - (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
 - (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
 - (a) Citations of the creator If your work uses existing assets. [Yes]
 - (b) The license information of the assets, if applicable. [Yes]
 - (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
 - (d) Information about consent from data providers/curators. [Not Applicable]
 - (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
 - (a) The full text of instructions given to participants and screenshots. [Not Applicable]
 - (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
 - (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Linear Strategic Classification with Endogenous Improvements: Supplementary Materials

A NOTATION

Notation	Description
$\mathcal{X} \subseteq \mathbb{R}^d$	Valid feature space; both original features and reported/post-response features lie in $\mathcal{X}$.
$\mathcal{X}_0 = \text{supp}(D) \subseteq \mathcal{X}$	Support of the original/pre-deployment feature distribution.
$\mathcal{A} \subseteq \mathcal{X}$	Feasible report/action space, with $\mathcal{X}_0 \subseteq \mathcal{A}$.
D	Distribution over original features, supported on $\mathcal{X}_0$.
$x \in \mathcal{X}_0$	Original feature vector sampled from D .
$z \in \mathcal{A}$	Reported/post-response feature vector.
$Y = \{0, 1\}$	Binary label space.
$\eta : \mathcal{X} \rightarrow [0, 1]$	Class-probability function, $\eta(x) = \mathbb{P}(Y = 1 \mid X = x)$.
$y_x \sim \text{Bernoulli}(\eta(x))$	Conditional label generated at feature vector $x \in \mathcal{X}$.
$\mathcal{F}$	Hypothesis class of linear classifiers $f_{w,b} : \mathcal{X} \rightarrow \{0, 1\}$.
$f_{w,b}(z) = \mathbb{1}\{w^\top z \geq b\}$	Linear classifier with weight vector $w \in \mathbb{R}_{\geq 0}^d$ and threshold $b \in \mathbb{R}$.
$c : \mathcal{X}_0 \times \mathcal{A} \rightarrow \mathbb{R}_+$	Cost of changing original feature $x \in \mathcal{X}_0$ to report $z \in \mathcal{A}$.
$c(x, z) = \sum_{i=1}^d \alpha_i (z_i - x_i)_+$	Linear-decomposable one-sided cost, for some $i \in [d]$, $\alpha_i > 0$.
$(a)_+ = \max\{a, 0\}$	Positive part of $a \in \mathbb{R}$.
$\beta > 0$	Utility gain from receiving positive classification.
$u_f(x, z) = \beta f(z) - c(x, z)$	Utility of reporting $z \in \mathcal{A}$ for an agent with original feature $x \in \mathcal{X}_0$.
$x^f \in \arg \max_{z \in \mathcal{A}} u_f(x, z)$	Best response/report under classifier $f_{w,b}$.
$\text{err}(f) = \mathbb{E}_{x \sim D} [\mathbb{1}\{f(x) \neq y_x\}]$	Standard non-strategic classification error.
$\text{err}_s(f) = \mathbb{E}_{x \sim D} [\mathbb{1}\{f(x^f) \neq y_x\}]$	Strategic error with immutable labels.
$\text{err}_{\text{Imp}}(f) = \mathbb{E}_{x \sim D} [\mathbb{1}\{f(x^f) \neq y_{x^f}\}]$	Strategic improvement error with post-response labels.
$\widehat{\text{err}}_{\text{Imp}}(f) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{f(x_i^f) \neq y_{x_i^f}\}$	Empirical strategic improvement error.
$g : \mathbb{R} \rightarrow [0, 1]$	Monotone link function in the single-index model.
$w^* \in \mathbb{R}_{\geq 0}^d$	True single-index direction, satisfying $\eta(x) = g((w^*)^\top x)$.
$b^* = \inf\{t \in \mathbb{R} : g(t) \geq 1/2\}$	Bayes threshold under the single-index model.
f^*	Non-strategic Bayes optimal classifier.
f_s^*	Strategic-optimal shifted classifier.
f_{Imp}^*	Improvement-aware optimal classifier.
$S(w) = \beta \max_{i \in [d]} w_i / \alpha_i$	Strategic shift induced by classifier $f_{w,b}$.
$b_s = b + S(w)$	Shifted threshold corresponding to $f_{w,b}$.
$M_{w,b}(x) = w^\top x - b + S(w)$	Strategic shifted margin used in the hinge surrogate.
$\rho_{\text{Imp}} = r + \beta / \min_{j \in [d]} \alpha_j$	Radius parameter used in the generalization bound, where $r = \sup_{z \in \mathcal{A}} \ z\ _2$.
$\hat{\eta}$	Plug-in estimator of η .
$\varepsilon_{\hat{\eta}} = \sup_{z \in \mathcal{A}} \hat{\eta}(z) - \eta(z) $	Uniform plug-in estimation error over feasible reports.
$\widehat{L}_{\text{hinge}}^{\hat{\eta}}(w, b)$	Empirical plug-in strategic improvement loss minimized by STRAT-IMP-AWARE.
$\mathcal{W}_{k,B}$	Norm- and bias-bounded parameter class $\{(w, b) : w \in \mathbb{R}_{\geq 0}^d, \ w\ _2 \leq k, b \leq B\}$.
$C_{k,B} = k\rho_{\text{Imp}} + B$	Uniform margin bound used in Theorem 5.3.

Table 2: Notation used in the paper.

B PROOFS FROM SECTION 2

Observation 2.4 (Existence of a linear Non-strategic Bayes optimal classifier). *Under 0–1 loss in the Single-index qualification model (Assumption 1), there exists a Non-strategic Bayes optimal classifier of the form*

$f^*(x) = \mathbf{1}\{w^{\star\top}x \geq b^*\}$ for some threshold $b^* \in \mathbb{R}$. In particular, if $b^* := \inf\{t \in \mathbb{R} : g(t) \geq 1/2\}$, then the classifier above is Bayes-optimal.

Proof. By Assumption 1,

$$\eta(x) = g(w^{\star\top}x).$$

Since g is nondecreasing, the set

$$\{x : \eta(x) \geq 1/2\} = \{x : g(w^{\star\top}x) \geq 1/2\}$$

is a threshold set of the form $\{x : w^{\star\top}x \geq b^*\}$, where

$$b^* := \inf\{t \in \mathbb{R} : g(t) \geq 1/2\}.$$

Hence the classifier $f^*(x) = \mathbf{1}\{w^{\star\top}x \geq b^*\}$ is Bayes optimal. $\square$

C PROOFS FROM SECTION 3

In this section, we present proofs from Section 3.

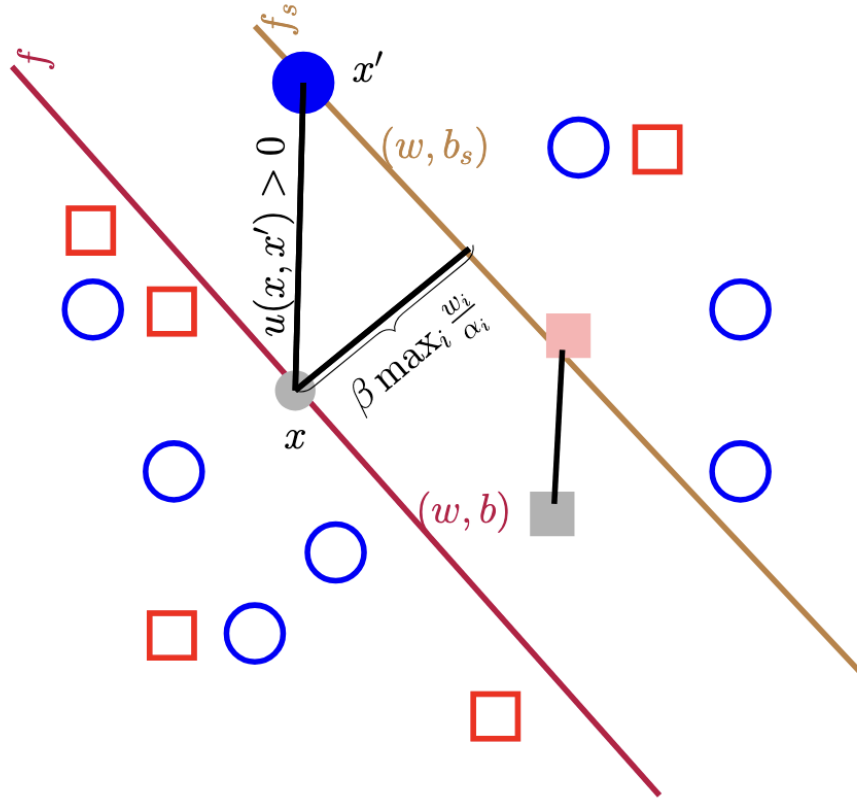


Figure 2: The strategic classifier f_s (brown) is obtained by shifting the linear classifier f (red) by a margin of $\max_{i \in [d]} \left(\frac{w_i}{\alpha_i} \right) \beta$. This translation anticipates utility-maximizing feature manipulation from x to x' by the agents.

Lemma 3.1. For $f_{w,b} \in \mathcal{F}$, let $S(w) := \beta \max_{j \in [d]} w_j / \alpha_j$ and let j^* be the smallest index in $\arg \max_{j \in [d]} w_j / \alpha_j$. Then for every $x \in \mathcal{X}_0$, the selected best response is

$$x^f = \begin{cases} x + \frac{b - w^\top x}{w_{j^*}} e_{j^*}, & \text{if } b - S(w) \leq w^\top x < b \\ x, & \text{otherwise.} \end{cases}$$

Proof. Let $\Delta := b - w^\top x$. Since $w \neq 0$ and $\alpha_j > 0$, we have $S(w) > 0$ and $w_{j^*} > 0$. For any $z \in \mathcal{A}$ with $w^\top z \geq b$, using $\alpha_j \geq \beta w_j / S(w)$, $(t)_+ \geq t$ and $w \geq 0$,

$$c(x, z) \geq \frac{\beta}{S(w)} \sum_{j=1}^d w_j (z_j - x_j)_+ \geq \frac{\beta}{S(w)} w^\top (z - x) \geq \frac{\beta \Delta}{S(w)}.$$

Hence every positively classified z has $u_f(x, z) \leq \beta(1 - \Delta/S(w))$, where $u_f(x, z) := \beta \mathbf{1}(w^\top z \geq b) - c(x, z)$, while every negatively classified z has $u_f(x, z) \leq 0$.

Case $\Delta \leq 0$. Staying at x is classified positively at zero cost, so $u_f(x, z) = \beta$ is maximal and $x^f = x$.

Case $0 < \Delta \leq S(w)$. Let $z^* := x + (\Delta/w_{j^*})e_{j^*}$. Since $\Delta/w_{j^*} \leq S(w)/w_{j^*} = \beta/\alpha_{j^*}$, $z^* \in \mathcal{A}$ by feasibility. Moreover $w^\top z^* = b$ and $c(x, z) = \alpha_{j^*} \Delta / w_{j^*} = \beta \Delta / S(w)$, so z^* attains the bound above with $u_f(x, z^*) \geq 0$. Thus, z^* is a positive best response; when $\Delta = S(w)$ it ties with staying, and tie-breaking selects it. Among point $x + te_{j^*}$, only $t = \Delta/w_{j^*}$ is both positive and cost-minimal, so $x^f = z^*$.

Case $\Delta > S(w)$. Every positively classified z has $u_f(x, z) < 0 = u_f(x, x)$, so no positive best response exists and $x^f = x$. $\square$

Theorem 3.3. For $f_{w,b} \in \mathcal{F}$, let $f_s(x) = \mathbf{1}\{w^\top x \geq b_s\}$ with $b_s := b + \beta \max_{j \in [d]} w_j / \alpha_j$. Then $f_s(x^{f_s}) = f_{w,b}(x)$ for every $x \in \mathcal{X}_0$. Consequently, $\text{err}_s(f_s) = \text{err}(f_{w,b})$.

Proof. Let $r(w) = \max_{j \in [d]} w_j / \alpha_j$, fix $f_{w,b} \in \mathcal{F}$. Let x^{f_s} denote the selected tie-broken best response to the shifted classifier f_s . For any selected report z in the range of the best-response map, define $\mathcal{Z}_z := \{x \in \mathcal{X}_0 : z \text{ is the selected best response to } f_s \text{ for agent } x\}$. These sets partition $\mathcal{X}_0$ according to the selected best-response rule. It is therefore enough to show that, for every selected report z and every $x \in \mathcal{Z}_z$, $f_s(z) = f_{w,b}(x)$.

We first record a useful bound. For any $x \in \mathcal{X}_0, y \in \mathcal{A}$,

$$\begin{aligned} w^\top (y - x) &\leq \sum_{i=1}^d w_i (y_i - x_i)_+ \leq r(w) \sum_{i=1}^d \alpha_i (y_i - x_i)_+ \\ &= r(w) c(x, y). \end{aligned}$$

We now consider three cases.

Case 1: $w^\top z > b_s$. Then $f_s(z) = 1$. Suppose, for contradiction, that $f_{w,b}(x) = 0$ for some $x \in \mathcal{Z}_z$. Then $w^\top x < b$, and hence $f_s(x) = 0$. Since staying at x gives utility 0, while reporting z gives utility $\beta - c(x, z)$, optimality of z implies $c(x, z) \leq \beta$. Using the bound above, $w^\top z \leq w^\top x + r(w)c(x, z) \leq w^\top x + \beta r(w) < b + \beta r(w) = b_s$, contradicting $w^\top z > b_s$. Hence $f_{w,b}(x) = 1 = f_s(z)$.

Case 2: $w^\top z = b_s$. Again $f_s(z) = 1$. Suppose, for contradiction, that $f_{w,b}(x) = 0$ for some $x \in \mathcal{Z}_z$. Then $w^\top x < b$ and $f_s(x) = 0$. Since z is a selected best response, we must have $c(x, z) \leq \beta$. Therefore, $b_s = w^\top z \leq w^\top x + r(w)c(x, z) \leq w^\top x + \beta r(w) < b + \beta r(w) = b_s$, a contradiction. Thus $f_{w,b}(x) = 1 = f_s(z)$.

Case 3: $w^\top z < b_s$. Then $f_s(z) = 0$. Suppose, for contradiction, that $f_{w,b}(x) = 1$ for some $x \in \mathcal{Z}_z$, i.e., $w^\top x \geq b$. If $w^\top x \geq b_s$, then staying at x gives positive classification with zero cost and utility β , whereas reporting z gives utility $-c(x, z) \leq 0$, contradicting optimality of z . Hence we may assume $b \leq w^\top x < b_s$. Let $\delta := b_s - w^\top x \in (0, \beta r(w)]$. Choose $i^* \in \arg \max_i w_i / \alpha_i$. Since $w \neq 0$ and $\alpha_i > 0$, we have $w_{i^*} > 0$. By the feasibility assumption, the point $y := x + \frac{\delta}{w_{i^*}} e_{i^*}$ belongs to $\mathcal{A}$. Moreover, $w^\top y = b_s$, so $f_s(y) = 1$, and $c(x, y) = \alpha_{i^*} \frac{\delta}{w_{i^*}} = \frac{\delta}{r(w)} \leq \beta$. Thus reporting y gives utility $\beta - c(x, y) \geq 0$. On the other hand, since $f_s(z) = 0$, reporting z gives utility $-c(x, z) \leq 0$. If $\beta - c(x, y) > 0$, this contradicts optimality of z . If $\beta - c(x, y) = 0$, then y is a positive best response with utility equal to that of any zero-utility negative response, so the tie-breaking rule selects a positive response rather than z . This again contradicts z being the selected report. Therefore $f_{w,b}(x) = 0 = f_s(z)$.

Combining the three cases, for every selected report z and every $x \in \mathcal{Z}_z$, $f_s(z) = f_{w,b}(x)$. Equivalently, for every $x \in \mathcal{X}_0$, $f_s(x^{f_s}) = f_{w,b}(x)$. Taking expectation over the data distribution yields $\text{err}_s(f_s) = \text{err}(f_{w,b})$. $\square$

Corollary 3.4. *Let $f_{w^*, b^*} \in \arg \min_{f \in \mathcal{F}} \mathbf{err}(f)$, i.e., it is a non-strategic Bayes-optimal classifier (Observation 2.4). Then, its shift $f_s^* := f_{w^*, b_s^*}$ is strategic optimal among linear classifiers in $\mathcal{F}$.*

Proof. Let f^* be the Non-strategic Bayes optimal linear classifier and f_s^* be the corresponding strategic classifier satisfying $\mathbf{err}_s(f_s^*) = \mathbf{err}(f^*)$. For contradiction, assume that there is a classifier f'_s with parameters (w', b') satisfying

$$\mathbf{err}_s(f'_s) < \mathbf{err}_s(f_s^*). \quad (4)$$

and consider classifier f' with parameters $w', b' - \max_{i \in [d]} \left(\frac{w'_i}{\alpha_i} \right) \beta$. The following set of inequalities result in a contradiction.

$$\mathbf{err}_s(f_s^*) = \mathbf{err}(f^*) \leq \mathbf{err}(f') = \mathbf{err}_s(f'_s) < \mathbf{err}_s(f_s^*).$$

The first inequality above follows from Bayes optimality of f^* and the second inequality follows from Equation 4. $\square$

Lemma C.1. *[Improvement threshold lies between Non-strategic Bayes and Strategic thresholds] Let $x \sim \mathcal{D}$ with $x \in \mathcal{X} \subseteq \mathbb{R}$, and the label associated with input x be $y_x \sim \text{Bernoulli}(\eta(x))$ where $\eta : \mathcal{X} \rightarrow [0, 1]$ is a strictly increasing function. Consider threshold classifiers of the form $f_\theta(x) = \mathbb{1}(x \geq \theta)$. Denote by θ^* the Non-strategic Bayes optimal threshold. Let θ_s be the optimal threshold under strategic manipulation with $\exists \alpha \in \mathbb{R}^+$ then suppose cost function $c(x, x') = \max(0, \alpha(x' - x))$ and let θ_{Imp} be the optimal threshold under our improvement-aware model. Then the thresholds satisfy:*

$$\theta^* \leq \theta_{\text{Imp}} \leq \theta_s.$$

Proof. We establish the inequality chain $\theta^* \leq \theta_{\text{Imp}} \leq \theta_s$ via proof by contradiction. We rely on the monotonicity of the class posterior $\eta(x)$ under the generalized linear model assumption. We proceed in two parts, First, suppose that $\theta_{\text{Imp}} > \theta_s$. We will show that this violates the optimality condition of the improvement-aware classifier. Next, suppose that $\theta_{\text{Imp}} < \theta^*$. We will again reach a contradiction by showing that the improvement risk can be strictly reduced by raising the threshold. For simplicity in notation, let $x' = x^{f_{\theta_{\text{Imp}}}}$ be strategic manipulated point under $f_{\theta_{\text{Imp}}}$ threshold classifier.

We know that,

$$\begin{aligned} \theta_{\text{Imp}} &= \arg \min_{\theta \in \mathbb{R}} \mathbf{err}_{\text{Imp}}(f_\theta) \\ &= \arg \min_{\theta \in \mathbb{R}} \mathbb{E}[\mathbb{1}(f_\theta(x^{f_\theta}) \neq y_{x^{f_\theta}})] \end{aligned}$$

We begin with a few supporting technical results.

Claim C.2. *Let us assume $\theta^* > \theta_{\text{Imp}}$. Then,*

$$\left\{ \begin{aligned} &\mathbb{P}(y_{\theta_{\text{Imp}}} = 0)(F(\theta_{\text{Imp}}) - F(\theta_{\text{Imp}} - \frac{\beta}{\alpha})) - \mathbb{P}(y_{\theta^*} = 0)(F(\theta^*) - F(\theta^* - \frac{\beta}{\alpha})) \\ &+ \int_{\theta_{\text{Imp}}}^{\theta^*} \mathbb{P}(y_x = 0) p_X(x) dx - \int_{\theta_{\text{Imp}} - \frac{\beta}{\alpha}}^{\theta^* - \frac{\beta}{\alpha}} \mathbb{P}(y_x = 1) p_X(x) dx \end{aligned} \right\} > 0$$

Proof.

$$\begin{aligned} &\mathbb{P}(y_{\theta_{\text{Imp}}} = 0)(F(\theta_{\text{Imp}}) - F(\theta_{\text{Imp}} - \frac{\beta}{\alpha})) - \mathbb{P}(y_{\theta^*} = 0)(F(\theta^*) - F(\theta^* - \frac{\beta}{\alpha})) \\ &+ \int_{\theta_{\text{Imp}}}^{\theta^*} \mathbb{P}(y_x = 0) p_X(x) dx - \int_{\theta_{\text{Imp}} - \frac{\beta}{\alpha}}^{\theta^* - \frac{\beta}{\alpha}} \mathbb{P}(y_x = 1) p_X(x) dx \\ &\geq \mathbb{P}(y_{\theta_{\text{Imp}}} = 0)(F(\theta_{\text{Imp}}) - F(\theta_{\text{Imp}} - \frac{\beta}{\alpha})) + \mathbb{P}(y_{\theta^*} = 0)(F(\theta^* - \frac{\beta}{\alpha}) - F(\theta_{\text{Imp}})) \end{aligned}$$

$$\begin{aligned}
 & - \int_{\theta_{\text{Imp}} - \frac{\beta}{\alpha}}^{\theta^* - \frac{\beta}{\alpha}} \mathbb{P}(y_x = 1) p_X(x) dx \\
 & \quad (\because \forall x \in [\theta_{\text{Imp}}, \theta^*], \quad \mathbb{P}(y_x = 0) \leq \mathbb{P}(y_{\theta_{\text{Imp}}} = 0)) \\
 & > \mathbb{P}(y_{\theta^*} = 0) (F(\theta^* - \frac{\beta}{\alpha}) - F(\theta_{\text{Imp}} - \frac{\beta}{\alpha})) - \int_{\theta_{\text{Imp}} - \frac{\beta}{\alpha}}^{\theta^* - \frac{\beta}{\alpha}} \mathbb{P}(y_x = 1) p_X(x) dx \\
 & \quad (\because \theta_{\text{Imp}} < \theta^* \implies \mathbb{P}(y_{\theta_{\text{Imp}}} = 0) > \mathbb{P}(y_{\theta^*} = 0)) \\
 & = (F(\theta^* - \frac{\beta}{\alpha}) - F(\theta_{\text{Imp}} - \frac{\beta}{\alpha}))(\mathbb{P}(y_{\theta^*} = 0) - 1) + \int_{\theta_{\text{Imp}} - \frac{\beta}{\alpha}}^{\theta^* - \frac{\beta}{\alpha}} \mathbb{P}(y_x = 0) p_X(x) dx \\
 & \geq (F(\theta^* - \frac{\beta}{\alpha}) - F(\theta_{\text{Imp}} - \frac{\beta}{\alpha}))(\mathbb{P}(y_{\theta^*} = 0) + \mathbb{P}(y_{\theta^* - \frac{\beta}{\alpha}} = 0) - 1) \\
 & \quad (\because \forall x \in [\theta_{\text{Imp}} - \frac{\beta}{\alpha}, \theta^* - \frac{\beta}{\alpha}], \quad \mathbb{P}(y_x = 0) \geq \mathbb{P}(y_{\theta^* - \frac{\beta}{\alpha}} = 0)) \\
 & > (F(\theta^* - \frac{\beta}{\alpha}) - F(\theta_{\text{Imp}} - \frac{\beta}{\alpha}))(2\mathbb{P}(y_{\theta^*} = 0) - 1) \\
 & = 0 \quad (\because \mathbb{P}(y_{\theta^*} = 0) = \frac{1}{2})
 \end{aligned}$$

Hence proved. $\square$

Claim C.3. Let us assume $\theta_{\text{Imp}} > \theta_s$. Then,

$$\left\{ \begin{aligned} & \mathbb{P}(y_{\theta_{\text{Imp}}} = 0)(F(\theta_{\text{Imp}}) - F(\theta_{\text{Imp}} - \frac{\beta}{\alpha})) - \mathbb{P}(y_{\theta_s} = 0)(F(\theta_s) - F(\theta_s - \frac{\beta}{\alpha})) \\ & - \int_{\theta_s}^{\theta_{\text{Imp}}} \mathbb{P}(y_x = 0) p_X(x) dx + \int_{\theta_s - \frac{\beta}{\alpha}}^{\theta_{\text{Imp}} - \frac{\beta}{\alpha}} \mathbb{P}(y_x = 1) p_X(x) dx \end{aligned} \right\} > 0$$

Proof.

$$\begin{aligned}
 & \mathbb{P}(y_{\theta_{\text{Imp}}} = 0)(F(\theta_{\text{Imp}}) - F(\theta_{\text{Imp}} - \frac{\beta}{\alpha})) - \mathbb{P}(y_{\theta_s} = 0)(F(\theta_s) - F(\theta_s - \frac{\beta}{\alpha})) \\
 & - \int_{\theta_s}^{\theta_{\text{Imp}}} \mathbb{P}(y_x = 0) p_X(x) dx + \int_{\theta_s - \frac{\beta}{\alpha}}^{\theta_{\text{Imp}} - \frac{\beta}{\alpha}} \mathbb{P}(y_x = 1) p_X(x) dx \\
 & \geq \mathbb{P}(y_{\theta_{\text{Imp}}} = 0)(F(\theta_{\text{Imp}}) - F(\theta_{\text{Imp}} - \frac{\beta}{\alpha})) + \mathbb{P}(y_{\theta_s} = 0)(F(\theta_s - \frac{\beta}{\alpha}) - F(\theta_{\text{Imp}})) \\
 & + \int_{\theta_s - \frac{\beta}{\alpha}}^{\theta_{\text{Imp}} - \frac{\beta}{\alpha}} \mathbb{P}(y_x = 1) p_X(x) dx \\
 & \quad (\because \forall x \in [\theta_s, \theta_{\text{Imp}}], \quad -\mathbb{P}(y_x = 0) \geq -\mathbb{P}(y_{\theta_s} = 0)) \\
 & > \mathbb{P}(y_{\theta_{\text{Imp}}} = 0)(F(\theta_s - \frac{\beta}{\alpha}) - F(\theta_{\text{Imp}} - \frac{\beta}{\alpha})) + \int_{\theta_s - \frac{\beta}{\alpha}}^{\theta_{\text{Imp}} - \frac{\beta}{\alpha}} \mathbb{P}(y_x = 1) p_X(x) dx \\
 & \quad (\because \theta_s < \theta_{\text{Imp}} \implies \mathbb{P}(y_{\theta_s} = 0) > \mathbb{P}(y_{\theta_{\text{Imp}}} = 0)) \\
 & = (F(\theta_{\text{Imp}} - \frac{\beta}{\alpha}) - F(\theta_s - \frac{\beta}{\alpha}))(1 - \mathbb{P}(y_{\theta_{\text{Imp}}} = 0)) - \int_{\theta_s - \frac{\beta}{\alpha}}^{\theta_{\text{Imp}} - \frac{\beta}{\alpha}} \mathbb{P}(y_x = 0) p_X(x) dx \\
 & \geq (F(\theta_{\text{Imp}} - \frac{\beta}{\alpha}) - F(\theta_s - \frac{\beta}{\alpha}))(1 - \mathbb{P}(y_{\theta_{\text{Imp}}} = 0) - \mathbb{P}(y_{\theta_s - \frac{\beta}{\alpha}} = 0))
 \end{aligned}$$

$$\begin{aligned}
 & (\because \forall x \in [\theta_s - \frac{\beta}{\alpha}, \theta_{\text{Imp}} - \frac{\beta}{\alpha}], \quad -\mathbb{P}(y_x = 0) \geq -\mathbb{P}(y_{\theta_s - \frac{\beta}{\alpha}} = 0)) \\
 & > (F(\theta_{\text{Imp}} - \frac{\beta}{\alpha}) - F(\theta_s - \frac{\beta}{\alpha}))(1 - 2\mathbb{P}(y_{\theta^*} = 0)) \quad (\because -\mathbb{P}(y_{\theta_{\text{Imp}}} = 0) > -\mathbb{P}(y_{\theta_s - \frac{\beta}{\alpha}} = 0) = -\mathbb{P}(y_{\theta^*} = 0)) \\
 & = 0 \quad (\because \mathbb{P}(y_{\theta^*} = 0) = \frac{1}{2})
 \end{aligned}$$

□

Case 1: Lower bound ($\theta_{\text{Imp}} \geq \theta^*$) Now, let us assume that $\theta_{\text{Imp}} < \theta^*$.

$$\begin{aligned}
 \text{err}_{\text{Imp}}(f_{\theta_{\text{Imp}}}) &= \mathbb{E}[\mathbb{1}(f_{\theta_{\text{Imp}}}(x') \neq y_{x'})] \\
 &= \mathbb{E}[\mathbb{1}(f_{\theta_{\text{Imp}}}(x') \neq y_{x'}, x \in R_{\theta_{\text{Imp}}})] + \mathbb{E}[\mathbb{1}(f_{\theta_{\text{Imp}}}(x') \neq y_{x'}, x \notin R_{\theta_{\text{Imp}}})] \\
 &= \mathbb{P}(f_{\theta_{\text{Imp}}}(x') \neq y_{x'}, x \in R_{\theta_{\text{Imp}}}) + \mathbb{P}(f_{\theta_{\text{Imp}}}(x') \neq y_{x'}, x \notin R_{\theta_{\text{Imp}}})
 \end{aligned}$$

$R_{\theta_{\text{Imp}}} := [\theta_{\text{Imp}} - \frac{\beta}{\alpha}, \theta_{\text{Imp}}]$ is the region such that agents in this region modify their features to get positive classification.

$$\begin{aligned}
 &= \mathbb{P}(y_{\theta_{\text{Imp}}} = 0, x \in R_{\theta_{\text{Imp}}}) + \mathbb{P}(f_{\theta_{\text{Imp}}}(x') \neq y_{x'}, x \notin R_{\theta_{\text{Imp}}}) \\
 &= \mathbb{P}(y_{\theta_{\text{Imp}}} = 0, x \in R_{\theta_{\text{Imp}}}) + \mathbb{P}(y_x = 0, x > \theta_{\text{Imp}}) + \mathbb{P}(y_x = 1, x < \theta_{\text{Imp}} - \frac{\beta}{\alpha}) \\
 &= \int_{\theta_{\text{Imp}} - \frac{\beta}{\alpha}}^{\theta_{\text{Imp}}} \mathbb{P}(y_{\theta_{\text{Imp}}} = 0) p_X(x) dx + \int_{\theta_{\text{Imp}}}^{\infty} \mathbb{P}(y_x = 0) p_X(x) dx \\
 &\quad + \int_{-\infty}^{\theta_{\text{Imp}} - \frac{\beta}{\alpha}} \mathbb{P}(y_x = 1) p_X(x) dx \\
 &= \text{err}_{\text{Imp}}(f_{\theta^*}) + \mathbb{P}(y_{\theta_{\text{Imp}}} = 0)(F(\theta_{\text{Imp}}) - F(\theta_{\text{Imp}} - \frac{\beta}{\alpha})) \\
 &\quad - \mathbb{P}(y_{\theta^*} = 0)(F(\theta^*) - F(\theta^* - \frac{\beta}{\alpha})) + \int_{\theta_{\text{Imp}}}^{\theta^*} \mathbb{P}(y_x = 0) p_X(x) dx \\
 &\quad - \int_{\theta_{\text{Imp}} - \frac{\beta}{\alpha}}^{\theta^* - \frac{\beta}{\alpha}} \mathbb{P}(y_x = 1) p_X(x) dx \\
 &> \text{err}_{\text{Imp}}(f_{\theta^*}) \quad (\text{Follows from the Lemma C.2})
 \end{aligned}$$

which contradicts improvement optimality condition of classifier, our assumption is false. Thus $\theta_{\text{Imp}} \geq \theta^*$

Case 2: Upper bound ($\theta_{\text{Imp}} \leq \theta_s$) Now, let us assume that $\theta_{\text{Imp}} > \theta_s$.

$$\begin{aligned}
 \text{err}_{\text{Imp}}(f_{\theta_{\text{Imp}}}) &= \mathbb{E}[\mathbb{1}(f_{\theta_{\text{Imp}}}(x') \neq y_{x'})] \\
 &= \mathbb{P}(y_{\theta_{\text{Imp}}} = 0, x \in R_{\theta_{\text{Imp}}}) + \mathbb{P}(y_x = 0, x > \theta_{\text{Imp}}) + \mathbb{P}(y_x = 1, x < \theta_{\text{Imp}} - \frac{\beta}{\alpha}) \\
 &= \int_{\theta_{\text{Imp}} - \frac{\beta}{\alpha}}^{\theta_{\text{Imp}}} \mathbb{P}(y_{\theta_{\text{Imp}}} = 0) p_X(x) dx + \int_{\theta_{\text{Imp}}}^{\infty} \mathbb{P}(y_x = 0) p_X(x) dx \\
 &\quad + \int_{-\infty}^{\theta_{\text{Imp}} - \frac{\beta}{\alpha}} \mathbb{P}(y_x = 1) p_X(x) dx \\
 &= \text{err}_{\text{Imp}}(f_{\theta_s}) + \mathbb{P}(y_{\theta_{\text{Imp}}} = 0)(F(\theta_{\text{Imp}}) - F(\theta_{\text{Imp}} - \frac{\beta}{\alpha}))
 \end{aligned}$$

$$\begin{aligned}
 & -\mathbb{P}(y_{\theta_s} = 0)(F(\theta_s) - F(\theta_s - \frac{\beta}{\alpha})) - \int_{\theta_s}^{\theta_{\text{Imp}}} \mathbb{P}(y_x = 0) p_X(x) dx \\
 & - \int_{\theta_s - \frac{\beta}{\alpha}}^{\theta_{\text{Imp}} - \frac{\beta}{\alpha}} \mathbb{P}(y_x = 1) p_X(x) dx \\
 & > \mathbf{err}_{\text{Imp}}(f_{\theta_s}) \quad (\text{Follows from the Lemma C.3})
 \end{aligned}$$

which contradicts the improvement optimality condition, our assumption is false. Thus $\theta_{\text{Imp}} \leq \theta_s$. $\square$

D GENERALIZATION BOUNDS WITH ORACLE ACCESS

Lemma D.1 (Uniform deviation bound for a VC class). *Let $\mathcal{H}$ be a class of binary functions $h : \mathcal{Z} \rightarrow \{0, 1\}$ with $\text{VC}(\mathcal{H}) \leq d$. Let $z_1, \dots, z_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{P}$, and define*

$$\Gamma(S) := \sup_{h \in \mathcal{H}} \left| \mathbb{E}[h(z)] - \frac{1}{n} \sum_{i=1}^n h(z_i) \right|.$$

Then

$$\mathbb{E}_S[\Gamma(S)] \leq \frac{2}{\sqrt{n}} + 4\sqrt{\frac{2d \log(en/d)}{n}}.$$

Moreover, with probability at least $1 - \delta$,

$$\Gamma(S) \leq \frac{2}{\sqrt{n}} + 4\sqrt{\frac{2d \log(en/d)}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Proof. Let $S' = \{z'_1, \dots, z'_n\}$ be an independent ghost sample drawn from $\mathcal{P}$, and let $\sigma_1, \dots, \sigma_n$ be independent Rademacher random variables. By symmetrization,

$$\mathbb{E}_S[\Gamma(S)] \leq 2\mathbb{E}_{S, \sigma} \left[\sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i h(z_i) \right| \right] = 2\mathbb{E}_S[\widehat{\mathfrak{R}}_S(\mathcal{H})].$$

Since $\mathcal{H}$ is binary and has VC dimension d , Sauer's lemma gives

$$\pi_H(n) \leq \left(\frac{en}{d} \right)^d, \quad n \geq d$$

where $\pi_H(n)$ is the growth function of $\mathcal{H}$. Combining this with the standard Rademacher bound using Massart's lemma for binary VC classes yields

$$\mathbb{E}_S[\widehat{\mathfrak{R}}_S(\mathcal{H})] \leq \frac{1}{\sqrt{n}} + 2\sqrt{\frac{2d \log(en/d)}{n}}.$$

Therefore,

$$\mathbb{E}_S[\Gamma(S)] \leq 2 \left(\frac{1}{\sqrt{n}} + 2\sqrt{\frac{2d \log(en/d)}{n}} \right) = \frac{2}{\sqrt{n}} + 4\sqrt{\frac{2d \log(en/d)}{n}}.$$

This proves the expectation bound.

For the high-probability part, observe that changing one sample point z_i changes $\frac{1}{n} \sum_{i=1}^n h(z_i)$ by at most $1/n$ for every $h \in \mathcal{H}$, and therefore changes $\Gamma(S)$ by at most $1/n$ (Lemma D.2). Hence $\Gamma(S)$ satisfies the bounded-difference condition. By McDiarmid's inequality, with probability at least $1 - \delta$,

$$\Gamma(S) \leq \mathbb{E}_S[\Gamma(S)] + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Substituting the expectation bound above proves the claim. $\square$

Lemma D.2 (Bounded difference for uniform deviations). *Let $\mathcal{H}$ be a class of functions $h : \mathcal{Z} \rightarrow [0, 1]$, and let*

$$\Gamma(S) := \sup_{h \in \mathcal{H}} \left| \mathbb{E}[h(z)] - \frac{1}{n} \sum_{i=1}^n h(z_i) \right|,$$

where $S = (z_1, \dots, z_n)$. If S and S' differ in exactly one coordinate, then

$$|\Gamma(S) - \Gamma(S')| \leq \frac{1}{n}.$$

Proof. Suppose S and S' differ only in the j -th coordinate. For every $h \in \mathcal{H}$,

$$\left| \frac{1}{n} \sum_{i=1}^n h(z_i) - \frac{1}{n} \sum_{i=1}^n h(z'_i) \right| = \frac{1}{n} |h(z_j) - h(z'_j)| \leq \frac{1}{n},$$

because $h \in [0, 1]$. Taking the supremum over $h \in \mathcal{H}$ preserves the same bound. $\square$

Lemma D.3 (Counting lemma). *Let $\{h_\theta : \theta \in \Theta\}$ be a class of $\{0, 1\}$ -valued functions on a set $\mathcal{Z}$, parameterized by $\Theta \subseteq \mathbb{R}^D$, and let $\Theta = \Theta_1 \cup \dots \cup \Theta_K$ be a partition that does not depend on the input. Suppose that for every $z \in \mathcal{Z}$ and $k \in [K]$ there are functions $\lambda_{z,k,1}, \dots, \lambda_{z,k,s} : \mathbb{R}^D \rightarrow \mathbb{R}$, each affine in θ , and a Boolean function $B_{z,k}$ such that $h_\theta(z) = B_{z,k}(\text{sgn } \lambda_{z,k,1}(\theta), \dots, \text{sgn } \lambda_{z,k,s}(\theta))$ for all $\theta \in \Theta_k$, where sgn takes values in $\{-, 0, +\}$. If m points are shattered and $sm \geq D$, then $2^m \leq K(8esm/D)^D$. In particular, if $D = d + 1$, $K \leq d$ and $s \leq 3$, then the VC-dimension of the class is less than $11(d + 1)$.*

Proof. Let $z_1, \dots, z_m$ be shattered. Padding with constant functions if necessary, we may assume that exactly s functions are attached to each pair (z_i, k) . Fix k . For $\theta \in \Theta_k$, the labeling $(h_\theta(z_1), \dots, h_\theta(z_m))$ is determined by the sign assignment to the sm affine functions $\{\lambda_{z_i,k,r}\}_{i \in [m], r \in [s]}$ at θ . Since $sm \geq D$, Corollary 2.1 of Goldberg and Jerrum (1993), applied to polynomials of degree one in D variables, bounds the number of consistent sign assignments by $(8esm/D)^D$. Summing over $k \in [K]$, at most $K(8esm/D)^D$ labelings are realizable, and shattering requires 2^m of them.

For the second claim, pad with constant functions so that $s = 3$. It suffices to consider $m \geq d + 1$, so that $3m \geq D$. With logarithms to base 2, the first claim gives

$$m \leq \log d + (d + 1) \log \frac{24em}{d + 1}.$$

Since $\log d \leq d + 1$, the ratio $u := m/(d + 1)$ satisfies $u \leq 1 + \log(24eu)$. The function $u \rightarrow 1 + \log(24eu) - u$ is decreasing for $u \geq 2$ and negative at $u = 11$, so $u < 11$, i.e., $m < 11(d + 1)$. $\square$

Theorem 4.2. *Let the sample and oracle labels be as in Definition 4.1, suppose the induced oracle-loss class $\mathcal{H} := \{h_f : f \in \mathcal{F}\}$, where*

$$h_f(x, y, u) := \mathbb{1}\{f(x^f) \neq \mathbb{1}\{y = 1 \vee u \leq p^f(x)\}\}$$

has finite VC dimension $\tilde{d} \geq 1$. Then for every $\delta \in (0, 1)$ and $n \geq \tilde{d}$, with probability at least $1 - \delta$ over sample $(x_i, y_i, u_i)_{i=1}^n$,

$$\sup_{f \in \mathcal{F}} |\text{err}_{\text{Imp}}(f) - \widehat{\text{err}}_{\text{Imp}}(f)| \leq \varepsilon_n(\delta),$$

$$\text{where } \varepsilon_n(\delta) = \frac{2}{\sqrt{n}} + 4\sqrt{\frac{2\tilde{d} \log(en/\tilde{d})}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Proof. For every $f \in \mathcal{F}$, define the induced oracle-loss function $h_f : \mathcal{X}_0 \times \{0, 1\} \times [0, 1] \rightarrow \{0, 1\}$ by

$$h_f(x, y, u) := \mathbb{1}\{f(x^f) \neq \mathbb{1}\{y = 1 \vee u \leq p^f(x)\}\}.$$

The tie-breaking rule ensures that x^f is well defined, and hence h_f is a well-defined binary function. Let $z = (x, y, u) \sim \mathcal{P}$, where $x \sim \mathcal{D}$, $y \mid x \sim \text{Bernoulli}(\eta(x))$, and $u \sim \text{Unif}[0, 1]$ independently, and let $z_i = (x_i, y_i, u_i)$ for $i \in [n]$. Since $u \sim \text{Unif}[0, 1]$ independently of y , and $\mathbb{P}(y = 1 \mid x) = \eta(x)$, conditioning on x gives

$\mathbb{P}(y = 1 \vee u \leq p^f(x) \mid x) = \eta(x) + (1 - \eta(x))p^f(x) = \eta(x^f)$, by the improvement model of Appendix I. Hence $\mathbb{1}(y = 1 \vee u \leq p^f(x)) \sim \text{Bernoulli}(\eta(x^f))$ conditionally on x . Therefore,

$$\mathbb{E}_{z \sim \mathcal{P}}[h_f(z)] = \mathbb{E}_{x \sim \mathcal{D}}[\mathbb{1}\{f(x^f) \neq y_{x^f}\}] = \text{err}_{\text{Imp}}(f).$$

Moreover, by the Definition 4.1 oracle-label construction $y_i^f := \mathbb{1}\{y_i = 1 \vee u_i \leq p^f(x_i)\}$, we have

$$\frac{1}{n} \sum_{i=1}^n h_f(z_i) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}[f(x_i^f) \neq y_i^f] = \widehat{\text{err}}_{\text{Imp}}(f).$$

Hence the uniform generalization gap in the theorem is exactly

$$\Gamma(S) := \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{z \sim \mathcal{P}}[h_f(z)] - \frac{1}{n} \sum_{i=1}^n h_f(z_i) \right|,$$

where $S = (z_1, \dots, z_n)$. The induced oracle-loss class $H = \{h_f : f \in \mathcal{F}\}$ has VC dimension $\tilde{d}$. Applying Lemma D.1 to this class gives

$$\mathbb{E}_S[\Gamma(S)] \leq \frac{2}{\sqrt{n}} + 4\sqrt{\frac{2\tilde{d} \log(en/\tilde{d})}{n}}.$$

Next, by Lemma D.2, changing one augmented sample point $z_i = (x_i, y_i, u_i)$ changes $\Gamma(S)$ by at most $1/n$. Therefore, McDiarmid's inequality implies that, with probability at least $1 - \delta$,

$$\Gamma(S) \leq \mathbb{E}_S[\Gamma(S)] + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Combining the last two displays yields

$$\sup_{f \in \mathcal{F}} |\text{err}_{\text{Imp}}(f) - \widehat{\text{err}}_{\text{Imp}}(f)| \leq \frac{2}{\sqrt{n}} + 4\sqrt{\frac{2\tilde{d} \log(en/\tilde{d})}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

This proves the theorem. $\square$

Lemma 4.3. *Let $\mathcal{X} \subseteq \mathbb{R}^d$ and let $\mathcal{F}$ be the class of linear classifiers of Section 2. Under Assumptions 1 and 2, the induced oracle-loss class $\mathcal{H}$ of Theorem 4.2 has VC dimension $\tilde{d} < 11(d+1) = \mathcal{O}(d)$.*

Proof. Write $\theta = (w, b)$ and $\eta(x) = g(w^{\star\top} x)$, where $w^{\star} \in \mathbb{R}_{\geq 0}^d$ is the single index parameter of Assumption 1; since $w^{\star}$, η and g are fixed, they enter only as coefficients. Let $g^{-1}(u) := \inf\{t \in \mathbb{R} : g(t) \geq u\}$ denote the generalized inverse of the single-index link function. Partition the parameter set into the regions $\Theta_j := \{\theta : j^{\star} = j\}$, $j \in [d]$, where $j^{\star}$ is the minimal maximizer of w_j/α_j . These are disjoint by the tie-breaking rule and do not depend on the input. Under Assumption 2, on the move branch, $w_{j^{\star}} > 0$ (Section 3).

Fix $j, \theta \in \Theta_j$ and input (x, y, u) . On Θ_j , $S(w) = \beta w_j/\alpha_j$, so by Theorem 3.3, $f(x^f) = \mathbb{1}\{\lambda_2(\theta) \geq 0\}$ with $\lambda_2(\theta) := w^{\top} x - b + \beta w_j/\alpha_j$. The agent moves iff $\lambda_1(\theta) := b - w^{\top} x > 0$ and $\lambda_2(\theta) \geq 0$, in which case $x^f = x + \frac{b-w^{\top}x}{w_j} e_j$; otherwise $x^f = x$.

If $\eta(x) = 1$, then $p^f(x) = 0$ by the convention of Appendix I, so the event $\{u \leq p^f(x)\}$ reduces to $\{u \leq 0\}$, which does not depend upon θ . Assume now that $\eta(x) < 1$. If the agent does not move, then $x^f = x$ and $p^f(x) = 0$, and the event again reduces to $\{u \leq 0\}$. If the agent moves, then $w^{\star\top} x^f \geq w^{\star\top} x$ because $w_j^{\star} \geq 0$, so $p^f(x) \in [0, 1]$. Since $1 - \eta(x) > 0$ and g is non-decreasing and continuous (Assumption 1), $u \leq p^f(x) \iff g(w^{\star\top} x^f) \geq u(1 - \eta(x)) + \eta(x) \iff \kappa(x, u) \leq w^{\star\top} x^f$, where $\kappa(x, u) := g^{-1}(u(1 - \eta(x)) + \eta(x))$. If $\kappa(x, u) = +\infty$ (resp. $-\infty$), this condition is false (resp. true) for every θ . Otherwise, substituting $w^{\star\top} x^f = w^{\star\top} x + w_j^{\star}(b - w^{\top} x)/w_j$ and multiplying by $w_j > 0$, it is equivalent to $\lambda_3(\theta) \geq 0$, with $\lambda_3(\theta) := w_j^{\star}(b - w^{\top} x) - w_j(\kappa(x, u) - w^{\star\top} x)$.

Consequently, on Θ_j , both $f(x^f)$ and the oracle label $\mathbb{1}\{y = 1 \vee u \leq p^f(x)\}$, and hence $h_f(x, y, u)$ are fixed Boolean functions of the signs of $\lambda_1(\theta), \lambda_2(\theta), \lambda_3(\theta)$ and of the input-only quantities y and $\mathbb{1}\{u \leq 0\}$. Each λ_i is affine in $\theta \in \mathbb{R}^{d+1}$, with coefficients depending only on (x, u) . Lemma D.3, applied with $D = d + 1$, $K = d$ and $s = 3$, gives $\text{VC}(\mathcal{H}) < 11(d + 1)$. $\square$

E GENERALIZATION BOUNDS FOR STRAT-IMP-AWARE

In this section we prove Theorem 5.3 for the plug-in version of STRAT-IMP-AWARE, in which the unknown class-probability function η is replaced by a plug-in estimator $\hat{\eta}$: each agent’s improvement probability $\hat{\pi}_{\text{Imp}}$ under the deployed model is computed from $\hat{\eta}$ by Algorithm 2, and the loss of Algorithm 3 is minimized. Throughout, let $r := \sup_{z \in \mathcal{A}} \|z\|_2 < \infty$, $\rho_{\text{Imp}} := r + \beta / \min_{j \in [d]} \alpha_j$, $\varepsilon_\eta := \sup_{z \in \mathcal{A}} |\hat{\eta}(z) - \eta(z)|$, and $\mathcal{W}_{k,B} := \{(w, b) : w \in \mathbb{R}_{\geq 0}^d, \|w\|_2 \leq k, |b| \leq B\}$; for $(w, b) \in \mathcal{W}_{k,B}$ define $S(w) := \beta \max_{j \in [d]} w_j / \alpha_j$, $M_{w,b}(x) := w^\top x - b + S(w)$, and $C_{k,B} := k\rho_{\text{Imp}} + B$. We take the radius over the feasible report set $\mathcal{A}$ rather than over $\mathcal{X}_0$ so that the same constant serves the estimation analysis of Appendix F, where $\hat{\eta}$ and η are compared at reported points; since $\mathcal{X}_0 \subseteq \mathcal{A}$, it follows that $|M_{w,b}(x)| \leq |w^\top x| + |b| + S(w) \leq C_{k,B}$ for every $x \in \mathcal{X}_0$. Here ε_η is treated as a given tolerance; a bound of order $\mathcal{O}(m^{-1/2})$ in the pre-deployment sample size m is established in Appendix F.

Improvement probability. For $x \in \mathcal{X}_0$, let x^f denote the selected best response to $f_{w,b}$, *i.e.*, the maximizer of $\beta \mathbb{1}(w^\top z \geq b) - \sum_j \alpha_j (z_j - x_j)_+$ over $z \in \mathcal{A}$ selected by the canonical tie-breaking rule of Theorem 3.3: $x^f = x + \frac{b - w^\top x}{w_{j^*}} e_{j^*}$ with $j^* = \min \arg \max_{j \in [d]} w_j / \alpha_j$ if $0 < b - w^\top x \leq S(w)$, and $x^f = x$ otherwise. Algorithm 2 computes the improvement probability

$$\hat{\pi}_{\text{Imp}}(x; w, b) = \text{clamp}_{[0,1]} \left(\frac{\hat{\eta}(x^f) - \hat{\eta}(x)}{\max\{1 - \hat{\eta}(x), \delta_g\}} \right),$$

the guard $\delta_g > 0$ prevents division by a vanishing denominator; we take $\delta_g \leq 1 - \sup_{z \in \mathcal{A}} \hat{\eta}(z)$, which is always possible since $\hat{\eta}(z) = g(\hat{w}^\top z) \leq g(rB_\eta) < 1$ on $\mathcal{A}$ (see Appendix F.1 for details). Under this choice the guard is never active, so it is a safeguard for finite-precision arithmetic rather than an analytical assumption. Algorithm 3 combines $\hat{\pi}_{\text{Imp}}$ with the agent’s observed pre-manipulation label y into

$$q_{w,b}(x, y) := y + (1 - y)\hat{\pi}_{\text{Imp}}(x; w, b) \in [0, 1],$$

the probability that the post-response label is positive. Equivalently, $q_{w,b}(x, y) = \mathbb{P}(y_{\text{Imp}} = 1 \mid x, y)$ for the improvement-aware label $y_{\text{Imp}} := \mathbb{1}(\{y = 1\} \vee \{u \leq \hat{\pi}_{\text{Imp}}(x; w, b)\})$ with $u \sim \text{Unif}[0, 1]$ independent of (x, y) . This construction enforces $\mathbb{P}(y_{\text{Imp}} = 1 \mid y = 1, x) = 1$ exactly and assigns $\mathbb{P}(y_{\text{Imp}} = 1 \mid y = 0, x) = \hat{\pi}_{\text{Imp}}(x; w, b)$, mirroring the improvement model of Appendix I, in which the true post-response label satisfies $\mathbb{P}(y^f = 1 \mid y = 1, x) = 1$ and $\mathbb{P}(y^f = 1 \mid y = 0, x) = p^f(x)$ with $p^f(x) = (\eta(x^f) - \eta(x)) / (1 - \eta(x))$. The label y_{Imp} serves only the analysis; the algorithm never samples it. Since $y \mid x \sim \text{Bernoulli}(\eta(x))$,

$$\mathbb{E}[q_{w,b}(x, y) \mid x] = \hat{q}_{w,b}(x) := \eta(x) + (1 - \eta(x))\hat{\pi}_{\text{Imp}}(x; w, b),$$

the estimator thus entering only through the conditional improvement probability for agents with $y = 0$. Lemma E.1 shows that $\hat{q}_{w,b}(x)$ deviates from the target $\eta(x^f)$ by at most $2\varepsilon_\eta$.

Plug-in strategic improvement loss. Writing $z := (x, y)$, we define the conditional loss and the per-point loss of Algorithm 3, a hinge surrogate of the improvement-aware error,

$$\begin{aligned} \ell_{w,b}^{\hat{q}}(x) &:= \hat{q}_{w,b}(x)(1 - M_{w,b}(x))_+ + (1 - \hat{q}_{w,b}(x))(1 + M_{w,b}(x))_+, \\ \ell_{w,b}(z) &:= q_{w,b}(x, y)(1 - M_{w,b}(x))_+ + (1 - q_{w,b}(x, y))(1 + M_{w,b}(x))_+, \end{aligned}$$

together with the empirical plug-in strategic improvement loss

$$\hat{L}_{\text{hinge}}^{\hat{q}}(w, b) := \frac{1}{n} \sum_{i=1}^n \ell_{w,b}(z_i),$$

where $z_i = (x_i, y_i)$ are i.i.d. with $x_i \sim \mathcal{D}$ and $y_i \mid x_i \sim \text{Bernoulli}(\eta(x_i))$. Since $\ell_{w,b}(z)$ is affine in $q_{w,b}(x, y)$ and $\mathbb{E}[q_{w,b}(x, y) \mid x] = \hat{q}_{w,b}(x)$, we have $\mathbb{E}[\ell_{w,b}(z) \mid x] = \ell_{w,b}^{\hat{q}}(x)$, so $\hat{L}_{\text{hinge}}^{\hat{q}}(w, b)$ is an unbiased estimate of $L_{\text{hinge}}^{\hat{q}}(w, b) := \mathbb{E}_{x \sim \mathcal{D}}[\ell_{w,b}^{\hat{q}}(x)]$. In addition, $\ell_{w,b}(z)$ is the expectation over u of the hinge loss evaluated at the label y_{Imp} , so weighting by $q_{w,b}$ rather than sampling y_{Imp} removes the sampling noise in u without changing the population risk. Moreover, $q_{w,b}(x, y) \in [0, 1]$ makes $\ell_{w,b}(z)$ a convex combination of the two hinges, so $0 \leq \ell_{w,b}(z) \leq 1 + |M_{w,b}(x)| \leq 1 + C_{k,B}$. Two features of this construction are used below: $\hat{L}_{\text{hinge}}^{\hat{q}}$ is the objective

minimized by STRAT-IMP-AWARE and is computable, whereas $\ell_{w,b}^{\hat{q}}$ involves the unknown η and is not; since $\hat{L}_{\text{hinge}}^{\hat{q}}$ involves no auxiliary randomization, it is a deterministic function of (w, b) given the sample, as the uniform convergence argument requires.

Lemma E.1 (Induced label deviation). *Let $(w, b) \in \mathcal{W}_{k,B}$ and $x \in \mathcal{X}_0$. Then*

$$|\hat{q}_{w,b}(x) - \eta(x^f)| \leq 2\varepsilon_\eta$$

Proof. By the choice of δ_g , the denominator of $\hat{\pi}_{\text{Imp}}(x; w, b)$ equals $1 - \hat{\eta}(x)$. We can write $\hat{\pi}_{\text{Imp}}(x; w, b) = \text{clamp}_{[0,1]}(\tilde{\pi})$ with $\tilde{\pi} := (\hat{\eta}(x^f) - \hat{\eta}(x))/(1 - \hat{\eta}(x))$; under Assumption F.1, $\hat{\eta}(x) \leq \hat{\eta}(x^f) < 1$, so $\tilde{\pi} \in [0, 1]$ and the clamp is inactive. Decomposing over the observed label, the improvement model of Appendix I gives $\eta(x^f) = \eta(x) + (1 - \eta(x))p^f(x)$ with true improvement probability $p^f(x) = (\eta(x^f) - \eta(x))/(1 - \eta(x)) \in [0, 1]$; thus $\hat{\pi}_{\text{Imp}}$ is the plug-in estimate of $p^f(x)$. Using $\hat{q}_{w,b}(x) = \eta(x) + (1 - \eta(x))\tilde{\pi}$ and $(1 - \hat{\eta}(x))\tilde{\pi} = \hat{\eta}(x^f) - \hat{\eta}(x)$,

$$\begin{aligned} \hat{q}_{w,b}(x) - \eta(x^f) &= \underbrace{\hat{\eta}(x) + (1 - \hat{\eta}(x))\tilde{\pi}}_{=\hat{\eta}(x^f)} - \eta(x^f) - (\hat{\eta}(x) - \eta(x))(1 - \tilde{\pi}) \\ &= [\hat{\eta}(x^f) - \eta(x^f)] - [\hat{\eta}(x) - \eta(x)](1 - \tilde{\pi}). \end{aligned}$$

Both x and x^f lie in $\mathcal{A}$, so each bracketed term is at most ε_η in absolute value, and $|1 - \tilde{\pi}| \leq 1$. The triangle inequality gives the claim. $\square$

Lemma 5.2. *Let $\mathfrak{d} := \text{Pdim}(\Pi_{k,B}) + \text{Pdim}(\mathcal{M}_{k,B})$, where Pdim denotes pseudo-dimension Anthony and Bartlett (2009) and $\Pi_{k,B}, \mathcal{M}_{k,B}$ be the classes of maps $x \mapsto \hat{\pi}_{\text{Imp}}(x; w, b)$ and $x \mapsto M_{w,b}(x)$ over $\mathcal{W}_{k,B}$. Then $\mathfrak{d} < 12(d+1)$, and the loss class of Algorithm 3 over $\mathcal{W}_{k,B}$ has the covering numbers of a class of pseudo-dimension $\mathfrak{d}$, up to absolute constants.*

Proof. Let $\mathcal{G}_{k,B} := \{\ell_{w,b} : (w, b) \in \mathcal{W}_{k,B}\}$, regarded as a class of real-valued functions on $\mathcal{Z} := \mathcal{X}_0 \times \{0, 1\}$. Condition on the estimation sample S_η . Since $\hat{\eta}$ is estimated independently of the training sample, $\mathcal{G}_{k,B}$ is then a fixed class of measurable functions on $\mathcal{Z}$, and the points $z_1, \dots, z_n$ remain i.i.d. Furthermore, each $\ell_{w,b}(z_i)$ is a deterministic function of (w, b) , since the loss involves no auxiliary randomization. All bounds below hold for every realization of $\hat{\eta}$ and therefore unconditionally. Write $R := 1 + C_{k,B}$ so that $0 \leq g \leq R$ for every $g \in \mathcal{G}_{k,B}$ by the setup.

Pseudo-dimension. By definition (Anthony and Bartlett, 2009), the pseudo-dimension of a class $\mathcal{F}$ of real-valued functions is the VC-dimension of $\{(x, \tau) \mapsto \text{sgn}(f(x) - \tau) : f \in \mathcal{F}\}$, with the threshold $\tau \in \mathbb{R}$ part of the input. For $\mathcal{M}_{k,B}$: since $S(w)$ does not depend on x , $M_{w,b}(x) = w^\top x - (b - S(w))$, so $\mathcal{M}_{k,B}$ is contained in the $(d+1)$ -dimensional vector space of affine functions on $\mathbb{R}^d$, whose pseudo-dimension is $d+1$ using Theorem 11.4 from Anthony and Bartlett (2009). Hence $\text{Pdim}(\mathcal{M}_{k,B}) \leq d+1$.

For $\Pi_{k,B}$: Under decomposable costs the selected best response is $x^f = x + \Delta e_{j^*}$, where $j^* \in \arg \max_{j \in [d]} w_j/\alpha_j$ with the stated tie-break and $\Delta = (b - w^\top x)/w_{j^*}$, when the agent moves and $x^f = x$ otherwise. The division is well defined as the move branch requires $w^\top x < b$ and $M_{w,b}(x) \geq 0$, which are incompatible at $w = 0$ since $M_{0,b}(x) = -b$, and $w \neq 0$ with $w \in \mathbb{R}_{\geq 0}^d$ gives $\max_{j \in [d]} w_j/\alpha_j > 0$, hence $w_{j^*} > 0$. Partition $\mathcal{W}_{k,B}$ into the regions $C_j := \{(w, b) \in \mathcal{W}_{k,B} : j^* = j\}$, $j \in [d]$; by the tie-breaking rule these are disjoint and defined by linear inequalities in w that do not involve the input. As g is non-decreasing and continuous (Assumption 1), and both the guard and the clamp are inactive (proof of Lemma E.1), the test $\hat{\pi}_{\text{Imp}} \geq \tau$ is equivalent to

$$\kappa(x, \tau) \leq \hat{w}^\top x^f, \quad \kappa(x, \tau) := g^{-1}(\tau(1 - \hat{\eta}(x)) + \hat{\eta}(x)),$$

where $g^{-1}(v) := \inf\{s : g(s) \geq v\}$; conditioning on S_η fixes $\hat{\eta}$ and $\hat{w}$, so $\kappa(x, \tau)$ and $\hat{w}^\top x$ are functions of the input alone. Fix j and $\theta = (w, b) \in C_j$. The agent moves iff $\ell_1(\theta) := b - w^\top x > 0$ and $\ell_2(\theta) := \beta w_j/\alpha_j - b + w^\top x \geq 0$, in which case $\hat{w}^\top x^f = \hat{w}^\top x + \hat{w}_j(b - w^\top x)/w_j$ and, multiplying by $w_j > 0$, the test becomes $\ell_3(\theta) := \hat{w}_j(b - w^\top x) - w_j(\kappa(x, \tau) - \hat{w}^\top x) \geq 0$; if $\kappa(x, \tau) = \infty$ (resp. $-\infty$), the test fails (resp. holds) for every θ , and we set $\ell_3 \equiv -1$ (resp. $\ell_3 \equiv 1$). Otherwise $x^f = x$ and the test reduces to $\tau \leq 0$. Each ℓ_i is affine in $\theta \in \mathbb{R}^{d+1}$ with coefficients depending only on (x, τ) , so on C_j the test is a fixed Boolean function of the signs of three affine functions of θ . Lemma D.3, applied with $D = d+1$, $K = d$ and $s = 3$, therefore gives $\text{Pdim}(\Pi_{k,B}) < 11(d+1)$, and together with $\text{Pdim}(\mathcal{M}_{k,B}) \leq d+1$, yields $\mathfrak{d} = \text{Pdim}(\Pi_{k,B}) + \text{Pdim}(\mathcal{M}_{k,B}) < 12(d+1) = \mathcal{O}(d)$.

Covering numbers. For a function f defined on $\mathcal{Z}$, define the empirical norms

$$\|f\|_{L_p(P_n)} := \left(\frac{1}{n} \sum_{i=1}^n |f(z_i)|^p \right)^{1/p}, \quad p \in \{1, 2\}$$

A function f defined on $\mathcal{X}_0$ is identified with a function on $\mathcal{Z}$ by setting $f(z_i) := f(x_i)$, so that all norms below are evaluated at the same sample points $z_1, \dots, z_n$. All covering numbers are internal, *i.e.*, cover elements belong to the class being covered, and all logarithms are natural. Write $\theta = (w, b)$, $\hat{\pi}_\theta := \hat{\pi}_{\text{Imp}}(\cdot; w, b)$, $M_\theta := M_{w,b}$, $q_\theta := q_{w,b}$ and $\ell_\theta := \ell_{w,b}$ and abbreviate $\|\cdot\|_{L_2(P_n)}$ by $\|\cdot\|_n$. We bound $\mathcal{N}(\epsilon, \mathcal{G}_{k,B}, L_2(P_n))$ for $0 < \epsilon \leq R$ below in three steps.

1. Lipschitz decomposition: Let $\psi(q, m) := q(1-m)_+ + (1-q)(1+m)_+$ so that $\ell_\theta(z) = \psi(q_\theta(x, y), M_\theta(x))$ with $q_\theta(x, y) = y + (1-y)\hat{\pi}_\theta(x)$. Since $|(1-m)_+ - (1+m)_+| \leq 2|m|$ for all $m \in \mathbb{R}$, the map $q \mapsto \psi(q, m)$ is $2R$ -Lipschitz as $|m| \leq C_{k,B} \leq R$. Moreover, for every $q \in [0, 1]$ the map $m \mapsto \psi(q, m)$ is 1-Lipschitz, being the sum of a q -Lipschitz and a $(1-q)$ -Lipschitz function. Fix $\theta, \theta' \in \mathcal{W}_{k,B}$ and apply the triangle inequality

$$|\psi(q, m) - \psi(q', m')| \leq |\psi(q, m) - \psi(q', m)| + |\psi(q', m) - \psi(q', m')|$$

with $(q, m) = (q_\theta(x, y), M_\theta(x))$ and $(q', m') = (q_{\theta'}(x, y), M_{\theta'}(x))$; the first term uses $|M_\theta(x)| \leq C_{k,B}$ and the second uses $q_{\theta'}(x, y) \in [0, 1]$. Together with $|q_\theta(x, y) - q_{\theta'}(x, y)| = (1-y)|\hat{\pi}_\theta(x) - \hat{\pi}_{\theta'}(x)| \leq |\hat{\pi}_\theta(x) - \hat{\pi}_{\theta'}(x)|$, this gives, pointwise, $|\ell_\theta(z) - \ell_{\theta'}(z)| \leq 2R|\hat{\pi}_\theta(x) - \hat{\pi}_{\theta'}(x)| + |M_\theta(x) - M_{\theta'}(x)|$ and Minkowski's inequality yields

$$\|\ell_\theta - \ell_{\theta'}\|_n \leq 2R\|\hat{\pi}_\theta - \hat{\pi}_{\theta'}\|_n + \|M_\theta - M_{\theta'}\|_n \quad (5)$$

2. Product cover: Take minimal covers $\{\pi_s\}_{s=1}^{N_\Pi}$ of $\Pi_{k,B}$ at scale $\epsilon/(8R)$ and $\{\mu_t\}_{t=1}^{N_\mathcal{M}}$ of $\mathcal{M}_{k,B}$ at scale $\epsilon/4$ in $\|\cdot\|_n$. By the cover property, each $\theta \in \mathcal{W}_{k,B}$ can be assigned to a pair (s, t) with $\|\hat{\pi}_\theta - \pi_s\|_n \leq \epsilon/(8R)$, $\|M_\theta - \mu_t\|_n \leq \epsilon/4$. For each occupied pair, select one representative $\theta_{s,t} \in \mathcal{W}_{k,B}$ assigned to it. If θ is assigned to (s, t) , the triangle inequality gives

$$\|\hat{\pi}_\theta - \hat{\pi}_{\theta_{s,t}}\|_n \leq \|\hat{\pi}_\theta - \pi_s\|_n + \|\pi_s - \hat{\pi}_{\theta_{s,t}}\|_n \leq \frac{\epsilon}{4R}, \quad \|M_\theta - M_{\theta_{s,t}}\|_n \leq \|M_\theta - \mu_t\|_n + \|\mu_t - M_{\theta_{s,t}}\|_n \leq \frac{\epsilon}{2},$$

and hence, by (5), $\|\ell_\theta - \ell_{\theta_{s,t}}\|_n \leq 2R\epsilon/(4R) + \epsilon/2 = \epsilon$. Thus the representative losses $\{\ell_{\theta_{s,t}}\}$ form an internal ϵ -cover of $\mathcal{G}_{k,B}$ with at most $N_\Pi N_\mathcal{M}$ elements, since there are at most $N_\Pi N_\mathcal{M}$ occupied index pairs (s, t) , proving

$$\mathcal{N}(\epsilon, \mathcal{G}_{k,B}, L_2(P_n)) \leq \mathcal{N}\left(\frac{\epsilon}{8R}, \Pi_{k,B}, L_2(P_n)\right) \mathcal{N}\left(\frac{\epsilon}{4}, \mathcal{M}_{k,B}, L_2(P_n)\right) \quad (6)$$

3. Bounding the components: Let $\mathcal{F}$ be a class of functions taking values in the interval of length $A > 0$, with pseudo-dimension $d_\mathcal{F}$. Rescaling $\mathcal{F}$ to $[0, 1]$ leaves its pseudo-dimension unchanged and scales distances by $1/A$, so applying Theorem 18.4 of Anthony and Bartlett (2009) which bounds the u -packing number and hence the internal u -covering number (a maximal u -packing is an internal u -cover), gives $\mathcal{N}(u, \mathcal{F}, L_1(P_n)) \leq e(d_\mathcal{F} + 1)\left(\frac{2eA}{u}\right)^{d_\mathcal{F}}$, where $0 < u \leq A$. Since $|f - g| \leq A$ pointwise for $f, g \in \mathcal{F}$, we have $\|f - g\|_n^2 \leq A\|f - g\|_{L_1(P_n)}$, so an internal (δ^2/A) -cover in $L_1(P_n)$ is an internal δ -cover in $L_2(P_n)$. Consequently, for $0 < \delta \leq A$,

$$\log \mathcal{N}(\delta, \mathcal{F}, L_2(P_n)) \leq \log(e(d_\mathcal{F} + 1)) + 2d_\mathcal{F} \log \left(\frac{A\sqrt{2e}}{\delta} \right). \quad (7)$$

Apply (7) to $\Pi_{k,B}$ which takes values in $[0, 1]$, with $(A, \delta) = (1, \epsilon/(8R))$, and to $\mathcal{M}_{k,B}$, which takes values in $[-C_{k,B}, C_{k,B}] \subseteq [-R, R]$, with $(A, \delta) = (2R, \epsilon/4)$; both choices satisfy $\delta \leq A$ since $\epsilon \leq R$, and in both cases the argument of the logarithm equals $8R\sqrt{2e}/\epsilon$. Writing $d_\Pi := \text{Pdim}(\Pi_{k,B})$ and $d_\mathcal{M} := \text{Pdim}(\mathcal{M}_{k,B})$, so that $\mathfrak{d} = d_\Pi + d_\mathcal{M}$, taking logarithms in (6) and using $d_\Pi, d_\mathcal{M} \leq \mathfrak{d}$ in the constant terms gives

$$\log \mathcal{N}(\epsilon, \mathcal{G}_{k,B}, L_2(P_n)) \leq 2 \log(e(\mathfrak{d} + 1)) + 2\mathfrak{d} \log \left(\frac{8R\sqrt{2e}}{\epsilon} \right). \quad (8)$$

Thus $\mathfrak{d}$ plays the role of a pseudo-dimension for $\mathcal{G}_{k,B}$: it controls the coefficient of $\log(1/\epsilon)$ in its log-covering number. $\square$

Lemma E.2 (Uniform convergence for plug-in strategic improvement loss). *Let $\hat{\eta}$ be estimated independently of the training sample, and let $\mathfrak{d}$ be as in Lemma 5.2. Then, for every $\delta \in (0, 1)$, with probability at least $1 - \delta$, uniformly over $(w, b) \in \mathcal{W}_{k,B}$,*

$$L_{\text{hinge}}^{\hat{q}}(w, b) \leq \hat{L}_{\text{hinge}}^{\hat{q}}(w, b) + c(1 + C_{k,B})\sqrt{\frac{\mathfrak{d}}{n}} + (1 + C_{k,B})\sqrt{\frac{\log(1/\delta)}{2n}}$$

where $c > 0$ is an absolute constant.

Proof. We work in the setting of the proof of Lemma 5.2: conditionally on S_{η} , $\mathcal{G}_{k,B}$ is a fixed class of measurable functions on $\mathcal{Z}$, the points $z_1, \dots, z_n$ are i.i.d., and every $g \in \mathcal{G}_{k,B}$ satisfies $0 \leq g \leq R$, with $R = 1 + C_{k,B}$; all bounds below hold for every realization of $\hat{\eta}$, and therefore unconditionally. By (8) in that proof, for $0 < \epsilon \leq R$,

$$\log \mathcal{N}(\epsilon, \mathcal{G}_{k,B}, L_2(P_n)) \leq 2 \log(e(\mathfrak{d} + 1)) + 2\mathfrak{d} \log \left(\frac{8R\sqrt{2e}}{\epsilon} \right).$$

Rademacher complexity. By Dudley's entropy integral at $\alpha = 0$ (Lemma A.3 of Srebro et al. (2010)), using $\sup_{g \in \mathcal{G}_{k,B}} \sqrt{\mathbb{E}[g^2]} \leq R$ for its upper limit,

$$\hat{\mathfrak{R}}_n(\mathcal{G}_{k,B}) \leq C_0 \int_0^R \sqrt{\frac{\log \mathcal{N}(\epsilon, \mathcal{G}_{k,B}, L_2(P_n))}{n}} d\epsilon$$

for an absolute constant $C_0 > 0$. Substituting $\log \mathcal{N}(\epsilon, \mathcal{G}_{k,B}, L_2(P_n))$ from (8), rescaling $\epsilon = Rs$, which cancels R inside the logarithm, and using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ twice together with $\int_0^1 \sqrt{\log(1/s)} ds = \sqrt{\pi}/2$,

$$\int_0^R \sqrt{\log \mathcal{N}(\epsilon, \mathcal{G}_{k,B}, L_2(P_n))} d\epsilon \leq R \left(\sqrt{2 \log(e(\mathfrak{d} + 1))} + \sqrt{2\mathfrak{d}} \left(\sqrt{\log(8\sqrt{2e})} + \frac{\sqrt{\pi}}{2} \right) \right) \leq c_1 R \sqrt{\mathfrak{d}},$$

where the last step uses $\log(e(\mathfrak{d} + 1)) \leq 1 + \mathfrak{d} \leq 2\mathfrak{d}$ (as $\mathfrak{d} \geq 1$) and $c_1 > 0$ is an absolute constant. Hence, with $c := 2C_0 c_1 > 0$, gives

$$\hat{\mathfrak{R}}_n(\mathcal{G}_{k,B}) \leq \frac{c}{2} R \sqrt{\frac{\mathfrak{d}}{n}}$$

This bound holds for every sample, hence also bounds $\mathfrak{R}_n(\mathcal{G}_{k,B}) := \mathbb{E}[\hat{\mathfrak{R}}_n(\mathcal{G}_{k,B})] \leq \frac{c}{2} R \sqrt{\frac{\mathfrak{d}}{n}}$.

Uniform convergence. Replacing one z_i alters $\sup_{g \in \mathcal{G}_{k,B}} (\mathbb{E}[g(z)] - \frac{1}{n} \sum_{i \in [n]} g(z_i))$ by at most R/n , so symmetrization together with McDiarmid's inequality gives, with probability at least $1 - \delta$, uniformly over $g \in \mathcal{G}_{k,B}$,

$$\mathbb{E}[g(z)] \leq \frac{1}{n} \sum_{i=1}^n g(z_i) + 2\mathfrak{R}_n(\mathcal{G}_{k,B}) + R \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Applying this to $g = \ell_{w,b}$: the empirical average is $\hat{L}_{\text{hinge}}^{\hat{q}}(w, b)$ by definition, while unbiasedness and the tower property give $\mathbb{E}_z[\ell_{w,b}(z)] = \mathbb{E}_{x \sim \mathcal{D}}[\ell_{w,b}^{\hat{q}}(x)] = L_{\text{hinge}}^{\hat{q}}(w, b)$. Substituting $R = 1 + C_{k,B}$ completes the proof. $\square$

Theorem 5.3. *Let $(\tilde{w}, \tilde{b}) \in \mathcal{W}_{k,B}$ be the model returned by STRAT-IMP-AWARE. Under Assumptions 1 and 2, for every $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\begin{aligned} \text{err}_{\text{Imp}}(f_{\tilde{w}, \tilde{b}}) &\leq \hat{L}_{\text{hinge}}^{\hat{q}}(\tilde{w}, \tilde{b}) + c(1 + C_{k,B})\sqrt{\mathfrak{d}/n} \\ &\quad + (1 + C_{k,B})\sqrt{\frac{\log(1/\delta)}{2n}} + 4C_{k,B}\varepsilon_{\eta}, \end{aligned}$$

where $\mathfrak{d}$ is as in Lemma 5.2 and $c > 0$ is an absolute constant.

Proof of Theorem 5.3. Recall $\mathcal{W}_{k,B} := \{(w, b) : w \in \mathbb{R}_{\geq 0}^d, \|w\|_2 \leq k, |b| \leq B\}$ and $C_{k,B} := k\rho_{\text{Imp}} + B$. Fix $(w, b) \in \mathcal{W}_{k,B}$, and let $x^f \in \mathcal{A}$ be the selected best response of $x \in \mathcal{X}_0$ to $f_{w,b}$. Write $S(w) := \beta \max_{j \in [d]} w_j / \alpha_j$ and $M_{w,b}(x) := w^\top x - b + S(w)$. By the strategic-shift characterization under decomposable costs (Theorem 3.3), the post-response prediction satisfies $f_{w,b}(x^f) = \mathbb{1}\{M_{w,b}(x) \geq 0\}$.

Under the improvement-aware model, the true post-response label satisfies $\mathbb{P}(y^f = 1 \mid x) = \eta(x^f)$, so the conditional improvement-aware 0/1 loss at x is $\mathbb{E}[\mathbb{1}\{f_{w,b}(x^f) \neq y^f\} \mid x] = \eta(x^f)\mathbb{1}\{M_{w,b}(x) < 0\} + (1 - \eta(x^f))\mathbb{1}\{M_{w,b}(x) \geq 0\}$. Since $\mathbb{1}\{M < 0\} \leq (1 - M)_+$ and $\mathbb{1}\{M \geq 0\} \leq (1 + M)_+$, taking expectations over $x \sim \mathcal{D}$ gives,

$$\text{err}_{\text{Imp}}(f_{w,b}) \leq L_{\text{hinge}}^\eta(w, b) \quad (9)$$

where $L_{\text{hinge}}^\eta(w, b) := \mathbb{E}_{x \sim \mathcal{D}}[\eta(x^f)(1 - M_{w,b}(x))_+ + (1 - \eta(x^f))(1 + M_{w,b}(x))_+]$. Recall the plug-in strategic improvement risk

$$L_{\text{hinge}}^{\hat{q}}(w, b) = \mathbb{E}_{x \sim \mathcal{D}}[\hat{q}_{w,b}(x)(1 - M_{w,b}(x))_+ + (1 - \hat{q}_{w,b}(x))(1 + M_{w,b}(x))_+].$$

The two risks differ only in the weight attached to the two hinges, so

$$\left| L_{\text{hinge}}^\eta(w, b) - L_{\text{hinge}}^{\hat{q}}(w, b) \right| \leq \mathbb{E}_{x \sim \mathcal{D}}[|\eta(x^f) - \hat{q}_{w,b}(x)| |(1 - M_{w,b}(x))_+ - (1 + M_{w,b}(x))_+|].$$

By Lemma E.1, $|\eta(x^f) - \hat{q}_{w,b}(x)| \leq 2\varepsilon_\eta$. Moreover, $|(1 - m)_+ - (1 + m)_+| \leq 2|m|$ for all $m \in \mathbb{R}$. Finally, for all $x \in \mathcal{X}_0 \subseteq \mathcal{A}$,

$$|M_{w,b}(x)| \leq |w^\top x| + |b| + S(w) \leq kr + B + \frac{\beta k}{\min_{j \in [d]} \alpha_j} = k\rho_{\text{Imp}} + B = C_{k,B}.$$

Hence

$$L_{\text{hinge}}^\eta(w, b) \leq L_{\text{hinge}}^{\hat{q}}(w, b) + 4C_{k,B}\varepsilon_\eta. \quad (10)$$

By Lemma 5.2 and Lemma E.2, with probability at least $1 - \delta$, uniformly over $(w, b) \in \mathcal{W}_{k,B}$,

$$L_{\text{hinge}}^{\hat{q}}(w, b) \leq \hat{L}_{\text{hinge}}^{\hat{q}}(w, b) + c(1 + C_{k,B})\sqrt{\frac{\delta}{n}} + (1 + C_{k,B})\sqrt{\frac{\log(1/\delta)}{2n}}. \quad (11)$$

Combining (9), (10) and (11), with probability at least $1 - \delta$, uniformly over $(w, b) \in \mathcal{W}_{k,B}$ gives,

$$\text{err}_{\text{Imp}}(f_{w,b}) \leq \hat{L}_{\text{hinge}}^{\hat{q}}(w, b) + c(1 + C_{k,B})\sqrt{\frac{\delta}{n}} + (1 + C_{k,B})\sqrt{\frac{\log(1/\delta)}{2n}} + 4C_{k,B}\varepsilon_\eta.$$

Since STRAT-IMP-AWARE returns $(\tilde{w}, \tilde{b}) \in \mathcal{W}_{k,B}$, and the above holds uniformly on $\mathcal{W}_{k,B}$, it applies to the returned model irrespective of how it was selected; substituting $(w, b) = (\tilde{w}, \tilde{b})$ proves the theorem. $\square$

Lemma E.3 (Gradient bound). *Let $\alpha_{\min} := \min_{j \in [d]} \alpha_j$ and $\hat{\eta} := g(\hat{w}^\top \cdot)$. Under Assumption F.1, for every (w, b) with $w \in \mathbb{R}_{\geq 0}^d \setminus \{0\}$ and every $z = (x, y)$ with $x \in \mathcal{X}_0$, every (sub)gradient of $\ell_{w,b}(z)$, computed with j^* and the manipulation status held fixed, satisfies $\|\nabla_w \ell_{w,b}(z)\|_2 \leq G_w$ and $|\partial_b \ell_{w,b}(z)| \leq G_b$, where*

$$G_w := r + \frac{\beta}{\alpha_{\min}} + \frac{2\beta L_g B_\eta r}{\alpha_{\min} \delta_g}, \quad G_b := 1 + \frac{2\beta L_g B_\eta}{\alpha_{\min} \delta_g}.$$

If g is the logistic function, the factor L_g/δ_g can be replaced by 1.

Proof. Write $\ell_{w,b}(z) = \Phi(q, M)$ with $q = q_{w,b}(x)$, $M = M_{w,b}(x)$ and $\Phi(q, M) := q(1 - M)_+ + (1 - q)(1 + M)_+$, so that $\nabla \ell = (\partial \ell / \partial M) \nabla M + (\partial \ell / \partial q) \nabla q$. Every subgradient satisfies $\partial \ell / \partial M \in [-q, 1 - q]$, hence $|\partial \ell / \partial M| \leq 1$; moreover $\nabla_w M = x + \beta s$ for a subgradient s of $w \rightarrow \max_j w_j / \alpha_j$, with $\|s\|_2 \leq 1/\alpha_{\min}$, and $\partial_b M = -1$. Since $\|x\| \leq r$, the first term contributes at most $r + \beta/\alpha_{\min}$ to $\|\nabla_w \ell\|_2$ and at most 1 to $|\partial_b \ell|$. The second term vanishes unless $y = 0$, the agent manipulates, and the clamp is inactive, since otherwise q is locally constant. In that case $0 \leq M < S(w) = \beta w_{j^*} / \alpha_{j^*}$, so $|\partial \ell / \partial q| = |(1 - M)_+ - (1 + M)_+| \leq 2M \leq 2\beta w_{j^*} / \alpha_{j^*}$. Further, $\nabla q = \frac{g'(\hat{w}^\top x^f) \hat{w}_{j^*}}{\max\{1 - \hat{\eta}(x), \delta_g\}} \nabla x_{j^*}^f$ with $\nabla_w x_{j^*}^f = -x^f / w_{j^*}$ and $\partial_b x_{j^*}^f = 1 / w_{j^*}$. From Assumption F.1, $g' \leq L_g$, $\hat{w}_{j^*} \leq B_\eta$, $\|x^f\|_2 \leq r$ (as $x^f \in \mathcal{A}$) and $\alpha_{j^*} \geq \alpha_{\min}$, the factor w_{j^*} cancels, and the second term contributes at most $2\beta L_g B_\eta r / (\alpha_{\min} \delta_g)$ to $\|\nabla_w \ell\|_2$ and at most $2\beta L_g B_\eta / (\alpha_{\min} \delta_g)$ to $|\partial_b \ell|$. For the logistic link, $g' = g(1 - g)$ and under Assumption F.1, $\hat{\eta}(x^f) \geq \hat{\eta}(x)$ give $g'(\hat{w}^\top x^f) / (1 - \hat{\eta}(x)) = \hat{\eta}(x^f)(1 - \hat{\eta}(x^f)) / (1 - \hat{\eta}(x)) \leq 1$, which replaces L_g/δ_g by 1. $\square$

F ORACLE-FREE ESTIMATION BOUNDS

Theorem 4.2 establishes a generalization guarantee under access to a true post-deployment label oracle. This section removes that assumption. We replace the oracle with the labeling rule of Algorithm 2, which is built from a class-probability estimator $\hat{\eta}$ learned entirely from pre-deployment data and quantify the resulting price. The analysis proceeds via Theorem 5.3, which bounds err_{imp} in terms of deviation of the label process from $\eta(\cdot^f)$; our task is to show that this deviation is controlled by the pre-deployment estimator error, and to bound the latter.

Estimation sample and independence. Let estimation sample be

$$S_\eta = \{x_i, y_i\}_{i=1}^m, \quad x_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{D}, \quad y_i \mid x_i \sim \text{Bernoulli}(\eta(x_i)),$$

drawn independently of the classifier’s sample $S_n = \mathcal{D}_{\text{tr}} \cup \mathcal{D}_{\text{val}}$, with $n := |S_n|$ and $m := |S_\eta|$. The estimation sample lies in $\mathcal{X}_0 := \text{supp}(\mathcal{D})$ — the estimator $\hat{\eta}$ is fit only on pre-deployment data, and no post-deployment or off-support label is used at any point. It is fit once and passed to Algorithm 3 as a frozen input, so it is not co-adapted to the classifier being learned. Independence of S_η and S_n is a hypothesis of Theorem 5.3, discharged in the proof of Lemma E.2 by conditioning on S_η . It is used again below, when the theorem is invoked for the random estimator $\hat{\eta}$: conditioning on S_η freezes $\hat{\eta}$ into a deterministic function while, by independence, S_n remains i.i.d. from $\mathcal{D}$; so the conclusion holds for every realization of S_η and hence, on averaging, unconditionally.

The estimator. We take $\hat{\eta}(x) = g(\hat{w}^\top x)$ with $\hat{w} \in \arg \min_{w \in \mathcal{W}} \hat{L}_m(w)$, the empirical cross-entropy risk on S_η . Writing $t := w^\top x$,

$$\ell(w; (x, y)) = -y \log g(w^\top x) - (1 - y) \log(1 - g(w^\top x)) = \varphi_y(t)$$

where $\varphi_1(t) = -\log g(t)$ and $\varphi_0(t) = -\log(1 - g(t))$. Since the loss depends on w only through the scalar $w^\top x$, the chain rule gives $\nabla_w \ell(w; (x, y)) = \varphi'_y(w^\top x)x$, with

$$\varphi'_1(t) = -\frac{g'(t)}{g(t)}, \quad \varphi'_0(t) = \frac{g'(t)}{1 - g(t)}, \quad \text{equivalently } \varphi'_y(t) = \frac{g'(t)(g(t) - y)}{g(t)(1 - g(t))},$$

and

$$\varphi''_1(t) = -\frac{g''(t)}{g(t)} + \left(\frac{g'(t)}{g(t)}\right)^2, \quad \varphi''_0(t) = \frac{g''(t)}{1 - g(t)} + \left(\frac{g'(t)}{1 - g(t)}\right)^2.$$

We write the population risk and the empirical risk for estimating η as $L(w) := \mathbb{E}_{(x,y)}[\ell(w; (x, y))]$ and $\hat{L}_m(w) := \frac{1}{m} \sum_{i=1}^m \ell(w; (x_i, y_i))$, set $\Delta_\eta(w) := \nabla \hat{L}_m(w) - \nabla L(w)$, and $\Phi := \sup_{w \in \mathcal{W}_\eta} \|\Delta_\eta(w)\|_2$. Recall from Appendix E that the classification-stage guarantee depends upon $\hat{\eta}$ only through $\varepsilon_\eta = \sup_{z \in \mathcal{A}} |\hat{\eta}(z) - \eta(z)|$, and not on the loss used to obtain it; cross entropy is therefore one sufficient choice, and any independently estimated $\hat{\eta}$ with $\varepsilon_\eta = \mathcal{O}(m^{-1/2})$ yields the same conclusion.

On-support estimation and off-support evaluation. Two regions appear in the analysis and are not interchangeable. All the estimation lemmas below are statements about $\mathcal{D}$ alone: the sample S_η , the risks L and $\hat{L}_m$, and the strong-convexity constant α are determined by the law of $x \sim \mathcal{D}$, hence depend on η only through its values on $\mathcal{X}_0$. Extrapolation to reported points enters at a single step, the final Lipschitz bound in the proof of Theorem F.6, which converts the on-support parameter error $\|\hat{w} - w^*\|_2$ into a bound on ε_η over $\mathcal{A}$, at the cost of a factor $L_g r$ with r the radius of the report set rather than of $\mathcal{X}_0$. Two ingredients license this step. Assumption 1 posits the single-index form on all of $\mathcal{X}$, hence on $\mathcal{A}$, so $\hat{\eta}$ and η share the same link g and differ off support only through the parameter $\hat{w} - w^*$; and condition F.1(5) — strong convexity of the population cross-entropy risk, i.e. no direction of feature variation is unobserved under $\mathcal{D}$ — makes the parameter identifiable from $\mathcal{D}$ (Section F.4). Accuracy at reported points is therefore a consequence of pre-deployment estimation rather than a separate hypothesis.

F.1 Assumptions

Throughout, $\mathcal{X}_0 := \text{supp}(\mathcal{D})$ and $\mathcal{A}$ denotes the feasible report set, so that $\mathcal{X}_0 \subseteq \mathcal{A} \subseteq \mathcal{X}$.

Assumption F.1 (Estimation regularity). In addition to Assumption 1:

1. (*Bounded reports.*) $r := \sup_{z \in \mathcal{A}} \|z\|_2 < \infty$.
2. (*Smooth link.*) g is twice differentiable on $\mathcal{I} := [-rB_\eta, rB_\eta]$, with $g'(t) \leq L_g < \infty$ and $|g''(t)| \leq L_{g'} < \infty$ for all $t \in \mathcal{I}$.
3. (*Bounded probabilities.*) There is $\tau \in (0, 1/2)$ such that $\tau \leq g(w^\top z) \leq 1 - \tau$ for all $z \in \mathcal{A}$ and $w \in \mathcal{W}_\eta$.
4. (*Parameter set.*) $\mathcal{W}_\eta = \{w \in \mathbb{R}_{\geq 0}^d : \|w\|_2 \leq B_\eta\}$, and $\|w^*\| \leq B_\eta$.
5. (*Population strong convexity.*) Population cross-entropy risk L is α -strongly convex on $\mathcal{W}_\eta$.

Conditions 1 – 4 are boundedness conditions yielding explicit constants; condition 5 is the identifiability condition (Section F.4), subsumes strict monotonicity of g on the observed range. We note that $\mathcal{X}$ itself need not be bounded: every point at which η , $\hat{\eta}$, or the loss is evaluated lies in $\mathcal{A}$, and $\mathcal{X}_0 \subseteq \mathcal{A}$ gives $\|x_i\| \leq r$ for the estimation sample automatically. Consequently, for all $w \in \mathcal{W}_\eta$ and $z \in \mathcal{A}$, $|w^\top z| \leq \|w\|_2 \|z\|_2 \leq rB_\eta$, so conditions 2 and 3 are required only on the interval $\mathcal{I}$ and the set $\mathcal{A}$ respectively.

Condition 4 constrains the estimator's parameter set, which is distinct from the classifier's $\mathcal{W}_{k,B}$: B_η must be large enough to contain true single-index parameter w^* , whereas k is the capacity-control parameter for the classifier. The sign constraint mirrors Assumption 1 and ensures that $\hat{\eta}$ is non-decreasing along the best response direction, as used in Lemma E.1. Condition 3 supplies the remaining hypothesis of that lemma: since $\hat{\eta}(z) = g(\hat{w}^\top z) \leq 1 - \tau$ for every $z \in \mathcal{A}$, any choice $\delta_g \leq \tau$ leaves the guard in $\hat{\pi}_{\text{Imp}}$ inactive. Define

$$A_1 := \frac{L_g}{\tau}, \quad A_2 := \frac{L_{g'}}{\tau} + \left(\frac{L_g}{\tau}\right)^2, \quad M := A_1 r, \quad H := A_2 r^2.$$

F.2 Estimation lemmas

The argument proceeds in four stages. Lemmas F.1 and F.2 bound the complexity of the gradient class of the cross-entropy loss; Lemma F.3 converts this into a uniform deviation bound between empirical and population gradients; Lemmas F.4 and F.5 convert that into a bound on the parameter error $\|\hat{w} - w^*\|$. Theorem F.6 combines them with Theorem 5.3.

Lemma F.1 (Derivative bounds.). *Under Assumption F.1 (1)–(3), for all $y \in \{0, 1\}$ and $t \in \mathcal{I}$:*

$$(a) \quad |\varphi'_y(t)| \leq A_1; \quad (b) \quad |\varphi''_y(t)| \leq A_2.$$

Consequently, for all $w_1, w_2 \in \mathcal{W}_\eta$, $z \in \mathcal{A}$:

$$(c) \quad \|\nabla_w \ell(w; (z, y))\|_2 \leq M \quad (d) \quad \|\nabla_w \ell(w_1; (z, y)) - \nabla_w \ell(w_2; (z, y))\|_2 \leq H \|w_1 - w_2\|_2.$$

Proof. By F.1, $t = w^\top z \in \mathcal{I}$ whenever $z \in \mathcal{A}$ and $w \in \mathcal{W}_\eta$, so conditions (2) and (3) apply.

(a) The two derivative formulas have the same shape and both denominators are at least τ using condition (3):

$$|\varphi'_1(t)| = \frac{g'(t)}{g(t)} \leq \frac{L_g}{\tau}, \quad |\varphi'_0(t)| = \frac{g'(t)}{1 - g(t)} \leq \frac{L_g}{\tau}.$$

(b) Applying $|g''| \leq L_{g'}$, $g' \leq L_g$ and $\min\{g, 1 - g\} \geq \tau$ to the two second-derivative formulas gives $|\varphi''_y(t)| \leq L_{g'}/\tau + L_g^2/\tau^2 = A_2$ for both y .

(c) We know $\|\nabla_w \ell(w; (z, y))\|_2 = |\varphi'_y(t)| \|z\|_2 \leq A_1 r = M$, by (a) and condition (1).

(d) Write $t_i = w_i^\top z \in \mathcal{I}$. By (b) and the mean value theorem, φ'_y is A_2 -Lipschitz on $\mathcal{I}$, so by Cauchy-Schwarz and condition (1),

$$\|\varphi'_y(t_1)z - \varphi'_y(t_2)z\|_2 = |\varphi'_y(t_1) - \varphi'_y(t_2)| \|z\|_2 \leq A_2 |t_1 - t_2| r \leq A_2 r^2 \|w_1 - w_2\|_2$$

□

Lemma F.2 (Rademacher complexity of the gradient class). *Let $\epsilon_1, \dots, \epsilon_m$ be i.i.d. Rademacher variables and $(z_t, y_t) \in \mathcal{A} \times \{0, 1\}, t \in [m]$, be arbitrary deterministic points. Under conditions (1)–(4) of Assumption F.1 ,*

$$\widehat{\mathfrak{R}}_{\|\cdot\|_2}(\nabla \ell \circ \mathcal{W}_\eta) = \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}_\eta} \left\| \sum_{t=1}^m \epsilon_t \nabla_w \ell(w; (z_t, y_t)) \right\|_2 \leq 2(HB_\eta + M)\sqrt{m}.$$

Proof. Write $\ell(w; (z_t, y_t)) = G_t(F_t(w))$ with $G_t := \varphi_{y_t}$ and $F_t(w) := \langle w, z_t \rangle$, so that $\nabla_w \ell(w; (z_t, y_t)) = G'_t(\langle w, z_t \rangle) z_t$. From Assumption F.1, $F_t(w) \in \mathcal{I}$ for every $w \in \mathcal{W}_\eta$, so Lemma F.1 applies throughout.

Chain rule. By Lemma F.1(a), $\sup_{s \in \mathcal{I}} |G'_t(s)| \leq A_1 =: L_G$, $\|\nabla F_t(w)\|_2 = \|z_t\|_2 \leq r =: L_F$, both uniformly in t and w . Applying the chain rule for normed Rademacher complexity using Theorem 1 of Foster et al. (2018) with $K = 1$, and noting that $\nabla F_t(w) = z_t$ does not depend on w ,

$$\widehat{\mathfrak{R}}_{\|\cdot\|_2} \leq 2r \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}_\eta} \sum_{t=1}^m \epsilon_t \varphi'_{y_t}(\langle w, z_t \rangle) + 2A_1 \mathbb{E}_\epsilon \left\| \sum_{t=1}^m \epsilon_t z_t \right\|_2$$

Scalar contraction. By Lemma F.1, each φ'_{y_t} is A_2 -Lipschitz on $\mathcal{I}$. The Ledoux-Talagrand contraction inequality (Mohri and Medina (2014), Lemma 8) applied for a fixed sample to the real-valued class $\{z \mapsto \langle w, z \rangle : w \in \mathcal{W}_\eta\}$, together with $\sup_{w \in \mathcal{W}_\eta} \langle w, v \rangle \leq B_\eta \|v\|_2$, gives

$$\mathbb{E}_\epsilon \sup_{w \in \mathcal{W}_\eta} \sum_{t=1}^m \epsilon_t \varphi'_{y_t}(\langle w, z_t \rangle) \leq A_2 \mathbb{E}_\epsilon \sup_{w \in \mathcal{W}_\eta} \langle w, \sum_{t=1}^m \epsilon_t z_t \rangle \leq A_2 B_\eta \mathbb{E}_\epsilon \left\| \sum_{t=1}^m \epsilon_t z_t \right\|_2$$

Rademacher average of the data. By Jensen's inequality, the expected norm of the Rademacher sum is at most the square root of its expected square norm. Expanding the squared norm and using independence with $\mathbb{E}_\epsilon[\epsilon_s \epsilon_t] = \mathbf{1}\{s = t\}$ eliminate all cross terms. Since $\epsilon_t^2 = 1$,

$$\mathbb{E}_\epsilon \left\| \sum_{t=1}^m \epsilon_t z_t \right\|_2 \leq \sqrt{\sum_{t=1}^m \|z_t\|_2^2} \leq r\sqrt{m}.$$

Combining the above bound yields,

$$\widehat{\mathfrak{R}}_{\|\cdot\|_2} \leq 2r \cdot A_2 B_\eta r \sqrt{m} + 2A_1 r \sqrt{m} = 2(A_2 r^2 B_\eta + A_1 r) \sqrt{m} = 2(HB_\eta + M) \sqrt{m}.$$

□

Lemma F.3 (Uniform gradient deviation.). *Under Assumption F.1 (1)–(4), for any $\delta' \in (0, 1)$, with probability at least $1 - \delta'$ over S_η ,*

$$\Phi \leq \Lambda_m(\delta') := \frac{4(HB_\eta + M)}{\sqrt{m}} + M \sqrt{\frac{2 \log(1/\delta')}{m}}$$

Proof. Since $x_i \in \mathcal{X}_0 \subseteq \mathcal{A}$, Lemma F.1 applies at every sample point. In particular $w \mapsto \Delta_\eta(w)$ is continuous on $\mathcal{W}_\eta$ by Lemma F.1 (d), so the supremum defining Φ can be restricted to a countable dense subset; hence Φ is measurable.

Concentration about the mean. Fix $i \in [m]$ and replace (x_i, y_i) by arbitrary $(x'_i, y'_i) \in \mathcal{A} \times \{0, 1\}$, writing $\Delta_\eta^{(i)}$ for the resulting deviation. The population term cancels, so for every $w \in \mathcal{W}_\eta$, by Lemma F.1 (c),

$$\|\Delta_\eta(w) - \Delta_\eta^{(i)}\|_2 = \frac{1}{m} \left\| \nabla_w \ell(w; (x_i, y_i)) - \nabla_w \ell(w; (x'_i, y'_i)) \right\|_2 \leq \frac{2M}{m}.$$

Since $|\sup_w a(w) - \sup_w b(w)| \leq \sup_w |a(w) - b(w)|$ for bounded a, b , together with the reverse triangle inequality, Φ has bounded differences $c_i = 2M/m$. McDiarmid's inequality with $\sum_i c_i^2 = 4M^2/m$ gives $\mathbb{P}(\Phi - \mathbb{E}[\Phi] \geq s) \leq \exp(-ms^2/2M^2)$, whence

$$\Phi \leq \mathbb{E}[\Phi] + M \sqrt{\frac{2 \log(1/\delta')}{m}} \quad \text{with probability at least } 1 - \delta' \quad (12)$$

Symmetrization. Let $(x'_i, y'_i)_{i \leq m}$ be an independent copy of the sample. Since $\nabla L(w) = \mathbb{E}[\frac{1}{m} \sum_i \nabla_w \ell(w; (x'_i, y'_i))]$, convexity of the norm and Jensen's inequality give

$$\mathbb{E}[\Phi] \leq \mathbb{E} \sup_{w \in \mathcal{W}_\eta} \left\| \frac{1}{m} \sum_{i=1}^m \left(\nabla_w \ell(w; (x_i, y_i)) - \nabla_w \ell(w; (x'_i, y'_i)) \right) \right\|_2$$

For any fixed $\epsilon \in \{\pm 1\}^m$, exchanging (x_i, y_i) with (x'_i, y'_i) on $\{i : \epsilon_i = -1\}$ leaves the joint law unchanged and flips the corresponding differences, so the expectation is unchanged upon inserting the signs ϵ_i ; averaging over Rademacher ϵ and applying the triangle inequality yields two terms equal in law, giving

$$\mathbb{E}[\Phi] \leq \frac{2}{m} \mathbb{E} \hat{\mathfrak{R}}_{\|\cdot\|}(\nabla \ell \circ \mathcal{W}_\eta).$$

Since Lemma F.2 holds for arbitrary deterministic points in $\mathcal{A} \times \{0, 1\}$, its bound remains valid under the outer expectation:

$$\mathbb{E}[\Phi] \leq \frac{4(HB_\eta + M)}{\sqrt{m}}. \quad (13)$$

Substituting inequality 13 into 12 completes the proof. $\square$

Lemma F.4 (Fischer Consistency). *Under Assumption 1 and F.1 (1)–(3) conditions, $\nabla L(w^*) = 0$.*

Proof. From the setup,

$$\nabla_w \ell(w; (x, y)) = \frac{g'(w^\top x)(g(w^\top x) - y)}{g(w^\top x)(1 - g(w^\top x))} x.$$

Under Assumption 1, $\mathbb{E}[y | x] = \eta(x) = g(w^{\star\top} x)$, so at $w = w^*$ the factor $g(w^{\star\top} x) - \mathbb{E}[y | x]$ vanishes and $\mathbb{E}[\nabla_w \ell(w^*; (x, y)) | x] = 0$. The integrand is bounded by M (Lemma F.1 (c)), so the tower property gives $\nabla L(w^*) = 0$. $\square$

Lemma F.5 (Parameter estimation.). *Let $\hat{w} \in \arg \min_{w \in \mathcal{W}_\eta} \hat{L}_m(w)$, which exists since $\mathcal{W}_\eta$ is compact and ℓ is continuous in w . Under Assumption F.1 (4)–(5) and $\nabla L(w^*) = 0$, deterministically,*

$$\|\hat{w} - w^*\| \leq \frac{\Phi}{\alpha}.$$

Proof. The set $\mathcal{W}_\eta$ is convex, being the intersection of a Euclidean ball with the non-negative orthant, and $\hat{w}$ minimizes the differential function $\hat{L}_m$ over it; hence $\hat{w}$ satisfies the variational inequality $\langle \nabla \hat{L}_m(\hat{w}), w - \hat{w} \rangle \geq 0$ for all $w \in \mathcal{W}_\eta$. Taking $w = w^*$, which is feasible by Assumption 1 and condition (4),

$$\langle \nabla \hat{L}_m(\hat{w}), \hat{w} - w^* \rangle \leq 0. \quad (14)$$

This requires no convexity of $\hat{L}_m$: any global minimizer over a convex set satisfies above bound. α -strong convexity of L on $\mathcal{W}_\eta$ is equivalent to strong monotonicity of ∇L , which with $\nabla L(w^*) = 0$ gives

$$\alpha \|\hat{w} - w^*\|_2^2 \leq \langle \nabla L(\hat{w}), \hat{w} - w^* \rangle. \quad (15)$$

Decomposing the right-hand side of 15 and applying Cauchy-Schwartz to the first term and 14 to the second,

$$\langle \nabla L(\hat{w}), \hat{w} - w^* \rangle = \langle -\Delta_\eta(\hat{w}), \hat{w} - w^* \rangle + \langle \nabla \hat{L}_m(\hat{w}), \hat{w} - w^* \rangle \leq \Phi \|\hat{w} - w^*\|_2.$$

Dividing by $\|\hat{w} - w^*\|_2$ when $\hat{w} \neq w^*$, the claim being trivial otherwise. $\square$

F.3 Main result

Theorem F.6 (Oracle-free guarantee). *Suppose the assumptions of Theorem 5.3 hold, and let Assumption F.1 hold. Let $\hat{\eta}(x) = g(\hat{w}^\top x)$ with $\hat{w} \in \arg \min_{w \in \mathcal{W}_\eta} \hat{L}_m(w)$ on the estimation sample S_η , drawn independently of the classifier's sample S_n , and let $\tilde{w}, \tilde{b}$ be the model returned by STRAT-IMP-AWARE. Then, for every $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\text{err}_{\text{Imp}}(f_{\tilde{w}, \tilde{b}}) \leq \hat{L}_{\text{hinge}}^{\tilde{q}}(\tilde{w}, \tilde{b}) + c(1 + C_{k,B})\sqrt{\frac{\mathfrak{d}}{n}} + (1 + C_{k,B})\sqrt{\frac{2\log(2/\delta)}{n}} + \frac{4C_{k,B}L_g r}{\alpha}\Lambda_m(\delta/2),$$

$$\text{where } \Lambda_m(\delta/2) = \frac{4(HB_\eta + M)}{\sqrt{m}} + M\sqrt{\frac{2\log(2/\delta)}{m}}.$$

Proof. Let E_1 denote the event on which the conclusion of Theorem 5.3 holds at level $\delta/2$, and E_2 the event of Lemma F.3 with $\delta' = \delta/2$.

Bounding $\mathbb{P}(E_1^c)$. Theorem 5.3 is a statement about a fixed estimator, whereas $\hat{\eta}$ depends upon S_η . Conditioning on S_η renders $\hat{\eta}$, and hence ε_η , deterministic, while by independence S_n remains i.i.d. from $\mathcal{D}$; the theorem therefore applies for every realization of S_η , giving $\mathbb{P}(E_1^c | S_\eta) \leq \delta/2$ pointwise, and hence $\mathbb{P}(E_1^c) \leq \delta/2$ by the law of total probability. On E_1 ,

$$\text{err}_{\text{Imp}}(f_{\tilde{w}, \tilde{b}}) \leq \hat{L}_{\text{hinge}}^{\tilde{q}}(\tilde{w}, \tilde{b}) + c(1 + C_{k,B})\sqrt{\frac{\mathfrak{d}}{n}} + (1 + C_{k,B})\sqrt{\frac{2\log(2/\delta)}{n}} + 4C_{k,B}\varepsilon_\eta, \quad (16)$$

the confidence term arising from Theorem 5.3 at level $\delta/2$.

Bounding ε_η on E_2 . By Lemma F.3, $\Phi \leq \Lambda_m(\delta/2)$; by Lemmas F.4 and F.5, $\|\hat{w} - w^*\| \leq \Lambda_m(\delta/2)/\alpha$. For any $z \in \mathcal{A}$, Assumption 1 gives $\eta(z) = g(w^{\star\top} z)$ and $\hat{\eta}(z) = g(\hat{w}^\top z)$ by construction; both arguments lie in $\mathcal{I}$ by Assumption F.1, so g is L_g -Lipschitz there, and by Cauchy-Schwartz with $\|z\|_2 \leq r$,

$$|\hat{\eta}(z) - \eta(z)| = |g(\hat{w}^\top z) - g(w^{\star\top} z)| \leq L_g |(\hat{w} - w^*)^\top z| \leq L_g r \|\hat{w} - w^*\|_2.$$

Taking the supremum over $z \in \mathcal{A}$ gives

$$\varepsilon_\eta \leq \frac{L_g r}{\alpha} \Lambda_m(\delta/2). \quad (17)$$

Combining the bounds. By union bound, $\mathbb{P}(E_1^c \cup E_2^c) \leq \delta/2 + \delta/2 = \delta$, so $E_1 \cap E_2$ holds with probability at least $1 - \delta$. Substituting 17 into 16 proves the theorem. $\square$

Remark F.7 (Rate). Treating $r, B_\eta, L_g, L_{g'}, \tau^{-1}$ and α^{-1} as constants, fixing δ , we have $\Lambda_m(\delta/2) = \mathcal{O}(m^{-1/2})$. Thus, the additional error due to oracle-free estimation is $\mathcal{O}(m^{-1/2})$, and the overall rate is $\mathcal{O}(\sqrt{\mathfrak{d}/n} + m^{-1/2})$, where $\mathfrak{d} = \mathcal{O}(d)$. In particular, $m \gtrsim n$, the estimation term is bounded by a constant multiple of the classifier term, yielding an overall rate of $\mathcal{O}(\sqrt{\mathfrak{d}/n})$. Hence, oracle-free estimation preserves the asymptotic rate of oracle-based bound.

F.4 Condition 5 as an identifiability condition

Under Assumption 1, $\eta(x) = g(w^{\star\top} x)$, so evaluating the Hessian of L at $w = w^*$ gives

$$\nabla^2 L(w^*) = \mathbb{E}_{x \sim \mathcal{D}} \left[\frac{g'(w^{\star\top} x)^2}{g(w^{\star\top} x)(1 - g(w^{\star\top} x))} xx^\top \right],$$

a product of a factor measuring whether η responds to feature variation and one measuring whether the features vary at all. Condition 5 asserts that this g' -weighted design is non-degenerate, and therefore excludes both ways in which w^* can fail to be identified from $\mathcal{D}$: if there is a vector $a \neq 0$ with $a^\top x = 0$ and $x \sim \mathcal{D}$ -a.s., then $g((w^* + \lambda a)^\top x) = \eta(x)$ for every $\lambda \in \mathbb{R}$ and every $x \in \mathcal{X}_0$, so w^* and $w^* + \lambda a$ are observationally indistinguishable at any sample size; and if $g' \equiv 0$ on a region of the observed range, the corresponding points contribute nothing to the Hessian. Quantitatively, since $g' \leq L_g$ and $g(1 - g) \geq \tau(1 - \tau)$ on $\mathcal{I}$,

$$\gamma := \lambda_{\min}(\mathbb{E}_{x \sim \mathcal{D}}[xx^\top]) \geq \frac{\alpha\tau(1 - \tau)}{L_g^2} > 0,$$

that is, $\text{supp}(\mathcal{D})$ spans $\mathbb{R}^d$. Conversely, if $\gamma > 0$ and $g'(t) \geq c > 0$ on $\mathcal{I}$, then $\nabla^2 L(w^*) \succeq 4c^2\gamma I$, using $g(1-g) \leq 1/4$; a lower bound on g' is thus sufficient for condition 5 at w^* but is not assumed, being implied on the observed range by the condition 5 itself. The same two quantities govern the regularity conditions used throughout the generalized-linear-model literature in Foster et al. (2018). Crucially, the identifiability content of condition 5 is verifiable on the pre-deployment sample: γ is estimated by the $\lambda_{\min}$ of the empirical second-moment matrix $\frac{1}{m} \sum_{i=1}^m x_i x_i^\top$; values for each dataset are reported in Section H.

F.5 Sufficient conditions for condition 5

Condition 5 imposes an additional restriction beyond Assumption 1 : for a general non-decreasing link, the cross-entropy risk need not be convex in w . We give two sufficient conditions with explicit strong-convexity constants. Write $t := w^\top x$ and suppress the arguments of g, η and φ_y where no ambiguity arises. Since $\mathbb{E}_{y|x}[\varphi_y''] = \eta\varphi_1'' + (1-\eta)\varphi_0''$ and $\nabla^2 L(w) = \mathbb{E}_{x \sim \mathcal{D}}[\mathbb{E}_{y|x}[\varphi_y'']xx^\top]$, it suffices to obtain a uniform lower bound on the conditional expected curvature. Let $c > 0$ satisfy $g' \geq c$ on $\mathcal{I}$, and suppose $\gamma := \lambda_{\min}(\mathbb{E}_{x \sim \mathcal{D}}[xx^\top]) > 0$; Section F.4 shows this is also necessary for Condition 5.

log-concave link. Suppose g and $1-g$ are log-concave on $\mathcal{I}$, equivalently $\varphi_1'' \geq 0$ and $\varphi_0'' \geq 0$, as for the logistic and probit links. Weighting the two second derivatives by g and $1-g$ cancels g'' terms, giving

$$g\varphi_1'' + (1-g)\varphi_0'' = \frac{(g')^2}{g(1-g)} \geq 4c^2,$$

where the inequality uses $g(1-g) \leq 1/4$. Non-negativity of the second derivatives then gives $\varphi_1'' + \varphi_0'' \geq 4c^2$. Since $\eta(x) = g(w^{*\top}x) \in [\tau, 1-\tau]$ by conditions 3 and 4, we obtain $\mathbb{E}_{y|x}[\varphi_y''] \geq \tau(\varphi_1'' + \varphi_0'') \geq 4\tau c^2$, uniformly over $w \in \mathcal{W}_\eta$ and $x \in \mathcal{X}_0$. Thus condition 5 holds with $\alpha = 4\tau c^2\gamma$; the factor τ is the price of a bound uniform in w rather than one at w^* alone, where $\eta = g$ and the constant improves to $4c^2\gamma$ (Section F.4). Strict log-concavity is unnecessary: since $g\varphi_1'' + (1-g)\varphi_0'' = (g')^2/[g(1-g)] > 0$ when $g' > 0$, the two second derivatives cannot vanish simultaneously.

For the special case of logistic link $g = \sigma$, where $\sigma(t) = (1 + e^{-t})^{-1}$, the identity $\sigma' = \sigma(1-\sigma)$ gives sharper constants. Specifically, $\varphi_y'(t) = \sigma(t) - y$ and $\varphi_1' = \varphi_0' = \sigma'$, so $\nabla_w^2 \ell(w; (x, y)) = \sigma'(w^\top x)xx^\top$, independently of the label. Since $|w^\top x| \leq rB_\eta$, we may take $\alpha = \sigma'(rB_\eta)\gamma$. Moreover $|\varphi_y'| \leq 1$ and $|\varphi_y''| \leq 1/4$, so the conclusions of Lemma F.1 hold with A_1 and A_2 replaced by 1 and $1/4$ respectively, with no dependence upon τ . This is the case that applies to our experiments.

Dominated curvature. The second sufficient condition follows by directly controlling the contribution of g'' . We have

$$\mathbb{E}_{y|x}[\varphi_y''] = (g')^2 \left[\frac{\eta}{g^2} + \frac{1-\eta}{(1-g)^2} \right] + g'' \frac{g-\eta}{g(1-g)} \geq \frac{c^2}{(1-\tau)^2} - \frac{L_{g'}(1-2\tau)}{\tau(1-\tau)}.$$

Indeed, $\eta(x), g(t) \in [\tau, 1-\tau]$ imply $|g-\eta| \leq 1-2\tau$ and $g(1-g) \geq \tau(1-\tau)$; the expression in brackets is a convex combination of $1/g^2$ and $1/(1-g)^2$, each at least $(1-\tau)^{-2}$; and the bound then follows from $g' \geq c$ and $|g''| \leq L_{g'}$. It is positive whenever $c^2 > L_{g'}(1-2\tau)(1-\tau)/\tau$, in which case condition 5 holds with α equal to the displayed lower bound times γ . This condition also accommodates links that are not log-concave, provided their second derivatives are sufficiently small relative to the lower bound on g' .

In both cases, the resulting α is proportional to γ , consistent with Section F.4.

F.6 Scope of the pre-deployment data

Assumption 1 posits the single-index form on all of $\mathcal{X}$, hence on $\mathcal{A}$, and this premise is not testable from pre-deployment data; we do not claim otherwise. The analysis shows that no additional untestable assumption about off-support extrapolation is required. Conditional on Assumption 1, the estimator and the target share the same link g and differ off support only through the parameter $\hat{w} - w^*$, so the accuracy of $\hat{\eta}$ at reported points is controlled entirely by $\|\hat{w} - w^*\|_2$ — a quantity estimated from $S_\eta \subseteq \mathcal{X}_0$ (Lemmas F.3 and F.5) and governed by condition 5, whose identifiability content is itself verifiable on the pre-deployment sample (Section F.4). The price of extrapolation is explicit and finite: the factor $L_{g'r}$ in Theorem F.6, with r radius of the report set $\mathcal{A}$ rather than $\mathcal{X}_0$. Direct validation of the label model would instead require observing post-response outcomes, which the oracle-free setting posits are unavailable; the improvement rates reported in Section 6 therefore do not constitute an independent test of Assumption 1.

G PROXY FOR IMPROVEMENT-AWARE CLASSIFIER

Theorem 7.1. *Under Assumptions 1 and 2, the feasibility and tie-breaking conditions from Section 3, the strategic-optimal classifier f_s^* and the Bayes-optimal classifier f^* satisfy*

$$\text{err}_{\text{Imp}}(f_s^*) \leq \text{err}_{\text{Imp}}(f^*).$$

Proof. Let $Z = w^{\star\top} X$. By Corollary 3.4, the strategic-optimal classifier is parallel to the Non-strategic Bayes optimal classifier and has threshold $b_s^* = b^* + \Delta$ where $\Delta := \beta \max_{i \in [d]} \frac{w_i^*}{\alpha_i}$. Thus we have

$$f^*(x) = \mathbb{1}\{w^{\star\top} x \geq b^*\} \text{ and } f_s^*(x) = \mathbb{1}\{w^{\star\top} x \geq b^* + \Delta\}.$$

Note that for any classifier of the form $f_b(x) = \mathbb{1}\{w^{\star\top} x \geq b\}$, an agent with score $z = w^{\star\top} x$ can reach the acceptance region if and only if $z + \Delta \geq b$. Under the feasibility condition over $\mathcal{A}$ and the canonical minimum-cost positive best-response rule, an agent with original score $z = w^{\star\top} x$ either remains at x if $z < b - \Delta$ or $z \geq b$, and otherwise moves to a minimum-cost accepted report in $\mathcal{A}$ with score exactly b . Therefore the improvement-aware error of f_b as

$$\begin{aligned} \text{err}_{\text{Imp}}(f_b) &= \int_{-\infty}^{b-\Delta} g(z) dP_Z(z) \\ &\quad + \int_{b-\Delta}^b (1 - g(b)) dP_Z(z) \\ &\quad + \int_b^{\infty} (1 - g(z)) dP_Z(z). \end{aligned}$$

Let $(I) := \text{err}_{\text{Imp}}(f_{b^*}) - \text{err}_{\text{Imp}}(f_{b^*+\Delta})$. Then,

$$\begin{aligned} (I) &= \int_{[b^*-\Delta, b^*)} (1 - g(b^*) - g(z)) dP_Z(z) \\ &\quad + \int_{[b^*, b^*+\Delta)} (g(b^* + \Delta) - g(z)) dP_Z(z) \\ &\geq \int_{b^*-\Delta}^{b^*} \left(\frac{1}{2} - g(z) \right) dP_Z(z). \end{aligned}$$

By Assumption 1, g is continuous and non-decreasing and $b^* = \inf\{t : g(t) \geq 1/2\}$ is finite; hence $g(b^*) = 1/2$. Also, since g is non-decreasing, for every $z < b^*$, we have $g(z) \leq g(b^*) = 1/2$, and for every $z \in [b^*, b^* + \Delta)$, $g(z) \leq g(b^* + \Delta)$. Hence both integrands are nonnegative, and therefore $\text{err}_{\text{Imp}}(f_s^*) \leq \text{err}_{\text{Imp}}(f^*)$. $\square$

Remark G.1. Reducing the integral lower bound in the proof of Theorem 7.1 to an explicit closed-form scalar rate would require a margin-type condition (Audibert and Tsybakov, 2007) governing how quickly $g(z)$ moves away from $1/2$ near b^* , together with a lower bound on the density $P_Z(z)$. We do not impose such conditions, as Theorem 7.1 is designed to hold under Assumptions 1 and 2.

H EXPERIMENTS: SUPPLEMENTARY DETAILS

Tie-breaking. When several coordinates attain $\max_{j \in [d]} w_j/\alpha_j$, the agent is indifferent among them, both in cost and in the classifier's decision. Moving along the coordinate j at such a tie raises the qualification index by $w_j^* \frac{(b-w^{\top}x)}{w_j} = \beta \frac{b-w^{\top}x}{S(w)} \cdot w_j^*/\alpha_j$, an amount proportional to w_j^*/α_j , as $\eta(x) = g(w^{\star\top} x)$, so we index features in decreasing order of w_j^*/α_j : the rule $j^* = \min \arg \max_{j \in [d]} w_j/\alpha_j$ then selects, among the equally cost-effective coordinates, the one yielding a largest improvement probability. In the plug-in algorithm, when w^* is unknown, we instead order features by $\hat{w}_j/\alpha_j$, where $\hat{\eta}(x) = g(\hat{w}^{\top} x)$ is the estimate of η , fitted on an independent sample, and keep this ordering fixed throughout training and evaluation.

Training. Because q depends on (w, b) , it is recomputed at every step and gradients propagate through it. For a manipulating point, $x^f = x + \frac{b - w^\top x}{w_{j^*}} e_{j^*}$ (so that $w^\top x^f = b$), which gives $\partial x_{j^*}^f / \partial b = 1/w_{j^*}$ and $\nabla_w x_{j^*}^f = -x^f / w_{j^*}$ (note $w_{j^*} > 0$); the gradient of q then follows by the chain rule through $\hat{\eta}$, which we therefore require to be differentiable. The discrete components of Algorithm 2 (the coordinate j^* and whether an agent manipulates) are piecewise constant in (w, b) and are held fixed when differentiating, which gives the exact gradient on each region where they are constant. With bounded inputs and bounded input derivatives of $\hat{\eta}$ on reachable responses, loss gradients within these regions are bounded uniformly in (w, b) for fixed $\beta, \bar{\alpha}, \delta_g > 0$: the movement branch, the bound $|\partial \ell / \partial q| \leq 2M(x) \leq 2\beta w_{j^*} / \alpha_{j^*}$ cancels the $1/w_{j^*}$ factor; elsewhere, $q = y$ locally. Across manipulation edges the loss is continuous: as $b - w^\top x \rightarrow 0^+$, $x^f \rightarrow x$ and q is continuous, while $b - w^\top x = S(w)$ the shifted score $M(x) = 0$, where the two hinge terms in (3) coincide. Across the tie hyperplanes $\{w_j / \alpha_j = w_k / \alpha_k\}$, by contrast, j^* switches and manipulating agents move along a different coordinate, so q , and with it the loss, can jump. This discontinuity is inherited from the agents’ best response, since an agent at a tie is indifferent between coordinates that lead to different outcomes, and err_{Imp} itself jumps across the same hyperplanes; it is confined to at most $d(d-1)/2$ hyperplanes that do not depend on the data and the best-validation selection in Algorithm 1 guards against degradation from crossing them. Differentiating through q distinguishes STRAT-IMP-AWARE from stochastic methods for performative prediction (Mendler-Dünner et al., 2020), which hold the induced data fixed within each update and target performatively stable points; we instead minimize the improvement-aware objective itself, including its dependence on the labels the model induces. Each update is a projected (sub)gradient step onto $\mathbb{R}_{\geq 0}^d \times \mathbb{R}$ (only w is clipped, as required by Theorem 3.3). After each epoch, improvement probabilities for $\mathcal{D}_{\text{val}}$ are computed under the current model, and the model with the lowest validation loss is returned. Algorithms 2 and 3 are applied row-wise to a batch, and the per-point losses are averaged over the batch; this batch mean is $\mathcal{L}_{\text{batch}}$ in training and the per-batch increment to $\mathcal{L}_{\text{val}}$ in validation, so $\mathcal{L}_{\text{val}}$ is the average batch loss over $\mathcal{D}_{\text{val}}$. The operator $(\cdot)_+$ is applied entrywise in Algorithm 1.

All experiments are produced by a single configuration-driven pipeline. Every dataset-specific choice (label, features, split, agent environment, and training hyperparameters) is stored in one YAML file per dataset (`configs/retiring_adult.yaml`, `configs/law_school.yaml`, `configs/heloc.yaml`, `configs/adult.yaml`, `configs/synthetic.yaml`); sweeps and comparison tables are configured in `configs/sweeps/`. All outputs (splits, models, logs, tables) are written to `outputs/<dataset>/`. Throughout, “ACSIncome” refers to the dataset configured as `retiring_adult`.

H.1 Experimental Setup and Environment

All experiments were executed on an isolated compute node running Ubuntu 20.04.6 LTS. The machine is equipped with an Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz and 256 GB of system RAM. The GPU was not used: the code contains no CUDA or MPS device placement and all tensors reside on the CPU; there is therefore no dedicated VRAM requirement.

Software. Python 3.11.15 with PyTorch 2.3.1 (CPU), scikit-learn 1.4.2, NumPy 1.26.4, pandas 2.2.3, SciPy 1.17.1, Matplotlib 3.10.9, PyYAML 6.0.3 and joblib 1.5.3. Our algorithm and the frozen logistic regression are implemented in PyTorch in `float64`; the neural/linear baselines of Attias et al. (2025) and SERM (Levanon and Rosenfeld, 2022) are trained in `float32`; the logistic regression and the linear SVM use scikit-learn.

Data partition (45/45/5/5). Each dataset is partitioned once, at random and without stratification, into four disjoint subsets with fractions 0.45/0.45/0.05/0.05 (`pipeline/data.py`, function `random_split`; driver `stage1_split_and_profile.py`). A single permutation of the row indices is drawn with `numpy.random.default_rng(42)`; the first three subsets receive $\lfloor 0.45n \rfloor$, $\lfloor 0.45n \rfloor$ and $\lfloor 0.05n \rfloor$ rows and the last subset receives the remainder. The row indices of every subset are stored in `outputs/<dataset>/splits/split_indices.json` together with the SHA-256 hash of the source file (`split_summary.json`), and all later stages read these fixed subsets. The roles are:

- **lr_train (45%):** fits the logistic regression $\hat{\eta}$ used as the differentiable estimate of $\Pr(y = 1 \mid x)$ in Algorithm 2; it is never seen by any classifier under comparison.
- **algo_train (45%):** trains STRAT-IMP-AWARE and all baselines.

- **algo_val** (5%): computes the validation loss used by Algorithm 1 for checkpoint selection, and is used for environment selection and for selecting baseline hyperparameters.
- **algo_test** (5%): unseen test set on which all reported numbers are computed.

Dataset	n	lr_train	algo_train	algo_val	algo_test
ACSIIncome	195,665	88,049 (0.408)	88,049 (0.412)	9,783 (0.411)	9,784 (0.412)
Law School	20,798	9,359 (0.891)	9,359 (0.889)	1,039 (0.868)	1,041 (0.909)
HELOC	9,682	4,356 (0.477)	4,356 (0.494)	484 (0.488)	486 (0.471)
Adult	48,842	21,978 (0.239)	21,978 (0.239)	2,442 (0.243)	2,444 (0.239)
Synthetic	20,000	9,000 (0.423)	9,000 (0.423)	1,000 (0.433)	1,000 (0.407)

Table 3: Subset sizes after preprocessing (from `outputs/<dataset>/splits/split_summary.json`). $\Pr(y=1)$ in parentheses.

Validity of the modelling assumptions. The following assumptions of STRAT-IMP-AWARE were checked empirically.

1. *Monotone, non-negative linear structure* ($w \in \mathbb{R}_{\geq 0}^d$). For every dataset, features were oriented so that larger values are favourable, and the fitted logistic regression has strictly positive coefficients on all selected features (`outputs/<dataset>/models/logistic_regression_meta.json`). Standardized coefficients: ACSIIncome (AGEP, WKHP, SCHL) = (0.690, 0.945, 1.235); Law School (lsat, ugpa, zgpa) = (0.624, 0.232, 1.305); HELOC (ExternalRiskEstimate, -NetFractionRevolvingBurden, -PercentTradesWBalance, PercentTradesNeverDelq) = (0.807, 0.231, 0.124, 0.199); Adult (age, hours-per-week, educational-num) = (0.651, 0.545, 0.892); Synthetic ($x_1, \dots, x_5$) = (1.496, 1.200, 0.880, 0.610, 0.257). Consequently $\hat{\eta}$ is increasing along every admissible improvement direction, so the improvement probability of Algorithm 2 is non-negative before clipping.
2. *Quality of the estimate $\hat{\eta}$* . The logistic regression is calibrated on every split (mean predicted probability within 0.03 of the empirical positive rate) with test AUC 0.834 (ACSIIncome), 0.844 (Law School), 0.752 (HELOC), 0.799 (Adult) and 0.862 (Synthetic). Per-feature profiles of $\hat{\eta}$ against the empirical positive rate are saved in `outputs/<dataset>/plots/(stage3_plot_lr_profiles.py)`. They reveal that the linear logit over-predicts $\Pr(y=1)$ for age > 60 and hours worked > 60 on ACSIIncome and Adult; all other profiles agree with the empirical rates.
3. *Exactness of the improvement probabilities (synthetic data)*. On the synthetic dataset, the true $\eta(x) = \sigma(w^{\star\top}x + b^{\star})$ is known. The fitted logistic regression recovers $w^{\star} = (1.5, 1.2, 0.9, 0.6, 0.3)$, $b^{\star} = -0.5$ as (1.504, 1.189, 0.888, 0.607, 0.252), -0.508; on `algo_test`, $\text{mean}|\hat{\eta} - \eta| = 0.006$, and for the agents that move, $\text{mean}|\hat{\pi}_{\text{Imp}} - \pi_{\text{Imp}}| = 0.0017$ (maximum 0.0068, correlation 0.9998) (`validate_true_eta.py`, `outputs/synthetic/evaluation/true_eta_validation_algo_test.json`). The expected improvement rate is 0.8740 with $\hat{\eta}$ versus 0.8733 with η , and the expected strategic improvement error is 0.0628 versus 0.0630.
4. *Feasible best responses*. An agent moves only if the boundary point of its cheapest coordinate lies within the largest value of that feature observed in `lr_train` (Section 3).
5. *Correctness of the implementation*. The PyTorch copy of the logistic regression reproduces scikit-learn probabilities exactly (maximum absolute difference 0). Agents that move land on the decision boundary (maximum $|w^{\top}x^f - b| = 1.7 \times 10^{-16}$). The autograd gradient of the batch loss, including the path through q , matches central finite differences (step 10^{-6}) to 9.0×10^{-11} .
6. *Positive agents keep their label*. In both training ($q = y + (1 - y)\hat{\pi}_{\text{Imp}}$, `pipeline/strat_imp.py`) and evaluation (`pipeline/evaluation.py`, function `sample_post_labels`), only agents with $y = 0$ can change label. It was verified on the saved per-agent records that no agent with $y = 1$ is assigned a post-response label of 0.

Dataset	Instances	Total features	Imp. features	Task
Adult	48,842	15	5	Income > 50K
HELOC	10,459	24	8	Credit risk
Law School	20,798	12	6	Bar exam
ACS Income	1,664,500	10	3	Income > 50K
Synthetic	20,000	8	8	+ve label

Table 4: Datasets used in experiments.

H.2 Comparison with benchmark algorithms

Both strategic baselines were re-implemented from their public code and papers (no third-party code was executed), so that every method is trained on the same `algo_train` rows and the same standardized features, and is evaluated on the same `algo_test` rows with the same agent model.

Attias et al. (2025). We implement a baseline for improvement aware classifier following the framework of PAC Learning with Improvements (Attias et al., 2025). The method models strategic agents as individuals who iteratively improve their feature vectors in order to obtain a positive classification from the deployed model.

Original setup:

- *Classifier.* A two-layer network $d \rightarrow 64 \rightarrow \text{ReLU} \rightarrow 1 \rightarrow \text{sigmoid}$, trained with Adam (learning rate 10^{-3}) for 100 epochs with batch size 64 and no shuffling.
- *Risk-averse loss.* A false-positive-weighted binary cross-entropy

$$\mathcal{L} = -\frac{1}{n} \sum_i [w_{\text{FP}}(1 - y_i) \log(1 - \hat{y}_i) + w_{\text{FN}} y_i \log \hat{y}_i],$$

with $w_{\text{FP}} = 2.0$, $w_{\text{FN}} = 1.33$ (default) and predictions clamped to $[10^{-6}, 1 - 10^{-6}]$.

- *Decision rule.* Positive iff $h(x) > \tau$, with $\tau = 0.9$ as the risk-averse default.
- *Agent model and labels.* Agents respond by projected signed-gradient ascent in an ℓ_∞ ball, and labels come from a decision tree fitted on all data.

Our adaptation (pipeline/baselines_pli.py).

- The architecture, loss, weights, optimizer, epochs, batch size and seed handling are reproduced exactly.
- The agent model and labels are replaced by ours. Agents best-respond to the decision rule $h(x) > \tau$ with the cost-based feasible response of Section 3 (computed numerically, since h is non-linear), and post-response labels are drawn from $\hat{\pi}_{\text{Imp}}$.
- *Threshold selection (HELOC and Adult only).* With $\tau = 0.9$, the network never predicts a positive and no agent can reach the threshold: the maximum score on HELOC `algo_test` is 0.864, and at the best attainable feature profile it is 0.830. For these two datasets we additionally report the same trained network with τ chosen on `algo_val` from the threshold grid of Attias et al. (2025), $\{0.5, 0.7, 0.8, 0.9\}$, by minimal expected strategic improvement error. This selects $\tau = 0.8$ on HELOC and $\tau = 0.7$ on Adult (`experiment_multiseed_table.py`, option `select_thresholds`).
- *Other datasets.* For Synthetic, Law School and ACSIncome, the paper default $\tau = 0.9$ is reported. On Synthetic, validation selection also returns $\tau = 0.9$.

SERM (Levanon and Rosenfeld, 2022): We implement another strategic learning baseline that explicitly accounts for feature manipulation under ℓ_2 costs.

Original setup:

- *Classifier.* A linear model $f(x) = w^\top x + b$, positive iff $f(x) \geq 0$.
- *Strategic hinge loss.*

$$\frac{1}{n} \sum_i \max\{0, 1 - y_i f(x_i) - 2 y_i \|w\|_2\} + \lambda \|w\|_2^2, \quad y_i \in \{-1, +1\},$$

i.e. users with ℓ_2 cost and budget 2 who all prefer a positive prediction ($r = +1$).

- *Optimization.* Adam with learning rate 0.05 and $\lambda \in \{0.01, 0.1, 1\}$.

Our adaptation (pipeline/baselines_gsc.py).

- The loss is reproduced exactly (shift $2\|w\|_2$).
- *Training.* 100 epochs with shuffled mini-batches of size 128 and Adam (learning rate 0.05); the model after the final epoch is returned.
- *Regularization.* $\lambda \in \{0.1, 1\}$ is selected per run on `algo_val` by minimal expected strategic improvement error (`configs/sweeps/<dataset>/multiseed_table.yaml`, `configs/sweeps/gsc_baseline.yaml`). $\lambda = 0.1$ was selected in every run reported in Table 1 (in the additional Adult seeds 0 and 1 the balanced variant selected $\lambda = 1$).
- *Class weighting (Law School and Adult only).* The hinge term of sample i is multiplied by $n/(2n_{y_i})$, which averages to one. This is required because the unweighted objective collapses on these two Law school and Adult datasets.
- A variant whose shift equals the reach of our cost model, $\beta \max_j |w_j|/\alpha_j$, was also evaluated. It is stored in the result files but not reported in the main table.

Linear SVM. `sklearn.svm.LinearSVC` with hinge loss, $C = 1.0$, `max_iter= 200,000` and `random_state` equal to the run seed. For Law School and Adult, `class_weight="balanced"` (`experiment_svm_baseline.py`, `experiment_multiseed_table.py`).

H.3 Datasets and Preprocessing

All loaders are implemented in `pipeline/data.py` (functions `load_raw`, `resolve_feature_types`, `prepare_dataset`).

Feature typing. Each column is classified automatically as follows:

- non-numeric or binary columns are *nominal*;
- integer-valued columns are *continuous* if they have more than 30 unique values and *ordinal* otherwise (`max_discrete_unique: 30`).

Nominal features are removed (`drop_types: [nominal]`), and the logistic regression and all classifiers use only the selected continuous and ordinal features. For features whose increase lowers the probability of a positive outcome, we take their negation.

Cleaning, normalization, label.

- *Missing values.* Configured missing-value codes are mapped to NaN, and rows with a missing value in any selected feature are dropped before the split.
- *Label.* The target is binarized as $y = \mathbb{1}[\text{target} = \text{positive class}]$.
- *Standardization.* Features are standardized with the mean and standard deviation of `lr_train` (`StandardScaler` inside the logistic-regression pipeline, `pipeline/lr_model.py`). The same affine map $u = (x - \mu)/\sigma$ is applied to all splits and used by every method (`FrozenLR` in `pipeline/strat_imp.py`), so that costs α_j are expressed per standard deviation.

H.4 Training Pipeline and Optimization

Stages. The pipeline consists of five scripts, each driven by `-config configs/<dataset>.yaml`.

1. `stage1_split_and_profile.py`: feature typing, cleaning and the fixed 45/45/5/5 split.
2. `stage2_train_lr.py`: logistic regression on `lr_train`; coefficients on the standardized and the original scale, and metrics on every split.
3. `stage3_plot_lr_profiles.py`: for each selected feature, the mean $\pm$ one standard deviation of $\hat{\eta}$ over the rows sharing each observed value (quantile bins if there are more than 100 unique values), overlaid with the empirical positive rate.
4. `stage4_train_strat_imp.py`: Algorithm 1 on `algo_train` with checkpoint selection on `algo_val`.
5. `stage5_improvement_rate.py`: improvement statistics on `algo_test`.

Comparison tables are produced by `experiment_multiseed_table.py` (Law School, HELOC, Adult, Synthetic) and by `experiment_svm_baseline.py`, `experiment_pli_paper.py` and `experiment_gsc_baseline.py` (AC-SIncome).

Implementation of Algorithms 1–3 (`pipeline/strat_imp.py`, `pipeline/strat_imp_clipped.py`).

- *Frozen estimate.* The logistic regression is copied into a PyTorch module whose parameters are registered as buffers. It therefore receives no updates, but gradients propagate through its input x^f .
- *Algorithm 2 (vectorized).*
 - The improvement direction is $j^\star = \min \arg \max_j w_j / \alpha_j$ (first maximal index).
 - Agents with $0 < b - w^\top x \leq S(w)$ move to $x^f = x + \frac{b - w^\top x}{w_{j^\star}} e_{j^\star}$.
 - *Feasible mode* additionally requires $x_{j^\star} + \frac{b - w^\top x}{w_{j^\star}} \leq \bar{x}_{j^\star}$, the largest value of feature $j^\star$ in `lr_train` (in standardized units).
 - The improvement probability is $\hat{\pi}_{\text{Imp}} = \text{clamp}_{[0,1]}((\hat{\eta}(x^f) - \hat{\eta}(x)) / \max\{1 - \hat{\eta}(x), \delta_g\})$.
 - The branch condition, the choice of $j^\star$ and saturated clamps carry no gradient. A safe divisor avoids NaN gradients when $w_{j^\star} = 0$.
- *Algorithm 3.* The per-sample loss is $q(1-s)_+ + (1-q)(1+s)_+$ with $q = y + (1-y)\hat{\pi}_{\text{Imp}}$ and $s = w^\top x - b + S(w)$. The gradient flows through q .
- *Algorithm 1.* Each step updates $w \leftarrow (w - \gamma \nabla_w)_+$ and $b \leftarrow b - \gamma \nabla_b$ (plain SGD, no momentum), with training batches reshuffled every epoch by a seeded generator.
- *Validation loss and checkpointing.* The validation loss is the mean of per-batch mean losses over `algo_val`, and the (w, b) with the lowest validation loss is returned.

Class imbalance and the need for class-weighted baselines. Unweighted hinge-type objectives are minimized by a constant classifier on the two imbalanced datasets.

- *Law School* (89% positive). The unweighted linear SVM and SERM converge to $w = 0$ and accept every agent. On `algo_train`, the hinge loss of the accept-all classifier (0.2229) is not improved by any rescaling of the logistic-regression direction, and doubling the training data (18,718 rows) does not change this.
- *Adult* (24% positive). The unweighted linear SVM and SERM converge to the reject-all classifier.

In both cases no agent moves and the strategic improvement error equals the minority-class share (0.0913 and 0.2390, respectively). Balanced class weights restore non-trivial boundaries, and these variants are reported (marked [†] in Table 1).

Variance and repeated runs. Table 1 reports training seed 42 and label-sampling seed 42 for every method. The following additional analyses quantify variability.

- *Label-sampling variance.* Stage 5 redraws the post-response labels 1,000 times and reports mean $\pm$ standard deviation, e.g. Synthetic improvement rate 0.873 ± 0.022 and Law School 0.645 ± 0.053 (`outputs/<dataset>/evaluation/improvement_rate_algo_test.json`).
- *Training-seed variance.* Every method was retrained with seeds $\{42, 0, 1\}$ on Law School and Adult (`configs/sweeps/law_school/multiseed_table.yaml`, `configs/sweeps/adult/multiseed_table_3seeds.yaml`). The seed affects initialization, batch order and label sampling. Mean $\pm$ sample standard deviation (`ddof=1`) of the strategic improvement error:
 - *Law School:* STRAT-IMP-AWARE $4.39 \pm 0.34\%$; Attias et al. (2025) $5.41 \pm 0.11\%$; balanced SVM $5.35 \pm 0.29\%$; balanced SERM $27.76 \pm 2.03\%$.
 - *Adult:* STRAT-IMP-AWARE $15.89 \pm 0.24\%$; balanced SERM $15.52 \pm 0.14\%$; Attias et al. (2025) ($\tau = 0.7$) $17.85 \pm 0.05\%$; balanced SVM $63.09 \pm 0.61\%$.
- *Sample size.* On ACSIncome, Algorithm 1 was trained on nested random subsets of `algo_train` of size $\{50, 200, 1000, 3000, 6000, 88049\}$ with 10 repetitions each (`experiment_improvement_error.py`, `configs/sweeps/improvement_error.yaml`). The expected strategic improvement error on `algo_test` decreases monotonically: 0.170 ± 0.028 , 0.159 ± 0.015 , 0.106 ± 0.008 , 0.083 ± 0.007 , 0.077 ± 0.005 and 0.064 ± 0.0001 .

No cross-validation was used; all model and environment selection relies on `algo_val`, and `algo_test` is used only for reporting.

Reproduction. For a dataset `<d>`, run in order: `stage1_split_and_profile.py`, `stage2_train_lr.py`, `stage3_plot_lr_profiles.py`, `stage4_train_strat_imp.py` and `stage5_improvement_rate.py`, each with `-config configs/<d>.yaml`. Then run `experiment_multiseed_table.py -sweep configs/sweeps/<d>/multiseed_table.yaml` (Law School additionally with `-from-runs -seeds 42`). Environment sweeps are run with `experiment_clipped_improvement.py -sweep configs/sweeps/<d>.env.yaml`.

I IMPROVEMENT PROBABILITY FORMULATION

Let $x \sim \mathcal{D}, y | x \sim \text{Bernoulli}(\eta(x))$. We model improvement-aware strategic classification, such that the following condition holds:

$$\mathbb{P}(y^f = 1 | x) = \eta(x^f)$$

where y^f denotes the realized label after moving to point x^f through strategic manipulation, further:

$$\mathbb{P}(y^f = 1 | y = 1, x) = 1, \quad \mathbb{P}(y^f = 1 | y = 0, x) = p^f(x) \in [0, 1]$$

Now, by the law of total probability

$$\mathbb{P}(y^f = 1 | x) = \mathbb{P}(y^f = 1, y = 1 | x) + \mathbb{P}(y^f = 1, y = 0 | x)$$

$$\begin{aligned}
 &= \mathbb{P}(y^f = 1 \mid y = 1, x) \mathbb{P}(y = 1 \mid x) + \mathbb{P}(y^f = 1 \mid y = 0, x) \mathbb{P}(y = 0 \mid x) \\
 &= \eta(x) + p^f(x) \cdot (1 - \eta(x))
 \end{aligned}$$

From our modelling assumption, it follows that the improvement probability is $p^f(x) = \frac{\eta(x^f) - \eta(x)}{1 - \eta(x)}$. When $\eta(x) = 1$ the event $\{y = 1\}$ has conditional probability one and $p^f(x)$ is set to 0; since $x^f \geq x$ coordinate wise under Assumption 2 and the tie-breaking rule of Section 3, Assumption 1 gives $\eta(x^f) \geq \eta(x)$, thus $\eta(x^f) = 1$, and the identity $\eta(x^f) = \eta(x) + p^f(x)(1 - \eta(x))$ continues to hold. To simulate the improvement probability in Algorithm 2, we utilize $p^f(x)$ as the probability that an unqualified agent ($y = 0 \mid x$) successfully flips their label, provided they strategically manipulated. The post-manipulation y^f is determined by a Bernoulli trial with parameter $p^f(x)$. Formally:

$$y^f \mid (y, x = 0) \sim \text{Bernoulli}(p^f(x))$$

This ensures that label flipping is intrinsically tied to the strategic transition from $x \rightarrow x^f$, maintaining the probabilistic consistency of the improvement-aware framework.

J ADDITIONAL THEORETICAL RESULTS

Proposition J.1. *Consider a scoring function $h(x) = \sum_{i=1}^d \alpha_i \psi(x_i)$ with $\alpha \in \mathbb{R}_{>0}^d$. Let c_{sep} and c_{dec} denote separable and decomposable cost functions with respect to h and ψ , respectively. Let the feature transformation functions ψ_i are non-decreasing i.e. if $x'_i \geq x_i$ coordinate-wise then i.e., $\psi_i(x'_i) \geq \psi_i(x_i)$ for all $i \in [d]$. Then, for any $x, x' \in \mathcal{X}$, the costs satisfy $c_{sep}(x, x') = c_{dec}(x, x')$.*

Proof. Let $x, x' \in \mathcal{X}$. Define the feature-wise deviation $\delta_i = \psi(x'_i) - \psi(x_i)$.

Recall the definition of the decomposable cost:

$$c_{dec}(x, x') = \sum_{i=1}^d \alpha_i |\psi(x'_i) - \psi(x_i)|_+ = \sum_{i=1}^d \alpha_i \max(0, \delta_i)$$

Similarly, recall the definition of the separable cost, substituting the scoring function structure:

$$\begin{aligned}
 c_{sep}(x, x') &= \max \left(0, \sum_{i=1}^d \alpha_i \psi(x'_i) - \sum_{i=1}^d \alpha_i \psi(x_i) \right) \\
 &= \max \left(0, \sum_{i=1}^d \alpha_i \delta_i \right)
 \end{aligned}$$

We now invoke the monotonicity condition, then $\psi(x'_i) \geq \psi(x_i)$ for all $i \in [d]$. This implies $\delta_i \geq 0$ for all i . Under this condition, we can clearly conclude that:

$$c_{dec}(x, x') = \sum_{i=1}^d \alpha_i \delta_i = c_{sep}(x, x')$$

□

Observation J.2. *Suppose that, for every $x \in \mathcal{X}_0$ and $z \in \mathcal{A}$, every point lying coordinate-wise between x and z belongs to $\mathcal{A}$. Let $f_{w,b} \in \mathcal{F}$ be a linear classifier with $w \in \mathbb{R}_{\geq 0}^d$ and $b \in \mathbb{R}$. Suppose a strategic best response exists for every $x \in \mathcal{X}_0$. Then, for every $x \in \mathcal{X}_0$, there exists a strategic best response x^f satisfying $x_i^f \geq x_i$ for all $i \in [d]$.*

Proof. For brevity, let $f = f_{w,b}$. Fix $x \in \mathcal{X}_0$, let $z \in \mathcal{A}$ be any strategic best response to f . Define $\tilde{z} \in \mathbb{R}^d$ by

$$\tilde{z}_i = \max\{x_i, z_i\}, \quad i \in [d]$$

$\tilde{z} \in \mathcal{A}$: For each i , $\tilde{z}_i \in \{x_i, z_i\}$, so $\tilde{z}_i$ lies between x_i and z_i . Hence $\tilde{z}$ lies coordinate-wise between x and z , and by the assumption on $\mathcal{A}$, $\tilde{z} \in \mathcal{A}$.

$\tilde{z}$ has the same cost as z : If $z_i \geq x_i$, then $\tilde{z}_i = z_i$, so $(\tilde{z}_i - x_i)_+ = (z_i - x_i)_+$. If $z_i < x_i$, then $\tilde{z}_i = x_i$, and both $(\tilde{z}_i - x_i)_+$ and $(z_i - x_i)_+$ equal to 0. Thus $(\tilde{z}_i - x_i)_+ = (z_i - x_i)_+$ for every $i \in [d]$, and therefore

$$c(x, \tilde{z}) = \sum_{i=1}^d \alpha_i (\tilde{z}_i - x_i)_+ = \sum_{i=1}^d \alpha_i (z_i - x_i)_+ = c(x, z) \quad (18)$$

$\tilde{z}$ is classified as least as favorably as z : Since $\tilde{z} \geq z$ coordinate-wise and $w \geq 0$,

$$w^\top \tilde{z} \geq w^\top z, \quad \text{hence} \quad f(\tilde{z}) \geq f(z) \quad (19)$$

$\tilde{z}$ is a best response: Using $\beta > 0$ together with (18) and (19),

$$u_f(x, \tilde{z}) = \beta f(\tilde{z}) - c(x, \tilde{z}) \geq \beta f(z) - c(x, z) = u_f(x, z).$$

Since z maximizes $u_f(x, \cdot)$ over $\mathcal{A}$ and $\tilde{z} \in \mathcal{A}$, the point $\tilde{z}$ is also a maximizer, *i.e.*, a strategic best response. Finally, $\tilde{z}_i = \max\{x_i, z_i\} \geq x_i$ for all $i \in [d]$ by construction. Taking $x^f = \tilde{z}$ proves the claim. $\square$