One-Shot Strategic Classification Under Unknown Costs

Elan Rosenfeld [1] **Nir Rosenfeld** [2]

Abstract

The goal of strategic classification is to learn decision rules which are robust to strategic input manipulation. Earlier works assume that these responses are known; while some recent works handle unknown responses, they exclusively study online settings with repeated model deployments. But there are many domains—particularly in public policy, a common motivating use case—where multiple deployments are infeasible, or where even one bad round is unacceptable. To address this gap, we initiate the formal study of *one-shot* strategic classification under unknown responses, which requires committing to a single classifier once. Focusing on uncertainty in the users' cost function, we begin by proving that for a broad class of costs, even a small mis-estimation of the true cost can entail trivial accuracy in the worst case. In light of this, we frame the task as a minimax problem, aiming to minimize worst-case risk over an uncertainty set of costs. We design efficient algorithms for both the full-batch and stochastic settings, which we prove converge (offline) to the minimax solution at the rate of $\tilde{\mathcal{O}}(T^{-1/2})$. Our analysis reveals important structure stemming from strategic responses, particularly the value of *dual norm regularization* with respect to the cost function.

1. Introduction

Across a multitude of domains, machine learning is increasingly being used to inform decisions about human users. But when users stand to gain from certain predictive outcomes, they may act to obtain favorable predictions by modifying their features. Since this can harm predictive performance, learning becomes susceptible to *Goodhart's law*, which states that "when a measure becomes a target, it ceases to be a good measure" (Goodhart, 1975). This natural tension between learning systems and their users applies broadly: loan approvals, admissions, hiring, insurance, and welfare benefits are all examples in which the system seeks predictions that are accurate, whereas users—irrespective of their true label—wish to be classified as positive.

The field of *strategic classification* (Brückner et al., 2012; Hardt et al., 2016) studies learning in such settings, with the principal aim of learning classifiers that are robust to strategic user behavior. However, most works in this field rely on the key assumption that the learner knows precisely how users would respond to any given classifier. This is typically modeled as knowledge of the underlying *cost function* $c(x, x')$ which defines the cost users incur for modifying features x to become x'. This assumption enables tractable learning, but it is unrealistic and it effectively takes the "sting" out of Goodhart's law: in a statistical sense, the power to anticipate user responses completely nullifies the effect of users' gaming behavior on predictive performance (for radial basis cost functions; see Appendix G).

Indeed, if we survey some well-known examples of Goodhart's law (e.g. Chrystal et al., 2003; Elton, 2004; Fire and Guestrin, 2019; Teney et al., 2020), it becomes apparent that much of the policy challenge arises precisely from *not* knowing how users would respond. This motivates us to instead focus on learning strategically robust classifiers under an *unknown* cost function. To cope with this uncertainty, we take a conservative approach and model the system as aiming to learn a classifier that is robust to all cost functions c in some uncertainty set $\mathcal{C}$, which we think of as including all cost functions which are believed to be plausible (alternatively, it is a set which we can be confident will contain the true, unknown c). Our approach is therefore doubly robust, providing guarantees under both the manipulation of inputs by strategic behavior and an adversarial choice of cost function. We argue this is necessary: we prove that if one optimizes for strategic classification under a single, fixed cost, then any discrepancy between that cost and the true cost can result in dramatically reduced accuracy.

While some works have studied strategic learning under unknown responses (Dong et al., 2018; Ahmadi et al., 2021; Shao et al., 2023; Lechner et al., 2023; Harris et al., 2023), their focus is entirely on sequential learning settings. This

[1]Carnegie Mellon University [2]Technion – Israel Institute of Technology. Correspondence to: Elan Rosenfeld <elan@cmu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

allows them to cope with response uncertainty via exploration: deploying a series of models over time, observing how users respond, and adapting accordingly. Algorithms for such online settings are designed to ensure that regret decays sufficiently fast with the number of rounds—but they provide no guarantees on worst-case outcomes for any single deployment. We observe that there are many realistic settings in which multiple deployments are too costly or technically impossible (e.g. financial regulation); in which arbitrary exploration is unrealistic or unethical (e.g. testing in education); in which there is need for immediately beneficial short-term outcomes (e.g. epidemic vaccination); or in which even a single bad round could be very harmful (e.g. environmental conservation).

Motivated by such examples, this work studies robust strategic learning in a "one-shot" setting where the learner must commit to one single model at deployment. In this setting, we recast our objective as optimizing for a classifier that minimizes the worst-case strategic risk over an uncertainty set $\mathcal{C}$. We devise two efficient algorithms, for the full-batch and stochastic settings, which provably converge to the minimax solution at the rate $\tilde{\mathcal{O}}(T^{-1/2})$. Notably, this rate is dimension independent for the typical cost functions—including ℓ_1 and ℓ_2 norms—and it is achieved despite the inner maximization being non-concave. A key step in our approach is to adapt the *strategic hinge loss* (Levanon and Rosenfeld, 2022) to properly handle a large set of possible costs beyond the ℓ_2 norm. As the strategic hinge loss is non-convex, we apply a regularization term that accounts for the unique structure of strategic response: our analysis uncovers that the "correct" form of regularization is in fact the *dual norm of β with respect to the cost*. We complement this with an updated generalization bound for this loss which corrects an error in that previous work.

More broadly, our approach relies on the observation that strategic responses induce a shift in the data distribution (Perdomo et al., 2020). Because different costs induce different shifts, uncertainty over cost functions translates to uncertainty over test distributions. This allows us to formulate our problem as one of *distributionally robust optimization* (DRO) (Namkoong and Duchi, 2016; Duchi and Namkoong, 2021), where the goal is to predict well on the worst-case distribution within some given set. Whereas the typical DRO formulation defines the uncertainty set with respect to a more traditional probability divergence, in our case the set of distributions is inherited from the structure of strategic responses and includes all shifts that can result from strategic behavior under any $c \in \mathcal{C}$. This means that the entire *set* of possible distributions becomes dependent on the classifier, which is one of the primary challenges our work addresses.

1.1. Related work

Strategic classification. Introduced in Hardt et al. (2016), and based on earlier works (Brückner and Scheffer, 2009; Brückner et al., 2012; Großhans et al., 2013), the literature on strategic classification has since been growing steadily. Focusing on supervised classification in the batch (offline) setting, here we list a relevant subset. Advances have been made on both statistical (Zhang and Conitzer, 2021; Sundaram et al., 2021) and algorithmic aspects of learning (Levanon and Rosenfeld, 2021), but the latter lacks guarantees. In contrast, our work provides efficient algorithms that are provably correct. Efforts have also been made to extend learning beyond the basic setting of Hardt et al. (2016). These include: accounting for users with noisy estimates (Jagadeesan et al., 2021), missing information (Ghalme et al., 2021; Bechavod et al., 2022), more general preferences (Sundaram et al., 2021; Levanon and Rosenfeld, 2022; Eilat et al., 2023); incorporating causal elements into learning (Miller et al., 2020; Chen et al., 2023; Horowitz and Rosenfeld, 2023; Mendler-Dünner et al., 2022), and considering societal implications (Milli et al., 2019; Levanon and Rosenfeld, 2021; Lechner and Urner, 2022).

(Initially) unknown user responses. Several works have considered the case of inferring unknown user responses under different forms of strategic learning, but in online or sequential settings. Dong et al. (2018) bound the Stackelberg regret of online learning when both costs and features are adversarial, but when only negatively-labeled users respond; Harris et al. (2023) bound a stronger form of regret when responses only come from positively-labeled users. Ahmadi et al. (2021) propose a strategic variant of the perceptron and provide mistake bounds for ℓ_1 norm with unknown diagonal scaling or ℓ_2 norm multiplied by an unknown constant. Shao et al. (2023) study learning under unknown ball manipulations with personalized radii, and give mistake bounds and (interactive) sample complexity bounds for different informational structures. Lechner et al. (2023) bound the sample and iteration complexity of learning under general, non-best response manipulation sets via repeated model deployments to infer the manipulation graph. Lin and Zrnic (2023) learn under a misspecified response model, inferred in a (nonadaptive) exploration phase from some class of models. The analyses in these works better reflect reality in that the exact response cannot be known. But online deployment and long-term regret minimization are not appropriate for many natural use cases of strategic classification, which motivates our investigation of the one-shot setting.

2. Preliminaries

Notation. We study the problem of linear strategic classification on a distribution over inputs $x \in \mathcal{X} \subseteq \mathbb{R}^d$ and labels $y \in \mathcal{Y} = \{\pm 1\}$ where the exact response of the test popula-

tion is unknown. We consider linear classifiers $\text{sign}(\beta^\top x)$, optimized over some bounded set $\mathcal{B} \subset \mathbb{R}^{d+1}$ (this includes a bias term which we leave implicit and which is not included in the vector norm). To model users' responses, we consider the typical setup of a *cost function* $c(x, x')$ which defines the cost for an agent to change their features from x to x'. Together with a utility $u \geq 0$ gained from a positive classification, this cost determines the strategic response of a rational agent to a classifier β via

$$x(\beta) := \arg\max_{x'} \left[\mathbf{1}\{\beta^\top x' \geq 0\} \cdot u - c(x, x')\right]. \quad (1)$$

Thus, an agent will move only if it would change their classification from negative to positive, and only if $c(x, x') < u$. We handle non-uniqueness of the argmax by breaking ties arbitrarily, but we follow convention by assuming no strategic response if the net utility is exactly 0. We let $\delta(x; \beta) := x(\beta) - x$ denote the movement of an agent, and we suppress dependence of δ on x, β where clear from context.

Form of the cost function. We study cost functions that can be written as $c(x, x') = \phi(\|x' - x\|)$ for a norm $\|\cdot\|$ and non-decreasing function $\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$. This includes the ℓ_2-norm (by far the most common cost, sometimes squared), but it is also much more general—it allows for different non-linear transformations which may better reflect real-world costs, such as $\phi(r) = \ln(1 + r)$, and we allow $\|\cdot\|$ to denote any differentiable and monotonic norm, which includes all p-norms (where $p \in [1, \infty]$). This significantly expands upon the costs discussed in the literature thus far.

Parameterizing the space of costs. Recall that we are interested in robust prediction when we cannot know the exact strategic response. Keeping the assumption of rational behavior, this naturally points to an unknown cost function as a primary source of this uncertainty. We parameterize this uncertainty via an unknown positive definite (PD) matrix $\Sigma \succ \mathbf{0}$ in $\mathbb{R}^{d \times d}$ which scales the relative costs per input dimension, denoting the induced norm as $\|\cdot\|_\Sigma$. For the 2-norm, this is the standard PD norm given by $\|x\|_\Sigma := \sqrt{x^\top \Sigma x}$, but we define it for general norms as $\|x\|_\Sigma := \|\Sigma^{1/2} x\|$. Hence, any cost function c is uniquely determined by its parameterization Σ as $c(x, x') = \phi\left(\|\Sigma^{1/2}(x' - x)\|\right)$; we will use these two variables interchangeably. The other two factors in strategic response are the positive utility u and the monotone transform ϕ: for reasons that will soon be made clear, we only require knowledge of the maximum value u_* such that $\phi(u_*) \leq u$ (this is clearly satisfied for known ϕ and u, as is typically assumed). Note this value need not be shared among users, and it suffices to know an interval in which it lies—but for simplicity we treat it as fixed.

Encoding uncertainty. Since the true cost is not known, and since we cannot estimate it in an online fashion, we instead assume a system-specified *uncertainty set* $\mathcal{C}$, defined as a compact, convex set of possible costs c which is expected to contain the true cost. The goal of our analysis will be to derive strategies for efficiently identifying a classifier which ensures optimal (and boundable) worst-case test performance over *all* costs $c \in \mathcal{C}$, and therefore also bounded error under the true cost. Notably, this also means that *even if the true cost changes over time*, our error bound will hold so long as the cost remains within $\mathcal{C}$. In practice we want $\mathcal{C}$ to be broad enough that we can be confident it contains the test-time cost—but we will also see that our convergence guarantees scale inversely with the diameter of this set, so it should be selected to be no larger than necessary.

2.1. Strategic Learning Under a Single, Known Cost

As a first step towards learning robustly under all costs in $\mathcal{C}$, it will be useful to first consider learning under a single fixed cost. For a *known* cost c, the typical goal would be to minimize the 0-1 loss under strategic response with this cost: $\ell^c_{0-1}(\beta^\top x, y) := \mathbf{1}\{\text{sign}(\beta^\top(x + \delta)) \neq y\}$. It is common to instead consider a more easily optimized surrogate such as the hinge loss $\ell_{\text{hinge}}(\beta^\top x, y) := \max\{0, 1 - y\beta^\top(x + \delta)\}$, but the discontinuous nature of $\delta(x; \beta)$ w.r.t. β means that optimization is still intractable. Instead, we make use of the recently proposed *strategic hinge loss* (Levanon and Rosenfeld, 2022) which augments the standard hinge loss with an additional term to account for this discontinuity: $\ell_{\text{s-hinge}}(\beta^\top x, y) := \max\{0, 1 - y(\beta^\top x + 2\|\beta\|_2)\}$ (note this does not explicitly include δ).

Unfortunately, even this relaxation poses difficulties. Firstly, the objective is non-convex—though Levanon and Rosenfeld (2022) show that for known costs it often learns reasonable classifiers in practice, once we transition to unknown costs it will become clear that having a guaranteed sub-optimality bound, which requires convexity, is important. Second, the additional term $2\|\beta\|_2$ captures the effect of strategic response only under the standard ℓ_2-norm cost.[1]

Cost-aware strategic hinge. For our setting, we derive a more general strategic hinge loss that applies to the broader class of costs. This loss admits a natural form which relies on the *dual norm* of β with respect to the cost function:

Definition 2.1. The Σ-*transformed dual norm* of β is denoted $\|\beta\|_{*,\Sigma} := \sup_{\|v\|_\Sigma = 1} \beta^\top v = \|\Sigma^{-1/2}\beta\|_*$. We may leave dependence on Σ implicit, writing simply $\|\beta\|_*$.

Definition 2.2. The *cost-dependent strategic hinge loss* is:

$$\ell^c_{\text{s-hinge}}(\beta^\top x, y) := \max\{0, 1 - y(\beta^\top x + u_* \|\beta\|_*)\}. \quad (2)$$

Note our proposed $\ell^c_{\text{s-hinge}}$ generalizes the previous $\ell_{\text{s-hinge}}$ since the ℓ_2-norm is its own dual. The appearance of the dual norm here is not by chance. This quantity captures the

[1] While Levanon and Rosenfeld (2022) do discuss more general costs, they give only a generic description.

dimension-wise sensitivity of our decision rule to changes in x, scaled *inversely proportionally* to the cost an agent incurs for moving in that dimension. This should be intuitive: for any given direction, the less it costs a user to modify their features, the more they can afford to move, and thus the greater the importance of reducing our classifier's sensitivity to it. We make this formal with the following lemma which bounds the maximal strategic change to inputs.

Lemma 2.3. *Fix β and let $c(x, x') = \phi(\|x - x'\|_\Sigma)$. The maximum possible change to a user's score due to strategic behavior is $u_* \|\beta\|_{*,\Sigma}$.*

Proofs of all lemmas can be found in Appendix C. Thus we see how augmenting the hinge loss with the dual norm serves as a natural approach to robust strategic classification, and we use this loss throughout. More generally, it will also be useful to define the *k-shifted strategic hinge loss* as $\ell_{\text{s-hinge}}(\beta; k) := \max\{0, 1 - y(\beta^\top x + k)\}$.

Though $\ell^c_{\text{s-hinge}}$ allows for more general costs, it remains non-convex—we will return to this point in Section 4.2. Also note that the "correct" transformation to use depends on the true cost. In our case this is unknown, but we show that it suffices for the dual norm to be bounded by a constant B: $\forall c \in \mathcal{C},\ \|\beta\|_* \leq B$. We also let $X \in \mathbb{R}$ denote the maximum of $\|x\|, \|x\|_2$ over the training examples. We treat both B and X as fixed constants. Finally, we denote by L the Lipschitz constant of the loss gradient, which appears in the convergence rates of the algorithms we derive. This can be generically bounded as $L \leq X + u_* L_*$, where L_* is the Lipschitz constant of the dual norm. For the common setting where $\|\cdot\|$ is a p-norm, we have $L_* = \max\left(1, d^{1/2 - 1/p}\right)$. Notably, this quantity is independent of the dimension d for $p \leq 2$ and scales no worse than $d^{1/2}$ otherwise.

Risk and generalization. Denote the overall strategic hinge risk by $R^c_{\text{s-hinge}}(\beta) := \mathbb{E}[\ell^c_{\text{s-hinge}}(\beta^\top x, y)]$, with the strategic 0-1 risk defined analogously as R^c_{0-1}.

Lemma 2.4. *For any cost $c \in \mathcal{C}$, $R^c_{0-1}(\beta) \leq R^c_{\text{s-hinge}}(\beta)$.*

Thus, our cost-dependent strategic hinge loss is an effective proxy for the 0-1 loss. We next establish its generalization.

Theorem 2.5. *With probability $\geq 1 - \delta$, for all $\beta \in \mathcal{B}$ and all cost functions $c \in \mathcal{C}$,*

$$R^c_{0-1}(\beta) \leq \hat{R}^c_{\text{s-hinge}}(\beta) + \frac{B(4X + u_*) + 3\sqrt{\ln 1/\delta}}{\sqrt{n}},$$

where $\hat{R}$ is the empirical risk over a training set of size n.

This result extends (and fixes an error in) the bound for ℓ_2-norm cost from Levanon and Rosenfeld (2022). The proof, found in Appendix D, applies standard Rademacher bounds by decomposing the strategic hinge loss and bounding the terms separately while accounting for general strategic responses. The fact that this bound holds uniformly for *all* cost functions is critical, as it allows us to apply it to the worst case cost even when that cost is unknown.

3. The Perils of Using a Wrong Cost Function

As the agents' movement depends intricately on the precise cost function, correctly anticipating strategic response requires a good estimate of that cost. In the one-shot setting, this is further complicated by the fact that we have no access to a mechanism by which to *infer* the cost (e.g., via online interaction): we must pick a single classifier and commit to it, without knowing the true cost function a priori. A natural approach would be to make use of existing machinery for single-cost strategic learning, as described above, using some reasonable choice for the cost. For example, one idea would be to simply pick a cost which we believe is reasonably "close" to the true cost, in the hope that predictive performance degrades gracefully with our error. Certainly, this is better than blindly proceeding with the default ℓ_2-norm. Unfortunately we find that the task of cost-robust strategic learning is much more difficult (learning theoretically) than is first apparent—it turns out that without more explicit assumptions on our guess's distance to the true cost *and* on the data distribution itself, providing any sort of robustness guarantee is impossible.

Hardness results. Our first result proves that if we must to commit to a single fixed cost, unless that cost is exactly correct, minimizing the empirical risk can never provide a non-trivial data-independent guarantee:

Theorem 3.1. *Consider the task of learning a norm-bounded binary linear classifier. Fix any two costs c_1, c_2 with non-equal cost matrices, and let $0 \leq \epsilon \leq \frac{1}{2}$. There exists a distribution q over $\mathcal{X} \times \mathcal{Y}$ such that:*

1. *For each of c_1 and c_2, there is a (different) classifier which achieves 0 error on q when facing strategic response under that cost; and*
2. *Any classifier which achieves 0 error on q under c_1 suffers error ϵ under c_2, and any classifier which achieves 0 error on q under c_2 suffers error $1 - \epsilon$ under c_2.*

Theorem 3.1 says that if there is *any error at all* in estimating Σ, then any classifier which achieves perfect accuracy under our assumed cost might do no better than random (or even worse than random) when deployed. The proof (see Appendix A) constructs a worst-case distribution that is chosen adversarially with respect to the error in cost estimation. This construction is quite robust in that there exists an entire space of such solutions and the lower bound decays smoothly for distributions which are close to the one we define (especially for larger error in the estimate of Σ). However, there is hope that perhaps for non-adversarial distributions, estimating a fixed cost may be reasonable. To investigate this possibility, we also study the more natural

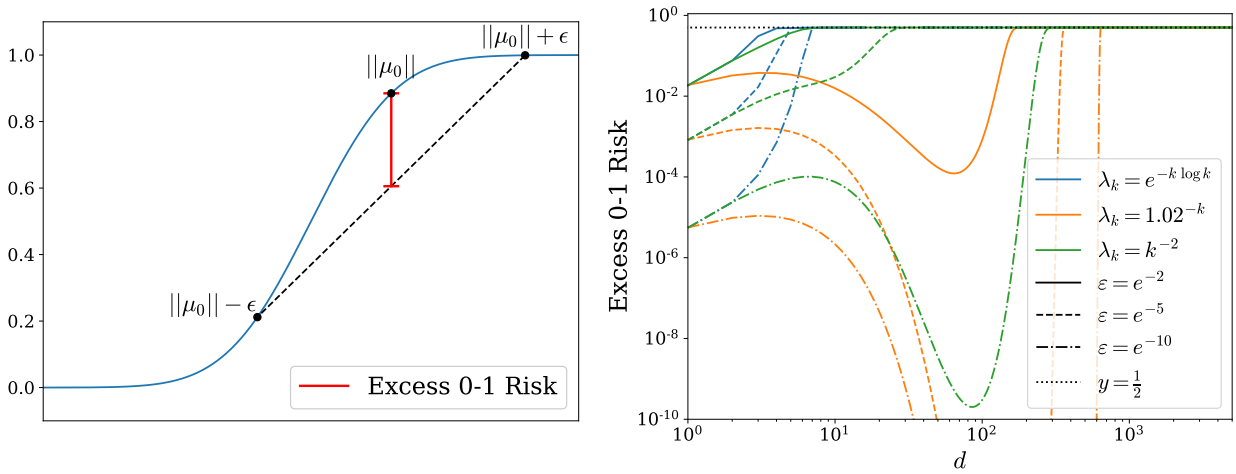

Figure 1. **Visualizing the excess 0-1 risk from Theorem 3.2. Left:** A toy illustration of the excess risk as a function of $\|\mu_0\|$ and ϵ and where they lie on the Gaussian CDF. **Right:** Excess risk curves as a function of the dimension d, where $\mu_0 = 1/\sqrt{d}\,\mathbf{1}$ and $\sigma^2 = 1/d$. Color indicates the function which generates the spectrum of Σ, where λ_k is the k-largest eigenvalue in the series. Line style indicates the error ε in estimating the eigenvalue in each dimension—precisely, we define $\hat{\Sigma} = (1-\varepsilon)\Sigma$. Even with all eigenvalues estimated to error $1 - e^{-10}$, excess risk grows rapidly with d towards $1/2$.

setting of isotropic Gaussian $q(x \mid y)$. We find once again that, even in the limit of infinite data, guessing the wrong cost can be quite harmful:

Theorem 3.2. *Define the distribution q over $\mathcal{X} \times \mathcal{Y}$ as $q(y=1) = q(y=-1) = 1/2$, $q(x \mid y) \overset{d}{=} \mathcal{N}(y \cdot \mu_0, \sigma^2 I)$. Denote by Φ the standard Normal CDF. Let the true cost be defined as $\|x - x'\|_\Sigma$ with unknown cost matrix Σ, and let β^* be the classifier which minimizes the strategic 0-1 risk under this cost.*

Suppose one instead learns a classifier $\hat{\beta}$ by assuming an incorrect cost $\hat{\Sigma}$ and minimizing the population strategic 0-1 risk under that cost: $\hat{\beta} := \arg\min_\beta \mathbb{E}_q[\ell^{\hat{c}}_{0-1}(\beta)]$. Then the excess 0-1 risk suffered by $\hat{\beta}$ is

$$\Phi\left(\frac{\|\mu_0\|}{\sigma}\right) - \frac{1}{2}\left(\Phi\left(\frac{\|\mu_0\| - \epsilon}{\sigma}\right) + \Phi\left(\frac{\|\mu_0\| + \epsilon}{\sigma}\right)\right),$$

where $\epsilon := \frac{u_* \left| \|\mu_0\|_{*,\hat{\Sigma}} - \|\mu_0\|_{*,\Sigma} \right|}{\|\mu_0\|}$.

Theorem 3.2 demonstrates that even for *much* more benign distributions, choosing a classifier based on a slightly incorrect cost can be very non-robust. Roughly, we should expect $\epsilon = \Theta(u_* \mathrm{tr}(\hat{\Sigma}^{-1} - \Sigma^{-1}))$, scaling as our error grows along directions where μ_0 is large. Crucially, this is with respect to the *inverse* cost matrix Σ^{-1}, which means that a tiny estimation error in the lower end of the eigen-spectrum can have a large effect. For example, imagine that the classes are sufficiently separated so that the optimal classifier *absent* strategic behavior can get accuracy close to 1 (which is trivial in high dimensions). Then under strategic behavior, the error in guessing the smallest eigenvalue of Σ need only be $\Omega(\|\mu_0\| \lambda_{\min}(\Sigma) \lambda_{\min}(\hat{\Sigma}))$ to induce excess risk of approximately $1/4$, and it will rapidly approach $1/2$ as d grows. To give a bit more intuition for these factors, Figure 1 visualizes a few simple examples of how the excess risk behaves as dimension increases, depending on the distribution of eigenvalues of Σ and our error in estimating them. More abstractly, we can see that if we only slightly err in estimating the cost of movement in a direction, it could cause a very large error in estimating how much agents will move in that direction.

Intentionally conservative guesses can still fail. It is not immediately obvious why guessing a single cost should always have poor worst-case performance. For example, it might appear that we could just select a cost function which *under*estimates the cost to move in any given direction, and therefore overestimates the movement of any given point. In being intentionally conservative, this choice seems like it should be reasonably robust to misspecification, even if not optimally so. However, this ignores the key fact that in strategic classification there is also the potential for *beneficial* strategic response: a point with true label $y = 1$ but incorrect prediction $\hat{y} = -1$ has the potential to correct this error—to the learner's benefit—by changing its features.

The above results show that it is not enough in strategic classification to be aware of the fact that users will move in response to the classifier, nor to have an educated guess for how they will move: it is essential to account for *uncertainty* in how they will move by anticipating the ways in which the true, unknown cost function may differ from what we expect. Our lower bounds thus motivate robustly learning a classifier

based not on a single cost, but on a *set* in which the true cost is expected to lie—this reduces potential misspecification of the response model (Lin and Zrnic, 2023) while remaining applicable in the one-shot setting.

4. Maximizing Risk for a Fixed Classifier and Minimizing Risk for a Fixed Cost

To identify a robust classifier when the true cost is unknown, we will optimize the worst-case strategic risk with respect to the learner-specified uncertainty set $\mathcal{C}$. Our objective is therefore to solve the min-max problem

$$\min_{\beta} \max_{c \in \mathcal{C}} R^{c}_{\text{s-hinge}}(\beta). \tag{3}$$

As we noted in the introduction, if we consider the cost function to be inducing a classifier-specific distribution, then Equation (3) becomes an instance of *distributionally robust optimization*. A typical approach to this type of problem would be to use a gradient-based method which converges to a Nash equilibrium between the learner and an adversary (who "chooses" the cost). To apply this in our setting, we must recast Equation (3) as two separate objectives: one for the learner minimizing β and the other for the adversary maximizing c. However, the idiosyncrasies of the strategic learning setting give rise to unique challenges which must be addressed for this approach to be successful.

4.1. Solving the Inner Max for a Fixed Classifier

The first step to solving Equation (3) is determining a way to solve the inner maximization problem, which will serve as an important subroutine for our eventual algorithm which solves the full min-max objective. However, the max objective is non-concave. We thus begin with an efficient algorithm to address this. Algorithm 1 gives pseudocode[2] for MaxLossCost which, for a given classifier β, maximizes the strategic hinge loss over costs $c \in \mathcal{C}$. Due to Equation (2), this amounts to finding the maximizing k-shift for $k = u_* \|\beta\|_*$ where $\|\beta\|_*$ is achievable by some $c \in \mathcal{C}$. We next prove correctness and runtime:

Lemma 4.1. *For any classifier β, dataset $(\mathcal{X} \times \mathcal{Y})^n$, and uncertainty set $\mathcal{C}$,* MaxLossCost *runs in $\mathcal{O}(nd + n \ln n)$ time and returns a cost function $c \in \mathcal{C}$ which maximizes the cost-dependent strategic hinge risk $\hat{R}^{c}_{\text{s-hinge}}(\beta)$.*

MaxLossCost takes advantage of the fact that the loss is piecewise linear in the *scalar value* $\|\beta\|_*$, carefully constructing a list of the training samples which are sorted according to a particular intermediate value and traversing this list while updating the risk at the boundary of each linear component. For a full description of the algorithm see Appendix E.

[2]The code calls for identifying $\|\beta\|_*^{\min}$ and $\|\beta\|_*^{\max}$. The specific way we parameterize $\mathcal{C}$ makes this simple (see Section 5.1).

Algorithm 1 Pseudocode for MaxLossCost

input: Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, Classifier β, Cost uncertainty set $\mathcal{C}$, Upper bound u_*.
define: $\|\beta\|_*^{\min} := \min_{\Sigma \in \mathcal{C}} \|\beta\|_*$, $\|\beta\|_*^{\max} := \max_{\Sigma \in \mathcal{C}} \|\beta\|_*$
initialize: $k \leftarrow \|\beta\|_*^{\min}$, $r \leftarrow \hat{R}_{\text{s-hinge}}(\beta; u_* k)$
  Sort training points in increasing order by $v_i := y_i(1 - y_i \beta^\top x_i)$.
  Set j as first index where $v_j - u_* k > 0$.
  for entry v_i in sorted list, starting from j **do**
    Set $k = v_i / u_*$.
    If $k > \|\beta\|_*^{\max}$, break.
    Update new risk $r = \hat{R}_{\text{s-hinge}}(\beta; u_* k)$. Maintain maximum $r_{\max}$ and $k_{\max}$ which induced it.
  end for
  If $\hat{R}_{\text{s-hinge}}(\beta; u_* \|\beta\|_*^{\max}) > r_{\max}$, return $\arg\max_{\Sigma \in \mathcal{C}} \|\beta\|_*$
  Otherwise, return $\Sigma \in \mathcal{C}$ such that $\|\beta\|_{*,\Sigma} = k_{\max}$.

Bounding the adversarial risk of any classifier. Independent of our main objective, MaxLossCost already allows us to derive adversarial strategic risk bounds for any classifier: simply evaluate its worst-case empirical strategic hinge loss and then apply Theorem 2.5. In principle, this means we could *try* minimizing the strategic hinge loss for a single cost and we would technically be able to derive an error bound for the solution. However, this approach is not ideal for two reasons: first, as noted earlier, the strategic hinge loss is non-convex in β, so direct optimization may be unsuccessful. Second, as shown in Section 3 there is little reason to believe that optimizing an objective which does not account for the min-max structure will achieve good adversarial risk, making the generalization bound correct but usually unhelpful (e.g., trivial). Still, it is useful to be able to give a valid bound on the adversarial risk of whatever classifier we may hope to evaluate.

4.2. Solving the Outer Min for a Fixed Cost

We now switch to the task of finding a classifier which minimizes worst-case risk. Note Equation (3) poses two key challenges: (i) the strategic hinge loss is non-convex, and (ii) optimization is further complicated by the inner max operator. Hence, we begin with the case of a single fixed cost, which will serve us as an intermediary step towards optimizing over the set of all costs. Even for a single cost, the non-convexity of the strategic hinge means that we would only be able to guarantee convergence to a stationary point. We address this via regularization.

Convexification via regularization. Regularization is a standard means to introduce (strong) convexity: for the hinge loss, applying (squared) ℓ_2 regularization makes the

objective strongly convex, improving optimization while simultaneously preventing overfitting. Since we would want to regularize our objective anyways, one possible solution would be to add just enough ℓ_2 regularization to convexify it. While sound in principle, this is inappropriate for our setting because it does not account for the cost-specific nature of strategic response, so the amount of regularization needed would be extreme. Specifically, we prove a lower bound on the required ℓ_2 regularization to ensure convexity of the strategic hinge loss:

Theorem 4.2. *Let $\|\cdot\|$ be a p-norm and fix a cost matrix Σ. There is a distribution q such that the ℓ_2-regularized loss $R^c_{\text{s-hinge}}(\beta) + \lambda u_* \|\beta\|_2$ is non-convex unless $\lambda \geq \|\Sigma^{-1}\|_2$.*

Thus, the necessary ℓ_2 regularization would be cost-dependent and quite large, scaling with the inverse of the smallest eigenvalue. Moreover, since we want to optimize this risk over all possible costs, this means that it would need to scale with the largest spectral norm among *all* inverse cost matrices in $\mathcal{C}$.

Dual norm regularization. Luckily, we observe that for a given cost c we can instead apply a small amount of *dual norm* regularization to β. This naturally accounts for the problem's structure and eliminates the need for excessive penalization. Our dual-regularized objective is $R^c_{\text{s-hinge}}(\beta) + \lambda u_* \|\beta\|_*$, where λ determines the regularization strength. Since the dual norm $\|\beta\|_*$ is implicitly defined via the cost matrix, the following is straightforward:

Proposition 4.3. *For any norm $\|\cdot\|$, cost matrix Σ, and distribution q, the dual-regularized loss $R^c_{\text{s-hinge}}(\beta) + \lambda u_* \|\beta\|_*$ is guaranteed to be convex for all $\lambda \geq q(y=1)$.*

In fact, we can show that $q(y=1)u_*\|\beta\|_*$ is the *minimum* regularization that ensures convexity: for any λ, if the objective in Theorem 4.2 is convex, then $q(y=1)\|\beta\|_* \leq \lambda \|\beta\|_2$. Proofs are in Appendix B. Proposition 4.3 shows that a small, cost-independent amount of dual norm regularization ensures convexity of the outer min problem for *any* cost function, making it the natural choice in this setting.

5. Efficiently Identifying the Minimax-Optimal Classifier

We have shown how to efficiently solve for the maximizing cost $c \in \mathcal{C}$ for any classifier β, allowing us to apply the corrected generalization bound in Theorem 2.5 and get an upper bound on the adversarial 0-1 strategic risk. We have also seen that with dual norm regularization, optimizing the strategic hinge loss becomes tractable and also accounts for the unique structure of strategic response. The remaining challenge is to combine these two methods to find the overall solution to the min-max objective.

Solving a different loss for each cost simultaneously. An interesting consequence of using the dual norm is that for any fixed cost, the "correct" regularization is a function of that cost. Since we want to minimize this objective with respect to the entirety of $\mathcal{C}$, we absorb a *separate* dual norm regularization term into the min-max objective for each possible cost, defining our new objective as

$$\min_{\beta} \max_{c \in \mathcal{C}} \left[R^c_{\text{s-hinge}}(\beta) + \lambda u_* \|\beta\|_{*,\Sigma} \right]. \tag{4}$$

Contrast this with typical regularization, which (e.g. for ℓ_2) would instead attempt to solve $\min_\beta \left[\max_{c\in\mathcal{C}} \left[R^c_{\text{s-hinge}}(\beta)\right] + \lambda\|\beta\|_2\right]$, where regularization does not depend on the inner maximization. For brevity, in the remainder of this work we leave the regularization implicit, writing simply $R^c_{\text{s-hinge}}(\beta)$. Proposition 4.3 shows that a single choice of λ always suffices: whereas the structure-aware regularizer depends on the cost, the ideal coefficient does not. By setting λ appropriately and optimizing this new objective we account for the cost uncertainty "for free", effectively solving the problem across all costs in $\mathcal{C}$ *simultaneously* with a separate, appropriate regularization term for each. Importantly, this means our solution will be optimal with respect to the regularized objective. But once a solution is found, we can evaluate the adversarial risk without the regularization term and Theorem 2.5 will still apply.

5.1. Solving the Objective in the Full-Batch Setting with the Subgradient Method

Given a train set $\{(x_i, y_i)\}_{i=1}^n$, perhaps the simplest idea for solving Equation (4) is to follow the gradient of the empirical adversarial risk. The validity of this approach is not trivial because of the min-max formulation, but we show that it is indeed correct by invoking the following result:

Theorem 5.1 (Danskin's Theorem (Bertsekas, 1997))**.** *Suppose $f : \mathbb{R}^n \times \mathcal{Z} \to \mathbb{R}$ is a continuous function, where $\mathcal{Z} \subset \mathbb{R}^m$ is a compact set. Define $g(x) := \max_{z\in\mathcal{Z}} f(x,z)$. Then $g(x)$ is convex in x if $f(x,z)$ is convex in x for every $z \in \mathcal{Z}$. Furthermore, $\partial_x g(x) = \text{Conv}\{\partial_x f(x,z) : z \in \arg\max_z f(x,z)\}$, where ∂ is the subdifferential and "Conv" indicates the convex hull.*

Note that this requires $\mathcal{Z} \subset \mathbb{R}^m$, whereas we introduced $\mathcal{C}$ as a set of PD matrices in $\mathbb{R}^{d\times d}$. Using the assumption that $\mathcal{C}$ is compact and convex, we resolve this by associating to each cost matrix Σ a vector of d eigenvalues $\sigma_1^2 \dots \sigma_d^2$ under a fixed basis, with each eigenvalue σ_i^2 constrained to lie in the set $[\sigma^2_{i\ell}, \sigma^2_{iu}]$. These lower and upper bounds allow us to re-parameterize the cost uncertainty set so that it is now a subset of $\mathbb{R}^d$. One remaining technicality is that this requires a fixed basis which may not be shared with the true cost. However, as a consequence of Lemma 2.3, our results will still be meaningful so long as there exists a $c \in \mathcal{C}$ which

induces the same dual norm as the true cost. Without loss of generality, we suppose this basis is the identity.

Thus, we let $\mathcal{Z}$ denote the reparameterized uncertainty set $\mathcal{C}$ and plug in $\hat{R}^{c=z}_{\text{s-hinge}}$ for f. We conclude that the subderivative of the worst-case strategic hinge loss is given by the subderivative of the loss evaluated at any cost in $\mathcal{C}$ which maximizes it. We can optimize the objective via the subgradient method (Algorithm 3 in the Appendix), where each subgradient evaluation involves a call to the maximization subroutine MaxLossCost. We then combine this with well-known convergence results for the subgradient method and bound the generalization error via Theorem 2.5, giving the complete result:

Theorem 5.2. *Suppose we run Algorithm 3 for T iterations and get classifier $\hat{\beta}$. With probability $\geq 1-\delta$, the worst-case 0-1 strategic loss over costs in $\mathcal{C}$ can be bounded by*

$$\max_{c\in\mathcal{C}} R^c_{0-1}(\hat{\beta}) \leq \min_{\beta} \max_{c\in\mathcal{C}} R^c_{\textit{s-hinge}}(\beta) + \mathcal{O}\left(\frac{LB}{\sqrt{T}} + (X+u_*)\frac{B+\sqrt{\ln 1/\delta}}{\sqrt{n}}\right).$$

The proof appears in Appendix E. Despite the difficulty posed by strategic response *and* an unknown cost (as made apparent in Section 3), Theorem 5.2 shows that robust one-shot learning is possible with sufficient consideration of the underlying structure.

5.2. Solving the Objective in the Minibatch Setting with Stochastic Mirror Descent-Ascent

Occasionally the number of training points n may be sufficiently large that gradient descent is intractable, or the full dataset may not be available all at once. We would still like to be able to solve Equation (3), even when we cannot evaluate the full gradient. Unfortunately, the stochastic subgradient method is not an option here: evaluating the subderivative requires that we identify the cost that maximizes the *population* objective, which cannot be done on the basis of a subsample. As an alternative, we turn to a method from convex-concave optimization known as *Stochastic Mirror Descent-Ascent* (SMDA) (Nemirovski et al., 2009) which iteratively optimizes the min and max players to converge to a Nash equilibrium.

Modifying mirror ascent to apply to our setting. Since the objective is convex in β, minimization is straightforward. However, recall that the maximization over costs is non-concave. Previously we got around this by solving for the maximizing cost in a non-differentiable manner and then invoking Danskin's theorem—but the guarantees given by Nemirovski et al. (2009) require that we take iterative steps of gradient ascent on the adversary's (assumed concave) objective. We address this via an *algorithmic ϵ-net*.

Algorithm 2 Stochastic Mirror Descent-Ascent on the regularized strategic hinge loss

Input: Batch size n, Iterations T, Costs $\mathcal{C}$, Upper bound u_*, Regularization λ, Discretization ϵ, Step sizes η_q, η_β.
define: $\beta^{(0)} \leftarrow \mathbf{0}$
$v \leftarrow [\epsilon(\sigma_{1\ell}^{-2} - \sigma_{1u}^{-2}), \ldots, \epsilon(\sigma_{d\ell}^{-2} - \sigma_{du}^{-2})]^\top$
$q^{(0)} \leftarrow \epsilon\mathbf{1}$
for $t = 1, \ldots, T$ **do**
  Draw samples $\{(x_i, y_i)\}_{i=1}^n$.
  $c_k \leftarrow \vec{\sigma_u}^{-2} + k \cdot v, \quad k \in \{1, \ldots, \lceil 1/\epsilon \rceil\}$
  $q' \leftarrow q^{(t-1)}$
  $q'_k \leftarrow q'_k \exp(\eta_q \hat{R}^{c_k}_{\text{s-hinge}}(\beta^{(t-1)}))$
  $q^{(t)} \leftarrow q' / \sum_k q'_k$
  $\beta^{(t)} \leftarrow \beta^{(t-1)} - \eta_\beta \left(\sum_k q_k^{(t)} \nabla \hat{R}^{c_k}_{\text{s-hinge}}(\beta^{(t-1)})\right)$
end for
return $\hat{\beta} := \frac{1}{T}\sum_{t=1}^T \beta^{(t)}$

Specifically, we can relax Equation (4) to a maximization over a *finite* set of costs $\mathcal{S} \subset \mathcal{C}$ such that the solution to this new objective is provably close to that of the original problem. Given such a set, the new objective becomes $\min_\beta \max_{c\in\mathcal{S}} R^c_{\text{s-hinge}}(\beta)$. Observe that the solution to this objective is equivalent to the solution to

$$\min_{\beta} \max_{q\in\Delta(|\mathcal{S}|)} \sum_{c_i\in\mathcal{S}} q_i R^{c_i}_{\text{s-hinge}}(\beta), \tag{5}$$

where $\Delta(|\mathcal{S}|)$ is the $|\mathcal{S}|$-simplex. In other words, we can solve this problem over all convex combinations of the risks under the different costs in $\mathcal{S}$ and arrive at the same solution (Rosenfeld et al., 2022b). Crucially, unlike our original objective, Equation (5) is convex-concave, so it can be efficiently optimized via SMDA! The correctness of this relaxation relies on the property proven in Lemma 2.3: strictly speaking, the strategic loss depends only the *scalar* dual norm, not on the full cost matrix.

The gap between the solution to the original objective and Equation (5) scales inversely with the fineness of discretization, so we want the maximum distance between any cost $c \in \mathcal{C}$ and its closest neighbor in $\mathcal{S}$ to be as small as possible. However, the memory and compute requirements of SMDA scale linearly with $|\mathcal{S}|$—and error scales as $\sqrt{\ln|\mathcal{S}|}$—so we also cannot let it grow too large. To balance these two considerations, we carefully construct a discretization over the "diagonal" of the space of eigenvalue intervals in $\mathcal{C}$, leading to a set $\mathcal{S}$ with cardinality $\Theta\left(\frac{TD}{\ln T}\right)$, where T is the number of iterations and $D := \max_i 1/\sigma_{i\ell}^2 - 1/\sigma_{iu}^2$ is the diameter of $\mathcal{C}$. The exact construction for $\mathcal{S}$ appears in Algorithm 2. With this careful discretization, the computational cost to achieve a fixed sub-optimality gap is *dimension independent* for p-norms with $p \leq 2$, and the worst-case scaling (when $p = \infty$) is $\mathcal{O}\left(d^2/\ln d\right)$. In contrast, a typical ϵ-net would

have $|\mathcal{S}| = \Theta(\exp(d))$, with memory and compute scaling commensurately.

Theorem 5.3. *Suppose we run Algorithm 2 for T rounds with discretization $\epsilon = \Theta\left(\frac{\ln T}{TD}\right)$ and get classifier $\hat{\beta}$. Then over the randomness of the stochastic gradients, its expected sub-optimality*[3] *is bounded by*

$$\mathbb{E}[\max_{c\in\mathcal{C}} R^c_{\textit{s-hinge}}(\hat{\beta})] \lesssim \min_{\beta} \max_{c\in\mathcal{C}} R^c_{\textit{s-hinge}}(\beta) + \mathcal{O}\left(\frac{LB}{\sqrt{T}} + B(X+u_*)\sqrt{\frac{\ln TD}{T}}\right).$$

The proof can be found in Appendix F. Remarkably, convergence to the population minimax solution occurs at the same rate as the full-batch setting up to logarithmic factors—the cost of stochasticity is surprisingly small. Once again, we can use MaxLossCost to evaluate the adversarial strategic risk and then apply Theorem 2.5 to get a generalization bound on the worst-case 0-1 error.

6. Conclusion

This paper studies robust strategic learning under unknown user costs in the challenging one-shot setting. Our results suggest that uncertainty in how users respond should be considered an integral aspect of strategic learning: motivated by realistic problem domains which permit only a single action (which must be immediately effective), we provide a learning framework based on distributionally robust optimization for modeling—and robustly handling—this uncertainty. We begin by showing that even a miniscule error in estimating the true cost can cause substantial error in deployment, motivating the use of an uncertainty set of possible costs. Next, we propose a natural proxy to the intractable min-max objective over this set, and we design two efficient algorithms for different settings that converge to the empirical solution, for which we also provide generalization guarantees.

Our approach highlights the value of dual norm regularization, which ensures good performance while accounting for the structure of strategic behavior coupled with cost uncertainty. Such structure can both harm accuracy (on negative examples) and help it (on positive examples). One important implication is that it often does not suffice to simply be more conservative, as this can fail to take advantage of settings where the learner's and agents' incentives are aligned. Rather, maintaining robustness without substantially sacrificing accuracy requires more careful consideration of and accounting for the interests and actions of future users.

[3] Exponential concentration of the adversarial risk to its expectation follows from Nemirovski et al. (2009), Proposition 3.2.

Impact Statement

This work studies the problem of strategic classification under unknown user costs. While our focus is entirely theoretical, it is important to acknowledge that, in a realistic setting, the learning task which we consider would have human beings as the targets for prediction. This is in fact our primary motivation: correctly accounting for strategic behavior in learning can benefit not only the system, but also the general population of users (Milli et al., 2019; Levanon and Rosenfeld, 2021; 2022). Questions of how to act under uncertainty or partial information are immanent in strategic learning; our work offers but one way of accounting for system uncertainty regarding user behavior, with a particular focus on uncertainty stemming from unknown user costs. Other sources of uncertainty therefore lie outside our scope, as does human behavior which deviates from our assumed rational response model (Eq. equation 1), as well as considerations beyond minimizing worst case risk. Although our analysis applies to a relatively large and flexible class of cost functions (broader than what is typically considered in the literature), a clear limitation of our work is that it is unclear if it applies to costs outside this class. A more subtle but nonetheless important limitation is that while our approach provides strong worst-case guarantees for any cost in the uncertainty set $\mathcal{C}$, it does not state what will happen if the 'true' cost c^* lies outside of the chosen $\mathcal{C}$. We conjecture that under appropriate smoothness assumptions, worst-case guarantees should degrade gracefully as c^* moves away from $\mathcal{C}$, but leave this as an open question for future work. Finally, although the algorithms we propose are theoretical, they may nonetheless be applicable in certain settings. Given this, we implore any future use of our algorithms to be carried out in a transparent and accountable manner—as should be expected from any tool developed under the general framework of strategic learning.

Acknowledgements

Thanks to Roni Rosenfeld for helpful discussions in developing the motivation for this work. This work is supported by the Israel Science Foundation grant no. 278/22.

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.

Yahav Bechavod, Chara Podimata, Steven Wu, and Juba Ziani. Information discrepancy in strategic learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 1691–1715. PMLR, 17–23 Jul 2022.

Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997.

Michael Brückner and Tobias Scheffer. Nash equilibria of static prediction games. In *Advances in neural information processing systems*, pages 171–179, 2009.

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

Yatong Chen, Jialu Wang, and Yang Liu. Linear classifiers that encourage constructive adaptation. *Transactions on Machine Learning Research*, 2023.

K Alec Chrystal, Paul D Mizen, and PD Mizen. Goodhart's law: its origins, meaning and implications for monetary policy. *Central banking, monetary theory and practice: Essays in honour of Charles Goodhart*, 1:221–243, 2003.

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.

John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

Itay Eilat, Ben Finkelshtein, Chaim Baskin, and Nir Rosenfeld. Strategic classification with graph neural networks. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

Lewis Elton. Goodhart's law and performance indicators in higher education. *Evaluation & Research in Education*, 18(1-2):120–128, 2004.

Michael Fire and Carlos Guestrin. Over-optimization of academic publishing metrics: observing goodhart's law in action. *GigaScience*, 8(6):giz053, 2019.

Ganesh Ghalme, Vineet Nair, Itay Eilat, Inbal Talgam-Cohen, and Nir Rosenfeld. Strategic classification in the dark. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Charles AE Goodhart. Problems of monetary management: The UK experience. In *Papers in Monetary Economics, vol. I*. Reserve Bank of Australia, 1975.

Michael Großhans, Christoph Sawade, Michael Brückner, and Tobias Scheffer. Bayesian games for adversarial regression problems. In *International Conference on Machine Learning*, pages 55–63, 2013.

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.

Keegan Harris, Chara Podimata, and Zhiwei Steven Wu. Strategic apple tasting. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Guy Horowitz and Nir Rosenfeld. Causal strategic classifiaction. In *International Conference on Machine Learning*. PMLR, 2023.

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

Tosca Lechner and Ruth Urner. Learning losses for strategic classification. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI*, pages 7337–7344. AAAI Press, 2022.

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

Sagi Levanon and Nir Rosenfeld. Strategic classification made practical. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Sagi Levanon and Nir Rosenfeld. Generalized strategic classification and the case of aligned incentives. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

Licong Lin and Tijana Zrnic. Plug-in performative optimization. *arXiv preprint arXiv:2305.18728*, 2023.

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

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.

Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in neural information processing systems*, 29, 2016.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

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

Elan Rosenfeld and Saurabh Garg. (Almost) provable error bounds under distribution shift via disagreement discrepancy. *Advances in Neural Information Processing Systems*, 36, 2024.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. Domain-adjusted regression or: Erm may already learn features sufficient for out-of-distribution generalization. *arXiv preprint arXiv:2202.06856*, 2022a.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. An online learning approach to interpolation and extrapolation in domain generalization. In *International Conference on Artificial Intelligence and Statistics*, pages 2641–2657. PMLR, 2022b.

Han Shao, Avrim Blum, and Omar Montasser. Strategic classification under unknown personalized manipulation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ravi Sundaram, Anil Vullikanti, Haifeng Xu, and Fan Yao. PAC-learning for strategic classification. In *International Conference on Machine Learning*, pages 9978–9988. PMLR, 2021.

Damien Teney, Ehsan Abbasnejad, Kushal Kafle, Robik Shrestha, Christopher Kanan, and Anton Van Den Hengel. On the value of out-of-distribution testing: An example of goodhart's law. *Advances in neural information processing systems*, 33:407–417, 2020.

Hanrui Zhang and Vincent Conitzer. Incentive-aware PAC learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

A. Proofs of Hardness Results

We restate the theorems for convenience.

A.1. Proof of Theorem 3.1

Theorem A.1. *Consider the task of learning a norm-bounded linear classifier. Fix any two costs c_1, c_2 with non-equal PD matrices, and let $0 \leq \epsilon \leq \frac{1}{2}$. There exists a distribution p over $\mathcal{X} \times \mathcal{Y}$ such that:*

1. *For each of c_1 and c_2, there is a (different) classifier which achieves 0 error on p when facing strategic response under that cost; and*
2. *Any classifier which achieves 0 error on p under cost c_1 suffers error ϵ under cost c_2, and any classifier which achieves 0 error on p under cost c_2 suffers error $1 - \epsilon$ under cost c_2.*

Proof. We will construct two distributions, one over the conditional $q(x \mid y = 1)$ and the other over $q(x \mid y = -1)$, and combine them via the mixture $q(y = 1) = \epsilon$, $q(y = -1) = 1 - \epsilon$. Let Σ_1, Σ_2 denote the cost matrices for c_1, c_2 respectively. Let B denote the upper bound on the classifier norm. Pick any β^* such that $\|\beta^*\|_{\Sigma_1} \neq \|\beta^*\|_{\Sigma_2}$, and define $r := \frac{u_* \|\beta^*\|}{3B}(\|\beta^*\|_{\Sigma_1} - \|\beta^*\|_{\Sigma_2})$. WLOG, suppose $r > 0$. Focusing on the negatively labeled portion, consider the $(d-1)$-dimensional plane in $\mathcal{X}$ defined by ${\beta^*}^\top x = -r$. We let $q(x \mid y = -1)$ be any distribution with full support over that plane. Similarly, define the conditional distribution $q(x \mid y = 1)$ as a full-support distribution over the plane defined by ${\beta^*}^\top x = r$. Thus the only classifier which achieves 0 error must have $\hat{\beta} = \alpha\beta^*$ for some scaling factor $B/\|\beta^*\| \geq \alpha > 0$. The only remaining degree of freedom is the bias term $\hat{\beta}_0$.

Consider the case where the true cost is c_1. If we wish to correctly classify the negative points under strategic response, we must classify the negatively labeled plane with margin greater than $u_*\|\beta^*\|_{\Sigma_1}$. However, if we wish to do the same with the positive points, the margin for that plane must be less than or equal to this same value. Formally, we must have

$$\begin{aligned} {\beta^*}^\top x = -r \implies \hat{\beta}^\top x + \hat{\beta}_0 < -u_*\|\beta^*\|_{\Sigma_1}, \quad (6)\\ {\beta^*}^\top x = r \implies \hat{\beta}^\top x + \hat{\beta}_0 \geq -u_*\|\beta^*\|_{\Sigma_1}, \quad (7)\end{aligned}$$

which, remembering that $\hat{\beta} = \alpha\beta^*$, immediately implies

$$\begin{aligned} \hat{\beta}_0 - \alpha r < -u_*\|\beta^*\|_{\Sigma_1}, \quad (8)\\ \hat{\beta}_0 + \alpha r \geq -u_*\|\beta^*\|_{\Sigma_1}. \quad (9)\end{aligned}$$

Thus, we have

$$-u_*\|\beta^*\|_{\Sigma_1} - \alpha r \leq \hat{\beta}_0 < -u_*\|\beta^*\|_{\Sigma_1} + \alpha r, \quad (10)$$

and since $\alpha r > 0$, this describes the non-empty set of all classifiers which achieve 0 error under cost c_1. By an analogous argument we can construct the set of classifiers which achieve 0 error under cost c_2.

However, observe that

$$\begin{aligned} 2\alpha r &= \frac{2}{3}\frac{\|\beta^*\|}{B} u_*(\|\beta^*\|_{\Sigma_1} - \|\beta^*\|_{\Sigma_2}) \quad (11)\\ &< u_*(\|\beta^*\|_{\Sigma_1} - \|\beta^*\|_{\Sigma_2}), \quad (12)\end{aligned}$$

and therefore

$$-u_*\|\beta^*\|_{\Sigma_1} + \alpha r < -u_*\|\beta^*\|_{\Sigma_2} - \alpha r. \quad (13)$$

This means that the upper bound for any $\hat{\beta}_0$ which achieves 0 error under cost c_1 cannot satisfy the lower bound for cost c_2, which means the positively labeled plane will be classified negatively with a margin that is too large for strategic response, causing error ϵ. By a symmetric argument, any $\hat{\beta}_0$ which achieves 0 error under cost c_2 cannot satisfy the upper bound for cost c_1, implying the negatively labeled plane will strategically shift and cause error $1 - \epsilon$. Finally, we remark that while the distribution defined here is not absolutely continuous on $\mathcal{X}$, this can be remedied by simply making the distribution a product of the constructed planar distribution and a uniform distribution along the direction β^* with width sufficiently small (e.g, width $c \cdot \alpha r$ for $c \ll 1$). □

A.2. Proof of Theorem 3.2

Theorem A.2. *Define the distribution q over $\mathcal{X} \times \mathcal{Y}$ as $q(y=1) = q(y=-1) = 1/2$, $q(x \mid y) \overset{d}{=} \mathcal{N}(y \cdot \mu_0, \sigma^2 I)$. Denote by Φ the standard Normal CDF. Let the true cost be defined as $\|x - x'\|_\Sigma$ with unknown cost matrix Σ, and let β^* be the classifier which minimizes the strategic 0-1 risk under this cost.*

Suppose one instead learns a classifier $\hat{\beta}$ by assuming an incorrect cost $\hat{\Sigma}$ and minimizing the population strategic 0-1 risk under that cost: $\hat{\beta} := \arg\min_\beta \mathbb{E}_q[\ell^{\hat{c}}_{0-1}(\beta)]$. Then the excess 0-1 risk suffered by $\hat{\beta}$ is

$$\Phi\left(\frac{\|\mu_0\|}{\sigma}\right) - \frac{1}{2}\left(\Phi\left(\frac{\|\mu_0\| - \epsilon}{\sigma}\right) + \Phi\left(\frac{\|\mu_0\| + \epsilon}{\sigma}\right)\right),$$

where $\epsilon := \frac{u_* |\|\mu_0\|_{*,\hat{\Sigma}} - \|\mu_0\|_{*,\Sigma}|}{\|\mu_0\|}$.

Proof. By the Neyman-Pearson Lemma, the classifier which minimizes non-strategic 0-1 risk will be the one which predicts $\operatorname{sign}(\ln \frac{p(y=1|x)}{p(y=0|x)})$, which gives $\beta^* = 2\mu_0$. To account for strategic response, we observe that as proven in Lemma 2.3, each user will have $\delta(x)$ in the direction of β^* which induces a change in their predicted value by at most $u_* \|\beta^*\|_* = 2u_* \|\mu_0\|_{*,\Sigma}$; since we are considering the 0-1 loss, this is equivalent to *every* user shifting in this way.

It is immediate that to find the corresponding minimizer of the population 0-1 *strategic* risk, we can simply add a negative bias equal to the change induced by this shift, because this maintains the same labels as before on all points, *after* they strategically shift. Therefore, the strategic risk minimizer will be $2\mu_0^\top x - 2u_* \|\mu_0\|_{*,\Sigma}$. By the same argument, the minimizer of the 0-1 strategic risk under the incorrect cost matrix $\hat{\Sigma}$ is $2\mu_0^\top x - 2u_* \|\mu_0\|_{*,\hat{\Sigma}}$. It remains to derive a lower bound on their difference in risk.

First note that as argued above, the 0-1 strategic risk of the minimizer for the correct cost is the same as the non-strategic risk of the non-strategic solution. For the both positively and negatively labeled points, this is equal to $\Phi\left(-\frac{\|\mu_0\|}{\sigma}\right)$ due to rotational symmetry. To determine the risk of the incorrect solution, we can identify the regions whose label will differ under that classifier and bound the measure of those regions. Define $\gamma = 2u_*(\|\mu_0\|_{*,\hat{\Sigma}} - \|\mu_0\|_{*,\Sigma})$ as the difference in the two solutions' predictions on all x. Due to the symmetry of the positive and negative conditional distributions, we can assume WLOG that $\gamma > 0$, i.e., the incorrect classifier assigns a smaller prediction to all x. This means it will have less risk on a region of negative points and more on positive. Specifically, the two classifiers will differ on all points for which the true strategic-optimal classifier assigns a value *before strategic response* which lies in $(-2u_* \|\mu_0\|_{*,\Sigma}, -2u_* \|\mu_0\|_{*,\Sigma} + \gamma)$; those assigned a value less than $-2u_* \|\mu_0\|_{*,\Sigma}$ will receive a negative prediction from both classifiers, and those assigned a value greater than $-2u_* \|\mu_0\|_{*,\Sigma} + \gamma$ will still be close enough to the decision boundary that they can shift to achieve a positive label prediction from $\hat{\beta}$. Formally, this region is

$$\{x \mid -2u_* \|\mu_0\|_{*,\Sigma} < \beta^{*\top} x < -2u_* \|\mu_0\|_{*,\Sigma} + \gamma\} = \{x \mid 0 < \mu_0^\top x < \gamma/2\}. \tag{14}$$

Observe this region depends only on the value $\mu_0^\top x$. Since the negative points are distributed as $\mathcal{N}(-\mu_0, \sigma^2 I)$, this term has the distribution $\mu_0^\top x \sim \mathcal{N}(-\|\mu_0\|^2, \sigma^2 \|\mu_0\|^2)$. Therefore, the measure of this region under $q(x \mid y = -1)$ is

$$\Phi\left(\frac{\gamma/2 + \|\mu_0\|^2}{\sigma \|\mu_0\|}\right) - \Phi\left(\frac{\|\mu_0\|}{\sigma}\right) \tag{15}$$

$$= \int_0^{\frac{\gamma}{2\|\mu_0\|}} \rho\left(\frac{\|\mu_0\| + z}{\sigma}\right) dz, \tag{16}$$

This is the amount by which the risk of the incorrect solution will *decrease* on the negative points. Likewise, the risk will *increase* on positive points in this region, which under $q(x \mid y = 1)$ has measure

$$\Phi\left(\frac{\gamma/2 - \|\mu_0\|^2}{\sigma \|\mu_0\|}\right) - \Phi\left(\frac{-\|\mu_0\|}{\sigma}\right) \tag{17}$$

$$= \int_0^{\frac{\gamma}{2\|\mu_0\|}} \rho\left(\frac{-\|\mu_0\| + z}{\sigma}\right) dz, \tag{18}$$

Therefore, the overall increase to risk will be

$$\frac{1}{2}\int_0^{\epsilon} \rho\left(\frac{-\|\mu_0\|+z}{\sigma}\right) - \rho\left(\frac{\|\mu_0\|+z}{\sigma}\right)\, dz, \tag{19}$$

where $\epsilon := \frac{u_*|\|\mu_0\|_{*,\hat{\Sigma}} - \|\mu_0\|_{*,\Sigma}|}{\|\mu_0\|}$. By applying the fundamental theorem of calculus and the fact that $\Phi(x) = 1 - \Phi(-x)$ we arrive at the stated equality. □

B. Proof of Theorem 4.2 and Proposition 4.3

Theorem B.1. *Let $\|\cdot\|$ be a p-norm and fix a cost matrix Σ. For any distribution q on $\mathcal{X}\times\mathcal{Y}$ with $q(y=1) =: \tau^+$, the dual-regularized loss $R^c_{\text{s-hinge}}(\beta) + \lambda u_*\|\beta\|_*$ is guaranteed to be convex for $\lambda \geq \tau^+$. In contrast, the ℓ_2-regularized loss $R^c_{\text{s-hinge}}(\beta) + \lambda u_*\|\beta\|_2$ is non-convex unless $\lambda \geq \tau^+\|\Sigma^{-1/2}\|_2$.*

Proof. Writing out the dual-regularized loss with the full definition of the strategic hinge,

$$\begin{aligned} R^c_{\text{s-hinge}}(\beta) + \lambda u_*\|\beta\|_* &= \mathbb{E}_{(x,y)\sim q}[\max\{0, 1 - y(\beta^\top x + u_*\|\beta\|_*)\}] + \lambda u_*\|\beta\|_* \\ &= \tau^+\mathbb{E}_{q(x|y=1)}[\max\{0, 1 - \beta^\top x - u_*\|\beta\|_*\} + {}^{\lambda}\!/_{\tau^+} u_*\|\beta\|_*] + \\ &\qquad (1-\tau^+)\mathbb{E}_{q(x|y=-1)}[\max\{0, 1 + \beta^\top x + u_*\|\beta\|_*\}]. \end{aligned}$$

The last term is already convex in β. Rewriting the first term, we get

$$\tau^+\mathbb{E}_{q(x|y=1)}[\max\{{}^{\lambda}\!/_{\tau^+} u_*\|\beta\|_*, 1 - \beta^\top x + u_*\left({}^{\lambda}\!/_{\tau^+} - 1\right)\|\beta\|_*\}] \tag{20}$$

For $\lambda \geq \tau^+$ this is the expectation of the maximum of two convex functions, and thus the full loss is convex.

To see why the ℓ_2 norm requires much stronger regularization, consider again the above term with regularization $\lambda u_*\|\beta\|_2$:

$$\tau^+\mathbb{E}_{x|y=1}\left[\max\left\{\frac{\lambda}{\tau^+}u_*\|\beta\|_2, 1 - \beta^\top x + u_*\left(\frac{\lambda}{\tau^+}\|\beta\|_2 - \|\beta\|_*\right)\right\}\right]. \tag{21}$$

The first term in the max is convex—so to show non-convexity, we will consider values where the second term is larger. Recalling that $\|v\|_* := \|\Sigma^{-1/2}v\|_q$, write the second term as a function $f(\beta) := 1 - \beta^\top x + u_*\left(\frac{\lambda}{\tau^+}\|\beta\|_2 - \|\Sigma^{-1/2}\beta\|_q\right)$, and thus

$$\nabla f(\beta) = -x + u_*\left(\frac{\lambda}{\tau^+}\frac{\beta}{\|\beta\|_2} - \Sigma^{-1/2}\left(\frac{\partial}{\partial v}\|v\|_q\Big|_{v=\Sigma^{-1/2}\beta_2}\right)\right). \tag{22}$$

Recall that a function is convex if and only if for all x, y in its domain,

$$f(x) - f(y) \geq \nabla f(y)^\top(x-y). \tag{23}$$

This means that f is convex only if for *all* vectors $\beta_1, \beta_2 \in \mathcal{B}$,

$$\frac{\lambda}{\tau^+}(\|\beta_1\|_2 - \|\beta_2\|_2) - \|\Sigma^{-1/2}\beta_1\|_q + \|\Sigma^{-1/2}\beta_2\|_q \geq \left(\frac{\lambda}{\tau^+}\frac{\beta_2}{\|\beta_2\|_2} - \Sigma^{-1/2}\left(\frac{\partial}{\partial v}\|v\|_q\Big|_{v=\Sigma^{-1/2}\beta_2}\right)\right)^\top(\beta_1 - \beta_2). \tag{24}$$

Without loss of generality, suppose Σ is diagonal. Let $v_i,\ i \in [d]$ denote the eigenvectors of Σ with *decreasing* eigenvalues $\sigma_1^2, \sigma_2^2, \ldots$. Choose any $\beta_2 = \sum_i \lambda_i v_i$ with non-negative λ_i and $\lambda_d = 0$. Thus $\Sigma^{-1/2}\beta_2 = \sum_i \left({}^{\lambda_i}\!/_{\sigma_i}\right) v_i$. Then

$$\beta_2^\top\Sigma^{-1/2}\frac{\partial}{\partial v}\|v\|_q\Big|_{v=\Sigma^{-1/2}\beta_2} = \sum_i \left({}^{\lambda_i}\!/_{\sigma_i}\right)\cdot\left(\frac{|{}^{\lambda_i}\!/_{\sigma_i}|}{\|\Sigma^{-1/2}\beta_2\|_q}\right)^{q-1} \tag{25}$$

$$= \frac{\|\Sigma^{-1/2}\beta_2\|_q^q}{\|\Sigma^{-1/2}\beta_2\|_q^{q-1}} \tag{26}$$

$$= \|\Sigma^{-1/2}\beta_2\|_q. \tag{27}$$

This allows us to simplify the above inequality and arrive at the condition

$$\frac{\lambda}{\tau^+}\|\beta_1\|_2 - \|\Sigma^{-1/2}\beta_1\|_q \geq \beta_1^\top \left(\frac{\lambda}{\tau^+}\frac{\beta_2}{\|\beta_2\|_2} - \Sigma^{-1/2}\frac{\partial}{\partial v}\|v\|_q\big|_{v=\Sigma^{-1/2}\beta_2}\right). \tag{28}$$

Now choose $\beta_1 = cv_d$ for some scalar $c \neq 0$. Then the RHS vanishes and $f(\beta)$ is convex only if

$$\frac{\lambda c}{\tau^+} - \frac{c}{\sigma_d} \geq 0 \tag{29}$$

$$\iff \lambda \geq \frac{\tau^+}{\sigma_d} = \tau^+\|\Sigma^{-1/2}\|_2. \tag{30}$$

Thus we've proven the required lower bound on λ. What remains is to show that this condition applies at a location in parameter space where the negatively labeled samples do not contribute to the gradient, and where the losses on the positively labeled samples are dominated by the second term of the maximum.

To do this, we scale down $\beta_2 \to 0$ and choose the bias as $\beta_0 = -(1 + u_*\|\beta_2\|_* + \epsilon(X+1))$ for some very small positive ϵ. It follows that for all x,

$$\beta_2^\top x + \beta_0 + u_*\|\beta_2\|_* \leq \epsilon X - (1 + u_*\|\beta_2\|_* + \epsilon(X+1)) + u_*\|\beta_2\|_* \tag{31}$$

$$\leq -(1+\epsilon). \tag{32}$$

This accomplishes both desiderata: first, it ensures that the loss on all negatively labeled points is $\max\{0, 1 + \beta^\top x + \beta_0 + u_*\|\beta\|_*\} \leq \max\{0, -\epsilon\} = 0$, with gradient equal to 0. Second, it ensures that on the positive examples, the second term in the loss dominates. We can see this by observing that the second term being larger is equivalent to

$$0 < 1 - \left(\beta_2^\top x + \beta_0 + u_*\|\beta_2\|_*\right), \tag{33}$$

and by construction the RHS is at least $2 + \epsilon$ for all x. □

As mentioned in the main body, we can also show the following additional result: for any λ which induces convexity with $\|\beta\|_2$, it must hold that $\|\beta\|_* \leq \|\beta\|_2$, and thus it imposes the least regularization.

Proposition B.2. *If the objective $R^c_{\text{s-hinge}}(\beta) + \lambda u_*\|\beta\|_2$ is convex, then the dual regularization is no greater than the ℓ_2 regularization: $\tau^+\|\beta\|_* \leq \lambda\|\beta\|_2$.*

Proof. Recall Equation (24) in the proof above: f is convex only if for all vectors $\beta_1, \beta_2 \in \mathcal{B}$,

$$\frac{\lambda}{\tau^+}(\|\beta_1\|_2 - \|\beta_2\|_2) - \|\Sigma^{-1/2}\beta_1\|_q + \|\Sigma^{-1/2}\beta_2\|_q \geq \left(\frac{\lambda}{\tau^+}\frac{\beta_2}{\|\beta_2\|_2} - \Sigma^{-1/2}\left(\frac{\partial}{\partial v}\|v\|_q\big|_{v=\Sigma^{-1/2}\beta_2}\right)\right)^\top (\beta_1 - \beta_2). \tag{34}$$

Choosing $\beta_1 = -\beta_2$ and simplifying, we get $\lambda||\beta_2||_2 \geq \tau^+||\beta_2||_*$ □

C. Proofs of Lemmas in Main Body

C.1. Proofs of Lemmas 2.3 and 2.4

Lemma C.1. *For any cost $c(x, x') = \phi(\|x - x'\|_\Sigma)$, the maximum change to a user's prediction score that can result from strategic behavior is given by*

$$\beta^\top x(\beta) - \beta^\top x \leq u_*\|\beta\|_* \tag{35}$$

where $\|\beta\|_* := \|\beta\|_{*,\Sigma} = \sup_{\|v\|_\Sigma=1}\beta^\top v$ is the Σ-transformed dual norm *of β and $u_* := \sup r \in \mathbb{R}_{\geq 0}$ s.t. $\phi(r) \leq u$.*

Proof. A user at x will move so as to maximize the inner product $\beta^\top x(\beta)$ so long as the cost of this move does not exceed the additional utility u (and only up until the point that they achieve a positive classification). In other words, the maximum logit they will feasibly achieve is given by the optimization problem

$$\sup_{x'} \beta^\top x' \quad \text{s.t. } c(x, x') \leq u. \tag{36}$$

We can reparameterize $x' = x + \delta$ to rewrite the objective as

$$\beta^\top x + \sup_{\{\delta \,:\, \phi(\|\delta\|_\Sigma) \leq u\}} \beta^\top \delta, \tag{37}$$

which, recalling the definition of u_* and monotonicity of ϕ, is equal to

$$\beta^\top x + \sup_{\{\delta \,:\, \|\delta\|_\Sigma \leq u_*\}} \beta^\top \delta \tag{38}$$

Here we recognize the variational formula for the dual norm, giving the solution $\beta^\top x + u_* \|\beta\|_*$. □

Lemma C.2. *For any cost $c \in \mathcal{C}$, $R^c_{0-1}(\beta) \leq R^c_{\text{s-hinge}}(\beta)$.*

Proof. For a fixed sample (x, y), recall the loss definitions:

$$\ell^c_{0-1}(\beta) := \mathbf{1}\{\text{sign}(\beta^\top(x+\delta)) \neq y\} \tag{39}$$

$$\ell^c_{\text{hinge}}(\beta) := \max\left(0, 1 - y\beta^\top(x+\delta)\right) \tag{40}$$

$$\ell^c_{\text{s-hinge}}(\beta) := \max\left(0, 1 - y(\beta^\top x + u_*\|\beta\|_*)\right). \tag{41}$$

Since strategic response is agnostic to the loss used (i.e., δ does not change) and the hinge loss upper bounds the 0-1 loss, it is immediate that $\ell^c_{0-1} \leq \ell^c_{\text{hinge}}$. Consider any point with true label $y = 1$. If the point is positively classified (whether it moves or not) then $\ell^c_{0-1} = 0 \leq \ell^c_{\text{s-hinge}}$. If the point is negatively classified and does not move, this means $\beta^\top x < -u_*\|\beta\|_* \implies \beta^\top x + u_*\|\beta\|_* < 0$, and therefore $\ell^c_{0-1} = 1 < \ell^c_{\text{s-hinge}}$. So the claim holds for any point with $y = 1$.

Next, if a point with true label $y = -1$ does not move, then neither loss changes as a result of strategic response, which means the strategic hinge loss is no less than the regular hinge loss. It remains to prove the inequality for points with $y = -1$ which move in response to the classifier. By Lemma 2.3 the classifier's output after strategic response will increase by no more than $u_*\|\beta\|_*$. We have

$$\begin{aligned}
\ell^c_{\text{hinge}}(\beta) &= \ell_{\text{hinge}}(\beta^\top(x+\delta), y = -1) && (42)\\
&= \max\{0, 1 + \beta^\top(x+\delta)\} && (43)\\
&\leq \max\{0, 1 + \beta^\top x + u_*\|\beta\|_*\} && (44)\\
&= \ell_{\text{s-hinge}}(\beta; u_*\|\beta\|_*) && (45)\\
&= \ell^c_{\text{s-hinge}}(\beta).
\end{aligned}$$

□

C.2. Proof of Lemma 4.1

Lemma C.3. *Fix some classifier β. Then for any dataset $(\mathcal{X} \times \mathcal{Y})^n$ and uncertainty set $\mathcal{C}$,* MaxLossCost *runs in $\mathcal{O}(nd + n \ln n)$ time and returns the value $k^* \in \mathbb{R}_{\geq 0}$ which maximizes the k-shifted strategic hinge loss $\hat{R}_{\text{s-hinge}}(\beta; k)$ subject to $k^* = u_*\|\beta\|_*$ for some cost $c \in \mathcal{C}$.*

Proof. Recall the regularized strategic hinge loss $\hat{R}_{\text{s-hinge}}(\hat{\beta}; u_*\|\hat{\beta}\|_*) = \frac{1}{n}\sum_{i=1}^n \max\{0, 1 - y_i(\hat{\beta}^\top x + u_*\|\hat{\beta}\|_*)\} + \lambda u_*\|\hat{\beta}\|_*$. As this function depends on c only through the dual norm, and since $\mathcal{C}$ is a convex set and the norm is continuous, the worst-case cost *scalar* can be reparameterized as $\arg\max_{k \in [\|\hat{\beta}\|_*^{\min}, \|\hat{\beta}\|_*^{\max}]} \hat{R}_{\text{s-hinge}}(\hat{\beta}; u_* k)$. This function is one-dimensional and piecewise linear in k, and therefore the maximum must occur either at an endpoint or at the boundary between two linear segments.

By sorting the $v_i := y_i(1 - y_i\hat{\beta}^\top x_i)$, we get the values $1 - y_i\hat{\beta}^\top x_i$ with $y = +1$ in increasing order and those with $y = -1$ in decreasing order. At each step, we maintain the condition that for all $j' < j$, $v_{j'} - u_* k \leq 0$. It follows that by increasing k to the boundary of the next linear segment at k', there are exactly c_{+1} points for which the loss will decrease by $u_*(k' - k)$ and c_{-1} points for which the loss will increase by that same amount, while the regularization term increases by $\lambda u_*(k' - k)$. Thus r tracks the induced risk for the current k, and we keep track of the k which so far induces the maximum risk. Finally, since we have moved to the next linear segment: either an example with $y = +1$ now has 0 loss and will not change for the remainder; or an example with $y = -1$ has > 0 loss and will contribute linearly to the risk for the remainder. We therefore update the appropriate count and iterate. The algorithm is complete when we reach the boundary $k = \|\beta\|_*^{\max}$. If before this point we reach the end of the sorted v_j, then we know that for the remaining possible increase $\|\beta\|_*^{\max} - k$, only the loss on the negative examples will change, growing linearly until the boundary. So we do one last evaluation and return the maximum. □

D. Proof of Rademacher Generalization Bound

Theorem D.1 (Strategic Hinge Generalization Bound)**.** *Fix a norm $\|\cdot\|$. Assume $\max_{x\in\mathcal{D}} \|x\| \leq X$ and $\|\beta\|_2, \|\beta\|_* \leq B,\ \forall \beta \in \mathcal{B}, c \in \mathcal{C}$. Then with probability $\geq 1 - \delta$, for all $\beta \in \mathcal{B}$ and all cost functions $c \in \mathcal{C}$,*

$$R^c_{0-1}(\beta) \leq \hat{R}^c_{\textit{s-hinge}}(\beta) + \frac{B(4X + u_*) + 3\sqrt{\ln 1/\delta}}{\sqrt{n}} \tag{46}$$

Lemma 2.4 shows that $R^c_{0-1}(\beta) \leq R^c_{\text{s-hinge}}(\beta)$. The result then follows from standard Rademacher bounds, requiring only the following additional Lemma:

Lemma D.2. *For any set of n samples,*

$$\hat{\mathcal{R}}_n(\ell_{\textit{s-hinge}} \circ \mathcal{B}) \leq \frac{B(4X + u_*)}{2\sqrt{n}}. \tag{47}$$

Proof. Define the function class $\mathcal{H} := \{x \mapsto \beta^\top x + z(y)u_*\|\beta\|_*\}$ ($z(y)$ follows notation from Levanon and Rosenfeld (2022)—in our case it is always equal to 1 but more generally we let it be a map from $\{\pm 1\} \mapsto \{\pm 1\}$). With the definition of Rademacher complexity,

$$\hat{\mathcal{R}}_n(\mathcal{H}) = \mathbb{E}_\sigma\left[\sup_{\beta\in\mathcal{B}, c\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^n \sigma_i \cdot (\beta^\top x + z(y)u_*\|\beta\|_*)\right] \tag{48}$$

$$\leq \mathbb{E}_\sigma\left[\sup_{\beta\in\mathcal{B}} \frac{1}{n}\sum_{i=1}^n \sigma_i \beta^\top x\right] + \mathbb{E}_\sigma\left[\sup_{\beta\in\mathcal{B}, c\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^n \sigma_i z(y) u_*\|\beta\|_*\right]. \tag{49}$$

The first term is the empirical Rademacher complexity of norm-bounded linear functions which is well known to be upper bounded by $\frac{2BX}{\sqrt{n}}$. An error in the proof by Levanon and Rosenfeld (2022) dropped the second term from the calculation of the Rademacher complexity. We observe that it is not zero, but we can bound it as follows:

Since $z(y) = \pm 1$ it can be dropped due to the symmetry of the Rademacher variables (since the y are fixed). Also, if the sum of the Rademacher variables is negative, the supremizing β will have dual norm 0, and if the sum is positive, it will have dual norm B. Thus,

$$\mathbb{E}_\sigma\left[\sup_{\beta\in\mathcal{B}, c\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^n \sigma_i u_*\|\beta\|_*\right] = \mathbb{P}\left(\sum \sigma_i > 0\right)\frac{u_* B}{n}\mathbb{E}_\sigma\left[\sum \sigma_i \middle| \sum \sigma_i > 0\right] \tag{50}$$

$$= \frac{u_* B}{2n}\mathbb{E}_\sigma\left[\left|\sum \sigma_i\right|\right] \tag{51}$$

$$\leq \frac{u_* B}{2n}\sqrt{\mathbb{E}_\sigma\left[\left(\sum \sigma_i\right)^2\right]} \tag{52}$$

$$= \frac{u_* B}{2\sqrt{n}}, \tag{53}$$

where the second equality is due to the symmetry of the distribution over σ and the inequality is by Jensen's. Since the function class $\ell_{\text{s-hinge}} \circ \mathcal{B}$ is generated by a 1-Lipschitz function applied to $\mathcal{H}$, the claim follows by Talagrand's contraction lemma. □

E. Proof for Full-Batch Subgradient Method

Theorem E.1. *Suppose we run the subgradient method on the regularized k-shifted strategic hinge loss as described in Algorithm 3 for T iterations and get classifier $\hat{\beta}$. Then with probability $\geq 1-\delta$, the worst-case 0-1 strategic loss under costs in $\mathcal{C}$ can be bounded by*

$$\max_{c\in\mathcal{C}} R^c_{0-1}(\hat{\beta}) \leq \max_{c\in\mathcal{C}} \hat{R}^c_{\textit{s-hinge}}(\hat{\beta}) + \frac{B(4X+u_*)+3\sqrt{\ln 2/\delta)}}{\sqrt{n}}. \tag{54}$$

Furthermore, the sub-optimality of $\hat{\beta}$ with respect to the population minimax solution is bounded by

$$\max_{c\in\mathcal{C}} \hat{R}^c_{\textit{s-hinge}}(\hat{\beta}) \leq \min_{\beta}\max_{c\in\mathcal{C}} R^c_{\textit{s-hinge}}(\beta) + B\left(\frac{L}{\sqrt{T}} + (X+u_*)\sqrt{\frac{\ln 2/\delta}{2n}}\right). \tag{55}$$

Proof. The first statement follows immediately with probability $\geq 1-\delta/2$ from Theorem D.1 since the bound holds uniformly for all $c\in\mathcal{C}$. Now we prove the second statement also holds with probability $\geq 1-\delta/2$, which we then combine via union bound. By Danskin's theorem, the function $r(\beta) := \max_{c\in\mathcal{C}} \hat{R}^c_{\text{s-hinge}}(\beta)$ is convex in β, and its subgradient is defined by $\partial_\beta \hat{R}^{c^*}_{\text{s-hinge}}(\beta)$ where $c^* := \arg\max_{c\in\mathcal{C}} \hat{R}^c_{\text{s-hinge}}(\beta)$.

A standard result says that if we run the subgradient method on $r(\beta)$ with step size $\eta = \frac{\epsilon}{L^2}$ for $T \geq \frac{B^2L^2}{\epsilon^2}$ steps, we will have $r(\beta^{t^*}) - \min_\beta r(\beta) \leq \epsilon$, which matches the hyperparameters of Algorithm 3 with $\epsilon = \frac{LB}{\sqrt{T}}$. The descent lemma in our setting is a bit different: we apply it to the $\Sigma_{\min}$-transformed gradient and show convergence in the norm $\|\cdot\|_{\Sigma_{\min}^{-1}}$. That is, our update is $\beta^{(t+1)} = \beta^{(t)} - \eta\Sigma_{\min} g_t$, and therefore

$$\|\beta^{(t+1)} - \beta^*\|^2_{\Sigma^{-1}_{\min}} = \|\beta^{(t)} - \beta^*\|^2_{\Sigma^{-1}_{\min}} - 2\eta g_t^\top(\beta^{(t)} - \beta^*) + \eta^2\|g_t\|^2_{\Sigma_{\min}} \tag{56}$$

$$\leq \|\beta^{(t)} - \beta^*\|^2_{\Sigma^{-1}_{\min}} - 2\eta\left(\hat{R}_{\text{s-hinge}}(\beta^{(t)}) - \hat{R}_{\text{s-hinge}}(\beta^*)\right) + \eta^2\|\Sigma^{1/2}_{\min} g_t\|_2^2. \tag{57}$$

Unrolling the chain of inequalities as in the typical convergence proof yields the same result, with the difference being that the Lipschitz constant will now be an upper bound on $\|\Sigma^{1/2}_{\min} g_t\|_2$. Here we observe that for all $\Sigma\in\mathcal{C}$ (and therefore for every max player cost choice Σ_t at each iteration),

$$\|\Sigma^{1/2}_{\min} g_t\|_2 \leq \left\|\Sigma^{1/2}_{\min}\left(x + u_*(1+\lambda)\frac{\partial\|\beta\|_{*,\Sigma_t}}{\partial\beta}\right)\right\|_2 \tag{58}$$

$$\leq \|\Sigma^{1/2}_{\min} x\|_2 + u_*(1+\lambda)\left\|\Sigma^{1/2}_{\min}\frac{\partial\|\Sigma_t^{-1/2}\beta\|_*}{\partial\beta}\right\|_2 \tag{59}$$

$$\leq X + u_*(1+\lambda)\left\|\Sigma^{1/2}_{\min}\Sigma_t^{-1/2}\frac{\partial\|\beta\|_*}{\partial\beta}\Big|_{\beta=\Sigma_t^{-1/2}\beta}\right\|_2 \tag{60}$$

$$\leq X + u_*(1+\lambda)\overbrace{\|\Sigma^{1/2}_{\min}\Sigma_t^{-1/2}\|_2}^{\leq 1}\left\|\frac{\partial\|\beta\|_*}{\partial\beta}\Big|_{\beta=\Sigma_t^{-1/2}\beta}\right\|_2 \tag{61}$$

$$\leq X + u_*(1+\lambda)L_* =: L, \tag{62}$$

where L_* is the Lipschitz constant of the gradient of the dual norm (recall for p-norms this is equal to $\max\left(1, d^{(p-2)/2p}\right)$). Now continuing with the standard proof of convergence of the subgradient method gives us the desired result.

It remains to show that Algorithm 3 successfully identifies the subgradient of r. We do so by invoking Lemma 4.1, which says that we can efficiently solve for c^* at each iteration. Thus we have that

$$\max_{c\in\mathcal{C}} \hat{R}^c_{\text{s-hinge}}(\hat{\beta}) \leq \min_\beta\max_{c\in\mathcal{C}} \hat{R}^c_{\text{s-hinge}}(\beta) + \frac{LB}{\sqrt{T}}, \tag{63}$$

noting that this bound is with respect to the minimax *empirical* risk. To complete the bound with respect to the minimax population risk, define $\beta^* := \arg\min_\beta \max_{c\in\mathcal{C}} R^c_{\text{s-hinge}}(\beta)$ as the population minimax solution. Then we have with probability $\geq 1 - \delta/2$

$$\min_\beta \max_{c\in\mathcal{C}} \hat{R}^c_{\text{s-hinge}}(\beta) \leq \max_{c\in\mathcal{C}} \hat{R}^c_{\text{s-hinge}}(\beta^*) \tag{64}$$

$$\leq \max_{c\in\mathcal{C}} R^c_{\text{s-hinge}}(\beta^*) + B(X+u_*)\sqrt{\frac{\ln 2/\delta}{2n}} \tag{65}$$

$$= \min_\beta \max_{c\in\mathcal{C}} R^c_{\text{s-hinge}}(\beta) + B(X+u_*)\sqrt{\frac{\ln 2/\delta}{2n}}, \tag{66}$$

where in the second inequality we've applied Hoeffding's between the empirical and population adversarial risk of β^* (which is bounded in $[0, B(X+u_*)]$) because that classifier does not depend on the training data. □

Algorithm 3 Subgradient method on k-shifted strategic hinge loss

input: Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, Iteration number T, Cost uncertainty set $\mathcal{C}$. Upper bound u_*, Regularization parameter λ.
define: $\beta^{(1)} \leftarrow \mathbf{0}$.
$\eta \leftarrow \frac{B}{L\sqrt{T}}$.
$\Sigma_{\min} \leftarrow \text{diag}([\sigma_{1\ell}, \ldots, \sigma_{d\ell}])$.
for $t = 1, \ldots, T$ **do**
1. $c_t \leftarrow \textsc{MaxLossCost}(\mathcal{D}, \beta^{(t)}, \mathcal{C}, u_*, \lambda)$.
2. Choose $g_t \in \partial_\beta[\hat{R}^{c_t}_{\text{s-hinge}}(\beta^{(t)}) + \lambda u_* \|\beta\|_*]$.
3. $\beta^{(t+1)} \leftarrow \beta^{(t)} - \eta \Sigma_{\min} g_t$
end for
return $\beta^{(t^*)}$, where $t^* := \arg\min_t \hat{R}^{c_t}_{\text{s-hinge}}(\beta^{(t)}) + \lambda u_* \|\beta^{(t)}\|_*$

subprocedure $\textsc{MaxLossCost}(\mathcal{D}, \beta, \mathcal{C}, u_*, \lambda)$
define: $\|\beta\|_*^{\min} := \min_{c \in \mathcal{C}} \|\beta\|_*$, $\|\beta\|_*^{\max} := \max_{c \in \mathcal{C}} \|\beta\|_*$.
initialize: $k \leftarrow \|\beta\|_*^{\min}$, $k_{\max} \leftarrow k$.
initialize: $r \leftarrow \hat{R}_{\text{s-hinge}}(\beta; u_* k) + \lambda u_* k$, $r_{\max} \leftarrow r$.
1. Evaluate $v_i = y_i(1 - y_i \beta^\top x_i)$ for all x_i.
2. Sort v_i in increasing order to get sorted indices j.
3. Initialize index $j \leftarrow \min j$ s.t. $v_j - u_* k > 0$.
4. Initialize counts $c_{+1} \leftarrow |\{j' \geq j : y_{j'} = +1\}|, c_{-1} \leftarrow |\{j' < j : y_{j'} = -1\}|$.
while $k < \|\beta\|_*^{\max}$ && $j < n$ **do**
$k' \leftarrow \min\{v_j / u_*, \|\beta\|_*^{\max}\}$.
$r \leftarrow r + u_*(k' - k)\left(\lambda + \frac{c_{-1} - c_{+1}}{n}\right)$.
$k \leftarrow k'$.
if $r > r_{\max}$ **then**
$r_{\max} \leftarrow r$, $k_{\max} \leftarrow k$.
end if
$c_{y_j} \leftarrow c_{y_j} - y_j$.
$j \leftarrow j + 1$.
end while

# *Found maximizing norm scalar, now need matrix Σ which induces it.*
if $j == n$ && $r + u_* [\|\beta\|_*^{\max} - k]\left(\lambda + \frac{n^-}{n}\right) > r_{\max}$ **then**
return $\arg\max_{c \in \mathcal{C}} \|\beta\|_* = \text{diag}([\sigma_{1\ell}, \ldots, \sigma_{d\ell}])$.
else
initialize: $\hat{\sigma} \leftarrow [\sigma_{1u}, \ldots, \sigma_{du}]$.
for $i = 1, \ldots, d$ **do**
Let $\hat{\Sigma} := \text{diag}([\hat{\sigma}_1, \ldots, \sigma_{i\ell}, \ldots, \hat{\sigma}_d])$.
if $\|\beta\|_{*,\hat{\Sigma}} < k_{\max}$ **then**
$\hat{\sigma}_i \leftarrow \sigma_{i\ell}$.
continue.
end if
Let $\hat{\Sigma}_{i=0} := \text{diag}([\hat{\sigma}_1, \ldots, 0_i, \ldots, \hat{\sigma}_d])$.
$\hat{\sigma}_i \leftarrow \frac{|\beta_i|}{\left(k_{\max}^p - \|\hat{\Sigma}_{i=0}^{-1/2} \beta\|_*^p\right)^{1/p}}$.
return $\text{diag}(\hat{\sigma})$.
end for
end if

F. Proof for Stochastic Mirror Descent-Ascent

We let $0 < \epsilon \leq 1$ denote the discretization parameter which tunes the size of the set $|\mathcal{S}|$ (for simplicity, assume $1/\epsilon$ is an integer). Specifically, we choose $1/\epsilon$ equally spaced points in each dimension's range of *inverse* eigenvalues and then define the elements of $\mathcal{S}$ to be the collection of smallest values in each dimension, then all the second-smallest values, etc. In this way we discretize the "diagonal" of the cost uncertainty set $\mathcal{C}$ to avoid an exponential dependence on the dimension.

Theorem F.1. *Suppose we run SMDA on the regularized strategic hinge loss as described in Algorithm 2 for T iterations and get averaged classifier iterates $\tilde{\beta}$. Define the convergence error*

$$\varepsilon_T := \max_{c \in \mathcal{C}} R^c_{\textit{s-hinge}}(\tilde{\beta}) - \min_{\beta} \max_{c \in \mathcal{C}} R^c_{\textit{s-hinge}}(\beta). \tag{67}$$

Then over the randomness of the optimization procedure it holds that

$$\mathbb{E}[\varepsilon_T] \lesssim B \left[u_* |1 - \lambda| \max_i \sqrt{\epsilon \left(\sigma_{i\ell}^{-2} - \sigma_{iu}^{-2} \right)} + \frac{L + (B^{-1} + X + u_*)\sqrt{\ln 1/\epsilon}}{\sqrt{T}} \right]. \tag{68}$$

Proof. From Proposition 4.3, we have that the regularized loss is convex in β. However, the loss is *not* concave in any parameterization of the cost function c. To resolve this, we discretize the space of cost functions. We parameterize the eigenvalues of the inverse cost matrix Σ^{-1} as a choice of eigenvalue for each eigenvector v_i from the compact set $[\sigma_{iu}^{-2}, \sigma_{i\ell}^{-2}]$. We do so by discretizing the space of choices linearly as $\sigma_k^{-2} = \sigma_{iu}^{-2} + k\epsilon(\sigma_{i\ell}^{-2} - \sigma_{iu}^{-2})$ for $k \in [1, 1/\epsilon]$. Now we can instead optimize

$$\min_{\beta} \max_{k \in [1/\epsilon]} R^{c(k)}_{\text{s-hinge}}(\beta), \tag{69}$$

where $R^{c(k)}_{\text{s-hinge}}$ denotes the loss under the cost defined by the eigenvalues induced by k. As this is a discrete set of $1/\epsilon$ choices, this objective is exactly equivalent to

$$\min_{\beta} \max_{\delta \in \Delta^{1/\epsilon}} \sum_{i=1}^{1/\epsilon} \delta_i R^{c(i)}_{\text{s-hinge}}(\beta), \tag{70}$$

where $\Delta^{1/\epsilon}$ is the simplex over $1/\epsilon$ values such that $\delta_i \geq 0 \; \forall i, \; \sum_i \delta_i = 1$. This objective *is* concave in δ, which means it can be solved via SMDA as described by Nemirovski et al. (2009). Let $\tilde{\beta}, \tilde{\delta}$ denote the two players' averaged iterates over choices β, δ. Define

$$\hat{\varepsilon}_T := \max_{k \in [1/\epsilon]} R^{c(k)}_{\text{s-hinge}}(\tilde{\beta}) - \min_{\beta} \sum_{i=1}^{1/\epsilon} \tilde{\delta}_i R^{c(i)}_{\text{s-hinge}}(\beta). \tag{71}$$

Note that this is the expected sub-optimality gap for a new optimization problem, whose solution is not the same as the one in the theorem statement. Nemirovski et al. (2009) prove that after T iterations of SMDA with the appropriate step size we have (in their original notation)

$$\mathbb{E}[\hat{\varepsilon}_T] \leq 2\sqrt{\frac{10[R_x^2 M_{*,x}^2 + M_{*,y}^2 \ln m]}{N}}. \tag{72}$$

Translating these terms into our notation,

$$m = 1/\epsilon, \tag{73}$$

$$N = T, \tag{74}$$

$$M_{*,x}^2 = \max_{1 \leq i \leq 1/\epsilon} \mathbb{E}[\|\nabla \ell^{c(i)}_{\text{s-hinge}}(\tilde{\beta})\|^2] \leq L^2, \tag{75}$$

$$M_{*,y}^2 = \mathbb{E}\left[\max_{1 \leq i \leq 1/\epsilon} |\ell^{c(i)}_{\text{s-hinge}}(\tilde{\beta})|^2 \right] \leq (1 + B(X + u_*))^2, \tag{76}$$

$$R_x^2 = \frac{1}{2} \max_{b_1, b_2 \in \mathcal{B}} \|b_1 - b_2\|_2^2 \lesssim B^2, \tag{77}$$

where the last inequality follows from the triangle inequality and the upper bound on $\|\beta\|_2$. Plugging these in gives the bound

$$\mathbb{E}[\hat{\varepsilon}_T] \lesssim \sqrt{\frac{B^2L^2 + (1 + B(X + u_*))^2 \ln 1/\epsilon}{T}} \tag{78}$$

$$\lesssim B\ \frac{L + (B^{-1} + X + u_*)\sqrt{\ln 1/\epsilon}}{\sqrt{T}}. \tag{79}$$

Next, we can rewrite the convergence error in the theorem statement in terms of this error as

$$\varepsilon_T := \max_{c\in\mathcal{C}} R^c_{\text{s-hinge}}(\tilde{\beta}) - \min_{\beta} \max_{c\in\mathcal{C}} R^c_{\text{s-hinge}}(\beta) \tag{80}$$

$$= \hat{\varepsilon}_T + \left(\max_{c\in\mathcal{C}} R^c_{\text{s-hinge}}(\tilde{\beta}) - \max_{k\in[1/\epsilon]} R^{c(k)}_{\text{s-hinge}}(\tilde{\beta})\right) \tag{81}$$

$$- \overbrace{\left(\min_{\beta} \max_{c\in\mathcal{C}} R^c_{\text{s-hinge}}(\beta) - \min_{\beta} \sum_{i=1}^{1/\epsilon} \tilde{\delta}_i R^{c(i)}_{\text{s-hinge}}(\beta)\right)}^{\geq 0} \tag{82}$$

$$\leq \hat{\varepsilon}_T + \left(\max_{c\in\mathcal{C}} R^c_{\text{s-hinge}}(\tilde{\beta}) - \max_{k\in[1/\epsilon]} R^{c(k)}_{\text{s-hinge}}(\tilde{\beta})\right). \tag{83}$$

This last term represents the error due to discretization. Revisiting the regularized risk definition, for any c and k we have

$$R^c_{\text{s-hinge}}(\tilde{\beta}) - R^{c(k)}_{\text{s-hinge}}(\tilde{\beta}) = \mathbb{E}\Big[\max\{0, 1 - y(\tilde{\beta}^\top x + u_*\|\Sigma_c^{-1/2}\tilde{\beta}\|_*)\} \tag{84}$$

$$- \max\{0, 1 - y(\tilde{\beta}^\top x + u_*\|\Sigma_{c(k)}^{-1/2}\tilde{\beta}\|_*)\}\Big] \tag{85}$$

$$+ \lambda u_*(\|\Sigma_c^{-1/2}\tilde{\beta}\|_* - \|\Sigma_{c(k)}^{-1/2}\tilde{\beta}\|_*) \tag{86}$$

$$\leq u_*|1-\lambda| \left|\|\Sigma_{c(k)}^{-1/2}\tilde{\beta}\|_* - \|\Sigma_c^{-1/2}\tilde{\beta}\|_*\right| \tag{87}$$

$$\leq u_*|1-\lambda| B \cdot \sigma_{\max}\left(\Sigma_{c(k)}^{-1/2} - \Sigma_c^{-1/2}\right) \tag{88}$$

by the reverse triangle inequality. Since these two matrices have the same eigenvectors, the maximum eigenvalue of their difference is simply the maximum absolute difference between their respective eigenvalues. By construction, for any choice Σ_c^{-1}, there is a choice $k \in [1/\epsilon]$ which differs in spectrum by no more than $\epsilon \cdot \max_i(\sigma_{i\ell}^{-2} - \sigma_{iu}^{-2})$ in any given direction, and therefore we have

$$\max_{c\in\mathcal{C}} R^c_{\text{s-hinge}}(\tilde{\beta}) - \max_{k\in[1/\epsilon]} R^{c(k)}_{\text{s-hinge}}(\tilde{\beta}) \leq u_* B \cdot |1-\lambda| \cdot \max_i \left|\sqrt{\sigma_i\left(\Sigma_{c(k)}^{-1}\right)} - \sqrt{\sigma_i\left(\Sigma_c^{-1}\right)}\right| \tag{89}$$

$$\leq u_* B \cdot |1-\lambda| \cdot \max_i \sqrt{\left|\sigma_i\left(\Sigma_{c(k)}^{-1}\right) - \sigma_i\left(\Sigma_c^{-1}\right)\right|} \tag{90}$$

$$\leq u_* B \cdot |1-\lambda| \cdot \max_i \sqrt{\epsilon\left(\sigma_{i\ell}^{-2} - \sigma_{iu}^{-2}\right)}. \tag{91}$$

Combining this bound with the one above on $\hat{\varepsilon}_T$ and taking expectations gives the result. □

Corollary F.2. *Recall $D := \max_i(\sigma_{i\ell}^{-2} - \sigma_{iu}^{-2})$. Choosing $\epsilon = \Theta\left(\frac{\ln T}{T\max(1,D)}\right)$, we have*

$$\mathbb{E}[\varepsilon_T] \lesssim \frac{LB}{\sqrt{T}} + B(X + u_*)\sqrt{\frac{\ln T + \max(0, \ln D)}{T}}. \tag{92}$$

G. Goodhart's Law Under Known Costs

We here show that when each user's strategic response to a classifier is known, then in principle (*information theoretically,* discarding optimization concerns), strategic response has no effect on predictive performance.

We start with a correspondence between classifiers operating on (non-strategic) inputs and appropriately modified classifiers operating on strategically modified inputs. In line with prior works—and unlike the other results in this paper—we make the additional assumption that u is fixed and ϕ is strictly monotone (thus $u_* = \phi^{-1}(u)$) We also explicitly parameterize the bias term in the classifiers because it plays an important role in the result.

Proposition G.1. *Let $f(x) = \mathbf{1}\{\beta^\top x + b \geq 0\}$ be the prediction of the classifier parameterized by (β, b), and let $f'(x) = \mathbf{1}\{\beta^\top x + b' \geq 0\}$ denote this classifier with a shifted bias $b' = b - u_*\|\beta\|_*$. Then for all x, it holds that $f(x) = f'(x')$, where $x' = x(\beta, b')$ is the new location of x after strategic response to the classifier (β, b').*

This results states that f outputs on every data point x exactly what f' (which only differs in its shifted bias term) outputs on x'—which represents the exact same "user" *after it has strategically responded*.

Proof. The proof is simple. We consider three separate cases:

- If $f(x) = 0$ and x is too far to cross the decision boundary of f (so $x'(\beta, b) = x$), then since $u_*\|\beta\|_* \geq 0$, x is also too far to cross the decision boundary of f', (that is, $x'(\beta, b') = x$). Therefore, $f'(x') = f'(x) = f(x)$.
- If $f(x) = 0$ and x is close enough to cross the decision boundary (so $x'(\beta, b) \neq x$), then Lemma 2.3 implies that the maximum amount by which x will change its linear prediction under β is $u_*\|\beta\|_*$. Since $f(x) = 0 \implies 0 > \beta^\top x + b \implies -u_*\|\beta\|_* > \beta^\top x + b'$, this means x cannot force $f'(x') = 1$ without increasing the prediction by more than $u_*\|\beta\|_*$, and thus it will not move. So, $f'(x') = f'(x) = f(x)$.
- If $f(x) = 1$, then $0 \leq \beta^\top x + b \implies -u_*\|\beta\|_* \leq \beta^\top x + b'$. Thus, either x will already get a positive classification from f', or it will be able to move enough to cross the decision boundary. Either way, $f'(x') = 1 = f(x)$. □

Note that this proof applies even if we *don't* know the user's cost function—it is sufficient to know for each user the maximum potential increase in their predicted logit under β after strategic response, i.e. $\beta^\top(x'(\beta) - x)$. Applying this insight to the optimal classifier gives the following result:

Corollary G.2. *Fix some distribution $\mathcal{D}$, and let f^* with parameters (β^*, b^*) be the classifier minimizing the* non-strategic *0-1 risk on $\mathcal{D}$. Denote this risk as α. Then for f' with $(\beta^*, b^* - u_*\|\beta^*\|_*)$, its expected* strategic *0-1 risk on $\mathcal{D}$ is exactly α.*

The take-away is that if we are able to minimize the standard 0-1 loss, and we know how users will respond (e.g. by knowing the exact cost function), then a trivial modification would provide us with an optimal classifier for the strategic 0-1 loss. Thus, strategic response does not pose any additional statistical difficulty over standard classification.

Importantly, however, this result does *not* imply that minimizing a *proxy* loss (such as the hinge or logistic loss) on non-strategic data and applying the transformation would give a good strategic classifier; this precisely why the strategic hinge loss is needed in the first place. Further, this result does not account for the social cost of the classifier f' versus some other classifier (Milli et al., 2019)—there could be a different classifier with similar accuracy under strategic response that induces a smaller cost to the users.