Strategic Feature Selection

Jivat Neet Kaur
UC Berkeley
jnkaur@berkeley.edu

Pratik Patil
UT Austin
pratikpatil@utexas.edu

Divya Shanmugam
Columbia University
dms2351@cumc.columbia.edu

Emma Pierson
UC Berkeley
emmapierson@berkeley.edu

Michael I. Jordan
UC Berkeley & Inria
jordan@cs.berkeley.edu

Nika Haghtalab
UC Berkeley
nika@berkeley.edu

Meena Jagadeesan*
University of Pennsylvania
meenaj@cis.upenn.edu

Ahmed Alaa*
UC Berkeley
amalaa@berkeley.edu

Serena Wang*
University of British Columbia
serenalw@cs.ubc.ca

Abstract

When algorithmic predictors inform resource allocation in high-stakes domains such as healthcare, these predictors must account for strategic manipulation of input features. The typical solution is to redesign the predictor itself to explicitly account for strategic interactions. In practice, however, decision makers are often constrained to adjusting coarser levers within existing prediction pipelines. For example, healthcare organizations often select which features to exclude based on perceived manipulability, while using standard regularization procedures to shrink the coefficients of retained features. In this work, we initiate a formal study of strategic classification through *feature selection* and its interaction with *ridge regularization*. Our main finding is that excluding individual features based on their manipulability alone is generally suboptimal. We provide a fine-grained characterization of the performance of a feature subset under optimal regularization, yielding new insights for policy design. Motivated by this characterization, we develop a practical algorithm for jointly choosing the feature set and the level of ridge regularization. Through a real-world case study on a healthcare payments benchmark, we illustrate how our algorithm can guide the design of coarse policy levers in practice. Our results provide a principled, practical framework for mitigating the effects of strategic behavior in algorithmic decision-making systems.

1 Introduction

Algorithmic predictions are increasingly used to inform the allocation of benefits and services in high-stakes domains. When these systems determine eligibility or allocation amounts from individuals' *reported features*, they create incentives for strategic behavior. In settings such as large-scale government-run resource allocation programs, organizations serving individuals may manipulate reported features in order to increase the predicted allocation. For instance, consider the Medicare Advantage program, in which the government pays private insurers to provide coverage for beneficiaries based on patient risk scores [12]. In this setting, insurers are known to report more or higher severity diagnoses for their enrollees in order to receive higher payments from the government [38]. This practice, known as "upcoding," has become increasingly controversial and is projected to cost the government $40 billion in additional spending in 2025, with no clear benefit to the beneficiaries [26].

*Equal contribution

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

These types of incentives are captured by the framework of strategic classification [18, 34], which models the strategic interaction between a decision maker who specifies the prediction rule and an agent who can modify features in response to that rule. In the Medicare Advantage setting, the agent is an organization that can modify the reported features of the individuals it serves in order to obtain higher payments. Organizations incur costs to manipulate features: for example, in health insurance, insurers pay chart review contractors to identify additional diagnosis codes. The typical approach in this framework is to design a prediction rule that is less vulnerable to strategic manipulation.

Much of the strategic classification literature focuses on computing the *strategic optimum*, i.e., prediction rules that are optimal once these strategic responses are taken into account. In large-scale policy settings, however, structural redesign faces significant institutional inertia and systems are rarely rebuilt from scratch. Models are embedded in regulatory and operational workflows, updated incrementally, and substantial departures from standard pipelines are difficult to justify and defend in regulatory rulemaking. As a result, decision makers often continue to use legacy prediction pipelines and instead leverage a small set of *coarse policy levers* to address strategic behavior. This raises a different question than the one typically studied in strategic learning: not how to redesign the entire predictor, but how to improve strategic performance using only the limited levers that institutions are constrained to use.

In practice, it is common to combat strategic behavior using feature selection together with regularization. One common intervention is to *exclude features* that are perceived to be especially manipulable from the model [4, 20]. For example, to counteract the effect of upcoding, Centers for Medicare & Medicaid Services (CMS) excludes features (i.e., diagnoses) that are at risk of inappropriate coding by health insurers [6, 7]. In 2024, this included removing the health conditions corresponding to Protein-Calorie Malnutrition and Angina Pectoris from the payment model to limit the model's sensitivity to higher coding intensity [7]. Another intervention is to *use regularization to shrink coefficients* on the retained features, often implemented through constrained regression. For instance, CMS recently constrained the coefficients on related diagnoses such as uncomplicated diabetes and diabetes with complications so that they carry the same weight in the payment model [7]. Given that such heuristics are used in practice, this raises a central question: *How should policymakers design feature selection together with regularization to combat strategic behavior?*

1.1 Our contributions

In this work, we initiate a formal study of strategic learning using *feature selection* and its interaction with *ridge regularization*. Our main finding is that excluding individual features based on their manipulability alone is generally suboptimal: instead, features should be selected *jointly* based on their manipulability and predictability relative to the other available features.

Theoretical results. We theoretically characterize the performance of a feature set under optimally tuned regularization, comparing it to the strategic optimum had the pipeline been redesigned. First, we establish a lower bound on this performance gap, identifying an irreducible loss that remains no matter which feature set is chosen (Theorem 1). Next, we establish an upper bound that characterizes when a feature subset is near-optimal up to this irreducible gap (Theorem 2). Specifically, our characterization reveals that the performance of a feature set is not determined by manipulability alone, but by the joint predictability-manipulability structure of its features and by how effectively regularization can leverage that structure. Our results also illustrate special cases where the coarse levers exactly achieve the irreducible gap.

Policy implications. Our theoretical results provide concrete insights for policymakers (Section 4). First, regularization can change the optimal feature set, so the subset that is best without regularization need not remain best under ridge regression; hence, feature selection and regularization must be tuned jointly (Figure 1). Second, a highly manipulable feature may still be kept if it is predictive and other retained features have similar manipulability (Figure 2). Finally, dropping a manipulable feature becomes desirable when a less manipulable proxy is available (Figure 5). These insights run counter to prevailing policy views that highly manipulable features should simply be excluded and that these two levers can be treated as separate design choices [8, 20].

Algorithmic and empirical results. To solve the corresponding design problem, we develop a practical algorithm for jointly choosing the feature set and regularization level in the strategic setting.

We propose a two-stage procedure that performs a continuous relaxation of the combinatorial support selection optimization problem, followed by local support refinement (Section 5). We complement our theoretical analysis with a healthcare payments case study with simulated upcoding data calibrated to realistic Medicare Advantage coding patterns documented in recent health policy work [20] (Section 6). Empirical evaluation shows that our proposed method substantially improves strategic robustness while preserving much of the predictive accuracy of full-support models. Notably, our results reveal that high manipulability alone need not necessitate exclusion.

Discussion. Altogether, our results provide a theoretically principled and practical framework for policymakers to design and evaluate feature selection and regularization policies in the presence of strategic behavior. They clarify when these coarse levers are sufficient to achieve near-optimal performance under strategic interactions, and how these levers should be designed jointly to optimize for strategic performance. At the same time, they highlight an important disconnect between existing heuristic practices and the true design problem, yielding nuanced implications for policy design. Finally, these levers offer robustness benefits under cost uncertainty and heterogeneity (Appendix H.1), illustrating their value as practical tools for strategic decision-making.

1.2 Related work

Strategic classification. We build on a growing line of work in strategic classification aimed at learning optimal prediction rules when decision subjects can manipulate their features at a cost [18, 10, 9, 36, 22]. Closely related to our work, Björkegren et al. [4] study manipulation-robust prediction in a setting similar to ours, under linear decision rules and quadratic manipulation costs. These works fall within a broader literature on incentive-aware machine learning (see Podimata [34] for a survey), which also studies strategic behavior from an improvement and causality perspective [28, 16, 19, 36], considering issues of incentive design when manipulations can lead to genuine improvement in individual outcomes and welfare. In contrast with all of these works, we study strategic classification through *coarse policy levers* within a fixed prediction pipeline rather than by redesigning the pipeline to achieve the strategically optimal solution.

Policy. Health policy also contends with strategic behavior, where payment models based on recorded diagnoses are used to allocate payments to insurers. In this setting, insurers have been shown to strategically engineer recorded diagnoses to increase payment allocations [38]. The health policy literature has put forth several proposals to combat such behavior. Rose [35] shows that a payment formula with a reduced set of variables can retain much of the predictive performance of richer formulas, potentially reducing incentives for aggressive upcoding. More broadly, this literature has proposed feature selection and coefficient restriction in such payment models. These proposals suggest, for example, excluding diagnoses added to a patient's health record via chart review [27], incorporating patient survey data [3, 25], or excluding diagnoses coded most differentially by insurers in order to reduce overpayments [20]. Closer to the levers we study, McGuire et al. [24] show how constrained regression and variable selection can improve the performance of risk adjustment systems. In a similar vein, Arias [2] proposes a LASSO-based method that balances predictive accuracy against potential adverse incentives, explicitly accounting for heterogeneity in the adverse incentives associated with each variable.

Related concerns arise beyond health policy in problems where the goal is to allocate limited resources to eligible individuals in order to maximize welfare. In public economics, proxy means tests are eligibility rules that use easier-to-observe variables such as household characteristics and asset ownership as proxies for income to target social programs to those in most need. Strategic misreporting of individuals' characteristics is an important policy concern in such settings [23, 5]. Work on proxy means tests emphasizes that the variables used to determine eligibility should be both predictive of welfare and verifiable in practice [14]. Niehaus et al. [30] study how rules for targeting benefits are designed and show that adding more criteria can create a tradeoff between statistical gains and enforceability.

Our work extends this agenda by providing a formal strategic framework for characterizing the performance of feature selection and coefficient restriction decisions, and in doing so contributes new perspectives on longstanding policy issues.

2 Model and Preliminaries

Building on Shavit et al. [36], Kleinberg and Raghavan [19], we study the strategic interaction between a decision maker who specifies a linear predictor and organizations that can modify the reported features of the individuals they serve. The decision maker publishes a predictor $f_{\theta,b}(x) = \theta^\top x + b$ parameterized by $\theta \in \mathbb{R}^d$ and $b \in \mathbb{R}$, which maps an individual's feature vector $x \in \mathcal{X} \subseteq \mathbb{R}^d$ to predicted outcomes. We model the population as a distribution $\mathcal{P}_{X,Y}$ over (X, Y) with $Y \in \mathbb{R}$, finite second moments, and feature covariance matrix $\Sigma := \mathbb{E}[XX^\top] \succ 0$. For simplicity we assume $\mathbb{E}[X] = 0$ and $\mathbb{E}[Y] = 0$ without loss of generality. Because our focus is on linear predictors, we define

$$\theta^* := \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}[(Y - \theta^\top X)^2]. \tag{1}$$

Since $\Sigma \succ 0$, this minimizer is unique and is given by $\theta^* = \Sigma^{-1}\mathbb{E}[XY]$. Writing $Y = \theta^{*\top} X + \varepsilon$, where $\mathbb{E}[X\varepsilon] = 0$, $\sigma^2 := \mathbb{E}[\varepsilon^2]$, we can interpret θ^* as the best linear projection coefficient, $\varepsilon := Y - \theta^{*\top} X$ as the projection residual, and σ^2 as the irreducible variance.

Following Hardt et al. [18], Björkegren et al. [4], we assume that the decision maker has access to *unmanipulated* samples from $\mathcal{P}_{X,Y}$. Such data can be obtained from a pre-deployment period, before the predictor $f_{\theta,b}$ is used, or from an alternative policy regime under which organizations are not incentivized to manipulate their features. Using these unmanipulated samples, θ^* can be estimated.

Organization's best response model. The organization observes an individual's features x and chooses an action $a \in \mathbb{R}^d$, which changes the reported features to $x + a$ at cost $C(a) \geq 0$, with strict positivity for nonzero a. Adopting the quadratic manipulation costs in earlier work [36, 4], we model the cost function as $C(a) = \frac{1}{2} a^\top H a$, where $H \in \mathbb{R}^{d \times d} \succ 0$ is the cost matrix. The matrix H allows manipulation costs to differ across features and to exhibit interactions across coordinates.

Given a published predictor, we assume that organizations *best-respond* to $f_{\theta,b}$ by choosing an action

$$a^*(x; \theta) \in \arg\max_{a \in \mathbb{R}^d} \left\{ \theta^\top (x + a) + b - \tfrac{1}{2} a^\top H a \right\}. \tag{2}$$

From here on, we suppress the dependence on x and write $a^*(\theta)$.

Decision maker's objective. The decision maker seeks to predict the true outcome as accurately as possible in the presence of strategic responses, which means minimizing prediction error under the deployed policy. Accordingly, the relevant loss is not the usual mean squared error but rather

$$\mathrm{MSE}_{\mathrm{strat}}(\theta, b) := \mathbb{E}\big[(\theta^\top (X + a^*(\theta)) + b - Y)^2\big]. \tag{3}$$

As a benchmark, we use the notion of the *strategic optimum* [18], allowing the decision maker to optimize over all coefficients in $\mathbb{R}^d$ and intercept in $\mathbb{R}$. The unrestricted strategic optimum is

$$\mathrm{OPT} := \min_{\theta \in \mathbb{R}^d,\, b \in \mathbb{R}} \mathrm{MSE}_{\mathrm{strat}}(\theta, b). \tag{4}$$

Coarse policy levers. Our analysis focuses on two *coarser* policy levers that are currently leveraged to combat strategic behavior in practice.

Feature selection. Intuitively, feature selection provides a direct way to limit strategic manipulation by excluding features whose vulnerability to gaming outweighs their predictive value. For a support $\mathsf{S} \subseteq [d]$, let $\mathsf{R} = [d] \setminus \mathsf{S}$ and define the coordinate subspace $\Theta(\mathsf{S}) := \{\theta \in \mathbb{R}^d : \mathrm{supp}(\theta) \subseteq \mathsf{S}\}$. The support-restricted least-squares estimator is:

$$(\theta^{\mathrm{LS}}(\mathsf{S}), b^{\mathrm{LS}}(\mathsf{S})) \in \arg\min_{\theta \in \Theta(\mathsf{S}), b \in \mathbb{R}} \mathbb{E}[(Y - \theta^\top X - b)^2]. \tag{5}$$

Ridge regularization. Regularization reduces the magnitude of model coefficients and thereby weakens incentives to manipulate the retained features. We study the canonical form of regularization given by ordinary ridge regression. For $\lambda \geq 0$, the ordinary ridge estimator is:

$$(\theta^{\mathrm{OR}}(\lambda), b^{\mathrm{OR}}(\lambda)) \in \arg\min_{\theta \in \mathbb{R}^d, b \in \mathbb{R}} \mathbb{E}[(Y - \theta^\top X - b)^2] + \lambda \|\theta\|_2^2. \tag{6}$$

To model the two levers jointly, we consider the *support-restricted ridge* estimator:

$$(\theta^{\mathrm{OR}}(\mathsf{S},\lambda), b^{\mathrm{OR}}(\mathsf{S},\lambda)) \in \underset{\theta\in\Theta(\mathsf{S}),b\in\mathbb{R}}{\arg\min}\ \mathbb{E}[(Y-\theta^\top X-b)^2]+\lambda\|\theta\|_2^2. \tag{7}$$

This estimator nests support-restricted least squares when $\lambda = 0$ and full-support ridge when $\mathsf{S} = [d]$.

3 Theoretical Results

We compare the strategic MSE achieved by the support-restricted ordinary ridge estimator to the strategic optimum OPT. Our goal is to characterize when the coarse policy levers of feature selection and ridge regularization can come close to this benchmark. In Section 3.1, we establish a lower bound on the optimality gap, which identifies an irreducible gap incurred by our estimator class. In Section 3.2, we establish an upper bound that gives interpretable conditions under which optimally tuned support-restricted ridge can nearly match OPT up to this irreducible loss.

3.1 Lower bound on the excess strategic MSE

We start by characterizing the fundamental limits of these coarse policy levers through a lower bound. As a tool for interpretation, we introduce the notation $\widetilde{\mathrm{OPT}} := \min_{\theta\in\mathbb{R}^d} \mathrm{MSE}_{\mathrm{strat}}(\theta, 0)$ and $\widetilde{\mathrm{OPT}}(\mathsf{S}) := \min_{\theta\in\Theta(\mathsf{S})} \mathrm{MSE}_{\mathrm{strat}}(\theta, 0)$.

Theorem 1 (Lower bound on excess strategic MSE)**.** *For any subset* $\mathsf{S} \subseteq [d]$ *and* $\lambda \geq 0$*,*

$$\mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S},\lambda), b^{\mathrm{OR}}(\mathsf{S},\lambda)) - \mathrm{OPT} \geq \widetilde{\mathrm{OPT}} - \mathrm{OPT}. \tag{8}$$

Proof sketch. Although OPT is defined by optimizing over both coefficients and an intercept, every estimator considered here is fit using a centered support-restricted ridge program and therefore has intercept 0 under our centering convention, i.e., $b^{\mathrm{OR}}(\mathsf{S},\lambda) = 0$ (see Lemma B.5). For a fixed support S, the estimator $\theta^{\mathrm{OR}}(\mathsf{S},\lambda)$ belongs to the class $\Theta(\mathsf{S})$, and therefore cannot outperform the support-restricted zero-intercept benchmark $\widetilde{\mathrm{OPT}}(\mathsf{S})$. Moreover, $\widetilde{\mathrm{OPT}}(\mathsf{S})$ is monotonically decreasing as the support S grows. □

Theorem 1 shows that the limits of the coarse levers can be characterized by a gap $\widetilde{\mathrm{OPT}} - \mathrm{OPT}$. This irreducible gap is small whenever the strategic shift induced by the best linear projection is small, for example when the manipulation costs are high in the predictive directions of θ^*: $0 \leq \widetilde{\mathrm{OPT}} - \mathrm{OPT} \leq (\theta^{*\top} H^{-1}\theta^*)^2$ (stated formally in Proposition C.1).

In the remainder of the section, we therefore upper bound the additional loss from support restriction and ridge regularization relative to $\widetilde{\mathrm{OPT}}$; the total excess relative to OPT is obtained by adding this irreducible gap. For notational simplicity, throughout the remainder of the paper we write $\mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S},\lambda))$ for $\mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S},\lambda), b^{\mathrm{OR}}(\mathsf{S},\lambda)) = \mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S},\lambda), 0)$, using that $b^{\mathrm{OR}}(\mathsf{S},\lambda) = 0$.

3.2 Upper bound on the excess strategic MSE

We now turn to upper bounds on the $\mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S},\lambda))$. Since $\widetilde{\mathrm{OPT}}$ serves as a lower bound per Theorem 1, we bound the gap between $\mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S},\lambda))$ and $\widetilde{\mathrm{OPT}}$. This upper bound illustrates that feature selection should be based on *relative* manipulability and predictability.

Theorem 2 (Upper bound on the excess strategic MSE)**.** *For any subset* $\mathsf{S} \subseteq [d]$*,*

$$\min_{\lambda\geq 0} \mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S},\lambda)) - \widetilde{\mathrm{OPT}} \leq -\underbrace{\big(\Gamma([d]) - \Gamma(\mathsf{S})\big)}_{\textit{manipulability gain}} + \underbrace{L_{\mathrm{pred}}(\mathsf{S})}_{\textit{predictive loss}} + \underbrace{C_H(\mathsf{S})\,\delta_H(\mathsf{S})^2}_{\textit{heterogeneity gap}}, \tag{9}$$

where the strategic burden Γ*, predictive loss* L_{pred}*, and heterogeneity defect* δ_H *for a support* S *are:*

1. $\Gamma(\mathsf{S}) := \min_{\eta\in\mathbb{R}^{|\mathsf{S}|}} \left\{(\eta - \theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S}))^\top \Sigma_{\mathsf{SS}}(\eta - \theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S})) + (\eta^\top H^{-1}_{\mathsf{SS}}\eta)^2\right\}$,
2. $L_{\mathrm{pred}}(\mathsf{S}) := (\theta^*_{\mathsf{R}})^\top \Sigma_{\mathsf{R}|\mathsf{S}}\theta^*_{\mathsf{R}}$ *and* $\Sigma_{\mathsf{R}|\mathsf{S}} := \Sigma_{\mathsf{RR}} - \Sigma_{\mathsf{RS}}\Sigma^{-1}_{\mathsf{SS}}\Sigma_{\mathsf{SR}}$ *for* $\mathsf{R} = [d] \setminus \mathsf{S}$*,*

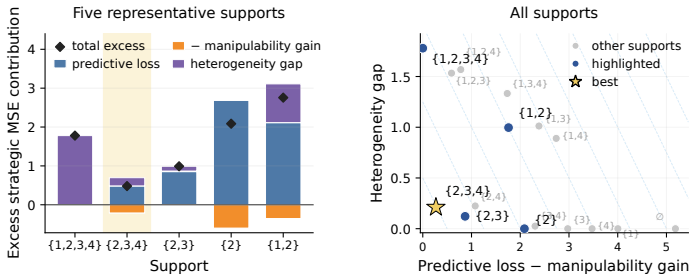

Figure 1: **The best subset depends on the joint predictability-manipulability structure.** In a $d = 4$ example, feature 1 is highly predictive but highly manipulable. Feature 2 is strongly correlated with feature 1 and therefore acts as a less manipulable proxy. Features 3, 4 are less manipulable features with moderate predictive value. (Left) Excess strategic MSE decomposition for five representative supports. (Right) All supports plotted by predictive loss minus manipulability gain and the gap after optimal ridge tuning; dashed lines are equal excess strategic MSE. The gap after optimal ridge tuning is bounded above by the heterogeneity gap $C_H(\mathsf{S})\delta_H(\mathsf{S})^2$. The optimal set $\{2, 3, 4\}$ is neither the full support nor the support with the largest manipulability gain, but the one that best balances predictive loss, manipulability gain, and cost heterogeneity. See Appendix D.1 for details.

3. $\delta_H(\mathsf{S}) := \inf_{\gamma \geq 0} \left\| \Sigma_{\mathsf{SS}}^{-1/2} (H_{\mathsf{SS}}^{-1} - \gamma I) \Sigma_{\mathsf{SS}}^{-1/2} \right\|_{\mathrm{op}}$,

and $C_H(\mathsf{S}) := 4\, L_H(\mathsf{S})\, \|\Sigma_{\mathsf{SS}}^{-1/2} H_{\mathsf{SS}}^{-1} \Sigma_{\mathsf{SS}}^{-1/2}\|_{\mathrm{op}}^2\, \|\Sigma_{\mathsf{SS}}^{1/2} \theta_{\mathsf{S}}^{\mathrm{LS}}(\mathsf{S})\|_2^6$, $L_H(\mathsf{S}) := 1 + 6\|\Sigma_{\mathsf{SS}}^{-1/2} H_{\mathsf{SS}}^{-1} \Sigma_{\mathsf{SS}}^{-1/2}\|_{\mathrm{op}}^2 \|\Sigma_{\mathsf{SS}}^{1/2} \theta_{\mathsf{S}}^{\mathrm{LS}}(\mathsf{S})\|_2^2$.

Theorem 2 gives a fine-grained characterization of the performance of a feature set under optimal regularization. Note here we write $\theta_{\mathsf{S}}^{\mathrm{LS}}(\mathsf{S}) := \Sigma_{\mathsf{SS}}^{-1} \Sigma_{\mathsf{S}y}$ for the retained coordinates of the support-restricted least squares estimator $\theta^{\mathrm{LS}}(\mathsf{S}) = (\theta_{\mathsf{S}}^{\mathrm{LS}}(\mathsf{S}), 0_{\mathsf{R}})$, where $\Sigma_{\mathsf{S}y} := \mathbb{E}[X_{\mathsf{S}} Y] = \Sigma_{\mathsf{SS}} \theta_{\mathsf{S}}^* + \Sigma_{\mathsf{SR}} \theta_{\mathsf{R}}^*$ for $\mathsf{R} = [d] \setminus \mathsf{S}$.

The first term, $-(\Gamma([d]) - \Gamma(\mathsf{S}))$, is the *manipulability gain*. For a given feature set, $\Gamma(\mathsf{S})$ captures how much performance is lost because the predictor must move away from the best-fitting coefficients on S in order to account for manipulability: the first term $(\eta - \theta_{\mathsf{S}}^{\mathrm{LS}}(\mathsf{S}))^\top \Sigma_{\mathsf{SS}} (\eta - \theta_{\mathsf{S}}^{\mathrm{LS}}(\mathsf{S}))$ measures the fit lost by deviating from the least squares fit on S, and the second term $(\eta^\top H_{\mathsf{SS}}^{-1} \eta)^2$ measures the loss due to manipulability. Thus, manipulability gain is the *strategic burden* relieved by restricting attention to S instead of the full feature set $[d]$. This reflects the existing intuition that excluding manipulable features lowers the strategic MSE. As we observe, however, manipulability gain alone does not determine excess strategic MSE. The performance of a feature set depends on the joint prediction-manipulation structure.

The second term, $L_{\mathrm{pred}}(\mathsf{S})$, is the usual non-strategic *predictive loss* from omitting features. It is small when the omitted features are redundant given the retained support. More precisely, it measures the predictive signal in the excluded coordinates R that cannot be linearly recovered from S.

The third term, $C_H(\mathsf{S})\, \delta_H(\mathsf{S})^2$, is the *heterogeneity gap* on feature set S. The quantity $\delta_H(\mathsf{S})$ is small when the features in S have sufficiently homogeneous manipulation costs (i.e., H_{SS}^{-1} is well approximated by a scalar multiple of the identity). In this regime, a single regularization level can shrink the retained coefficients in a coordinated way, so optimally tuned ridge regularization reduces the heterogeneity gap and thereby lowers excess strategic MSE. Figure 1 illustrates the three terms.

Proof sketch. There are two steps. First, for any fixed λ, we add and subtract the benchmark $\widetilde{\mathrm{OPT}}(\mathsf{S})$:

$$\mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda)) - \widetilde{\mathrm{OPT}} = \underbrace{\left(\widetilde{\mathrm{OPT}}(\mathsf{S}) - \widetilde{\mathrm{OPT}}\right)}_{\textit{support restriction cost}} + \underbrace{\left(\mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda)) - \widetilde{\mathrm{OPT}}(\mathsf{S})\right)}_{\textit{scalar-ridge expressivity cost}}.$$

The support restriction identity in Proposition C.2 gives $\widetilde{\mathrm{OPT}}(\mathsf{S}) - \widetilde{\mathrm{OPT}} = L_{\mathrm{pred}}(\mathsf{S}) - (\Gamma([d]) - \Gamma(\mathsf{S}))$. Second, for any fixed support S, we control the scalar-ridge expressivity cost. Lemma C.2 shows that the benchmark $\widetilde{\mathrm{OPT}}(\mathsf{S})$ corresponds to a generalized ridge solution with penalty proportional to H_{SS}^{-1}, whereas ordinary ridge uses a scalar penalty λI. The fixed-support approximation bound in Theorem C.1 then gives $\min_{\lambda \geq 0} \mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda)) - \widetilde{\mathrm{OPT}}(\mathsf{S}) \leq C_H(\mathsf{S}) \delta_H(\mathsf{S})^2$. □

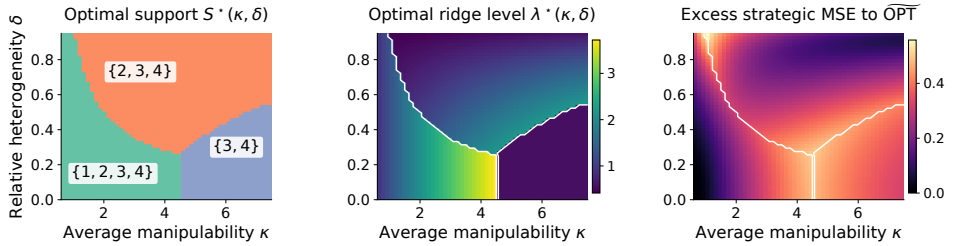

Figure 2: **Manipulable feature groups with homogeneous costs can be retained and regularized.** We compute the optimal support and regularization level exactly for a $d = 4$ example with diagonal costs. Features $\{1, 2\}$ are a manipulable pair with average manipulability κ and relative cost heterogeneity δ within the pair; features 3 and 4 have low manipulability. We vary κ on the x-axis and δ on the y-axis. (Left) The exact optimal support $\mathsf{S}^\star(\kappa, \delta)$ after joint optimization over S and λ. (Middle) The corresponding optimal ridge level $\lambda^\star(\kappa, \delta)$, where white contours indicate support boundaries. (Right) Excess strategic MSE relative to $\widetilde{\mathrm{OPT}}$. When the manipulable pair is homogeneous (δ small), it is often retained even at high manipulability, with the optimal response being to increase λ. The resulting estimator remains close to optimal over a broad range. See Appendix D.2 for details.

When are the coarse levers near-optimal? Theorem 2 gives an immediate ϵ-optimality condition. If there exists a support S such that $L_{\mathrm{pred}}(\mathsf{S}) - (\Gamma([d]) - \Gamma(\mathsf{S})) + C_H(\mathsf{S})\delta_H(\mathsf{S})^2 \le \epsilon$, then optimally tuned support-restricted ridge satisfies $\min_{\mathsf{T}\subseteq[d],\,\lambda\ge 0} \mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{T}, \lambda)) \le \widetilde{\mathrm{OPT}} + \epsilon$. Thus, relative to the unrestricted strategic optimum, the coarse levers are near-optimal up to the irreducible gap $\widetilde{\mathrm{OPT}} - \mathrm{OPT}$ whenever there exists a support with little predictive loss, sufficient reduction in strategic burden, and sufficiently homogeneous retained manipulation costs.

The heterogeneity gap can vanish in natural cases: if the retained inverse-cost matrix is isotropic such that $H_{\mathsf{SS}}^{-1} = \gamma I$, then scalar ridge exactly attains the support-restricted zero-intercept oracle: $\min_{\lambda\ge 0} \mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda)) = \widetilde{\mathrm{OPT}}(\mathsf{S})$. Thus a highly manipulable group need not be dropped if it is predictive and internally homogeneous; it can instead be retained and regularized. Corollary C.1 states this exactness result, and Appendix C.9 formalizes the same intuition in a two-level cost model: supports contained in a single cost class have zero heterogeneity gap, while mixed supports incur a heterogeneity gap that grows with the separation between cost levels.

4 Policy Implications for Feature Selection under Optimal Regularization

Our characterization yields practical implications for policymakers seeking to make algorithmic decision-making systems more robust to strategic behavior using these coarser levers.

Feature selection and regularization must be tuned jointly. The support S appears in all three terms of Theorem 2. Changing S changes the predictive loss, the strategic burden, and the cost heterogeneity of the retained features. Therefore, the best feature set cannot generally be determined by ranking features only by predictive value or manipulability. In particular, regularization changes which subset is optimal, so feature selection and ridge tuning must be designed jointly (Figure 1).

Predictive manipulable groups can be retained when their manipulation costs are homogeneous. Contrary to common intuition, our characterization shows that high manipulability alone does not imply that a predictive feature group should be excluded (see Figure 2). When the subgroup is sufficiently predictive and its manipulation costs are sufficiently homogeneous, so that $\delta_H(\mathsf{S})$ is small, scalar ridge can shrink the group. In this regime, the optimal response is often to retain the subgroup and increase regularization, rather than remove it outright. Feature exclusion becomes the preferred lever if the subgroup is insufficiently predictive, or its manipulation geometry is too heterogeneous for a single scalar parameter to regularize.

This suggests a broader policy implication: when certain features cannot be excluded because of legal or institutional constraints, the appropriate response need not be all-or-nothing exclusion. Rather, the decision maker can use interventions such as targeted audits, verification, or tighter reporting requirements to raise or equalize manipulation costs on those features, making those features easier to control through regularization.

Less manipulable correlated proxies can replace manipulable features. Our characterization shows that correlation can fundamentally change which support is optimal. When a less manipulable feature is sufficiently correlated with a manipulable one, the predictive loss $L_{\mathrm{pred}}(\mathsf{S})$ from dropping the latter can be small because much of its signal is recoverable from the proxy. If the resulting retained support also lowers the strategic burden or reduces cost heterogeneity, then the support that keeps the proxy and drops the manipulable feature can dominate the full support. Figure 5 illustrates this support switch. The key insight is that a decision maker can preserve much of the predictive value while reducing strategic vulnerability by using correlated alternative features as proxies.

Taken together, these results clarify the distinct role of the two policy levers. Scalar ridge regularization is most effective on supports whose retained manipulation geometry is close to homogeneous, while feature selection is most effective when it removes strategically heterogeneous or easily manipulated features at limited predictive loss. The best design therefore balances the two levers jointly rather than relying on either in isolation. In the next section, we turn to the corresponding algorithmic problem. Appendix C.10 gives complementary local results showing that infinitesimal ridge regularization improves strategic MSE over support-restricted least squares, and that the locally optimal ridge level increases with manipulation intensity.

5 Computing the Best Subset and Regularization Level

The theory identifies the tradeoffs governing support choice and regularization. We now turn to the corresponding algorithmic problem. For a support budget $s \in \{1, \dots, d\}$, the exact design problem is

$$(\mathsf{S}^\star, \lambda^\star) \in \arg\min_{\substack{\mathsf{S}\subseteq[d]\\ |\mathsf{S}|\le s}} \min_{\lambda\ge 0} \Phi(\mathsf{S}, \lambda), \qquad \Phi(\mathsf{S}, \lambda) := \mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda), 0).$$

For a fixed support S, the optimization over λ is one-dimensional (and smooth); the computational challenge is the combinatorial search over supports.

A weighted relaxation. To avoid exhaustive search, we introduce a continuous weight vector $w \in [0,1]^d$ with support budget $\mathbf{1}^\top w \le s$. For a fixed $\lambda > 0$, define the weighted ridge estimator:

$$\theta_\lambda(w) \in \arg\min_{\theta\in\mathbb{R}^d} \left\{ \mathbb{E}[(Y - \theta^\top X)^2] + \lambda \sum_{i=1}^{d} \varphi(\theta_i, w_i) \right\}, \tag{10}$$

where $\varphi(t,u) = t^2/u$ for $u > 0$, while $\varphi(0,0) = 0$ and $\varphi(t,0) = +\infty$ for $t \neq 0$. Thus $w_i = 0$ forces $\theta_i = 0$. When all $w_i > 0$, writing $W = \mathrm{diag}(w)$ gives $\theta_\lambda(w) = (\Sigma + \lambda W^{-1})^{-1}\Sigma\theta^*$ (see Lemma E.2 for more details). At fixed λ, we minimize the exact strategic objective $\mathrm{MSE}_{\mathrm{strat}}(\theta_\lambda(w), 0)$ over the weight polytope $\mathcal{W}_s := \{w \in [0,1]^d : \mathbf{1}^\top w \le s\}$.

Proposition 1 (Binary weights recover exact support-restricted ridge)**.** *Let $w \in \{0,1\}^d$ and define $\mathsf{S}(w) := \{i \in [d] : w_i = 1\}$. Then $\theta_\lambda(w) = \theta^{\mathrm{OR}}(\mathsf{S}(w), \lambda)$, and hence the weighted objective $\mathrm{MSE}_{\mathrm{strat}}(\theta_\lambda(w), 0)$ exactly matches the fixed-λ support-restricted objective $\Phi(\mathsf{S}(w), \lambda)$ on binary supports.*

The continuous problem is therefore a relaxation of the exact fixed-λ subset selection problem. In particular, if

$$\Phi^\star_{\lambda,s} := \min_{\substack{\mathsf{S}\subseteq[d]\\ |\mathsf{S}|\le s}} \Phi(\mathsf{S}, \lambda), \qquad L^{\mathrm{wt}}_{\lambda,s} := \min_{w\in\mathcal{W}_s} \mathrm{MSE}_{\mathrm{strat}}(\theta_\lambda(w), 0),$$

then $L^{\mathrm{wt}}_{\lambda,s} \le \Phi^\star_{\lambda,s}$ (see Proposition E.1 for a formal treatment and derivation). We use the relaxed solution as a soft initialization, then return to the discrete support-restricted ridge by rounding, refining, and refitting. We provide these details next.

Our procedure. Towards outlining the procedure, let us define some useful notation. Given a weight vector w, let $\mathrm{Top}_s(w)$ denote the indices of the s largest entries of w. For a current support S, let $\mathcal{N}_s(\mathsf{S})$ denote the add/drop/swap neighborhood of S, consisting of supports obtained by adding one feature, dropping one feature, or swapping one retained feature for one excluded feature, while respecting the budget s. Our procedure is then summarized in Algorithm 1.

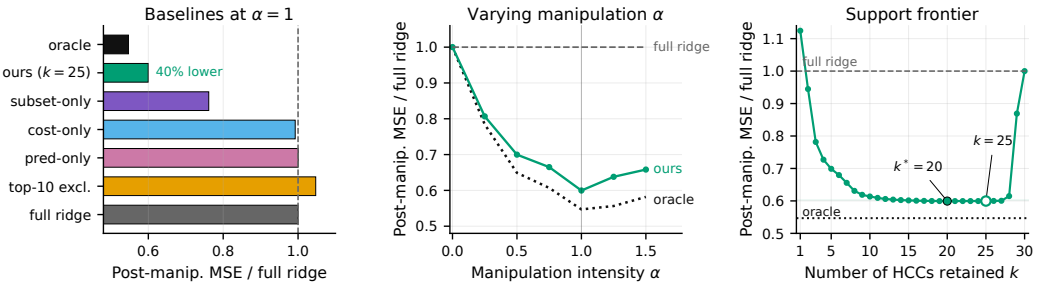

Figure 3: **Feature selection under optimal regularization retains many features while keeping post-manipulation error low.** (Left) Post-manipulation MSE at manipulation intensity $\alpha = 1$ for all methods. (Middle) Post-manipulation MSE under varying α for the oracle $\widetilde{\mathrm{OPT}}$ and the support-restricted ridge design with $k = 25$. (Right) Cardinality frontier of support-restricted ridge at $\alpha = 1$, showing post-manipulation MSE across support sizes k. Post-manipulation MSE is normalized by the full ridge value.

The weighted relaxation plays a purely algorithmic role: it provides a soft initialization for support search at fixed λ. The final estimator is always a standard support-restricted ridge fit, so the procedure remains faithful to the estimator class studied in Section 3. Each accepted local refinement move strictly decreases the exact fixed-λ objective $\Phi(\mathsf{S}, \lambda)$, so the refinement step terminates after finitely many moves and never worsens the rounded support (see Proposition E.2). We provide implementation details and a convex reformulation for the diagonal-covariance case in Appendix E and synthetic benchmarks in Appendix F.

6 Case Study: Strategic Feature Selection in Healthcare Payments

We study healthcare payments in the Medicare Advantage (MA) program, as introduced earlier in the paper. In MA, the government determines payments to health insurers for the patients they cover using a least squares regression based on the diagnoses reported for their enrollees. Here, diagnosis codes are the model features, which creates incentives to increase payment by assigning additional or more severe diagnosis codes to a patient, a practice known as upcoding. The payment model includes 115 variables corresponding to health conditions, termed as hierarchical condition categories (HCCs). We base our study on recent policy work that recommends excluding the *top ten diagnosis groups* that contributed the most to higher coding intensity in Medicare Advantage [20]. We show that this proposal is suboptimal in strategic MSE, and that our method can preserve both predictive and strategic performance while retaining some of these diagnoses. Code to reproduce our experiments is available at `https://github.com/jivatneet/strategic-feature-selection`.

Experimental setup. We use the `upcoding` package, which simulates realistic co-occurring HCCs in the Medicare population using self-reported health conditions from a large national survey that are unaffected by coding incentives [11]. The baseline dataset contains 5,000 beneficiaries and 115 V28 HCC indicators, of which 30 have nonzero prevalence. We upcode selected HCCs among the top-ten diagnosis groups identified by Kronick et al. [20] to match their reported prevalence shifts. We use the published CMS-HCC V28 coefficients as θ^* and generate outcomes from the baseline covariates. All methods are fit using only baseline data, following the MA setting, and evaluation reports post-manipulation MSE under the strategic response model. We compare our method against cost-only (feature selection based on manipulability alone), prediction-only (feature selection based on predictive value alone), top-ten exclusion (removing HCCs in the top-ten diagnosis groups), subset-only (support-restricted least squares), and full ridge (ordinary ridge on full support). All baselines except subset-only tune ridge regularization. We also include the full-support zero-intercept oracle $\widetilde{\mathrm{OPT}}$ as a benchmark. We report post-manipulation MSE normalized by full ridge at the same manipulation intensity. Further details on data and experimental setup are deferred to Appendix G.

Proposed method substantially reduces post-manipulation MSE relative to baselines and remains competitive with the oracle. We compare the proposed support-restricted ridge estimator to the baselines and the oracle $\widetilde{\mathrm{OPT}}$. The left panel of Figure 3 shows that the support-restricted ridge design with $k = 25$ retained features reduces post-manipulation MSE by about 40% relative to full

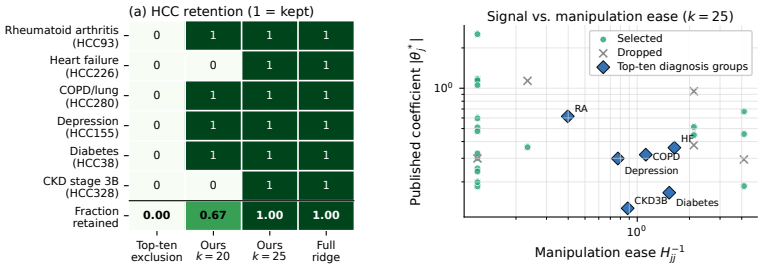

Figure 4: **The selected model reflects the joint predictability-manipulability structure, rather than dropping all intensely coded features.** (Left) HCC groups with nonzero prevalence in the data among the top-ten diagnosis groups identified by Kronick et al. [20]. The top-ten exclusion policy drops all of these groups, whereas the unconstrained optimum at $k = 20$ retains some of them and the broader $k = 25$ design retains all six such groups. (Right) Each HCC with nonzero prevalence in the data is plotted by its published CMS-HCC coefficient magnitude and manipulation-ease score H_{jj}^{-1}; marker type indicates whether the broader $k = 25$ support-restricted ridge design retains that HCC.

ridge and outperforms the top-ten exclusion policy, prediction-only selection, cost-only selection, and subset-only selection without ridge shrinkage. This illustrates the importance of tuning feature selection and regularization jointly. Further, the middle panel of Figure 3 shows that the support-restricted ridge estimator remains competitive with the oracle across a range of manipulation intensities.

Low strategic error does not require aggressive exclusion. The right panel of Figure 3 showcases the support frontier of the proposed estimator, illustrating that low post-manipulation prediction error can be achieved while retaining a larger feature set. Specifically, while the unconstrained optimum is attained at $k^* = 20$, expanding the model to include $k = 25$ HCCs yields nearly identical performance. For this reason, we use $k = 25$ as a representative operating point in the main results. As shown in the left panel of Figure 4, the support-restricted ridge estimator with $k = 25$ retains all the HCCs (with nonzero baseline prevalence) that belong to the top-ten diagnosis groups while outperforming both the full ridge estimator and the top-ten exclusion policy. The support frontier confirms that the best design is often an interior solution rather than either the full feature set or aggressive exclusion. Importantly, these gains are not driven by feature selection alone: the strongest performance comes from combining support restriction with regularization.

Selected support is not determined by coding vulnerability alone. The selected support is not obtained by simply ranking diagnoses by coding vulnerability: the $k = 25$ support-restricted ridge design retains several HCC groups identified by Kronick et al. [20], while excluding other HCCs (Figure 4 (right)). The plot further illustrates there is no straightforward pattern based on a feature's direct predictiveness and its manipulation ease that determines inclusion or exclusion. This matches our theory, which predicts that feature selection should balance the predictive loss from excluding features against the strategic burden and cost geometry of the retained set. Figure 4 also contrasts our design with the top-ten exclusion policy, which removes all such groups by construction, whereas support-restricted ridge can retain some of them and control strategic exposure through regularization.

7 Conclusion

We develop a framework for studying strategic classification through the coarse policy levers of feature selection and regularization. Our main message is that features should be selected jointly based on their relative manipulability and predictability, rather than on manipulability of individual features alone. Through theoretical analysis and a real-world case study, we show how these levers interact to yield nuanced implications for policy design that broaden the prevailing narrative. More broadly, given the ubiquity of prediction systems in practice, we envision a growing need for principled and practical frameworks that bridge theoretical foundations with the policy and institutional contexts in which these systems are deployed. Our analysis assumes a one-shot interaction and known, quadratic manipulation costs that are homogeneous across organizations. Important directions for future work include studying long-term dynamics, richer and heterogeneous cost structures, and robust policy design under cost misspecification. We defer a broader discussion of extensions and limitations to Appendices H and I.

Acknowledgments and Disclosure of Funding

We would like to thank Oana M. Enache, Keegan Harris, Richard Kronick, Lydia T. Liu, Juan C. Perdomo, Sherri Rose, Manolis Zampetakis, Anna Zink, and Tijana Zrnic for insightful conversations on our work. Ida Sim's guest lecture on healthcare payment models introduced the authors to this problem, and we are grateful for several helpful discussions. Computing support is partially provided by the Texas Advanced Computing Center (TACC). This work was supported in part by a Stanford AI Lab postdoctoral fellowship and by the European Union (ERC-2022-SYG-OCEAN-101071601). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] All of Us Research Program Investigators. The "All of Us" research program. *New England Journal of Medicine*, 381(7):668–676, 2019.

[2] Camilo Arias. Selecting risk adjusters with penalized regression and expert judgment: evidence from Colombia. *Health Economics Review*, 16(1):6, 2025.

[3] Meghan Bellerose, Hannah O. James, Jay Shroff, Andrew M. Ryan, and David J. Meyers. Combining patient survey data with diagnosis codes improved Medicare Advantage risk-adjustment accuracy. *Health Affairs*, 44(1):58–65, 2025.

[4] Daniel Björkegren, Joshua E. Blumenstock, and Samsun Knight. Manipulation-robust prediction. *American Economic Review*, 2025. Forthcoming.

[5] Adriana Camacho and Emily Conover. Manipulation of social program eligibility. *American Economic Journal: Economic Policy*, 3(2):41–65, 2011.

[6] Centers for Medicare & Medicaid Services. Advance notice of methodological changes for calendar year (CY) 2014 for medicare advantage (MA) capitation rates and part C and part D payment policies. Technical report, Centers for Medicare & Medicaid Services, 2013. URL `https://www.cms.gov/medicare/health-plans/medicareadvtgspecratestats/downloads/advance2014.pdf`.

[7] Centers for Medicare & Medicaid Services. Advance notice of methodological changes for calendar year (CY) 2024 for medicare advantage (MA) capitation rates and part C and part D payment policies. Technical report, Centers for Medicare & Medicaid Services, 2023. URL `https://www.cms.gov/files/document/2024-advance-notice-pdf.pdf`.

[8] Centers for Medicare & Medicaid Services. Report to Congress: Risk adjustment in Medicare Advantage. Technical report, Centers for Medicare & Medicaid Services, 2024. URL `https://www.cms.gov/files/document/report-congress-risk-adjustment-medicare-advantage-december-2024.pdf`.

[9] Yiling Chen, Yang Liu, and Chara Podimata. Learning strategy-aware linear classifiers. In *Advances in Neural Information Processing Systems*, 2020.

[10] Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the ACM Conference on Economics and Computation*, 2018.

[11] Oana M. Enache and Sherri Rose. Time-to-event estimation with unreliably reported events in medicare health plan payment. arXiv: 2602.04092, 2026.

[12] Michael Geruso and Timothy Layton. Upcoding: Evidence from Medicare on squishy risk adjustment. *Journal of Political Economy*, 128(3):984–1026, 2020.

[13] António Góis, Sophia Günlük, Nir Rosenfeld, Nidhi Hegde, Simon Lacoste-Julien, and Dhanya Sridhar. The role of causal features in strategic classification for robustness and alignment. In *International Conference on Artificial Intelligence and Statistics*, 2026.

[14] Margaret E. Grosh and Judy L. Baker. Proxy means tests for targeting social programs: Simulations and speculation. LSMS Working Paper 118, The World Bank, 1995. URL https://digitallibrary.un.org/record/194918.

[15] Laszlo Gyorfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

[16] Nika Haghtalab, Nicole Immorlica, Brendan Lucier, and Jack Z. Wang. Maximizing welfare with incentive-aware evaluation mechanisms. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2020.

[17] Tinashe Handina and Eric Mazumdar. Understanding model selection for learning in strategic environments. In *Advances in Neural Information Processing Systems*, 2024.

[18] Moritz Hardt, Nimrod Megiddo, Christos H. Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the ACM Conference on Innovations in Theoretical Computer Science*, 2016.

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

[20] Richard Kronick, F. Michael Chua, Ramona Krauss, Logan Johnson, and Daniel Waldo. Are fewer diagnoses better? Assessing a proposal to improve the Medicare Advantage payment system. *Health Affairs*, 44(1):66–74, 2025.

[21] Richard Kronick, F. Michael Chua, Ramona Krauss, Logan Johnson, and Daniel Waldo. Insurer-level estimates of revenue from differential coding in Medicare Advantage. *Annals of Internal Medicine*, 178(5):655–662, 2025.

[22] Sagi Levanon and Nir Rosenfeld. Strategic classification made practical. In *Proceedings of the International Conference on Machine Learning*, 2021.

[23] César Martinelli and Susan Wendy Parker. Deception and misreporting in a social program. *Journal of the European Economic Association*, 7(4):886–908, 2009.

[24] Thomas G. McGuire, Anna L. Zink, and Sherri Rose. Improving the performance of risk adjustment systems: Constrained regressions, reinsurance, and variable selection. *American Journal of Health Economics*, 7(4):497–521, 2021.

[25] J. Michael McWilliams, Gabe Weinreb, Mary Beth Landrum, and Michael E. Chernew. Use of patient health survey data for risk adjustment to limit distortionary coding incentives in Medicare. *Health Affairs*, 44(1):48–57, 2025.

[26] Medicare Payment Advisory Commission. March 2025 report to the Congress: Medicare payment policy. Technical report, Medicare Payment Advisory Commission (MedPAC), 2025. URL https://www.medpac.gov/document/march-2025-report-to-the-congress-medicare-payment-policy/.

[27] David J. Meyers and Amal N. Trivedi. Medicare Advantage chart reviews are associated with billions in additional payments for some plans. *Medical Care*, 59(2):96–100, 2021.

[28] John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *Proceedings of the International Conference on Machine Learning*, 2020.

[29] Andrew Y. Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance. In *Proceedings of the International Conference on Machine Learning*, 2004.

[30] Paul Niehaus, Antonia Atanassova, Marianne Bertrand, and Sendhil Mullainathan. Targeting with agents. *American Economic Journal: Economic Policy*, 5(1):206–38, 2013.

[31] Pratik Patil, Yuting Wei, Alessandro Rinaldo, and Ryan J. Tibshirani. Uniform consistency of cross-validation for high-dimensional ridge regression. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2021.

[32] Pratik Patil, Alessandro Rinaldo, and Ryan J. Tibshirani. Estimating functionals of the out-of-sample error distribution in high-dimensional ridge regression. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2022.

[33] Pratik Patil, Jin-Hong Du, and Ryan J. Tibshirani. Optimal ridge regularization for out-of-distribution prediction. In *Proceedings of the International Conference on Machine Learning*, 2024.

[34] Chara Podimata. Incentive-aware machine learning: Robustness, fairness, improvement & causality. arXiv: 2505.05211, 2025.

[35] Sherri Rose. A machine learning framework for plan payment risk adjustment. *Health Services Research*, 51(6):2358–2374, 2016.

[36] Yonadav Shavit, Benjamin L. Edelman, and Brian Axelrod. Causal strategic linear regression. In *Proceedings of the International Conference on Machine Learning*, 2020.

[37] Yong Shi, Jianyu Miao, Zhengyu Wang, Peng Zhang, and Lingfeng Niu. Feature selection with $\ell_{2,1-2}$ regularization. *IEEE Transactions on Neural Networks and Learning Systems*, 29(10): 4967–4982, 2018.

[38] U.S. Department of Justice. Kaiser Permanente affiliates pay $556M to resolve False Claims Act allegations. Office of Public Affairs, 2026. URL `https://www.justice.gov/opa/pr/kaiser-permanente-affiliates-pay-556m-resolve-false-claims-act-allegations`.

[39] Huang Xiao, Battista Biggio, Gavin Brown, Giorgio Fumera, Claudia Eckert, and Fabio Roli. Is feature selection secure against training data poisoning? In *Proceedings of the International Conference on Machine Learning*, 2015.

[40] Fei Zhang, Patrick P. K. Chan, Battista Biggio, Daniel S. Yeung, and Fabio Roli. Adversarial feature selection against evasion attacks. *IEEE Transactions on Cybernetics*, 46(3):766–777, 2016.

[41] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

A Additional Related Work

Strategic classification. Closely related in spirit to our work, Handina and Mazumdar [17] show that in strategic environments, larger or more expressive model classes need not improve or may even worsen equilibrium performance. Our contribution is conceptually and mathematically distinct: rather than comparing abstract model classes, we study the concrete and widely used design levers of feature selection and regularization, and characterize the performance of feature selection under optimal regularization.

Góis et al. [13] study strategic classification in a setting with causal features that are invariant predictors of the outcome and spurious features that do not cause the outcome of interest. They present a decomposition for the cross-entropy loss post-adaptation into incomplete information error, transfer error, and entropy of post-adapt distribution. Classifiers that only use causal features can eliminate transfer error but can incur incomplete information error by discarding the predictive information in spurious features. In contrast, classifiers that use all features can avoid some incomplete information error but incur transfer error when strategic adaptation changes the relationship of spurious features with the outcome. This tradeoff is related in spirit to the excess strategic MSE decomposition in our paper. The settings and results are nevertheless distinct. We do not assume the true outcome changes when reported features change, and we also do not assume that features are partitioned into causal or non-causal. Our decomposition of the optimality gap separates into the predictive loss from omitting features, the change in strategic burden, and the heterogeneity gap from using scalar ridge on the retained support.

Feature selection. Ng [29] and Shi et al. [37] study feature selection induced directly through the regularization penalty. Ng [29] shows that L_1-regularized logistic regression can perform effectively in the presence of many irrelevant features, specifically even if there are exponentially many irrelevant features as there are training examples. Shi et al. [37] propose a new nonconvex regularizer that induces feature sparsity in supervised and unsupervised learning. Zou and Hastie [41] propose the elastic net that performs feature selection and shrinkage, while encouraging correlated features to be selected together. In contrast, we use an L_2 ridge penalty, which does not itself induce sparsity, and treat feature selection and regularization as distinct policy levers. We therefore jointly select the support S and regularization level λ according to the strategic MSE objective.

Adversarial robustness. Our work is also related to adversarially robust feature selection and poisoning-robust feature selection [40, 39] which show that adversary-aware feature selection can improve robustness to evasion and poisoning attacks. Our setting differs, however, in the threat and perturbation models: these papers study exogenous attacks on the data, whereas we study endogenous manipulation induced by the deployed predictor itself.

B Theoretical Details and Proofs for Section 2

B.1 Strategic optimum characterization

Lemma B.1 (Strategic optimum characterization)**.** *The strategic optimum* (4) *is attained at* $(\theta^{\mathrm{OPT}}, b^{\mathrm{OPT}}) = (\theta^*, -\theta^{*\top} H^{-1}\theta^*)$ *and its value is* $\mathrm{OPT}_{\theta^*}(\mathcal{P}_{X,Y}, H) = \sigma^2$.

Proof. By Lemma B.4, we can decompose the strategic MSE (3) into:

$$\mathrm{MSE}_{\mathrm{strat}}(\theta, b) = (\theta - \theta^*)^\top \Sigma (\theta - \theta^*) + \left(b + \theta^\top H^{-1}\theta\right)^2 + \sigma^2.$$

For any fixed θ, the optimal intercept is $b(\theta) = -\theta^\top H^{-1}\theta$, which eliminates the strategic shift term. Therefore

$$\min_{b\in\mathbb{R}} \mathrm{MSE}_{\mathrm{strat}}(\theta, b) = (\theta - \theta^*)^\top \Sigma (\theta - \theta^*) + \sigma^2.$$

Since $\Sigma \succ 0$, this expression is uniquely minimized at $\theta = \theta^*$. Hence the strategic optimum (4) is attained at

$$\theta^{\mathrm{OPT}} = \theta^*, \qquad b^{\mathrm{OPT}} = -\theta^{*\top} H^{-1}\theta^*.$$

The optimal value is $\mathrm{OPT}_{\theta^*}(\mathcal{P}_{X,Y}, H) = \sigma^2$. □

B.2 Centering and projection identity

We work under the centering convention $\mathbb{E}[X] = 0$ and $\mathbb{E}[Y] = 0$. This convention is without loss of generality for the linear models studied in the paper: after replacing X and Y by $X - \mathbb{E}[X]$ and $Y - \mathbb{E}[Y]$, any shift in means can be absorbed into the intercept of the deployed linear model. Moreover, since the organization's action is additive in the features, replacing X by $X - \mathbb{E}[X]$ does not change the action a or the quadratic manipulation cost.

We next record a projection identity implied by the projection parameter θ^* as defined in (1). This is a standard result (see, e.g., [15]), whose derivation we collect below for completeness.

Lemma B.2 (Projection decomposition)**.** *Suppose $\mathbb{E}[X] = 0$, $\mathbb{E}[Y] = 0$, and $\Sigma := \mathbb{E}[XX^\top] \succ 0$. Let θ^* be the coefficients of the best linear projection of Y onto X as in* (1)*. Defining $\varepsilon := Y - \theta^{*\top} X$, we have $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[X\varepsilon] = 0$. Moreover, for every $\theta \in \mathbb{R}^d$, we can decompose $\mathbb{E}[(Y - \theta^\top X)^2] = (\theta - \theta^*)^\top \Sigma(\theta - \theta^*) + \sigma^2$, where $\sigma^2 := \mathbb{E}[\varepsilon^2]$.*

Proof. Since $\mathbb{E}[X] = 0$ and $\mathbb{E}[Y] = 0$, $\mathbb{E}[\varepsilon] = \mathbb{E}[Y] - \theta^{*\top}\mathbb{E}[X] = 0$. Moreover, since $\Sigma\theta^* = \mathbb{E}[XY]$ because $\theta^* = \Sigma^{-1}\mathbb{E}[XY]$ (or from the first-order condition for the objective in (1)), we have

$$\mathbb{E}[X\varepsilon] = \mathbb{E}[XY] - \mathbb{E}[XX^\top]\theta^* = \mathbb{E}[XY] - \Sigma\theta^* = 0.$$

Now, for any $\theta \in \mathbb{R}^d$, $Y - \theta^\top X = (\theta^* - \theta)^\top X + \varepsilon$. Therefore

$$\begin{aligned}\mathbb{E}[(Y - \theta^\top X)^2] &= \mathbb{E}[((\theta^* - \theta)^\top X)^2] + 2\mathbb{E}[((\theta^* - \theta)^\top X)\varepsilon] + \mathbb{E}[\varepsilon^2] \\ &= (\theta - \theta^*)^\top \Sigma(\theta - \theta^*) + \sigma^2,\end{aligned}$$

where the cross term vanishes because $\mathbb{E}[X\varepsilon] = 0$. □

B.3 Best response and strategic MSE

Below we collect basic identities used in the proofs of results in Section 2 and throughout the paper. All results in this section are standard and are only reproduced for completeness.

Lemma B.3 (Organization's best response)**.** *For any linear predictor $f_{\theta,b}(x) = \theta^\top x + b$, the organization's best response in* (2) *is unique and equals $a^*(x;\theta) = H^{-1}\theta$. In particular, under the unconstrained quadratic-cost model, the optimal action does not depend on the realized feature vector x.*

Proof. For fixed x, θ, and b, the organization's objective is:

$$\theta^\top(x + a) + b - \tfrac{1}{2}a^\top Ha.$$

The terms $\theta^\top x + b$ do not depend on a, so the optimization problem is equivalent to:

$$\max_{a \in \mathbb{R}^d} \left\{\theta^\top a - \tfrac{1}{2}a^\top Ha\right\}.$$

Since $H \succ 0$, this objective is strictly concave in a. Its first-order condition is $\theta - Ha = 0$. Thus, the unique maximizer is $a^*(x;\theta) = H^{-1}\theta$. Observe that this optimal manipulation depends on the announced coefficient vector θ but not on the realized feature vector x. □

Under this best response per Lemma B.3, the deployed prediction is

$$\theta^\top(X + a^*(X;\theta)) + b = \theta^\top X + \theta^\top H^{-1}\theta + b.$$

Thus strategic manipulation induces an endogenous deterministic shift $\theta^\top H^{-1}\theta$ in the prediction.

Lemma B.4 (Strategic MSE identity)**.** *For any $\theta \in \mathbb{R}^d$ and $b \in \mathbb{R}$,*

$$\mathrm{MSE}_{\mathrm{strat}}(\theta, b) = (\theta - \theta^*)^\top \Sigma(\theta - \theta^*) \;+\; (b + \theta^\top H^{-1}\theta)^2 \;+\; \sigma^2. \tag{11}$$

Proof. By Lemma B.3, $a^*(X;\theta) = H^{-1}\theta$. Therefore

$$\theta^\top(X + a^*(X;\theta)) + b - Y = \theta^\top X + \theta^\top H^{-1}\theta + b - Y.$$

Using $Y = \theta^{*\top} X + \varepsilon$, this becomes

$$(\theta - \theta^*)^\top X + \theta^\top H^{-1}\theta + b - \varepsilon.$$

Let $c(\theta, b) := \theta^\top H^{-1}\theta + b$. Then

$$\theta^\top (X + a^*(X;\theta)) + b - Y = (\theta - \theta^*)^\top X + c(\theta, b) - \varepsilon.$$

Squaring and taking expectations lets us decompose the strategic MSE (3) as:

$$\begin{aligned}
\mathrm{MSE}_{\mathrm{strat}}(\theta, b) &= \mathbb{E}\big[((\theta - \theta^*)^\top X + c(\theta, b) - \varepsilon)^2\big] \\
&= \mathbb{E}[((\theta - \theta^*)^\top X)^2] + c(\theta, b)^2 + \mathbb{E}[\varepsilon^2] \\
&\quad - 2\mathbb{E}[((\theta - \theta^*)^\top X)\varepsilon] + 2c(\theta, b)\mathbb{E}[(\theta - \theta^*)^\top X] - 2c(\theta, b)\mathbb{E}[\varepsilon].
\end{aligned}$$

The cross terms vanish because $\mathbb{E}[X\varepsilon] = 0$, $\mathbb{E}[X] = 0$, and $\mathbb{E}[\varepsilon] = 0$. Hence

$$\mathrm{MSE}_{\mathrm{strat}}(\theta, b) = (\theta - \theta^*)^\top \Sigma (\theta - \theta^*) + c(\theta, b)^2 + \sigma^2,$$

as claimed. □

We can interpret the three terms in the strategic MSE decomposition (11) as follows: The first term is the loss from using coefficients θ instead of the best linear projection coefficients θ^*. The second term is the deterministic shift induced jointly by the intercept and the organization's equilibrium manipulation. The final term σ^2 is the irreducible residual variance.

B.4 Closed forms for the policy levers

We record here closed-form expressions for the estimators ((5), (6), and (7)) introduced in Section 2. The results here are again standard and are only reported for completeness, in our notation.

Recall from Section 2 that for a subset $\mathsf{S} \subseteq [d]$, we let $\mathsf{R} = [d] \setminus \mathsf{S}$. We write $\Sigma_{\mathsf{S}y} := \mathbb{E}[X_{\mathsf{S}} Y]$. Note that since $Y = \theta^{*\top} X + \varepsilon$ and $\mathbb{E}[X\varepsilon] = 0$, we have

$$\Sigma_{\mathsf{S}y} = \Sigma_{\mathsf{SS}}\theta^*_{\mathsf{S}} + \Sigma_{\mathsf{SR}}\theta^*_{\mathsf{R}}. \tag{12}$$

Lemma B.5 (Closed forms for support-restricted least squares and support-restricted ridge)**.** *Under the centering convention $\mathbb{E}[X] = 0$, $\mathbb{E}[Y] = 0$, all three estimators* (5)*,* (6)*, and* (7)*, introduced in Section 2 have intercept* 0*. Moreover:*

1. *For any $\mathsf{S} \subseteq [d]$, the support-restricted least squares estimator is:*

$$\theta^{\mathrm{LS}}(\mathsf{S}) = (\theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S}), 0_{\mathsf{R}}),$$

   where

$$\theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S}) = \Sigma^{-1}_{\mathsf{SS}}\Sigma_{\mathsf{S}y} = \theta^*_{\mathsf{S}} + \Sigma^{-1}_{\mathsf{SS}}\Sigma_{\mathsf{SR}}\theta^*_{\mathsf{R}}.$$

2. *For any $\lambda \geq 0$, the ordinary ridge estimator is:*

$$\theta^{\mathrm{OR}}(\lambda) = (\Sigma + \lambda I_d)^{-1}\Sigma\theta^*.$$

3. *For any $\mathsf{S} \subseteq [d]$ and $\lambda \geq 0$, the support-restricted ridge estimator is:*

$$\theta^{\mathrm{OR}}(\mathsf{S}, \lambda) = (\theta^{\mathrm{OR}}_{\mathsf{S}}(\mathsf{S}, \lambda), 0_{\mathsf{R}}),$$

   where

$$\theta^{\mathrm{OR}}_{\mathsf{S}}(\mathsf{S}, \lambda) = (\Sigma_{\mathsf{SS}} + \lambda I_{|\mathsf{S}|})^{-1}\Sigma_{\mathsf{S}y}.$$

   Equivalently,

$$\theta^{\mathrm{OR}}_{\mathsf{S}}(\mathsf{S}, \lambda) = (\Sigma_{\mathsf{SS}} + \lambda I_{|\mathsf{S}|})^{-1} (\Sigma_{\mathsf{SS}}\theta^*_{\mathsf{S}} + \Sigma_{\mathsf{SR}}\theta^*_{\mathsf{R}}).$$

Proof. We first verify that the optimal intercept is zero in all three programs. For any fixed θ, the unregularized squared-error part satisfies

$$\mathbb{E}[(Y - \theta^\top X - b)^2] = \mathbb{E}[(Y - \theta^\top X)^2] + b^2.$$

Thus, the optimal unique intercept is zero. The ridge penalty does not involve b, so the same conclusion holds for the ridge programs. We next write out the coefficient expressions.

Support-restricted least squares. For $\theta \in \Theta(\mathsf{S})$, write $\theta = (\eta, 0_{\mathsf{R}})$ with $\eta \in \mathbb{R}^{|\mathsf{S}|}$. The support-restricted least squares solution reduces to

$$\min_{\eta \in \mathbb{R}^{|\mathsf{S}|}} \mathbb{E}[(Y - \eta^\top X_{\mathsf{S}})^2].$$

The objective is strictly convex because $\Sigma_{\mathsf{SS}} \succ 0$. Hence, the solution is unique given by $\theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S}) = \Sigma^{-1}_{\mathsf{SS}}\Sigma_{\mathsf{S}y}$, and using (12) gives

$$\theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S}) = \Sigma^{-1}_{\mathsf{SS}}\left(\Sigma_{\mathsf{SS}}\theta^*_{\mathsf{S}} + \Sigma_{\mathsf{SR}}\theta^*_{\mathsf{R}}\right) = \theta^*_{\mathsf{S}} + \Sigma^{-1}_{\mathsf{SS}}\Sigma_{\mathsf{SR}}\theta^*_{\mathsf{R}}.$$

Ridge. By Lemma B.2, we can decompose

$$\mathbb{E}[(Y - \theta^\top X)^2] = (\theta - \theta^*)^\top \Sigma(\theta - \theta^*) + \sigma^2.$$

Thus ordinary ridge minimizes

$$\min_{\theta \in \mathbb{R}^d} (\theta - \theta^*)^\top \Sigma(\theta - \theta^*) + \lambda\|\theta\|_2^2.$$

Since $\Sigma + \lambda I_d \succ 0$, the solution is unique and equals

$$\theta^{\mathrm{OR}}(\lambda) = (\Sigma + \lambda I_d)^{-1}\Sigma\theta^*.$$

Support-restricted ridge. Again write $\theta = (\eta, 0_{\mathsf{R}})$. The support-restricted ridge problem reduces to

$$\min_{\eta \in \mathbb{R}^{|\mathsf{S}|}} \left\{\mathbb{E}[(Y - \eta^\top X_{\mathsf{S}})^2] + \lambda\|\eta\|_2^2\right\}.$$

Since $\Sigma_{\mathsf{SS}} + \lambda I_{|\mathsf{S}|} \succ 0$, the solution is unique again and after substituting (12) equals

$$\theta^{\mathrm{OR}}_{\mathsf{S}}(\mathsf{S}, \lambda) = (\Sigma_{\mathsf{SS}} + \lambda I_{|\mathsf{S}|})^{-1}\Sigma_{\mathsf{S}y} = (\Sigma_{\mathsf{SS}} + \lambda I_{|\mathsf{S}|})^{-1}(\Sigma_{\mathsf{SS}}\theta^*_{\mathsf{S}} + \Sigma_{\mathsf{SR}}\theta^*_{\mathsf{R}}),$$

as desired. □

The formula $\theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S}) = \theta^*_{\mathsf{S}} + \Sigma^{-1}_{\mathsf{SS}}\Sigma_{\mathsf{SR}}\theta^*_{\mathsf{R}}$ shows that omitted features can still affect the fitted coefficients on retained features through the covariance structure of X. Thus, the value of dropping a feature depends not only on its own manipulability, but also on how its signal loads onto the features that remain. This observation becomes important in the support restriction decomposition in Section 3.

Corollary B.1 (Nesting of the policy levers)**.** *The support-restricted ordinary ridge estimator satisfies* $\theta^{\mathrm{OR}}(\mathsf{S}, 0) = \theta^{\mathrm{LS}}(\mathsf{S})$ *and, for* $\mathsf{S} = [d]$, $\theta^{\mathrm{OR}}([d], \lambda) = \theta^{\mathrm{OR}}(\lambda)$.

Proof. Both identities follow immediately from the closed forms in Lemma B.5. □

C Theoretical Details and Proofs for Section 3

Throughout this appendix, we use the same convention as in the main text and write $\mathrm{MSE}_{\mathrm{strat}}(\theta)$ as shorthand for $\mathrm{MSE}_{\mathrm{strat}}(\theta, 0)$, as defined in (3). Thus, by Lemma B.4,

$$\mathrm{MSE}_{\mathrm{strat}}(\theta) = (\theta - \theta^*)^\top \Sigma(\theta - \theta^*) + (\theta^\top H^{-1}\theta)^2 + \sigma^2.$$

For a support $\mathsf{S} \subseteq [d]$,

$$\Theta(\mathsf{S}) := \{\theta \in \mathbb{R}^d : \mathrm{supp}(\theta) \subseteq \mathsf{S}\},$$

and

$$\widetilde{\mathrm{OPT}}(\mathsf{S}) := \min_{\theta \in \Theta(\mathsf{S})} \mathrm{MSE}_{\mathrm{strat}}(\theta), \qquad \widetilde{\mathrm{OPT}} := \min_{\theta \in \mathbb{R}^d} \mathrm{MSE}_{\mathrm{strat}}(\theta).$$

C.1 Proof of Theorem 1

Fix any $\mathsf{S} \subseteq [d]$ and $\lambda \geq 0$. By construction, $\theta^{\mathrm{OR}}(\mathsf{S}, \lambda) \in \Theta(\mathsf{S})$. Therefore,

$$\mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda)) \geq \widetilde{\mathrm{OPT}}(\mathsf{S}).$$

Moreover, since $\Theta(\mathsf{S}) \subseteq \mathbb{R}^d$,

$$\widetilde{\mathrm{OPT}}(\mathsf{S}) \geq \widetilde{\mathrm{OPT}}.$$

Combining the two inequalities gives

$$\mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda)) \geq \widetilde{\mathrm{OPT}}.$$

Subtracting OPT from both sides and minimizing over (S, λ) yields

$$\widetilde{\mathrm{OPT}} - \mathrm{OPT} \leq \min_{\mathsf{S} \subseteq [d],\, \lambda \geq 0} \left\{ \mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda)) - \mathrm{OPT} \right\}.$$

This proves the claimed lower bound.

C.2 Proof of Proposition C.1

Proposition C.1 (Irreducible error bounds)**.** *We have* $0 \leq \widetilde{\mathrm{OPT}} - \mathrm{OPT} \leq (\theta^{*\top} H^{-1} \theta^*)^2$.

Proof. Since $\widetilde{\mathrm{OPT}}$ minimizes over zero-intercept linear rules, while OPT minimizes over both coefficients and intercepts, we have $\mathrm{OPT} \leq \widetilde{\mathrm{OPT}}$. Hence $0 \leq \widetilde{\mathrm{OPT}} - \mathrm{OPT}$. For the upper bound, evaluate the zero-intercept benchmark at $\theta = \theta^*$: $\widetilde{\mathrm{OPT}} \leq \mathrm{MSE}_{\mathrm{strat}}(\theta^*)$. By the strategic MSE identity,

$$\mathrm{MSE}_{\mathrm{strat}}(\theta^*) = (\theta^* - \theta^*)^\top \Sigma (\theta^* - \theta^*) + (\theta^{*\top} H^{-1} \theta^*)^2 + \sigma^2 = \sigma^2 + (\theta^{*\top} H^{-1} \theta^*)^2.$$

By Lemma B.1, $\mathrm{OPT} = \sigma^2$. Therefore $\widetilde{\mathrm{OPT}} - \mathrm{OPT} \leq (\theta^{*\top} H^{-1} \theta^*)^2$. □

C.3 Proof of Proposition C.2

Proposition C.2 (Support restriction identity)**.** *Fix a subset of features* $\mathsf{S} \subseteq [d]$ *and let* $\mathsf{R} := [d] \setminus \mathsf{S}$. *Define the predictive loss from omitting the coordinates in* R

$$L_{\mathrm{pred}}(\mathsf{S}) := (\theta^*_{\mathsf{R}})^\top \Sigma_{\mathsf{R}|\mathsf{S}} \theta^*_{\mathsf{R}}, \qquad \Sigma_{\mathsf{R}|\mathsf{S}} := \Sigma_{\mathsf{RR}} - \Sigma_{\mathsf{RS}} \Sigma_{\mathsf{SS}}^{-1} \Sigma_{\mathsf{SR}},$$

and the support-restricted strategic burden

$$\Gamma(\mathsf{S}) := \min_{\eta \in \mathbb{R}^{|\mathsf{S}|}} \left\{ (\eta - \theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S}))^\top \Sigma_{\mathsf{SS}} (\eta - \theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S})) + (\eta^\top H^{-1}_{\mathsf{SS}} \eta)^2 \right\}. \tag{13}$$

Then

$$\widetilde{\mathrm{OPT}}(\mathsf{S}) - \widetilde{\mathrm{OPT}} = L_{\mathrm{pred}}(\mathsf{S}) \; - \; \big(\Gamma([d]) - \Gamma(\mathsf{S})\big). \tag{14}$$

We first record a useful support-restricted decomposition.

Lemma C.1 (Support-restricted Pythagorean identity)**.** *Fix* $\mathsf{S} \subseteq [d]$ *and let* $\mathsf{R} = [d] \setminus \mathsf{S}$. *For any* $\theta \in \Theta(\mathsf{S})$, *write* $\theta = (\eta, 0_{\mathsf{R}})$, *where* $\eta \in \mathbb{R}^{|\mathsf{S}|}$. *Let*

$$\theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S}) = \Sigma^{-1}_{\mathsf{SS}} \Sigma_{\mathsf{S}y} = \theta^*_{\mathsf{S}} + \Sigma^{-1}_{\mathsf{SS}} \Sigma_{\mathsf{SR}} \theta^*_{\mathsf{R}}.$$

Then

$$(\theta - \theta^*)^\top \Sigma (\theta - \theta^*) = L_{\mathrm{pred}}(\mathsf{S}) + (\eta - \theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S}))^\top \Sigma_{\mathsf{SS}} (\eta - \theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S})),$$

where

$$L_{\mathrm{pred}}(\mathsf{S}) = (\theta^*_{\mathsf{R}})^\top \Sigma_{\mathsf{R}|\mathsf{S}} \theta^*_{\mathsf{R}}, \qquad \Sigma_{\mathsf{R}|\mathsf{S}} = \Sigma_{\mathsf{RR}} - \Sigma_{\mathsf{RS}} \Sigma^{-1}_{\mathsf{SS}} \Sigma_{\mathsf{SR}}.$$

Proof. Let $z := \eta - \theta^*_{\mathsf{S}}$. Since $\theta = (\eta, 0_{\mathsf{R}})$, we have $\theta - \theta^* = (z, -\theta^*_{\mathsf{R}})$. Expanding the quadratic form blockwise gives

$$(\theta - \theta^*)^\top \Sigma(\theta - \theta^*) = z^\top \Sigma_{\mathsf{SS}} z - 2z^\top \Sigma_{\mathsf{SR}} \theta^*_{\mathsf{R}} + (\theta^*_{\mathsf{R}})^\top \Sigma_{\mathsf{RR}} \theta^*_{\mathsf{R}}.$$

Now define $z_0 := \eta - \theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S}) = z - \Sigma^{-1}_{\mathsf{SS}} \Sigma_{\mathsf{SR}} \theta^*_{\mathsf{R}}$. Completing the square yields

$$z^\top \Sigma_{\mathsf{SS}} z - 2z^\top \Sigma_{\mathsf{SR}} \theta^*_{\mathsf{R}} = z_0^\top \Sigma_{\mathsf{SS}} z_0 - (\theta^*_{\mathsf{R}})^\top \Sigma_{\mathsf{RS}} \Sigma^{-1}_{\mathsf{SS}} \Sigma_{\mathsf{SR}} \theta^*_{\mathsf{R}}.$$

Substituting this into the previous display gives

$$(\theta - \theta^*)^\top \Sigma(\theta - \theta^*) = z_0^\top \Sigma_{\mathsf{SS}} z_0 + (\theta^*_{\mathsf{R}})^\top \left(\Sigma_{\mathsf{RR}} - \Sigma_{\mathsf{RS}} \Sigma^{-1}_{\mathsf{SS}} \Sigma_{\mathsf{SR}}\right) \theta^*_{\mathsf{R}},$$

which is the claimed identity. □

Proof of Proposition C.2. Fix $\mathsf{S} \subseteq [d]$ and write $\mathsf{R} = [d] \backslash \mathsf{S}$. For any $\theta = (\eta, 0_{\mathsf{R}}) \in \Theta(\mathsf{S})$, Lemma C.1 gives

$$(\theta - \theta^*)^\top \Sigma(\theta - \theta^*) = L_{\mathrm{pred}}(\mathsf{S}) + (\eta - \theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S}))^\top \Sigma_{\mathsf{SS}} (\eta - \theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S})).$$

Moreover, $\theta^\top H^{-1} \theta = \eta^\top H^{-1}_{\mathsf{SS}} \eta$. Therefore

$$\mathrm{MSE}_{\mathrm{strat}}(\theta) = \sigma^2 + L_{\mathrm{pred}}(\mathsf{S}) + (\eta - \theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S}))^\top \Sigma_{\mathsf{SS}} (\eta - \theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S})) + (\eta^\top H^{-1}_{\mathsf{SS}} \eta)^2.$$

Minimizing over $\eta \in \mathbb{R}^{|\mathsf{S}|}$ yields

$$\widetilde{\mathrm{OPT}}(\mathsf{S}) = \sigma^2 + L_{\mathrm{pred}}(\mathsf{S}) + \Gamma(\mathsf{S}),$$

where

$$\Gamma(\mathsf{S}) = \min_{\eta \in \mathbb{R}^{|\mathsf{S}|}} \left\{ (\eta - \theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S}))^\top \Sigma_{\mathsf{SS}} (\eta - \theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S})) + (\eta^\top H^{-1}_{\mathsf{SS}} \eta)^2 \right\}.$$

For the full support $[d]$, we have $L_{\mathrm{pred}}([d]) = 0$ and $\theta^{\mathrm{LS}}([d]) = \theta^*$, so

$$\widetilde{\mathrm{OPT}} = \widetilde{\mathrm{OPT}}([d]) = \sigma^2 + \Gamma([d]).$$

Subtracting gives

$$\widetilde{\mathrm{OPT}}(\mathsf{S}) - \widetilde{\mathrm{OPT}} = L_{\mathrm{pred}}(\mathsf{S}) + \Gamma(\mathsf{S}) - \Gamma([d]) = L_{\mathrm{pred}}(\mathsf{S}) - \big(\Gamma([d]) - \Gamma(\mathsf{S})\big),$$

as claimed. □

C.4 Proof of Theorem C.1

Theorem C.1 (Near-optimality of ridge on a fixed support)**.** *For every* $\mathsf{S} \subseteq [d]$*,*

$$0 \le \min_{\lambda \ge 0} \left\{ \mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda)) - \widetilde{\mathrm{OPT}}(\mathsf{S}) \right\} \le C_H(\mathsf{S})\, \delta_H(\mathsf{S})^2,$$

where $C_H(\mathsf{S}) := 4\, L_H(\mathsf{S})\, \|\Sigma^{-1/2}_{\mathsf{SS}} H^{-1}_{\mathsf{SS}} \Sigma^{-1/2}_{\mathsf{SS}}\|^2_{\mathrm{op}}\, \|\Sigma^{1/2}_{\mathsf{SS}} \theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S})\|^6_2$*, with* $L_H(\mathsf{S}) := 1 + 6\|\Sigma^{-1/2}_{\mathsf{SS}} H^{-1}_{\mathsf{SS}} \Sigma^{-1/2}_{\mathsf{SS}}\|^2_{\mathrm{op}} \|\Sigma^{1/2}_{\mathsf{SS}} \theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S})\|^2_2$*.*

We now prove the fixed-support ridge approximation theorem. Fix a support $\mathsf{S} \subseteq [d]$. Let $\eta^\star_{\mathsf{S}}$ denote the unique minimizer in the definition of $\Gamma(\mathsf{S})$:

$$\eta^\star_{\mathsf{S}} \in \arg\min_{\eta \in \mathbb{R}^{|\mathsf{S}|}} \left\{ (\eta - \theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S}))^\top \Sigma_{\mathsf{SS}} (\eta - \theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S})) + (\eta^\top H^{-1}_{\mathsf{SS}} \eta)^2 \right\}.$$

For $\lambda \ge 0$, let $\theta^{\mathrm{OR}}_{\mathsf{S}}(\lambda)$ denote the retained-coordinate vector of $\theta^{\mathrm{OR}}(\mathsf{S}, \lambda)$:

$$\theta^{\mathrm{OR}}_{\mathsf{S}}(\lambda) = (\Sigma_{\mathsf{SS}} + \lambda I)^{-1} \Sigma_{\mathsf{SS}} \theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S}).$$

Lemma C.2 (Support-restricted oracle is generalized ridge)**.** *The minimizer* $\eta^\star_{\mathsf{S}}$ *is unique. Moreover, defining* $\alpha_{\mathsf{S}} := 2\, \eta^{\star\top}_{\mathsf{S}} H^{-1}_{\mathsf{SS}} \eta^\star_{\mathsf{S}}$*, we have*

$$(\Sigma_{\mathsf{SS}} + \alpha_{\mathsf{S}} H^{-1}_{\mathsf{SS}}) \eta^\star_{\mathsf{S}} = \Sigma_{\mathsf{SS}} \theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S}).$$

Equivalently,

$$\eta^\star_{\mathsf{S}} = (\Sigma_{\mathsf{SS}} + \alpha_{\mathsf{S}} H^{-1}_{\mathsf{SS}})^{-1} \Sigma_{\mathsf{SS}} \theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S}).$$

Proof. The Hessian of the support-restricted objective in η is

$$2\Sigma_{\mathsf{SS}} + 8H_{\mathsf{SS}}^{-1}\eta\eta^\top H_{\mathsf{SS}}^{-1} + 4(\eta^\top H_{\mathsf{SS}}^{-1}\eta)H_{\mathsf{SS}}^{-1} \succeq 2\Sigma_{\mathsf{SS}} \succ 0.$$

Thus the objective is strongly convex and has a unique minimizer. The first-order condition at $\eta_{\mathsf{S}}^\star$ is

$$2\Sigma_{\mathsf{SS}}(\eta_{\mathsf{S}}^\star - \theta_{\mathsf{S}}^{\mathrm{LS}}(\mathsf{S})) + 4(\eta_{\mathsf{S}}^{\star\top} H_{\mathsf{SS}}^{-1}\eta_{\mathsf{S}}^\star)H_{\mathsf{SS}}^{-1}\eta_{\mathsf{S}}^\star = 0.$$

Dividing by 2 and using the definition of α_{S} gives

$$\Sigma_{\mathsf{SS}}(\eta_{\mathsf{S}}^\star - \theta_{\mathsf{S}}^{\mathrm{LS}}(\mathsf{S})) + \alpha_{\mathsf{S}} H_{\mathsf{SS}}^{-1}\eta_{\mathsf{S}}^\star = 0.$$

Equivalently,

$$(\Sigma_{\mathsf{SS}} + \alpha_{\mathsf{S}} H_{\mathsf{SS}}^{-1})\eta_{\mathsf{S}}^\star = \Sigma_{\mathsf{SS}}\theta_{\mathsf{S}}^{\mathrm{LS}}(\mathsf{S}).$$

□

We use whitened coordinates on the support. Define

$$u_{\mathsf{S}}^\star := \Sigma_{\mathsf{SS}}^{1/2}\eta_{\mathsf{S}}^\star, \qquad u_{\mathsf{S}}^{\mathrm{LS}} := \Sigma_{\mathsf{SS}}^{1/2}\theta_{\mathsf{S}}^{\mathrm{LS}}(\mathsf{S}), \qquad u_{\lambda,\mathsf{S}} := \Sigma_{\mathsf{SS}}^{1/2}\theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda).$$

The support-restricted oracle satisfies

$$\left(I + \alpha_{\mathsf{S}}\Sigma_{\mathsf{SS}}^{-1/2}H_{\mathsf{SS}}^{-1}\Sigma_{\mathsf{SS}}^{-1/2}\right)u_{\mathsf{S}}^\star = u_{\mathsf{S}}^{\mathrm{LS}},$$

while the ordinary ridge path satisfies

$$(I + \lambda\Sigma_{\mathsf{SS}}^{-1})u_{\lambda,\mathsf{S}} = u_{\mathsf{S}}^{\mathrm{LS}}.$$

Lemma C.3 (Distance from the oracle to the ordinary ridge path)**.** *For every $\gamma \geq 0$, if $\lambda = \alpha_{\mathsf{S}}\gamma$, then*

$$\|u_{\lambda,\mathsf{S}} - u_{\mathsf{S}}^\star\|_2 \leq \alpha_{\mathsf{S}} \left\|\Sigma_{\mathsf{SS}}^{-1/2}(H_{\mathsf{SS}}^{-1} - \gamma I)\Sigma_{\mathsf{SS}}^{-1/2}u_{\mathsf{S}}^\star\right\|_2.$$

Consequently,

$$\inf_{\lambda\geq 0}\|u_{\lambda,\mathsf{S}} - u_{\mathsf{S}}^\star\|_2 \leq \alpha_{\mathsf{S}}\,\delta_H(\mathsf{S})\,\|u_{\mathsf{S}}^\star\|_2.$$

Proof. Subtracting the whitened normal equations gives

$$(I + \lambda\Sigma_{\mathsf{SS}}^{-1})(u_{\lambda,\mathsf{S}} - u_{\mathsf{S}}^\star) = \left(\alpha_{\mathsf{S}}\Sigma_{\mathsf{SS}}^{-1/2}H_{\mathsf{SS}}^{-1}\Sigma_{\mathsf{SS}}^{-1/2} - \lambda\Sigma_{\mathsf{SS}}^{-1}\right)u_{\mathsf{S}}^\star.$$

Therefore

$$u_{\lambda,\mathsf{S}} - u_{\mathsf{S}}^\star = (I + \lambda\Sigma_{\mathsf{SS}}^{-1})^{-1}\left(\alpha_{\mathsf{S}}\Sigma_{\mathsf{SS}}^{-1/2}H_{\mathsf{SS}}^{-1}\Sigma_{\mathsf{SS}}^{-1/2} - \lambda\Sigma_{\mathsf{SS}}^{-1}\right)u_{\mathsf{S}}^\star.$$

Choosing $\lambda = \alpha_{\mathsf{S}}\gamma$ gives

$$u_{\alpha_{\mathsf{S}}\gamma,\mathsf{S}} - u_{\mathsf{S}}^\star = \alpha_{\mathsf{S}}(I + \alpha_{\mathsf{S}}\gamma\Sigma_{\mathsf{SS}}^{-1})^{-1}\Sigma_{\mathsf{SS}}^{-1/2}(H_{\mathsf{SS}}^{-1} - \gamma I)\Sigma_{\mathsf{SS}}^{-1/2}u_{\mathsf{S}}^\star.$$

Since $\Sigma_{\mathsf{SS}}^{-1} \succ 0$,

$$\|(I + \alpha_{\mathsf{S}}\gamma\Sigma_{\mathsf{SS}}^{-1})^{-1}\|_{\mathrm{op}} \leq 1.$$

Thus

$$\|u_{\alpha_{\mathsf{S}}\gamma,\mathsf{S}} - u_{\mathsf{S}}^\star\|_2 \leq \alpha_{\mathsf{S}}\left\|\Sigma_{\mathsf{SS}}^{-1/2}(H_{\mathsf{SS}}^{-1} - \gamma I)\Sigma_{\mathsf{SS}}^{-1/2}u_{\mathsf{S}}^\star\right\|_2.$$

Taking the infimum over $\gamma \geq 0$ and using the definition of $\delta_H(\mathsf{S})$ gives

$$\inf_{\lambda\geq 0}\|u_{\lambda,\mathsf{S}} - u_{\mathsf{S}}^\star\|_2 \leq \alpha_{\mathsf{S}}\delta_H(\mathsf{S})\|u_{\mathsf{S}}^\star\|_2.$$

□

Lemma C.4 (Curvature bound on the fixed support)**.** *Let*

$$L_H(\mathsf{S}) := 1 + 6\left\|\Sigma_{\mathsf{SS}}^{-1/2}H_{\mathsf{SS}}^{-1}\Sigma_{\mathsf{SS}}^{-1/2}\right\|_{\mathrm{op}}^2\left\|\Sigma_{\mathsf{SS}}^{1/2}\theta_{\mathsf{S}}^{\mathrm{LS}}(\mathsf{S})\right\|_2^2.$$

Then

$$\mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S},\lambda)) - \widetilde{\mathrm{OPT}}(\mathsf{S}) \leq L_H(\mathsf{S})\|u_{\lambda,\mathsf{S}} - u_{\mathsf{S}}^\star\|_2^2$$

for every $\lambda \geq 0$.

Proof. After subtracting the support-dependent constants $\sigma^2 + L_{\mathrm{pred}}(\mathsf{S})$, the support-restricted zero-intercept objective in whitened coordinates is

$$u \mapsto \|u - u_{\mathsf{S}}^{\mathrm{LS}}\|_2^2 + \left(u^\top \Sigma_{\mathsf{SS}}^{-1/2} H_{\mathsf{SS}}^{-1} \Sigma_{\mathsf{SS}}^{-1/2} u\right)^2.$$

The Hessian of this function is

$$2I + 8\Sigma_{\mathsf{SS}}^{-1/2} H_{\mathsf{SS}}^{-1} \Sigma_{\mathsf{SS}}^{-1/2} u u^\top \Sigma_{\mathsf{SS}}^{-1/2} H_{\mathsf{SS}}^{-1} \Sigma_{\mathsf{SS}}^{-1/2} + 4\left(u^\top \Sigma_{\mathsf{SS}}^{-1/2} H_{\mathsf{SS}}^{-1} \Sigma_{\mathsf{SS}}^{-1/2} u\right) \Sigma_{\mathsf{SS}}^{-1/2} H_{\mathsf{SS}}^{-1} \Sigma_{\mathsf{SS}}^{-1/2}.$$

Hence its operator norm is at most

$$2 + 12\left\|\Sigma_{\mathsf{SS}}^{-1/2} H_{\mathsf{SS}}^{-1} \Sigma_{\mathsf{SS}}^{-1/2}\right\|_{\mathrm{op}}^2 \|u\|_2^2.$$

Both $u_{\mathsf{S}}^\star$ and $u_{\lambda,\mathsf{S}}$ lie in the ball $\|u\|_2 \le \|u_{\mathsf{S}}^{\mathrm{LS}}\|_2$. Indeed,

$$u_{\mathsf{S}}^\star = \left(I + \alpha_{\mathsf{S}} \Sigma_{\mathsf{SS}}^{-1/2} H_{\mathsf{SS}}^{-1} \Sigma_{\mathsf{SS}}^{-1/2}\right)^{-1} u_{\mathsf{S}}^{\mathrm{LS}},$$

and

$$u_{\lambda,\mathsf{S}} = (I + \lambda \Sigma_{\mathsf{SS}}^{-1})^{-1} u_{\mathsf{S}}^{\mathrm{LS}},$$

and both resolvents have operator norm at most 1. Therefore, along the line segment between $u_{\mathsf{S}}^\star$ and $u_{\lambda,\mathsf{S}}$, the Hessian operator norm is at most $2L_H(\mathsf{S})$. Since $u_{\mathsf{S}}^\star$ is the minimizer, Taylor's theorem gives

$$\mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda)) - \widetilde{\mathrm{OPT}}(\mathsf{S}) \le L_H(\mathsf{S}) \|u_{\lambda,\mathsf{S}} - u_{\mathsf{S}}^\star\|_2^2.$$

□

Proof of Theorem C.1. By Lemma C.4 and Lemma C.3,

$$\min_{\lambda \ge 0} \left\{\mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda)) - \widetilde{\mathrm{OPT}}(\mathsf{S})\right\} \le \alpha_{\mathsf{S}}^2 L_H(\mathsf{S}) \delta_H(\mathsf{S})^2 \|u_{\mathsf{S}}^\star\|_2^2.$$

Now

$$\|u_{\mathsf{S}}^\star\|_2 \le \|u_{\mathsf{S}}^{\mathrm{LS}}\|_2 = \left\|\Sigma_{\mathsf{SS}}^{1/2} \theta_{\mathsf{S}}^{\mathrm{LS}}(\mathsf{S})\right\|_2.$$

Also,

$$\alpha_{\mathsf{S}} = 2\eta_{\mathsf{S}}^{\star\top} H_{\mathsf{SS}}^{-1} \eta_{\mathsf{S}}^\star = 2u_{\mathsf{S}}^{\star\top} \Sigma_{\mathsf{SS}}^{-1/2} H_{\mathsf{SS}}^{-1} \Sigma_{\mathsf{SS}}^{-1/2} u_{\mathsf{S}}^\star,$$

so

$$\alpha_{\mathsf{S}} \le 2\left\|\Sigma_{\mathsf{SS}}^{-1/2} H_{\mathsf{SS}}^{-1} \Sigma_{\mathsf{SS}}^{-1/2}\right\|_{\mathrm{op}} \|u_{\mathsf{S}}^\star\|_2^2 \le 2\left\|\Sigma_{\mathsf{SS}}^{-1/2} H_{\mathsf{SS}}^{-1} \Sigma_{\mathsf{SS}}^{-1/2}\right\|_{\mathrm{op}} \|u_{\mathsf{S}}^{\mathrm{LS}}\|_2^2.$$

Therefore

$$\alpha_{\mathsf{S}}^2 \|u_{\mathsf{S}}^\star\|_2^2 \le 4\left\|\Sigma_{\mathsf{SS}}^{-1/2} H_{\mathsf{SS}}^{-1} \Sigma_{\mathsf{SS}}^{-1/2}\right\|_{\mathrm{op}}^2 \|u_{\mathsf{S}}^{\mathrm{LS}}\|_2^6.$$

Combining the displays gives

$$\min_{\lambda \ge 0} \left\{\mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda)) - \widetilde{\mathrm{OPT}}(\mathsf{S})\right\} \le C_H(\mathsf{S}) \delta_H(\mathsf{S})^2,$$

with the valid choice

$$C_H(\mathsf{S}) := 4\, L_H(\mathsf{S}) \left\|\Sigma_{\mathsf{SS}}^{-1/2} H_{\mathsf{SS}}^{-1} \Sigma_{\mathsf{SS}}^{-1/2}\right\|_{\mathrm{op}}^2 \left\|\Sigma_{\mathsf{SS}}^{1/2} \theta_{\mathsf{S}}^{\mathrm{LS}}(\mathsf{S})\right\|_2^6.$$

□

C.5 Proof of Theorem 2

For every support S and every $\lambda \ge 0$, we have the exact decomposition

$$\mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda)) - \widetilde{\mathrm{OPT}} = \left(\widetilde{\mathrm{OPT}}(\mathsf{S}) - \widetilde{\mathrm{OPT}}\right) + \left(\mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda)) - \widetilde{\mathrm{OPT}}(\mathsf{S})\right).$$

By Proposition C.2,

$$\widetilde{\mathrm{OPT}}(\mathsf{S}) - \widetilde{\mathrm{OPT}} = L_{\mathrm{pred}}(\mathsf{S}) - \big(\Gamma([d]) - \Gamma(\mathsf{S})\big).$$

By Theorem C.1,

$$\min_{\lambda \ge 0} \left\{\mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda)) - \widetilde{\mathrm{OPT}}(\mathsf{S})\right\} \le C_H(\mathsf{S}) \delta_H(\mathsf{S})^2.$$

Therefore, for every S,

$$\min_{\lambda \ge 0} \left\{\mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda)) - \widetilde{\mathrm{OPT}}\right\} \le L_{\mathrm{pred}}(\mathsf{S}) - \big(\Gamma([d]) - \Gamma(\mathsf{S})\big) + C_H(\mathsf{S}) \delta_H(\mathsf{S})^2.$$

C.6 Structure of the zero-intercept oracle

The next proposition records the generalized-ridge structure of the zero-intercept oracle over the full feature set. It is the full-support analogue of Lemma C.2.

Proposition C.3 (Zero-intercept oracle over $\mathbb{R}^d$)**.** *The objective* $\mathrm{MSE}_{\mathrm{strat}}(\theta)$ *is strongly convex and therefore admits a unique minimizer* $\theta^{\widetilde{\mathrm{OPT}}} \in \mathbb{R}^d$*. Moreover,* $\theta^{\widetilde{\mathrm{OPT}}}$ *is characterized by*

$$\Sigma(\theta^{\widetilde{\mathrm{OPT}}} - \theta^*) + 2(\theta^{\widetilde{\mathrm{OPT}}\top} H^{-1} \theta^{\widetilde{\mathrm{OPT}}}) H^{-1} \theta^{\widetilde{\mathrm{OPT}}} = 0.$$

Equivalently, defining

$$\lambda^{\widetilde{\mathrm{OPT}}} := 2\, \theta^{\widetilde{\mathrm{OPT}}\top} H^{-1} \theta^{\widetilde{\mathrm{OPT}}},$$

we have

$$(\Sigma + \lambda^{\widetilde{\mathrm{OPT}}} H^{-1}) \theta^{\widetilde{\mathrm{OPT}}} = \Sigma \theta^*,$$

and hence

$$\theta^{\widetilde{\mathrm{OPT}}} = (\Sigma + \lambda^{\widetilde{\mathrm{OPT}}} H^{-1})^{-1} \Sigma \theta^*.$$

Thus the zero-intercept oracle is a generalized-ridge shrinkage of θ^**, with the regularization level determined endogenously by*

$$\lambda^{\widetilde{\mathrm{OPT}}} = 2\, \theta^{\widetilde{\mathrm{OPT}}\top} H^{-1} \theta^{\widetilde{\mathrm{OPT}}}.$$

Proof. Ignoring the additive constant σ^2, the zero-intercept objective is

$$\mathrm{MSE}_{\mathrm{strat}}(\theta) - \sigma^2 = (\theta - \theta^*)^\top \Sigma (\theta - \theta^*) + (\theta^\top H^{-1} \theta)^2.$$

Its gradient is

$$\nabla_\theta \mathrm{MSE}_{\mathrm{strat}}(\theta) = 2\Sigma(\theta - \theta^*) + 4(\theta^\top H^{-1} \theta) H^{-1} \theta,$$

and its Hessian is

$$\nabla^2_\theta \mathrm{MSE}_{\mathrm{strat}}(\theta) = 2\Sigma + 8 H^{-1} \theta \theta^\top H^{-1} + 4(\theta^\top H^{-1} \theta) H^{-1}.$$

Since $\Sigma \succ 0$ and $H^{-1} \succ 0$,

$$\nabla^2_\theta \mathrm{MSE}_{\mathrm{strat}}(\theta) \succeq 2\Sigma \succ 0.$$

Thus $\mathrm{MSE}_{\mathrm{strat}}(\theta)$ is strongly convex and has a unique minimizer.

At the minimizer $\theta^{\widetilde{\mathrm{OPT}}}$, the first-order condition gives

$$2\Sigma(\theta^{\widetilde{\mathrm{OPT}}} - \theta^*) + 4(\theta^{\widetilde{\mathrm{OPT}}\top} H^{-1} \theta^{\widetilde{\mathrm{OPT}}}) H^{-1} \theta^{\widetilde{\mathrm{OPT}}} = 0.$$

Dividing by 2 and defining

$$\lambda^{\widetilde{\mathrm{OPT}}} = 2\, \theta^{\widetilde{\mathrm{OPT}}\top} H^{-1} \theta^{\widetilde{\mathrm{OPT}}}$$

yields

$$(\Sigma + \lambda^{\widetilde{\mathrm{OPT}}} H^{-1}) \theta^{\widetilde{\mathrm{OPT}}} = \Sigma \theta^*.$$

Since $\Sigma + \lambda^{\widetilde{\mathrm{OPT}}} H^{-1} \succ 0$, this is equivalent to

$$\theta^{\widetilde{\mathrm{OPT}}} = (\Sigma + \lambda^{\widetilde{\mathrm{OPT}}} H^{-1})^{-1} \Sigma \theta^*.$$

□

C.7 Ridge optimality under isotropic costs

The next corollary formalizes the case in which scalar ridge is expressive enough to attain the support-restricted zero-intercept oracle exactly.

Corollary C.1 (Zero heterogeneity gap under isotropy)**.** *Fix a support* $\mathsf{S} \subseteq [d]$*. If* $H_{\mathsf{SS}}^{-1} = \gamma I$ *for some* $\gamma \geq 0$*, then*

$$\min_{\lambda \geq 0} \mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda)) = \widetilde{\mathrm{OPT}}(\mathsf{S}).$$

Proof of Corollary C.1. Let $\eta^\star_{\mathsf{S}}$ denote the support-restricted zero-intercept oracle on S. By Lemma C.2, there exists $\alpha_{\mathsf{S}} := 2\,\eta^{\star\top}_{\mathsf{S}} H^{-1}_{\mathsf{SS}} \eta^\star_{\mathsf{S}}$ such that

$$(\Sigma_{\mathsf{SS}} + \alpha_{\mathsf{S}} H^{-1}_{\mathsf{SS}})\eta^\star_{\mathsf{S}} = \Sigma_{\mathsf{SS}} \theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S}).$$

If $H^{-1}_{\mathsf{SS}} = \gamma I$, then

$$(\Sigma_{\mathsf{SS}} + \alpha_{\mathsf{S}} \gamma I)\eta^\star_{\mathsf{S}} = \Sigma_{\mathsf{SS}} \theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S}).$$

Thus $\eta^\star_{\mathsf{S}}$ is exactly the ordinary ridge estimator on support S with ridge parameter $\lambda = \alpha_{\mathsf{S}}\gamma$. Therefore the ordinary ridge path contains the support-restricted zero-intercept oracle, and so

$$\min_{\lambda \geq 0} \mathrm{MSE}_{\mathrm{strat}}\big(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda)\big) = \widetilde{\mathrm{OPT}}(\mathsf{S}).$$

□

Corollary C.1 shows that scalar ridge is exact when the retained manipulation costs are isotropic. This condition is sufficient but not necessary. Proposition C.4 below gives the exact necessary-and-sufficient condition: the fixed-support ridge gap is zero if and only if the support-restricted oracle lies in an eigenspace of the retained inverse-cost matrix.

C.8 Necessary and sufficient condition for ridge optimality

The isotropy condition in Corollary C.1 is sufficient but not necessary. The exact necessary-and-sufficient condition is that the support-restricted oracle lies in an eigenspace of the inverse-cost matrix on the retained support.

Proposition C.4 (Necessary and sufficient condition for exact fixed-support ridge optimality)**.** *Fix a support $\mathsf{S} \subseteq [d]$. Let $\eta^\star_{\mathsf{S}}$ denote the support-restricted zero-intercept oracle on S. Then*

$$\min_{\lambda \geq 0} \mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda)) = \widetilde{\mathrm{OPT}}(\mathsf{S})$$

if and only if there exists $\gamma \geq 0$ such that $(H^{-1}_{\mathsf{SS}} - \gamma I)\eta^\star_{\mathsf{S}} = 0$.

Proof. The support-restricted oracle $\eta^\star_{\mathsf{S}}$ satisfies

$$(\Sigma_{\mathsf{SS}} + \alpha_{\mathsf{S}} H^{-1}_{\mathsf{SS}})\eta^\star_{\mathsf{S}} = \Sigma_{\mathsf{SS}} \theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S}),$$

where $\alpha_{\mathsf{S}} = 2\eta^{\star\top}_{\mathsf{S}} H^{-1}_{\mathsf{SS}} \eta^\star_{\mathsf{S}}$. The ordinary ridge estimator at level λ satisfies

$$(\Sigma_{\mathsf{SS}} + \lambda I)\theta^{\mathrm{OR}}_{\mathsf{S}}(\lambda) = \Sigma_{\mathsf{SS}} \theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S}).$$

Suppose first that the fixed-support ridge gap is zero. Since the support-restricted objective is strongly convex, its minimizer is unique. Hence there exists $\lambda \geq 0$ such that $\theta^{\mathrm{OR}}_{\mathsf{S}}(\lambda) = \eta^\star_{\mathsf{S}}$. Subtracting the two normal equations gives $(\alpha_{\mathsf{S}} H^{-1}_{\mathsf{SS}} - \lambda I)\eta^\star_{\mathsf{S}} = 0$. If $\alpha_{\mathsf{S}} > 0$, set $\gamma = \lambda/\alpha_{\mathsf{S}}$. Then $(H^{-1}_{\mathsf{SS}} - \gamma I)\eta^\star_{\mathsf{S}} = 0$. If $\alpha_{\mathsf{S}} = 0$, then $\eta^\star_{\mathsf{S}} = 0$ because $H^{-1}_{\mathsf{SS}} \succ 0$. In this case, $(H^{-1}_{\mathsf{SS}} - \gamma I)\eta^\star_{\mathsf{S}} = 0$ holds for every $\gamma \geq 0$.

Conversely, suppose there exists $\gamma \geq 0$ such that $(H^{-1}_{\mathsf{SS}} - \gamma I)\eta^\star_{\mathsf{S}} = 0$. If $\eta^\star_{\mathsf{S}} = 0$, then the first-order condition implies $\Sigma_{\mathsf{SS}} \theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S}) = 0$, and hence $\theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S}) = 0$. Ordinary ridge also returns 0 for every $\lambda \geq 0$, so the ridge gap is zero.

If $\eta^\star_{\mathsf{S}} \neq 0$, then $\alpha_{\mathsf{S}} > 0$. Set $\lambda = \alpha_{\mathsf{S}}\gamma$. Then

$$(\Sigma_{\mathsf{SS}} + \lambda I)\eta^\star_{\mathsf{S}} = (\Sigma_{\mathsf{SS}} + \alpha_{\mathsf{S}}\gamma I)\eta^\star_{\mathsf{S}} = (\Sigma_{\mathsf{SS}} + \alpha_{\mathsf{S}} H^{-1}_{\mathsf{SS}})\eta^\star_{\mathsf{S}} = \Sigma_{\mathsf{SS}} \theta^{\mathrm{LS}}_{\mathsf{S}}(\mathsf{S}).$$

Therefore $\eta^\star_{\mathsf{S}}$ is the ordinary ridge estimator at level λ, so ordinary ridge attains $\widetilde{\mathrm{OPT}}(\mathsf{S})$. □

C.9 Two-level cost model

This subsection formalizes the intuition that high manipulability alone does not force exclusion. What matters for scalar ridge is whether the retained features have manipulation costs that are homogeneous enough to be controlled by a single regularization parameter.

Partition the features as $[d] = \mathsf{S}_m \cup \mathsf{S}_b$ with $\mathsf{S}_m \cap \mathsf{S}_b = \emptyset$, where S_m denotes a more manipulable feature group and S_b denotes a baseline feature group. For $\kappa > \mu > 0$, consider the two-level inverse-cost model

$$H_{\kappa,\mu}^{-1} = \begin{pmatrix} \kappa I_{\mathsf{S}_m} & 0 \\ 0 & \mu I_{\mathsf{S}_b} \end{pmatrix},$$

where the rows and columns are ordered with the features in S_m first. The special case $\mu = 1$ corresponds to a highly manipulable group and a baseline group; the limiting regime $\mu \to 0$ corresponds to a baseline group that is essentially non-manipulable.

Proposition C.5 (Homogeneity and heterogeneity in the two-level cost model)**.** *Assume $\Sigma = I$ and $H^{-1} = H_{\kappa,\mu}^{-1}$ with $\kappa > \mu > 0$. For any support $\mathsf{T} \subseteq [d]$,*

$$\delta_H(\mathsf{T}) = \begin{cases} 0, & \mathsf{T} \subseteq \mathsf{S}_m \text{ or } \mathsf{T} \subseteq \mathsf{S}_b, \\ (\kappa - \mu)/2, & \mathsf{T} \cap \mathsf{S}_m \neq \emptyset \text{ and } \mathsf{T} \cap \mathsf{S}_b \neq \emptyset. \end{cases}$$

Consequently, if $\mathsf{T} \subseteq \mathsf{S}_m$ or $\mathsf{T} \subseteq \mathsf{S}_b$, then

$$\min_{\lambda \geq 0} \left\{ \mathrm{MSE}_{\mathrm{strat}}\big(\theta^{\mathrm{OR}}(\mathsf{T}, \lambda)\big) - \widetilde{\mathrm{OPT}}(\mathsf{T}) \right\} = 0.$$

Proof. Under the isotropic covariance $\Sigma = I$, we have

$$\delta_H(\mathsf{T}) = \inf_{\gamma \geq 0} \left\| H_{\mathsf{T}\mathsf{T}}^{-1} - \gamma I \right\|_{\mathrm{op}}.$$

If $\mathsf{T} \subseteq \mathsf{S}_m$, then $H_{\mathsf{T}\mathsf{T}}^{-1} = \kappa I_{\mathsf{T}}$, so choosing $\gamma = \kappa$ gives $\delta_H(\mathsf{T}) = 0$. Similarly, if $\mathsf{T} \subseteq \mathsf{S}_b$, then $H_{\mathsf{T}\mathsf{T}}^{-1} = \mu I_{\mathsf{T}}$, so choosing $\gamma = \mu$ gives $\delta_H(\mathsf{T}) = 0$.

Now suppose that $\mathsf{T} \cap \mathsf{S}_m \neq \emptyset$ and $\mathsf{T} \cap \mathsf{S}_b \neq \emptyset$. Then $H_{\mathsf{T}\mathsf{T}}^{-1}$ has eigenvalue κ on $\mathsf{T} \cap \mathsf{S}_m$ and eigenvalue μ on $\mathsf{T} \cap \mathsf{S}_b$. Therefore

$$\left\| H_{\mathsf{T}\mathsf{T}}^{-1} - \gamma I \right\|_{\mathrm{op}} = \max\{|\kappa - \gamma|,\ |\mu - \gamma|\}.$$

The minimizer is $\gamma = (\kappa + \mu)/2$, and the minimum value is $(\kappa - \mu)/2$. This proves the expression for $\delta_H(\mathsf{T})$.

The zero scalar-ridge gap for supports contained in a single group follows from $\delta_H(\mathsf{T}) = 0$ together with Theorem C.1. It also follows directly from Proposition C.5, since $H_{\mathsf{T}\mathsf{T}}^{-1}$ is a scalar multiple of the identity on such supports. □

Proposition C.5 leads to a near-optimality result for the optimal support-restricted ridge estimator:

Corollary C.2 (Near-optimality of support-restricted ridge in the two-level cost model)**.** *Assume the setup of Proposition C.5. If there exists a support $\mathsf{T} \subseteq \mathsf{S}_m$ or $\mathsf{T} \subseteq \mathsf{S}_b$ such that*

$$L_{\mathrm{pred}}(\mathsf{T}) - \big(\Gamma([d]) - \Gamma(\mathsf{T})\big) \leq \varepsilon,$$

then

$$\min_{\mathsf{S} \subseteq [d],\, \lambda \geq 0} \mathrm{MSE}_{\mathrm{strat}}\big(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda)\big) \leq \widetilde{\mathrm{OPT}} + \varepsilon.$$

Proof. By Proposition C.5, any support contained entirely in S_m or entirely in S_b has zero scalar-ridge gap:

$$\min_{\lambda \geq 0} \mathrm{MSE}_{\mathrm{strat}}\big(\theta^{\mathrm{OR}}(\mathsf{T}, \lambda)\big) = \widetilde{\mathrm{OPT}}(\mathsf{T}).$$

Therefore

$$\min_{\lambda \geq 0} \mathrm{MSE}_{\mathrm{strat}}\big(\theta^{\mathrm{OR}}(\mathsf{T}, \lambda)\big) - \widetilde{\mathrm{OPT}} = \widetilde{\mathrm{OPT}}(\mathsf{T}) - \widetilde{\mathrm{OPT}}.$$

Using the support restriction identity from Proposition C.2,

$$\widetilde{\mathrm{OPT}}(\mathsf{T}) - \widetilde{\mathrm{OPT}} = L_{\mathrm{pred}}(\mathsf{T}) - \big(\Gamma([d]) - \Gamma(\mathsf{T})\big).$$

By assumption, this quantity is at most ε. Taking the minimum over all supports and all ridge parameters can only improve the value, which proves the claim. □

In this two-level cost model, scalar ridge is exact on any support drawn from a single cost class, even when that class is highly manipulable. The obstacle is therefore not high manipulability by itself, but combining features with very different manipulation costs. On a mixed support, the oracle would shrink the two cost classes at different rates, whereas ordinary ridge has only one scalar shrinkage parameter. As a result, the heterogeneity gap grows with the separation $\kappa - \mu$. This formalizes the policy insight in Figure 2: predictive manipulable groups can be retained when their manipulation costs are internally homogeneous, whereas feature selection becomes more important when retained costs are heterogeneous.

C.10 Additional properties of ridge tuning

This subsection records two elementary facts about ridge regularization in the strategic setting. They are not needed for our main approximation theorem, but they support the qualitative message that regularization directly reduces strategic MSE.

Fix a support S. Let

$$\theta_{\mathsf{S}}^{\mathrm{LS}} := \theta_{\mathsf{S}}^{\mathrm{LS}}(\mathsf{S}), \qquad \theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda) = (\Sigma_{\mathsf{SS}} + \lambda I)^{-1}\Sigma_{\mathsf{SS}}\theta_{\mathsf{S}}^{\mathrm{LS}}.$$

Ignoring constants independent of λ, the strategic MSE along the support-restricted ridge path is

$$R_{\mathsf{S}}(\lambda) = (\theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda) - \theta_{\mathsf{S}}^{\mathrm{LS}})^\top \Sigma_{\mathsf{SS}}(\theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda) - \theta_{\mathsf{S}}^{\mathrm{LS}}) + \big(\theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda)^\top H_{\mathsf{SS}}^{-1}\theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda)\big)^2.$$

Proposition C.6 (Infinitesimal ridge improves strategic MSE)**.** *If $\theta_{\mathsf{S}}^{\mathrm{LS}\top} H_{\mathsf{SS}}^{-1}\Sigma_{\mathsf{SS}}^{-1}\theta_{\mathsf{S}}^{\mathrm{LS}} > 0$, then there exists $\lambda_0 > 0$ such that for all $\lambda \in (0, \lambda_0)$,*

$$\mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda)) < \mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S}, 0)).$$

Proof. The derivative of the ridge path is

$$\frac{d}{d\lambda}\theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda) = -(\Sigma_{\mathsf{SS}} + \lambda I)^{-1}\theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda).$$

Thus

$$\frac{d}{d\lambda}\theta_{\mathsf{S}}^{\mathrm{OR}}(0) = -\Sigma_{\mathsf{SS}}^{-1}\theta_{\mathsf{S}}^{\mathrm{LS}}.$$

Write $G(\lambda) := (\theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda) - \theta_{\mathsf{S}}^{\mathrm{LS}})^\top \Sigma_{\mathsf{SS}}(\theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda) - \theta_{\mathsf{S}}^{\mathrm{LS}})$ and $Q(\lambda) := \big(\theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda)^\top H_{\mathsf{SS}}^{-1}\theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda)\big)^2$. Since $\theta_{\mathsf{S}}^{\mathrm{OR}}(0) = \theta_{\mathsf{S}}^{\mathrm{LS}}$, we have $G'(0) = 0$. Next, define $q(\lambda) := \theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda)^\top H_{\mathsf{SS}}^{-1}\theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda)$. Then $Q'(\lambda) = 2q(\lambda)q'(\lambda)$, and $q'(0) = -2\theta_{\mathsf{S}}^{\mathrm{LS}\top} H_{\mathsf{SS}}^{-1}\Sigma_{\mathsf{SS}}^{-1}\theta_{\mathsf{S}}^{\mathrm{LS}}$. Therefore $Q'(0) = -4(\theta_{\mathsf{S}}^{\mathrm{LS}\top} H_{\mathsf{SS}}^{-1}\theta_{\mathsf{S}}^{\mathrm{LS}})(\theta_{\mathsf{S}}^{\mathrm{LS}\top} H_{\mathsf{SS}}^{-1}\Sigma_{\mathsf{SS}}^{-1}\theta_{\mathsf{S}}^{\mathrm{LS}})$.

Since $H_{\mathsf{SS}}^{-1} \succ 0$, the assumed condition implies $\theta_{\mathsf{S}}^{\mathrm{LS}} \neq 0$, and hence $\theta_{\mathsf{S}}^{\mathrm{LS}\top} H_{\mathsf{SS}}^{-1}\theta_{\mathsf{S}}^{\mathrm{LS}} > 0$. Thus $R'_{\mathsf{S}}(0) = G'(0) + Q'(0) < 0$. By continuity, there exists $\lambda_0 > 0$ such that $R_{\mathsf{S}}(\lambda) < R_{\mathsf{S}}(0)$ for every $\lambda \in (0, \lambda_0)$. Since $R_{\mathsf{S}}(\lambda)$ differs from $\mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda))$ only by constants independent of λ, the claim follows. □

Proposition C.7 (Scaling inverse-cost intensity increases the locally optimal ridge level)**.** *Fix a support S. For $\tau > 0$, consider scaling the inverse-cost matrix from H_{SS}^{-1} to $\tau H_{\mathsf{SS}}^{-1}$. Ignoring constants independent of λ, define the strategic* MSE *along the support-restricted ridge path:*

$$R_{\mathsf{S},\tau}(\lambda) = G(\lambda) + \tau^2 Q(\lambda),$$

where $G(\lambda)$ and $Q(\lambda)$ are defined in the proof of Proposition C.6. Suppose that for each τ in an open interval $I \subset (0, \infty)$, $R_{\mathsf{S},\tau}$ admits a finite interior minimizer $\lambda^*(\tau) \in (0, \infty)$, and that this minimizer is strict, i.e.,

$$\frac{\partial R_{\mathsf{S},\tau}}{\partial \lambda}(\lambda^*(\tau)) = 0, \qquad \frac{\partial^2 R_{\mathsf{S},\tau}}{\partial \lambda^2}(\lambda^*(\tau)) > 0.$$

Then $\lambda^*(\tau)$ is differentiable on I, and

$$\frac{d}{d\tau}\lambda^*(\tau) > 0.$$

Proof. Let

$$\mathcal{F}(\lambda, \tau) := \frac{\partial R_{\mathsf{S},\tau}}{\partial \lambda}(\lambda) = G'(\lambda) + \tau^2 Q'(\lambda).$$

At $\lambda = \lambda^*(\tau)$, $\mathcal{F}(\lambda^*(\tau), \tau) = 0$. Moreover,

$$\frac{\partial \mathcal{F}}{\partial \lambda}(\lambda^*(\tau), \tau) = \frac{\partial^2 R_{\mathsf{S},\tau}}{\partial \lambda^2}(\lambda^*(\tau)) > 0.$$

By the implicit function theorem, $\lambda^*(\tau)$ is differentiable and

$$\frac{d}{d\tau}\lambda^*(\tau) = -\frac{\partial \mathcal{F}/\partial \tau}{\partial \mathcal{F}/\partial \lambda} = -\frac{2\tau Q'(\lambda^*(\tau))}{R''_{\mathsf{S},\tau}(\lambda^*(\tau))}.$$

It remains to show that $Q'(\lambda^*(\tau)) < 0$.

At the optimum, $G'(\lambda^*(\tau)) + \tau^2 Q'(\lambda^*(\tau)) = 0$. We now show that $G'(\lambda) > 0$ for every $\lambda > 0$ in the nontrivial case. Using

$$(\Sigma_{\mathsf{SS}} + \lambda I)\theta^{\mathrm{OR}}_{\mathsf{S}}(\lambda) = \Sigma_{\mathsf{SS}}\theta^{\mathrm{LS}}_{\mathsf{S}},$$

we have

$$\theta^{\mathrm{OR}}_{\mathsf{S}}(\lambda) - \theta^{\mathrm{LS}}_{\mathsf{S}} = -\lambda\Sigma^{-1}_{\mathsf{SS}}\theta^{\mathrm{OR}}_{\mathsf{S}}(\lambda).$$

Also,

$$\frac{d}{d\lambda}\theta^{\mathrm{OR}}_{\mathsf{S}}(\lambda) = -(\Sigma_{\mathsf{SS}} + \lambda I)^{-1}\theta^{\mathrm{OR}}_{\mathsf{S}}(\lambda).$$

Hence

$$G'(\lambda) = 2(\theta^{\mathrm{OR}}_{\mathsf{S}}(\lambda) - \theta^{\mathrm{LS}}_{\mathsf{S}})^\top \Sigma_{\mathsf{SS}} \frac{d}{d\lambda}\theta^{\mathrm{OR}}_{\mathsf{S}}(\lambda) = 2\lambda\, \theta^{\mathrm{OR}}_{\mathsf{S}}(\lambda)^\top (\Sigma_{\mathsf{SS}} + \lambda I)^{-1}\theta^{\mathrm{OR}}_{\mathsf{S}}(\lambda).$$

For $\lambda > 0$ and $\theta^{\mathrm{OR}}_{\mathsf{S}}(\lambda) \neq 0$, this is strictly positive. Under the stated strict interior-minimizer assumption, the nontrivial case has $\theta^{\mathrm{OR}}_{\mathsf{S}}(\lambda^*(\tau)) \neq 0$. Thus $G'(\lambda^*(\tau)) > 0$. The first-order condition then implies

$$Q'(\lambda^*(\tau)) = -\frac{G'(\lambda^*(\tau))}{\tau^2} < 0.$$

Therefore

$$\frac{d}{d\tau}\lambda^*(\tau) = -\frac{2\tau Q'(\lambda^*(\tau))}{R''_{\mathsf{S},\tau}(\lambda^*(\tau))} > 0,$$

because $\tau > 0$, $Q'(\lambda^*(\tau)) < 0$, and $R''_{\mathsf{S},\tau}(\lambda^*(\tau)) > 0$. □

D Simulation Details and Illustrations for Section 4

D.1 Details for Figure 1

We consider the following $d = 4$ example:

$$\Sigma = \begin{pmatrix} 1 & 0.9 & 0.2 & 0.1 \\ 0.9 & 1 & 0.2 & 0.1 \\ 0.2 & 0.2 & 1 & 0.5 \\ 0.1 & 0.1 & 0.5 & 1 \end{pmatrix}, \qquad \theta^* = (1.6, 1.2, 1.0, 0.7), \qquad H^{-1} = \mathrm{diag}(8.0, 1.5, 0.8, 0.5).$$

Feature 1 is highly predictive but also highly manipulable. Feature 2 is strongly correlated with feature 1, and therefore acts as a less manipulable proxy. Features 3 and 4 are less manipulable, moderately predictive features.

We enumerate all $2^4 = 16$ supports $\mathsf{S} \subseteq [4]$. For each support, we compute the support-restricted zero-intercept oracle value

$$\widetilde{\mathrm{OPT}}(\mathsf{S}) = \min_{\theta \in \Theta(\mathsf{S})} \mathrm{MSE}_{\mathrm{strat}}(\theta, 0),$$

and the best support-restricted ridge value

$$\min_{\lambda \geq 0} \mathrm{MSE}_{\mathrm{strat}}\big(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda), 0\big)$$

by one-dimensional optimization over λ. The plotted quantities are predictive loss $L_{\mathrm{pred}}(\mathsf{S})$, negative manipulability gain $-(\Gamma([d]) - \Gamma(\mathsf{S}))$, the gap after optimal ridge tuning

$$\min_{\lambda \geq 0} \mathrm{MSE}_{\mathrm{strat}}\big(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda), 0\big) - \widetilde{\mathrm{OPT}}(\mathsf{S}),$$

and total excess relative to $\widetilde{\mathrm{OPT}}$, given by the sum of these three quantities. The gap after optimal ridge tuning is bounded above by the heterogeneity gap $C_H(\mathsf{S})\delta_H(\mathsf{S})^2$; the figure plots the former, not its upper bound. The support restriction identity in Proposition C.2 gives the support restriction cost $\widehat{\mathrm{OPT}}(\mathsf{S}) - \widehat{\mathrm{OPT}} = L_{\mathrm{pred}}(\mathsf{S}) - (\Gamma([d]) - \Gamma(\mathsf{S}))$.

The left panel illustrates five representative supports. The full support $\{1, 2, 3, 4\}$ has zero support restriction cost but a large gap after optimal ridge tuning because it combines features with very different manipulation costs. The support $\{2, 3, 4\}$ drops the most manipulable feature while retaining its less manipulable proxy and the remaining less manipulable predictive features; it has a larger support restriction cost than the full support but a much smaller gap after optimal ridge tuning, and is best overall. The support $\{2, 3\}$ retains the proxy and one less manipulable feature but loses additional predictive signal. The singleton support $\{2\}$ retains only the proxy and therefore has zero gap after optimal ridge tuning but incurs substantial predictive loss. The pair $\{1, 2\}$ retains the manipulable feature and its proxy, and hence incurs smaller predictive loss than the singleton support, but has a large gap after optimal ridge tuning.

The right panel plots every support by support restriction cost and the gap after optimal ridge tuning. The light labels show all supports, while the five representative supports are highlighted. Because $\Theta(\mathsf{S}) \subseteq \Theta(\mathsf{T})$ whenever $\mathsf{S} \subseteq \mathsf{T}$, support restriction cost is monotone with respect to set inclusion: enlarging the support cannot increase support restriction cost. This monotonicity is visible in the placement of the full support near the left edge and smaller supports farther to the right. By contrast, the gap after optimal ridge tuning is not monotone in support size: adding features can either improve or worsen this gap, depending on whether the added features make the retained inverse-cost geometry more heterogeneous and how the retained inverse-cost geometry interacts with the covariance and predictive signal.

D.2 Details for Figure 2

We consider the following $d = 4$ example:

$$\Sigma = I_4, \qquad \theta^* = (a, a, 1.0, 0.9), \qquad H^{-1} = \mathrm{diag}\big(\kappa(1+\delta),\, \kappa(1-\delta),\, 0.6,\, 0.6\big),$$

where we fix $a = 1.2$. Thus features 1 and 2 form a manipulable pair with average manipulability κ and relative cost heterogeneity δ, while features 3 and 4 are less manipulable baseline features. We vary (κ, δ) on a 51×51 grid, with $\kappa \in [0.6, 7.5]$ and $\delta \in [0, 0.95]$.

For each grid point (κ, δ), we enumerate all 2^4 supports $\mathsf{S} \subseteq [4]$. For each support, we compute the exact support-restricted ridge optimum

$$\min_{\lambda \geq 0} \mathrm{MSE}_{\mathrm{strat}}\big(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda), 0\big),$$

and the exact support-restricted zero-intercept oracle

$$\widetilde{\mathrm{OPT}}(\mathsf{S}) := \min_{\theta \in \Theta(\mathsf{S})} \mathrm{MSE}_{\mathrm{strat}}(\theta, 0).$$

We also compute the full-support zero-intercept oracle

$$\widetilde{\mathrm{OPT}} := \min_{\theta \in \mathbb{R}^4} \mathrm{MSE}_{\mathrm{strat}}(\theta, 0).$$

These are computed using the generalized-ridge fixed-point characterization (see Proposition C.3 for more details):

$$\theta(q) = (\Sigma + 2qH^{-1})^{-1}\Sigma\theta^*, \qquad q = \theta(q)^\top H^{-1}\theta(q). \tag{15}$$

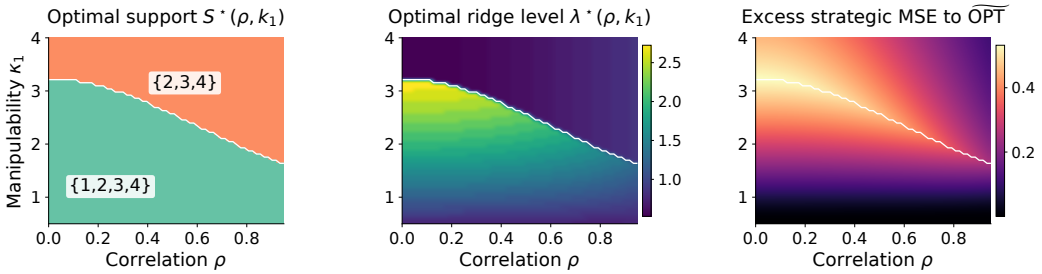

Figure 5: **Less manipulable correlated proxies can replace manipulable features.** We compute the optimal support and regularization level exactly for a $d = 4$ example with correlated features. Feature 1 is predictive but manipulable, feature 2 is a correlated less manipulable proxy for feature 1 with zero direct coefficient in θ^*, and features 3 and 4 are less manipulable baseline features. We vary the correlation ρ between features 1 and 2 on the x-axis, and the manipulability k_1 of feature 1 on the y-axis. (Left) The optimal support $\mathsf{S}^*(\rho, k_1)$ after joint optimization over S and λ. (Middle) The corresponding optimal ridge level $\lambda^*(\rho, k_1)$, where white contours indicate support boundaries. (Right) Excess strategic MSE relative to $\widetilde{\mathrm{OPT}}$. As correlation increases, the less manipulable feature becomes a more effective substitute for the manipulable feature. When k_1 is sufficiently large, the optimal support switches from $\{1, 2, 3, 4\}$ to $\{2, 3, 4\}$, dropping the manipulable feature while retaining its proxy. The resulting support-restricted ridge estimator remains close to $\widetilde{\mathrm{OPT}}$ over a broad region of the parameter space. See Appendix D.3 for details.

The support-restricted ridge estimator is then chosen by joint optimization over the support and ridge level:

$$(\mathsf{S}^\star, \lambda^\star) \in \operatorname*{arg\,min}_{\mathsf{S}\subseteq[4],\, \lambda\geq 0} \mathrm{MSE}_{\mathrm{strat}}\big(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda), 0\big).$$

The left, middle, and right panels show, respectively, the optimal support $\mathsf{S}^\star(\kappa, \delta)$, the corresponding optimal ridge level $\lambda^\star(\kappa, \delta)$, and the excess strategic MSE relative to $\widetilde{\mathrm{OPT}}$: $\mathrm{MSE}_{\mathrm{strat}}\big(\theta^{\mathrm{OR}}(\mathsf{S}^\star, \lambda^\star), 0\big) - \widetilde{\mathrm{OPT}}$.

Only three supports appear on the grid: $\{1, 2, 3, 4\}$, $\{2, 3, 4\}$, and $\{3, 4\}$. The transition from $\{1, 2, 3, 4\}$ to $\{2, 3, 4\}$ shows that once the manipulable pair becomes sufficiently heterogeneous, the optimizer drops the more manipulable coordinate while retaining the less manipulable one. The transition from $\{2, 3, 4\}$ to $\{3, 4\}$ shows that when average manipulability becomes large enough, the entire manipulable pair is excluded and only the less manipulable baseline features are retained.

D.3 Less manipulable correlated proxies can replace manipulable features

We consider the following $d = 4$ example in Figure 5:

$$\Sigma(\rho) = \begin{pmatrix} 1 & \rho & 0.1 & 0 \\ \rho & 1 & 0 & 0.1 \\ 0.1 & 0 & 1 & 0.2 \\ 0 & 0.1 & 0.2 & 1 \end{pmatrix}, \qquad \theta^* = (1.6, 0, 1.0, 0.9), \qquad H^{-1} = \mathrm{diag}(k_1, 0.6, 0.5, 0.5).$$

Feature 1 is manipulable and directly predictive, feature 2 is a correlated less manipulable proxy whose coefficient in θ^* is zero, and features 3 and 4 are less manipulable baseline features. We vary $\rho \in [0, 0.95]$ and $k_1 \in [0.5, 4.0]$ on a 61×61 grid.

For each grid point (ρ, k_1), we enumerate all 2^4 supports $\mathsf{S} \subseteq [4]$. For each support, we compute the exact support-restricted ridge optimum

$$\min_{\lambda\geq 0} \mathrm{MSE}_{\mathrm{strat}}\big(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda), 0\big).$$

We then record the minimizing support $\mathsf{S}^*(\rho, k_1)$ and ridge level $\lambda^*(\rho, k_1)$. The full-support zero-intercept oracle

$$\widetilde{\mathrm{OPT}} = \min_{\theta\in\mathbb{R}^4} \mathrm{MSE}_{\mathrm{strat}}(\theta, 0)$$

is computed using the generalized-ridge fixed-point equation (15) described above.

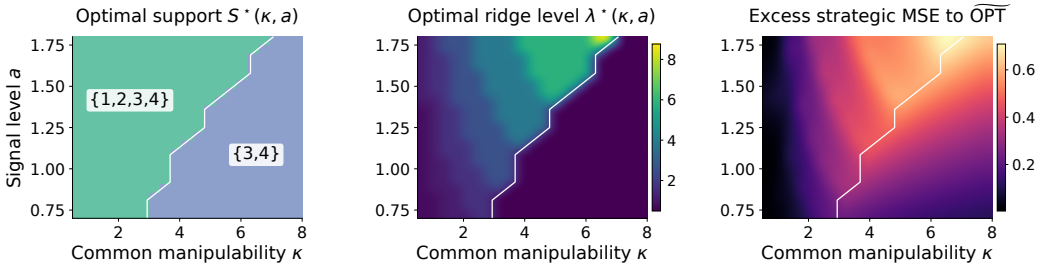

Figure 6: **A homogeneous manipulable pair can be retained and regularized when sufficiently predictive.** We compute the optimal support and regularization level for $d = 4$ features with diagonal costs. Features $\{1, 2\}$ are a homogeneous manipulable pair with common signal level a and common manipulability κ; features 3 and 4 have low manipulability. (Left) The optimal support $\mathsf{S}^*(\kappa, a)$ after joint optimization over S and λ. (Middle) The corresponding optimal ridge level $\lambda^*(\kappa, a)$, where white contour indicates the support boundary. (Right) Excess strategic MSE relative to $\widetilde{\mathrm{OPT}}$. As the manipulable pair becomes more predictive, it is retained rather than excluded; as manipulability increases, the optimal response is to increase regularization. The resulting support-restricted ridge estimator remains close to $\widetilde{\mathrm{OPT}}$ over a broad range of the parameter space.

The left, middle, and right panels show, respectively, the optimal support $\mathsf{S}^\star(\rho, k_1)$, the corresponding optimal ridge level $\lambda^\star(\rho, k_1)$, and the excess strategic MSE relative to $\widetilde{\mathrm{OPT}}$: $\mathrm{MSE}_{\mathrm{strat}}\big(\theta^{\mathrm{OR}}(\mathsf{S}^\star, \lambda^\star), 0\big) - \widetilde{\mathrm{OPT}}$.

The support map in the left panel shows that only two supports appear on the grid: $\{1, 2, 3, 4\}$ and $\{2, 3, 4\}$. The phase transition is therefore especially transparent: as the proxy becomes more correlated with the manipulable feature, and as that feature becomes more manipulable, the optimal design drops feature 1 while retaining feature 2. The middle and right panels show that this support transition is accompanied by a systematic change in the preferred regularization level, while the resulting support-restricted ridge estimator remains close to the oracle benchmark over a broad region of the parameter space.

D.4 Predictiveness versus manipulability with homogeneous costs

We consider the following $d = 4$ example in Figure 6:

$$\Sigma = I_4, \qquad \theta^* = (a, a, 1.0, 0.9), \qquad H^{-1} = \mathrm{diag}(\kappa, \kappa, 0.6, 0.6).$$

Features 1 and 2 form a homogeneous manipulable pair with common signal level a and common manipulability κ, while features 3 and 4 are less manipulable baseline features. We vary $a \in [0.7, 1.8]$ and $\kappa \in [0.5, 8]$ on a 21×21 grid.

For each grid point (a, κ), we compute the optimal support-restricted ridge choice $(\mathsf{S}^*, \lambda^*)$ and evaluate the resulting strategic MSE. To assess near-optimality, we compute the full-support zero-intercept oracle

$$\widetilde{\mathrm{OPT}} = \min_{\theta \in \mathbb{R}^4} \mathrm{MSE}_{\mathrm{strat}}(\theta, 0).$$

Because Σ and H^{-1} are diagonal here, the oracle can be computed from the scalar fixed-point equation (15):

$$q = \sum_{j=1}^{4} \frac{H_{jj}^{-1}(\theta_j^*)^2}{(1 + 2qH_{jj}^{-1})^2}, \qquad \theta_j^{\star,\mathrm{oracle}} = \frac{\theta_j^*}{1 + 2qH_{jj}^{-1}}.$$

The left, middle, and right panels show, respectively, the optimal support $\mathsf{S}^*(\kappa, a)$, the corresponding optimal ridge level $\lambda^*(\kappa, a)$, and the excess strategic MSE relative to $\widetilde{\mathrm{OPT}}$: $\mathrm{MSE}_{\mathrm{strat}}\big(\theta^{\mathrm{OR}}(\mathsf{S}^*, \lambda^*), 0\big) - \widetilde{\mathrm{OPT}}$.

Only two supports appear on the grid: the full support $\{1, 2, 3, 4\}$ and the baseline support $\{3, 4\}$. The support switch therefore cleanly illustrates the message: a homogeneous manipulable pair is retained when it is sufficiently predictive and otherwise excluded; conditional on retaining it, the preferred regularization strength increases with manipulability. The resulting support-restricted ridge

estimator remains close to $\widetilde{\mathrm{OPT}}$ over a broad region of the parameter space, showing that coarse levers are often near-optimal in this setting.

E Theoretical Details and Proofs for Section 5

As Appendix C, throughout this appendix, we use the shorthand

$$\mathrm{MSE}_{\mathrm{strat}}(\theta) = \mathrm{MSE}_{\mathrm{strat}}(\theta, 0) = (\theta - \theta^*)^\top \Sigma(\theta - \theta^*) + (\theta^\top H^{-1}\theta)^2 + \sigma^2.$$

For fixed $\lambda \geq 0$, we also write

$$\Phi_\lambda(\mathsf{S}) := \Phi(\mathsf{S}, \lambda) = \mathrm{MSE}_{\mathrm{strat}}\big(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda)\big).$$

Recall that for a support budget s, the exact fixed-λ support-restricted ridge optimum is

$$\Phi^\star_{\lambda,s} := \min_{\substack{\mathsf{S}\subseteq[d] \\ |\mathsf{S}|\leq s}} \Phi(\mathsf{S}, \lambda).$$

E.1 Proof of Proposition 1

Let $w \in \{0,1\}^d$ and define $\mathsf{S}(w) = \{i \in [d] : w_i = 1\}$. By the definition of the perspective function φ, if $w_i = 0$, then $\varphi(\theta_i, w_i) = +\infty$ unless $\theta_i = 0$. Thus every finite-valued feasible vector in the weighted ridge problem must satisfy $\mathrm{supp}(\theta) \subseteq \mathsf{S}(w)$. On coordinates $i \in \mathsf{S}(w)$, we have $w_i = 1$, so $\varphi(\theta_i, w_i) = \theta_i^2$. Therefore the weighted ridge problem becomes:

$$\operatorname*{arg\,min}_{\theta\in\mathbb{R}^d:\mathrm{supp}(\theta)\subseteq\mathsf{S}(w)} \left\{\mathbb{E}[(Y - \theta^\top X)^2] + \lambda\|\theta\|_2^2\right\}.$$

This is exactly the support-restricted ordinary ridge problem defining $\theta^{\mathrm{OR}}(\mathsf{S}(w), \lambda)$. Hence, $\theta_\lambda(w) = \theta^{\mathrm{OR}}(\mathsf{S}(w), \lambda)$. Evaluating the same exact strategic MSE on both sides gives $\mathrm{MSE}_{\mathrm{strat}}(\theta_\lambda(w)) = \Phi(\mathsf{S}(w), \lambda)$. Finally, the constraint $\mathbf{1}^\top w \leq s$ is equivalent to $|\mathsf{S}(w)| \leq s$ for binary w, so

$$\Phi^\star_{\lambda,s} = \min_{\substack{\mathsf{S}\subseteq[d] \\ |\mathsf{S}|\leq s}} \Phi(\mathsf{S}, \lambda) = \min_{\substack{w\in\{0,1\}^d \\ \mathbf{1}^\top w\leq s}} \mathrm{MSE}_{\mathrm{strat}}(\theta_\lambda(w)).$$

E.2 Weighted relaxation as a lower bound

For $\lambda > 0$, define the weight polytope $\mathcal{W}_s := \{w \in [0,1]^d : \mathbf{1}^\top w \leq s\}$. The weighted relaxation at fixed λ is $L^{\mathrm{wt}}_{\lambda,s} := \min_{w\in\mathcal{W}_s} \mathrm{MSE}_{\mathrm{strat}}(\theta_\lambda(w))$. Because the binary feasible set is contained in $\mathcal{W}_s$, this relaxation gives a lower bound on the exact fixed-λ support-restricted ridge optimum.

Proposition E.1 (Weighted relaxation lower bound)**.** *For every $\lambda > 0$ and support budget s, $L^{\mathrm{wt}}_{\lambda,s} \leq \Phi^\star_{\lambda,s}$.*

Proof. By Proposition 1, $\Phi^\star_{\lambda,s} = \min_{\substack{w\in\{0,1\}^d \\ \mathbf{1}^\top w\leq s}} \mathrm{MSE}_{\mathrm{strat}}(\theta_\lambda(w))$. Since $\{w \in \{0,1\}^d : \mathbf{1}^\top w \leq s\} \subseteq \mathcal{W}_s$, minimizing the same objective over $\mathcal{W}_s$ can only decrease the value:

$$L^{\mathrm{wt}}_{\lambda,s} = \min_{w\in\mathcal{W}_s} \mathrm{MSE}_{\mathrm{strat}}(\theta_\lambda(w)) \leq \Phi^\star_{\lambda,s}.$$

□

This also yields an a posteriori certificate whenever the weighted relaxation is solved globally, or whenever a certified lower bound on it is available.

Corollary E.1 (A posteriori certificate at fixed λ)**.** *Fix $\lambda > 0$ and suppose $\widehat{\mathsf{S}}_\lambda$ is any support with $|\widehat{\mathsf{S}}_\lambda| \leq s$. Then*

$$0 \leq \Phi(\widehat{\mathsf{S}}_\lambda, \lambda) - \Phi^\star_{\lambda,s} \leq \Phi(\widehat{\mathsf{S}}_\lambda, \lambda) - L^{\mathrm{wt}}_{\lambda,s}.$$

More generally, the same bound holds if $L^{\mathrm{wt}}_{\lambda,s}$ is replaced by any certified lower bound on $\Phi^\star_{\lambda,s}$.

Proof. Since $\Phi^\star_{\lambda,s}$ is the exact optimum over all supports of size at most s,

$$0 \le \Phi(\widehat{\mathsf{S}}_\lambda, \lambda) - \Phi^\star_{\lambda,s}.$$

By Proposition E.1, $L^{\mathrm{wt}}_{\lambda,s} \le \Phi^\star_{\lambda,s}$. Therefore

$$\Phi(\widehat{\mathsf{S}}_\lambda, \lambda) - \Phi^\star_{\lambda,s} \le \Phi(\widehat{\mathsf{S}}_\lambda, \lambda) - L^{\mathrm{wt}}_{\lambda,s}.$$

The same argument applies to any lower bound on $\Phi^\star_{\lambda,s}$. □

E.3 Local support refinement

The weighted relaxation produces a fractional vector, which must be converted into a discrete support. For $w \in [0,1]^d$, let $\mathrm{Top}_s(w)$ denote the indices of the s largest coordinates of w, with ties broken arbitrarily. After rounding, we refine the support using the exact fixed-λ objective $\Phi(\mathsf{S}, \lambda)$.

For a current support S with $|\mathsf{S}| \le s$, define the add/drop/swap neighborhood

$$\mathcal{N}_s(\mathsf{S}) := \big\{\mathsf{S} \cup \{j\} : j \notin \mathsf{S},\ |\mathsf{S}| < s\big\} \cup \big\{\mathsf{S} \setminus \{i\} : i \in \mathsf{S}\big\} \cup \big\{(\mathsf{S} \setminus \{i\}) \cup \{j\} : i \in \mathsf{S},\ j \notin \mathsf{S}\big\}.$$

Starting from an initial support $\mathsf{S}^{(0)}$, the refinement step repeatedly replaces the current support by the best strictly improving support in $\mathcal{N}_s(\mathsf{S}) \cup \{\mathsf{S}\}$.

Proposition E.2 (Termination and monotonicity of local refinement)**.** *Fix $\lambda \ge 0$ and support budget s. Starting from any support $\mathsf{S}^{(0)}$ with $|\mathsf{S}^{(0)}| \le s$, the add/drop/swap local refinement procedure terminates after finitely many accepted moves. If $\widehat{\mathsf{S}}_\lambda$ is its final output, then $\Phi(\widehat{\mathsf{S}}_\lambda, \lambda) \le \Phi(\mathsf{S}^{(0)}, \lambda)$. Moreover, $\widehat{\mathsf{S}}_\lambda$ is locally optimal with respect to $\mathcal{N}_s(\cdot)$: $\Phi(\widehat{\mathsf{S}}_\lambda, \lambda) \le \Phi(\mathsf{T}, \lambda)$ for every $\mathsf{T} \in \mathcal{N}_s(\widehat{\mathsf{S}}_\lambda)$.*

Proof. Each accepted move replaces the current support S by a support T satisfying $\Phi(\mathsf{T}, \lambda) < \Phi(\mathsf{S}, \lambda)$. Thus the exact fixed-$\lambda$ objective strictly decreases at every accepted move. Because there are only finitely many supports of size at most s, the procedure must terminate after finitely many accepted moves.

The monotonicity statement follows by chaining the strict improvements along the accepted moves: $\Phi(\widehat{\mathsf{S}}_\lambda, \lambda) \le \Phi(\mathsf{S}^{(0)}, \lambda)$. At termination, there is no neighbor $\mathsf{T} \in \mathcal{N}_s(\widehat{\mathsf{S}}_\lambda)$ with $\Phi(\mathsf{T}, \lambda) < \Phi(\widehat{\mathsf{S}}_\lambda, \lambda)$, which is precisely the stated local optimality condition. □

E.4 A convex reformulation under diagonal covariance

The weighted relaxation is generally nonconvex. However, it becomes convex after a change of variables when the feature covariance is diagonal. Assume throughout this section that the feature covariance is diagonal: $\Sigma = \mathrm{diag}(v_1, \dots, v_d)$ with $v_i > 0$ for all $i \in [d]$. Write $\theta^* = (\theta^*_1, \dots, \theta^*_d)$. Fix $\lambda > 0$ and recall that $\mathcal{W}_s = \{w \in [0,1]^d :\ \mathbf{1}^\top w \le s\}$.

Under diagonal covariance, the weighted ridge coefficients decouple coordinatewise. Indeed, for any $w \in \mathcal{W}_s$,

$$\theta_{\lambda,i}(w) = \frac{v_i w_i}{v_i w_i + \lambda}\, \theta^*_i, \qquad i = 1, \dots, d. \tag{16}$$

This motivates the reparameterization

$$r_i := \frac{v_i w_i}{v_i w_i + \lambda}, \qquad i = 1, \dots, d. \tag{17}$$

Let $r = (r_1, \dots, r_d)$. The variables r_i have a simple interpretation. They are exactly the fraction of the original signal coefficients that survive after weighted shrinkage, since (16) becomes $\theta_{\lambda,i}(w) = \theta^*_i\, r_i$.

The change of variables in (17) is one-to-one for $\lambda > 0$, with inverse

$$w_i = \frac{\lambda r_i}{v_i(1 - r_i)}, \qquad 0 \le r_i \le \frac{v_i}{v_i + \lambda}. \tag{18}$$

Under this reparameterization, the support-budget constraint $\mathbf{1}^\top w \le s$ becomes

$$\sum_{i=1}^{d} \frac{\lambda r_i}{v_i(1 - r_i)} \le s. \tag{19}$$

The next proposition gives the exact weighted relaxation in these variables:

Proposition E.3 (Convex reformulation under diagonal covariance)**.** *Under the diagonal covariance setup, the weighted relaxation is exactly equivalent to the optimization problem:*

$$\begin{aligned} \min_{r\in\mathbb{R}^d} \quad & \sum_{i=1}^{d} v_i(\theta_i^*)^2(1-r_i)^2 + \Big((\theta^*\odot r)^\top H^{-1}(\theta^*\odot r)\Big)^2 \\ \text{subject to} \quad & 0\le r_i\le \frac{v_i}{v_i+\lambda}, \qquad i=1,\dots,d, \\ & \sum_{i=1}^{d}\frac{\lambda r_i}{v_i(1-r_i)}\le s, \end{aligned} \tag{20}$$

where $\odot$ denotes coordinatewise multiplication. Moreover, (20) *is a convex optimization problem.*

Proposition E.3 is especially useful because it turns the inner weighted problem into a globally solvable convex program. The first term in the objective is the prediction loss from shrinking the retained fraction of each signal coordinate away from its full value. The second term is the strategic burden induced by the retained signal. The budget constraint in (19) plays the role of a soft support-size constraint.

When H is also diagonal, say $H=\mathrm{diag}(h_1,\dots,h_d)$ with $h_i>0$ for $i\in[d]$, the strategic term simplifies further to $(\sum_{i=1}^d h_i^{-1}(\theta_i^*)^2 r_i^2)^2$. This is the form that appears in our exact diagonal examples. It makes especially transparent how weighted shrinkage trades off predictive signal against exposure to manipulation.

There is also an exact discrete simplification in the diagonal case. When both Σ and H^{-1} are diagonal, the fixed-λ binary support-selection objective can be written as a modular predictive benefit minus a quadratic total-exposure penalty. The resulting gain function is submodular, though not necessarily monotone; see Propositions E.4 and E.5. This explains why greedy local search is a natural baseline, while also motivating the exact local-refinement step in Algorithm 1.

Proof. Before we begin, note that the statement in the main text omits the additive constant σ^2, which does not affect the minimizer.

Assume $\Sigma=\mathrm{diag}(v_1,\dots,v_d)$ with $v_i>0$. For $w\in\mathcal{W}_s$, the weighted ridge problem decouples coordinatewise:

$$\theta_{\lambda,i}(w)=\frac{v_i w_i}{v_i w_i+\lambda}\theta_i^*.$$

Define

$$r_i=\frac{v_i w_i}{v_i w_i+\lambda}.$$

Then $\theta_{\lambda,i}(w)=\theta_i^* r_i$. Since $w_i\in[0,1]$, we have

$$0\le r_i\le\frac{v_i}{v_i+\lambda}.$$

For $\lambda>0$, the map $w_i\mapsto r_i$ is one-to-one, with inverse

$$w_i=\frac{\lambda r_i}{v_i(1-r_i)}.$$

Thus the budget constraint $\mathbf{1}^\top w\le s$ is equivalent to

$$\sum_{i=1}^{d}\frac{\lambda r_i}{v_i(1-r_i)}\le s.$$

Now, we compute the strategic MSE. Since $\theta_\lambda(w)=\theta^*\odot r$, we have

$$(\theta_\lambda(w)-\theta^*)^\top\Sigma(\theta_\lambda(w)-\theta^*)=\sum_{i=1}^{d}v_i(\theta_i^*)^2(1-r_i)^2.$$

Also,

$$\theta_\lambda(w)^\top H^{-1}\theta_\lambda(w)=(\theta^*\odot r)^\top H^{-1}(\theta^*\odot r).$$

Therefore

$$\mathrm{MSE}_{\mathrm{strat}}(\theta_\lambda(w)) = \sigma^2 + \sum_{i=1}^{d} v_i(\theta_i^*)^2(1-r_i)^2 + \left((\theta^* \odot r)^\top H^{-1}(\theta^* \odot r)\right)^2.$$

Dropping the additive constant σ^2 gives the stated optimization problem.

It remains to prove convexity. The first term $\sum_{i=1}^{d} v_i(\theta_i^*)^2(1-r_i)^2$ is convex in r. The map $r \mapsto (\theta^* \odot r)^\top H^{-1}(\theta^* \odot r)$ is a nonnegative convex quadratic function because $H^{-1} \succ 0$. Hence the composition $r \mapsto \left((\theta^* \odot r)^\top H^{-1}(\theta^* \odot r)\right)^2$ is convex.

The box constraints are convex. Finally, for each i, the function

$$r_i \mapsto \frac{\lambda r_i}{v_i(1-r_i)}$$

is convex on $0 \le r_i < 1$, since

$$\frac{d^2}{dr_i^2}\frac{\lambda r_i}{v_i(1-r_i)} = \frac{2\lambda}{v_i(1-r_i)^3} > 0.$$

Because $r_i \le v_i/(v_i+\lambda) < 1$, the budget constraint is convex. Thus the entire problem is a convex optimization problem. □

E.5 Exact strategic MSE of support-restricted ridge

We record here an explicit expression for the exact strategic MSE of a support-restricted ridge estimator. This expression is useful for implementation and for exact local search.

Lemma E.1 (Exact strategic MSE on a support)**.** *Fix* $\mathsf{S} \subseteq [d]$ *and let* $\mathsf{R} = [d] \setminus \mathsf{S}$. *Let*

$$\theta_{\mathsf{S}}^{\mathrm{LS}}(\mathsf{S}) = \Sigma_{\mathsf{SS}}^{-1}\Sigma_{\mathsf{S}y}, \qquad \theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda) = (\Sigma_{\mathsf{SS}} + \lambda I)^{-1}\Sigma_{\mathsf{S}y}.$$

Then

$$\Phi(\mathsf{S},\lambda) = \sigma^2 + L_{\mathrm{pred}}(\mathsf{S}) + \lambda^2 \theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda)^\top \Sigma_{\mathsf{SS}}^{-1}\theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda) + \left(\theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda)^\top H_{\mathsf{SS}}^{-1}\theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda)\right)^2.$$

Proof. By Lemma C.1, for every $\theta = (\eta, 0_{\mathsf{R}}) \in \Theta(\mathsf{S})$,

$$(\theta-\theta^*)^\top \Sigma (\theta-\theta^*) = L_{\mathrm{pred}}(\mathsf{S}) + (\eta - \theta_{\mathsf{S}}^{\mathrm{LS}}(\mathsf{S}))^\top \Sigma_{\mathsf{SS}}(\eta - \theta_{\mathsf{S}}^{\mathrm{LS}}(\mathsf{S})).$$

Substituting $\eta = \theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda)$ gives

$$\Phi(\mathsf{S},\lambda) = \sigma^2 + L_{\mathrm{pred}}(\mathsf{S}) + (\theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda) - \theta_{\mathsf{S}}^{\mathrm{LS}}(\mathsf{S}))^\top \Sigma_{\mathsf{SS}}(\theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda) - \theta_{\mathsf{S}}^{\mathrm{LS}}(\mathsf{S})) + \left(\theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda)^\top H_{\mathsf{SS}}^{-1}\theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda)\right)^2.$$

The ridge normal equation is

$$(\Sigma_{\mathsf{SS}} + \lambda I)\theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda) = \Sigma_{\mathsf{SS}}\theta_{\mathsf{S}}^{\mathrm{LS}}(\mathsf{S}).$$

Equivalently,

$$\theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda) - \theta_{\mathsf{S}}^{\mathrm{LS}}(\mathsf{S}) = -\lambda\Sigma_{\mathsf{SS}}^{-1}\theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda).$$

Therefore

$$(\theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda) - \theta_{\mathsf{S}}^{\mathrm{LS}}(\mathsf{S}))^\top \Sigma_{\mathsf{SS}}(\theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda) - \theta_{\mathsf{S}}^{\mathrm{LS}}(\mathsf{S})) = \lambda^2 \theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda)^\top \Sigma_{\mathsf{SS}}^{-1}\theta_{\mathsf{S}}^{\mathrm{OR}}(\lambda).$$

This proves the claim. □

E.6 Closed form for the weighted ridge

Lemma E.2 (Closed form for the weighted ridge estimator)**.** *Suppose* $w_i > 0$ *for all* $i \in [d]$ *and let* $W = \mathrm{diag}(w)$. *Then* $\theta_\lambda(w) = (\Sigma + \lambda W^{-1})^{-1}\Sigma\theta^*$.

Proof. The weighted ridge objective is $\mathbb{E}[(Y - \theta^\top X)^2] + \lambda\theta^\top W^{-1}\theta$. Using $Y = \theta^{*\top}X + \varepsilon$ and $\mathbb{E}[X\varepsilon] = 0$, the first-order condition is $2\Sigma(\theta - \theta^*) + 2\lambda W^{-1}\theta = 0$. Thus $(\Sigma + \lambda W^{-1})\theta = \Sigma\theta^*$. Since $\Sigma \succ 0$ and $\lambda W^{-1} \succ 0$, the matrix $\Sigma + \lambda W^{-1}$ is positive definite, and hence $\theta_\lambda(w) = (\Sigma + \lambda W^{-1})^{-1}\Sigma\theta^*$. □

E.7 Exact set function in the diagonal case

Assume

$$\Sigma = \mathrm{diag}(v_1, \dots, v_d), \qquad H^{-1} = \mathrm{diag}(\omega_1, \dots, \omega_d), \qquad \theta^* = (\theta_1^*, \dots, \theta_d^*),$$

where $v_i > 0$ and $\omega_i > 0$. Fix $\lambda \geq 0$. For a subset $\mathsf{S} \subseteq [d]$,

$$\theta_i^{\mathrm{OR}}(\mathsf{S}, \lambda) = \begin{cases} \dfrac{v_i}{v_i + \lambda} \theta_i^*, & i \in \mathsf{S}, \\ 0, & i \notin \mathsf{S}. \end{cases}$$

Define

$$b_i(\lambda) := \frac{v_i^2 (v_i + 2\lambda)}{(v_i + \lambda)^2} (\theta_i^*)^2, \qquad c_i(\lambda) := \omega_i \left(\frac{v_i}{v_i + \lambda} \right)^2 (\theta_i^*)^2.$$

Then define

$$G_\lambda(\mathsf{S}) := \sum_{i \in \mathsf{S}} b_i(\lambda) - \left(\sum_{i \in \mathsf{S}} c_i(\lambda) \right)^2.$$

Proposition E.4 (Exact diagonal fixed-λ objective)**.** *Under the diagonal setup above,*

$$\Phi_\lambda(\mathsf{S}) = \Phi_\lambda(\varnothing) - G_\lambda(\mathsf{S}),$$

where

$$\Phi_\lambda(\varnothing) = \sigma^2 + \sum_{i=1}^{d} v_i (\theta_i^*)^2.$$

Consequently,

$$\operatorname*{arg\,min}_{\substack{\mathsf{S} \subseteq [d] \\ |\mathsf{S}| \leq s}} \Phi_\lambda(\mathsf{S}) = \operatorname*{arg\,max}_{\substack{\mathsf{S} \subseteq [d] \\ |\mathsf{S}| \leq s}} G_\lambda(\mathsf{S}).$$

Proof. Let

$$r_i(\lambda) = \frac{v_i}{v_i + \lambda}.$$

For $i \in \mathsf{S}$, $\theta_i^{\mathrm{OR}}(\mathsf{S}, \lambda) = r_i(\lambda) \theta_i^*$, and for $i \notin \mathsf{S}$, $\theta_i^{\mathrm{OR}}(\mathsf{S}, \lambda) = 0$. The prediction component of the strategic MSE is

$$\sum_{i \notin \mathsf{S}} v_i (\theta_i^*)^2 + \sum_{i \in \mathsf{S}} v_i (\theta_i^*)^2 (1 - r_i(\lambda))^2.$$

Equivalently, relative to the empty support, retaining coordinate i gives prediction gain

$$v_i (\theta_i^*)^2 - v_i (\theta_i^*)^2 (1 - r_i(\lambda))^2 = v_i (\theta_i^*)^2 \left(2 r_i(\lambda) - r_i(\lambda)^2 \right).$$

Since

$$2 r_i(\lambda) - r_i(\lambda)^2 = \frac{v_i (v_i + 2\lambda)}{(v_i + \lambda)^2},$$

this gain is exactly $b_i(\lambda)$.

The strategic term is

$$\left(\sum_{i \in \mathsf{S}} \omega_i r_i(\lambda)^2 (\theta_i^*)^2 \right)^2 = \left(\sum_{i \in \mathsf{S}} c_i(\lambda) \right)^2.$$

Thus

$$\Phi_\lambda(\mathsf{S}) = \Phi_\lambda(\varnothing) - \sum_{i \in \mathsf{S}} b_i(\lambda) + \left(\sum_{i \in \mathsf{S}} c_i(\lambda) \right)^2,$$

which is equivalent to the stated identity. □

Proposition E.5 (Submodularity of the diagonal gain function)**.** *For fixed $\lambda \geq 0$, the set function G_λ is submodular.*

Proof. For $j \notin \mathsf{S}$, the marginal gain from adding j is

$$\begin{aligned}\Delta_j^\lambda(\mathsf{S}) &:= G_\lambda(\mathsf{S} \cup \{j\}) - G_\lambda(\mathsf{S}) \\ &= b_j(\lambda) - \left(\sum_{i\in\mathsf{S}} c_i(\lambda) + c_j(\lambda)\right)^2 + \left(\sum_{i\in\mathsf{S}} c_i(\lambda)\right)^2 \\ &= b_j(\lambda) - 2c_j(\lambda)\sum_{i\in\mathsf{S}} c_i(\lambda) - c_j(\lambda)^2.\end{aligned}$$

Now let $\mathsf{S} \subseteq \mathsf{T}$ and $j \notin \mathsf{T}$. Since $c_i(\lambda) \geq 0$ for all i,

$$\sum_{i\in\mathsf{S}} c_i(\lambda) \leq \sum_{i\in\mathsf{T}} c_i(\lambda).$$

Therefore $\Delta_j^\lambda(\mathsf{S}) \geq \Delta_j^\lambda(\mathsf{T})$. This is exactly the diminishing-returns characterization of submodularity. □

E.8 Exact bilinear formulation and McCormick formulation

The weighted problem can also be written as an exact nonlinear constrained program. This formulation is useful for deriving convex lower bounds.

Fix $\lambda > 0$ and define $r := \Sigma(\theta^* - \theta)$. For interior weights, the weighted ridge normal equation is $(\Sigma + \lambda W^{-1})\theta = \Sigma\theta^*$. Equivalently,

$$r_i = \lambda\frac{\theta_i}{w_i}, \qquad i = 1, \ldots, d,$$

or

$$\lambda\theta_i = w_i r_i, \qquad i = 1, \ldots, d.$$

Proposition E.6 (Exact bilinear formulation)**.** *The weighted relaxation is equivalent to*

$$\begin{aligned}\min_{\theta,r,w} \quad & (\theta - \theta^*)^\top\Sigma(\theta - \theta^*) + (\theta^\top H^{-1}\theta)^2 + \sigma^2 \\ \text{s.t.} \quad & r = \Sigma(\theta^* - \theta), \\ & \lambda\theta_i = w_i r_i, \qquad i = 1, \ldots, d, \\ & 0 \leq w_i \leq 1, \qquad i = 1, \ldots, d, \\ & \mathbf{1}^\top w \leq s.\end{aligned} \tag{21}$$

Proof. For any w with strictly positive coordinates, the constraints $r = \Sigma(\theta^* - \theta)$ and $\lambda\theta_i = w_i r_i$ for $i = 1, \ldots, d$, are exactly the weighted ridge normal equations. Therefore the feasible θ associated with w is precisely $\theta_\lambda(w)$. The objective is exactly $\mathrm{MSE}_{\mathrm{strat}}(\theta)$. At boundary points with some $w_i = 0$, the equality $\lambda\theta_i = w_i r_i$ forces $\theta_i = 0$, which matches the perspective convention in the weighted ridge definition. Thus the formulation remains exact on the boundary as well. □

We next describe a convex lower-bound relaxation. Let $B_0 := \theta^{*\top}\Sigma\theta^*$. Since $w = 0$ is feasible and gives $\theta = 0$, any minimizer of the weighted relaxation has prediction error at most B_0:

$$(\theta - \theta^*)^\top\Sigma(\theta - \theta^*) \leq B_0.$$

Therefore, for each coordinate,

$$|r_i| = |e_i^\top\Sigma(\theta^* - \theta)| \leq \sqrt{\Sigma_{ii}B_0} =: R_i.$$

The bound follows from Cauchy's inequality:

$$|e_i^\top\Sigma(\theta^* - \theta)| = |(\Sigma^{1/2}e_i)^\top\Sigma^{1/2}(\theta^* - \theta)| \leq \sqrt{\Sigma_{ii}}\sqrt{(\theta - \theta^*)^\top\Sigma(\theta - \theta^*)}.$$

For $0 \leq w_i \leq 1$ and $-R_i \leq r_i \leq R_i$, the convex hull relaxation of $\lambda\theta_i = w_i r_i$ is given by the McCormick inequalities:

$$\lambda\theta_i \geq -R_i w_i, \tag{22}$$

$$\lambda\theta_i \leq R_i w_i, \tag{23}$$

$$\lambda\theta_i \geq r_i - R_i(1 - w_i), \tag{24}$$

$$\lambda\theta_i \leq r_i + R_i(1 - w_i). \tag{25}$$

E.9 Our proposed algorithm

Algorithm 1 Support-restricted ridge design by weighted relaxation and local refinement

Input: Support budget s and ridge grid $\Lambda \subset (0, \infty)$.

for each $\lambda \in \Lambda$ **do**
  Solve or approximately solve the weighted relaxation: $\widehat{w}_\lambda \approx \arg\min_{w \in \mathcal{W}_s} \mathrm{MSE}_{\mathrm{strat}}(\theta_\lambda(w), 0)$.
  Initialize the support by rounding: $\mathsf{S} \leftarrow \mathrm{Top}_s(\widehat{w}_\lambda)$.
  while there exists $\mathsf{T} \in \mathcal{N}_s(\mathsf{S})$ with $\Phi(\mathsf{T}, \lambda) < \Phi(\mathsf{S}, \lambda)$ **do**
    Update to the best improving neighbor: $\mathsf{S} \leftarrow \arg\min_{\mathsf{T} \in \mathcal{N}_s(\mathsf{S}) \cup \{\mathsf{S}\}} \Phi(\mathsf{T}, \lambda)$.
  end while
  Set $\widehat{\mathsf{S}}_\lambda \leftarrow \mathsf{S}$ and refit exact support-restricted ridge: $\widehat{\theta}_\lambda := \theta^{\mathrm{OR}}(\widehat{\mathsf{S}}_\lambda, \lambda)$.
  Evaluate $\widehat{R}_\lambda := \mathrm{MSE}_{\mathrm{strat}}(\widehat{\theta}_\lambda, 0)$.
end for
Choose $\widehat{\lambda} \in \arg\min_{\lambda \in \Lambda} \widehat{R}_\lambda$.

Output: $(\widehat{\lambda}, \widehat{\mathsf{S}}_{\widehat{\lambda}}, \widehat{\theta}_{\widehat{\lambda}})$.

To reiterate a point mentioned in the main paper, the weighted relaxation is used only to initialize the support search. The output of Algorithm 1 is always a standard support-restricted ridge estimator of the form $\theta^{\mathrm{OR}}(\widehat{\mathsf{S}}_{\widehat{\lambda}}, \widehat{\lambda})$. Thus, the algorithm always deploys an estimator from exactly the class analyzed in Section 3.

F Algorithmic Details and Illustrations for Section 5

The synthetic benchmarks in this section validate the algorithmic pipeline from Section 5 in settings where a support-constrained oracle can be computed exactly. The main question behind these benchmarks is whether, given population objects (Σ, θ^*, H) and a support budget s, weighted relaxation followed by rounding and exact local refinement recovers the best support-restricted ridge rule. Because the benchmark is low-dimensional, we can enumerate all supports and directly compare every method to the exact grid oracle.

Problem size and oracle. All instances use

$$d = 18, \qquad s = 5, \qquad \Lambda = \{0.05,\ 0.20,\ 0.50,\ 1.00,\ 2.00\}.$$

For each instance, the exact grid oracle is

$$(\mathsf{S}^\star, \lambda^\star) \in \arg\min_{\substack{\mathsf{S} \subseteq [d] \\ |\mathsf{S}| \le s}} \min_{\lambda \in \Lambda} \Phi(\mathsf{S}, \lambda), \qquad \Phi(\mathsf{S}, \lambda) = \mathrm{MSE}_{\mathrm{strat}}\big(\theta^{\mathrm{OR}}(\mathsf{S}, \lambda)\big).$$

The oracle is computed by exhaustive search over all supports of size at most s and all ridge levels over the grid Λ. All reported values are population strategic MSE under the synthetic (Σ, θ^*, H). We suppress the additive noise variance σ^2, since it is constant across methods for a fixed instance.

Synthetic regimes. We consider three covariance regimes and two signal–cost regimes. The covariance regimes are: identity covariance where $\Sigma = I_d$; AR(1) covariance where $\Sigma_{ij} = \rho^{|i-j|}$ with $\rho = 0.65$; and block-correlated covariance where the 18 coordinates are split into three blocks of size 6, with pairwise correlation 0.80 within each block and 0.05 across blocks, and diagonal entries normalized to one. Throughout, the cost matrix H is diagonal, and we generate inverse costs H_{jj}^{-1} directly.

We consider two regimes for the joint distribution of signal and manipulation costs. In the independent regime, signal and inverse costs are drawn independently as: $\theta_j^* = |\mathcal{N}(0.9, 0.45^2)| + 0.05$ with $H_{jj}^{-1} = \exp(\mathcal{N}(0, 0.8^2))$. In the tradeoff regime, higher-signal coordinates also tend to be more manipulable. Specifically, we draw $z_j, \varepsilon_j, \eta_j \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, 1)$ and set $\theta_j^* = |0.95 + 0.60 z_j + 0.20 \varepsilon_j| + 0.05$, with $H_{jj}^{-1} = \exp(0.95 z_j + 0.35 \eta_j)$. For each of the 3×2 regime combinations, we generate 5 independent replications, giving 30 total benchmark instances.

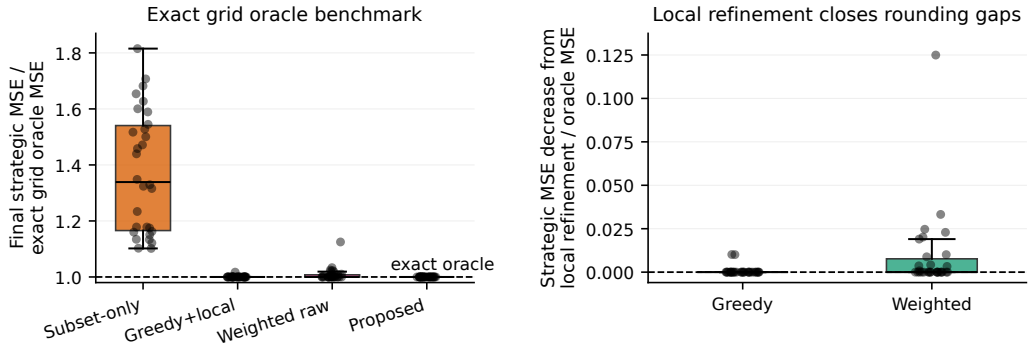

Figure 7: **Low-dimensional synthetic benchmarks validate our proposed weighted screen-and-refit method.** (Left) Final strategic MSE normalized by the exact grid oracle MSE on the same instance, $\mathrm{MSE}_{\mathrm{strat}}(\widehat{\theta})/\mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S}^\star, \lambda^\star))$. A value of 1 means that the method attains the exact grid oracle. The proposed weighted screen-and-refit method attains the oracle on all 30 instances; greedy with local refinement attains it on 29 of 30; weighted rounding alone attains it on 18 of 30 and can have a small rounding gap; subset selection without regularization remains consistently farther from the oracle. (Right) Decrease in strategic MSE produced by local refinement, normalized by the exact grid oracle MSE. Local refinement rarely changes the greedy solution, but it repairs several rounded weighted supports. This motivates its inclusion in our default algorithm.

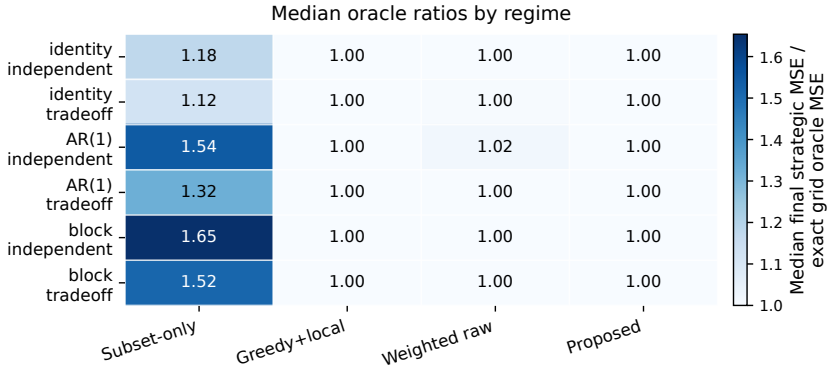

Figure 8: **Median oracle ratios by synthetic regime.** Each entry reports the median final strategic MSE normalized by the exact grid oracle MSE, $\mathrm{MSE}_{\mathrm{strat}}(\widehat{\theta})/\mathrm{MSE}_{\mathrm{strat}}(\theta^{\mathrm{OR}}(\mathsf{S}^\star, \lambda^\star))$, across the five replications in the corresponding regime. The proposed weighted screen-and-refit method attains median ratio 1.00 in every regime. The subset-only baseline is farthest from the oracle, especially in the correlated regimes. This illustrates the value of combining support selection with regularization stressed throughout the paper.

Methods. We compare four support-budget-feasible methods: (1) Best subset without regularization, which exhaustively searches over supports of size at most s at $\lambda = 0$. This baseline isolates the value of support selection alone, without ridge shrinkage. (2) Greedy plus local refinement, in which for each fixed $\lambda \in \Lambda$, we start from the empty support and repeatedly add the coordinate whose inclusion gives the largest decrease in the exact fixed-λ objective $\Phi(\mathsf{S}, \lambda)$, until the support budget is reached. The resulting greedy support is then refined by the same add/drop/swap local search used in our method (see Algorithm 1 for more details), and the best refined candidate is chosen over $\lambda \in \Lambda$. (3) Weighted rounding only, in which we solve the weighted relaxation over $w \in [0,1]^d$ with $\mathbf{1}^\top w \le s$, round to the top-s coordinates of the resulting weight vector, refit exact support-restricted ridge on that rounded support, and then choose the best λ. (4) Weighted screen-and-refit, which starts from the same weighted rounded support but then applies exact add/drop/swap local refinement before the final support-restricted ridge refit.

The local-refinement neighborhood contains all one-step additions, deletions, and swaps that respect the support budget. The refinement procedure repeatedly accepts the best improving move under the exact fixed-λ objective $\Phi(\mathsf{S}, \lambda)$ until no improving move remains. This step can only improve the rounded support and terminates after finitely many moves by Proposition E.2. We do not include

full-support ridge in the oracle-ratio figures because it violates the support budget and therefore is not a feasible method for this support-constrained benchmark.

Results. Figure 7 shows that support selection and regularization must be tuned jointly: the exact subset-only baseline at $\lambda = 0$ has median normalized MSE 1.34 and maximum normalized MSE 1.82, relative to the exact grid oracle. Figure 8 further shows that the subset-only baseline remains farthest from the oracle across covariance and signal–cost regimes. Weighted relaxation is useful as a screening device, but the rounded support should be refined. Weighted rounding alone attains the exact grid oracle on 18 out of 30 instances and has maximum normalized MSE 1.125, whereas the weighted screen-and-refit pipeline after exact local refinement attains the exact grid oracle on all 30 instances. Local refinement improves the rounded weighted solution on 12 of 30 instances, with maximum normalized MSE decrease 0.125.

These synthetic benchmarks motivate the final algorithmic steps used in the experiments: weighted screening to obtain a strong support initializer, exact local refinement to remove rounding artifacts, and exact support-restricted ridge refitting on the refined support. Greedy local search is a strong discrete baseline in these low-dimensional problems, but the weighted relaxation provides the more general methodology: it directly implements the continuous relaxation introduced in Section 5 and admits the convex reformulation for the diagonal-covariance case in Proposition E.3.

G Experimental Details and Illustrations for Section 6

G.1 Data description

The `upcoding` package[2] simulates realistic co-occurring HCCs for American adults aged 65 years and older (i.e., Medicare eligible) based on self-reported health conditions from All of Us survey data [1]. Because these self-reported conditions are unaffected by coding incentives, the resulting baseline data provide an unmanipulated representation of patients' underlying health conditions. The package allows users to modify the baseline data by specifying either upcoding or undercoding of existing baseline diagnoses. We use the package to construct realistic baseline and manipulated CMS-HCC Version 28 (V28) feature vectors. The resulting baseline data contain $n = 5000$ beneficiaries and 115 binary V28 HCC indicators. The HCC features are denoted `HCC1`, `HCC2`, ... and correspond to CMS-HCC condition categories. For example, HCC238 corresponds to specified heart arrhythmias, while HCCs 125–127 correspond to severe, moderate, and mild or unspecified dementia, respectively.

The upcoding targets are the ten V28 diagnosis groups identified in recent work [20] as the groups that contributed the most to differential coding between traditional Medicare (TM) and Medicare Advantage (MA): chronic obstructive pulmonary disease (COPD), acute or chronic heart failure, diabetes, morbid obesity, drug use disorder, chronic kidney disease, major depression, rheumatoid arthritis, hemiplegia/hemiparesis, and alcohol use disorder. These groups are reported to account for almost all of the V28 MA–TM measured-risk difference in 2021, while the remaining 100 HCCs contribute little aggregate difference. Related work [21] also shows that differential persistence and new incidence in the top ten diagnosis groups explain most of the MA–TM risk-score gap, with little contribution from the remaining HCCs.

The ten diagnosis groups identified by [20] as most intensely coded in MA (which we refer to as the top-ten diagnosis groups throughout) are used to calibrate the direction and magnitude of the simulated coding shifts. For each targeted group g, the desired post-upcoding prevalence is constructed as

$$p_{g,m}^{\mathrm{target}} = p_g^{(0)} + m\,(p_g^{\mathrm{MA}} - p_g^{\mathrm{TM}}),$$

where $p_g^{(0)}$ is the baseline simulator prevalence, p_g^{MA} and p_g^{TM} are the Medicare Advantage and traditional Medicare prevalences reported in Exhibit 2 of Kronick et al. [20], and $m \in \{0.5, 0.75, 1.0\}$ indexes the low, main, and high upcoding scenarios. The simulator converts these target prevalences into upcoding proportions, clips them to $[0, 1]$, and uses one of two upcoding approaches, any-available or severity-based, depending on whether the targeted group has lower-severity conditions within the same clinical family. Under any-available upcoding, any beneficiary not already coded for a target HCC is eligible to be upcoded into that HCC; under severity-based upcoding, only beneficiaries already coded for a lower-severity related HCC are eligible. Applying the CMS-HCC V28 hierarchy

[2] `https://github.com/StanfordHPDS/upcoding`

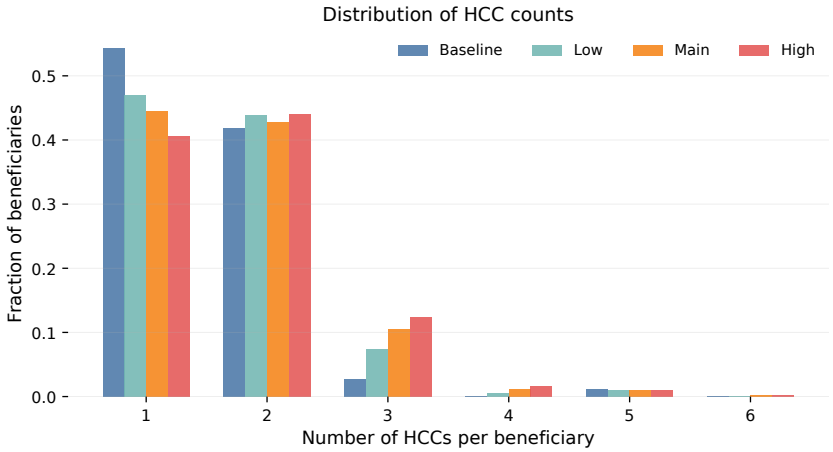

Figure 9: **Distribution of HCC counts per beneficiary.** The figure shows the number of V28 HCCs per beneficiary in the baseline data and in the low, main, and high upcoding scenarios. Only HCCs with nonzero baseline prevalence are included. The baseline is sparse, with mean 1.52 HCCs, median 1, and maximum 5. The upcoding scenarios increase coding intensity modestly.

then sets lower-severity HCC indicators to zero whenever a higher-severity HCC is present for a beneficiary. The three scenarios are generated independently.

A target-mapping table is constructed as part of this upcoding calibration step. Each row corresponds to a diagnosis group from the top-ten diagnosis groups and records: (i) the target HCC whose prevalence is targeted in the simulator, (ii) the V28 HCCs belonging to the same diagnosis group, and (iii) the group-level calibration score r_g, which is the group-level contribution to the MA–TM risk-score gap. The simulator then uses this table, together with the baseline HCC matrix, to generate the low, main, and high upcoding scenarios.

We treat the baseline covariates as the pre-deployment feature distribution. The manipulated covariates are used to calibrate and validate the coding shifts. We use the full baseline cohort of $n = 5000$ beneficiaries to compute the empirical mean and covariance of the HCCs with nonzero baseline prevalence, construct the semi-synthetic signal vector and manipulation-ease matrix, and evaluate each method using the closed-form post-manipulation MSE described in Section G.3.

G.2 Exploratory data analysis and calibration diagnostics

The baseline is sparse. In the baseline data, the mean number of HCCs per beneficiary is 1.52, the median is 1, and the maximum is 5. In the high-upcoding scenario, the corresponding mean is 1.79. Figure 9 reports the beneficiary-level distribution of these HCC counts across the baseline and calibrated upcoding scenarios.

For each target HCC, the achieved prevalence after applying the V28 hierarchy closely matches the intended prevalence. Across the low, main, and high scenarios, the mean absolute target-prevalence errors are 0.0002, 0.0001, and 0.0003, respectively, and the largest absolute error over all targets and scenarios is 0.0010. The left panel of Figure 10 compares baseline, target, and achieved prevalences, while the right panel reports the realized prevalence shifts across scenarios.

G.3 Semi-synthetic outcome, strategic response, and metrics

Let $X_i^{(0)} \in \{0,1\}^{115}$ denote the baseline HCC vector for beneficiary i. We restrict the benchmark optimization to the 30 HCCs with nonzero baseline prevalence and center these features using the empirical baseline mean. Writing the resulting centered feature vector as $\widetilde{X}_i^{(0)}$, we construct the semi-synthetic outcome as

$$Y_i = \widetilde{X}_i^{(0)\top}\theta^* + \varepsilon_i, \qquad \varepsilon_i \sim (0, \sigma^2),$$

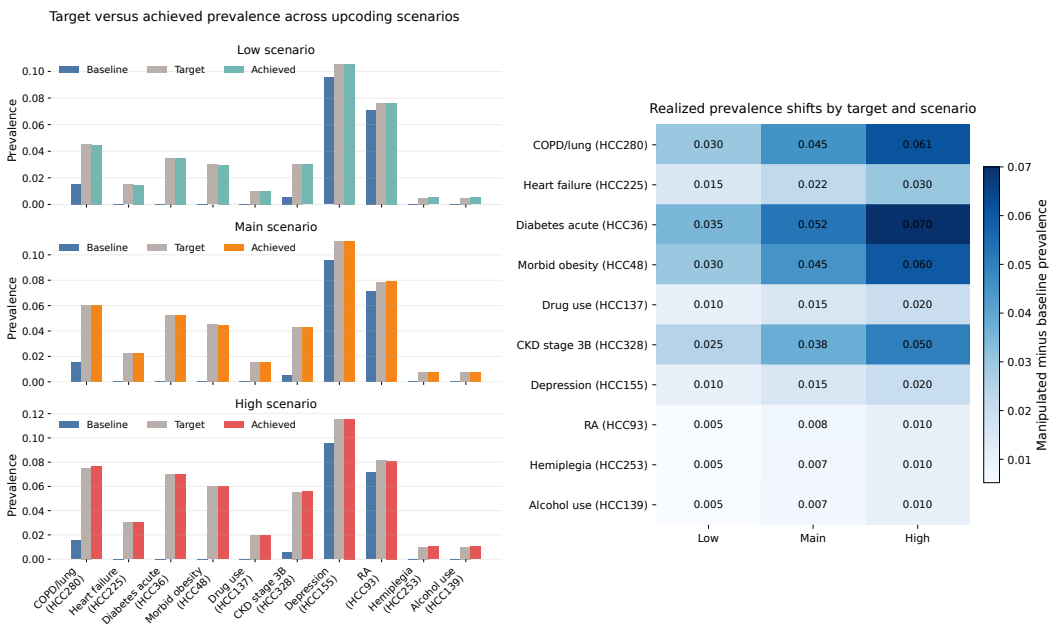

Figure 10: **Calibration of simulated upcoding across target HCCs and scenarios.** (Left) Baseline, target, and achieved prevalence after applying the V28 hierarchy for each target HCC under the low, main, and high scenarios. The achieved prevalences closely track the calibration targets. (Right) Realized manipulated-minus-baseline prevalence shifts after applying the V28 hierarchy. The low, main, and high scenarios scale the same diagnosis-level calibration targets, and the realized shifts are concentrated in the targeted top-ten diagnosis groups.

where θ^* is the published CMS-HCC V28 Community, NonDual, Aged coefficient vector restricted to these 30 HCCs. We set $\sigma^2 = 0.25\,\widehat{\mathrm{Var}}(\widetilde{X}^{(0)\top}\theta^*)$. The same θ^* is used both to generate outcomes and to compute the plug-in strategic MSE.

For a fitted coefficient vector $\widehat{\theta}$, evaluation is based on the endogenous strategic response model studied in the paper:

$$\widetilde{X}_i^{(\alpha,\widehat{\theta})} = \widetilde{X}_i^{(0)} + \alpha H^{-1}\widehat{\theta},$$

where $\alpha \geq 0$ is the manipulation-intensity parameter. The main figures report the plug-in post-manipulation MSE

$$\mathrm{MSE}_\alpha(\widehat{\theta}) = (\widehat{\theta} - \theta^*)^\top \widehat{\Sigma}(\widehat{\theta} - \theta^*) + \alpha^2(\widehat{\theta}^\top H^{-1}\widehat{\theta})^2 + \sigma^2. \tag{26}$$

This is the exact empirical analogue of the theoretical strategic MSE studied in the paper. We normalize most reported quantities by the full ridge value at the same manipulation intensity, so values below one indicate an improvement over full-support ridge.

G.4 Cost matrix construction

The manipulation cost matrix H is not directly observed. We therefore construct a proxy using available evidence from health policy on differential coding. Since the strategic best response $a^*(\theta) = H^{-1}\theta$ depends on H^{-1}, it is convenient to build the inverse-cost matrix first. We refer to H^{-1} as the manipulation-ease matrix. Larger entries of H^{-1} mean that a fixed coefficient vector induces a larger strategic response, or equivalently, that manipulation is less costly in the corresponding direction.

The cost construction is fixed before fitting any model or choosing any support. It uses three inputs: (i) the CMS-HCC V28 feature list; (ii) the baseline HCC matrix; and (iii) the target-mapping table, which links each calibrated top-ten diagnosis group to its V28 HCC members and records a group-level calibration score. (Recall from Section G.1 that the realized upcoded matrices are used to check that these calibration targets are achieved.)

The final manipulation-ease matrix has the form:

$$H^{-1} = \tau\{D + \rho B\} + \eta I. \tag{27}$$

Table 1: Cost-construction parameters for the experiments in Section 6.

Parameter	Description	Value
τ	global manipulation-intensity scale	0.5
ρ	correlated-cost scale	0.5
η	positive-definiteness ridge	10^{-6}
d_{floor}	baseline diagonal manipulation ease for every HCC	0.03
ξ	moderation factor for HCCs in the top-ten diagnosis groups	0.25

Here D is diagonal and controls the marginal ease of manipulating each HCC. The matrix B is a block matrix that introduces correlated manipulation within pre-specified HCC blocks based on the CMS-HCC V28 hierarchy. We refer to D as the diagonal component and B as the block component of the inverse-cost matrix. The scalar τ controls the overall manipulation-intensity scale, ρ controls the strength of the block component, and ηI is a small ridge term that ensures strict positive definiteness. These scalar parameters are fixed across all experiments and summarized in Table 1. We next describe the construction of the matrices D and B.

Construction of matrix D. We first construct diagonal scores over all V28 HCC columns and then restrict to the HCC columns with nonzero prevalence. We initialize every HCC at the floor: $d_j^{\text{raw}} = d_{\text{floor}}$. For HCCs belonging to a diagnosis group in the target-mapping table, we increase its diagonal manipulation-ease score according to the normalized group-level calibration score. Specifically, if HCC j belongs to one of the top-ten diagnosis groups g, we update $d_j^{\text{raw}} \leftarrow \max\{d_j^{\text{raw}}, d_{\text{floor}} + s_g\}$, where $s_g := r_g / \max_{g'} r_{g'} \in [0, 1]$ is the normalized (group-level) calibration score for that group g. Thus the diagonal manipulation-ease scores are tied to the same diagnosis groups used to generate the upcoded scenarios.

Finally, we apply a moderation factor to the HCCs belonging to the top-ten diagnosis groups (i.e., `HCC155`, `HCC280`, `HCC38`, `HCC93`, `HCC328`, `HCC226`, which are also highlighted in Figure 4). Specifically, for HCC j in this group, we set $d_j = d_{\text{floor}} + \xi(d_j^{\text{raw}} - d_{\text{floor}})$, where ξ is the moderation factor applied to the excess manipulation score. The purpose of this moderation factor is to avoid making the conclusions degenerate. The top-ten group membership is coarse group-level evidence of coding sensitivity, but not a direct estimate of an infinite coordinate-level manipulation ease. If every HCC belonging to the top-ten diagnosis groups were assigned an extreme unmoderated ease score, the cost matrix would mechanically favor exclusion of these HCCs. The moderation factor instead creates a non-degenerate selection problem: these HCCs remain easier to manipulate than baseline HCCs, but their predictive value and covariance structure can still justify retention under ridge shrinkage. For all other HCCs, we set $d_j = d_j^{\text{raw}}$.

Construction of matrix B. Upcoding practices for a single diagnosis are likely to affect closely related diagnoses as well. This intuition is captured in the block matrix B, which captures correlated manipulation within pre-specified HCC blocks. These blocks are based on CMS-HCC V28 hierarchy branches and represent groups of HCCs that correspond to the same underlying condition or severity levels within the same condition. Table 2 lists the blocks used in our construction. For each block, we add a positive-semidefinite outer-product term and then normalize the resulting block matrix over the HCCs with nonzero prevalence. More precisely, for each listed HCC block $\mathcal{G}_\ell$ for $\ell = 1, \ldots, L$, define a vector v_ℓ over all HCC columns by

$$(v_\ell)_j = \begin{cases} \sqrt{\max\{d_j, d_{\text{floor}}\}}, & j \in \mathcal{G}_\ell, \\ 0, & j \notin \mathcal{G}_\ell. \end{cases}$$

We then set $B_0 = \sum_{\ell=1}^{L} v_\ell v_\ell^\top$. This construction makes B_0 positive semidefinite. Whenever B_0 is nonzero, we normalize it by its largest diagonal entry $B = \frac{B_0}{\max_j (B_0)_{jj}}$. Finally, we restrict to the HCC columns with nonzero baseline prevalence to get the manipulation-ease matrix. Since $D, B \succeq 0$ and $\eta > 0$, the final matrix is positive definite. Figure 11 shows the resulting diagonal scores in the left panel and the full manipulation-ease matrix in the right panel.

It is worth remarking that the construction of the manipulation-ease matrix in (27) should be interpreted as a proxy, not as an estimated structural inverse cost matrix. Its purpose in this paper is to

Table 2: Pre-specified diagnosis group blocks used in the block component B.

Diagnosis group blocks	HCCs
Diabetes	HCC36, HCC37, HCC38
Acute or chronic heart failure, except end stage	HCC224, HCC225, HCC226
Chronic kidney disease, moderate, stage 3	HCC328, HCC329
Substance-use disorder	HCC137, HCC138, HCC139
Mental health	HCC151, HCC152, HCC153, HCC154, HCC155
Vascular disease	HCC263, HCC264, HCC267
Cancer	HCC20, HCC21, HCC22, HCC23

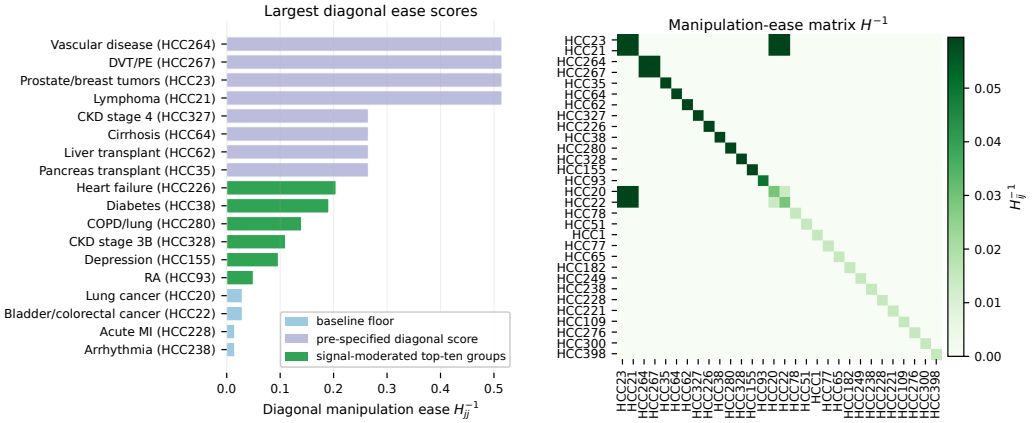

Figure 11: **Visualization of the manipulation-ease matrix.** (Left) Largest HCC diagonal entries of the final H^{-1}, grouped by the source of the diagonal score. (Right) The final inverse-cost matrix over HCCs with nonzero baseline prevalence, including the block component.

encode three qualitative features of the Medicare Advantage coding setting: coding sensitivity is concentrated in a small number of diagnosis groups; coding opportunities can be correlated within certain diagnosis group blocks; and coding-sensitive diagnoses can still carry crucial predictive signal. The various choices above are made with respect to these considerations.

G.5 Baselines and tuning

All methods are evaluated using the same plug-in post-manipulation MSE_α objective (26). The main comparisons are summarized below.

1. Full ridge: ordinary ridge on all HCCs with nonzero baseline prevalence, with the ridge level chosen to minimize MSE_α.
2. Top-ten exclusion: ridge after removing all HCCs with nonzero baseline prevalence that belong to the calibrated top-ten differential-coding groups.
3. Prediction-only: retain the k HCCs with nonzero baseline prevalence that have the largest absolute covariance with the outcome, $|\widehat{\Sigma}\theta^*|$, and then tune ridge on that support.
4. Cost-only: retain the k HCCs with nonzero baseline prevalence that have the smallest diagonal manipulation-ease scores H_{jj}^{-1}, and then tune ridge on that support.
5. Subset-only: use the same support as the proposed support-restricted ridge design but remove ridge shrinkage by setting $\lambda \approx 0$.
6. Ours: jointly choose a support of size k and a ridge level using the MSE_α objective, and then refit support-restricted ridge on the chosen support.
7. Oracle $\widetilde{\mathrm{OPT}}$: the full-support zero-intercept oracle over all HCCs with nonzero baseline prevalence. It is not a deployable method, since it lies outside the restricted estimator class, but serves as a benchmark for that class (see discussion in Section 3.1).

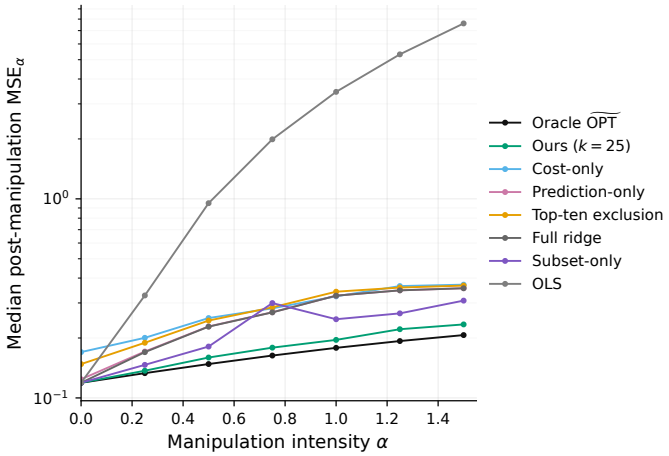

Figure 12: **Post-manipulation** MSE **under varying manipulation intensity** α **for all baselines.** Each curve reports median post-manipulation MSE as the manipulation intensity α varies, shown on a log scale. The proposed support-restricted ridge estimator with $k = 25$ remains close to the zero-intercept oracle $\widetilde{\text{OPT}}$ and improves substantially over full ridge, top-ten exclusion, prediction-only selection, cost-only selection, and subset-only selection. OLS is highly sensitive to manipulation and becomes unstable as α increases.

For the oracle at manipulation intensity α, the first-order condition has the generalized-ridge form (15), which we solve by a scalar iterative fixed-point procedure. All ridge-based methods are tuned over the common grid $\{10^{-6}, 10^{-3}, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 3, 10\}$. Because our case study is a deterministic plug-in benchmark, we tune ridge levels by minimizing the plug-in strategic objective over this grid. In finite-sample deployments, this tuning could instead be implemented by cross-validation. In particular, for ridge regression, variants of cross-validation have consistency guarantees in non-strategic settings, even when these settings are high-dimensional and model-free; see, e.g., Patil et al. [31, 32]. Extending these guarantees to strategic settings, along with joint support selection, is an interesting direction for future work.

G.6 Supporting illustrations

Figure 12 extends the comparison under varying manipulation intensity in Figure 3 to all baselines. The plot illustrates that unregularized subset selection and OLS-style estimators become unstable once manipulation is nontrivial, while prediction-only and cost-only rules fail to achieve the same tradeoff as joint support selection and ridge tuning.

Figure 4 shows that, while the top-ten exclusion policy removes all intensely coded HCCs by construction, the proposed method retains some of these HCCs at the unconstrained optimum $k = 20$ and all of them at the broader $k = 25$ operating point. This does not mean that all intensely coded variables should always be retained. Rather, it demonstrates that the optimal decision is joint: an HCC with higher coding intensity can be retained when its predictive value and covariance structure make it sufficiently valuable, and when ridge shrinkage sufficiently controls the induced strategic exposure. Figure 13 shows the retention path as the support budget is varied. As can be observed, the retention pattern is not a discontinuous artifact of a particular budget choice: several intensely coded HCCs enter before $k = 25$, and all six are retained by the broader design. Figure 14 shows the same in the signal–manipulability plane.

G.7 Experiments compute resources

The source code for generating the figures in our paper is included with the supplementary material. The experiments are run on a mix of: a local machine with 12 CPU cores and a cluster with 48 CPU cores. Our experiments had less than 24 hours of total run time.

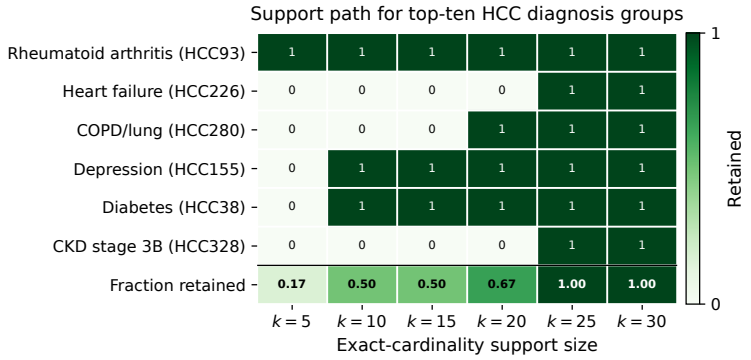

Figure 13: **Support path for top-ten diagnosis groups with nonzero prevalence.** Rows show the HCC groups that belong to the top-ten diagnosis groups identified by Kronick et al. [20], and columns show exact-cardinality support-restricted ridge solutions at different support sizes. The final row reports the fraction of these groups retained. The unconstrained optimum at $k = 20$ already retains several of these groups, while the broader $k = 25$ design retains all six groups with nonzero prevalence.

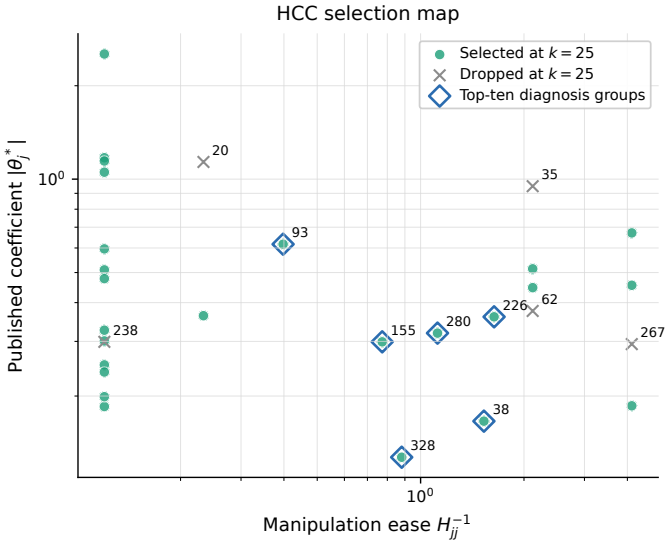

Figure 14: **Signal–manipulability selection map.** Each point is an HCC with nonzero prevalence in the data. Diamonds mark HCCs among the top-ten diagnosis groups identified by Kronick et al. [20], and labels are shown for the diagnosis groups that are dropped or belong to the top-ten set. The selected support is not determined by manipulation ease alone: the $k = 25$ design retains several of these top-ten diagnosis groups with nontrivial signal while dropping other HCCs.

H Extensions

H.1 Inherent robustness under cost uncertainty and heterogeneity

Our main results assume that the cost matrix H is known and common across organizations. We now consider uncertainty about this matrix and, under a common baseline distribution, heterogeneity in costs across organizations.

The goal in this section is to study the *inherent robustness* of various methods, rather than designing a new robust optimization procedure. We show that regularization and feature selection provide additional robustness guarantees relative to other strategies. For simplicity, we reparameterize and consider the inverse cost matrix $K = H^{-1}$.

To demonstrate this, we first revisit the case of known costs and consider the benchmark predictor that attains the strategic optimum OPT. The decision maker can attain OPT if they are able to *directly* tune the intercept based on the costs according to $b^{\mathrm{OPT}} = -\theta^{*\top} K\theta^*$, achieving the irreducible error

σ^2 (see Lemma B.1). Thus, with exact knowledge of costs and the flexibility to choose an intercept, the unrestricted intercept correction (along with the true signal model θ^*) cannot be improved upon.

However, the situation changes in the practical scenario when the decision maker only has partial information about the inverse-cost matrix K. To make this precise, we introduce some additional notation. For a set $\mathcal{K}$ of plausible inverse-cost matrices, define the worst-case excess strategic MSE of a fixed rule (θ, b) by

$$\sup_{K\in\mathcal{K}} \{\mathrm{MSE}_K(\theta, b) - \mathrm{OPT}\}, \quad \text{where} \quad \mathrm{MSE}_K(\theta, b) = (\theta - \theta^*)^\top \Sigma(\theta - \theta^*) + (b + \theta^\top K\theta)^2 + \sigma^2.$$

This expression makes the robustness issue transparent. An intercept correction can remove the strategic shift exactly when K is known: the strategic optimum uses the (coefficients, intercept) tuple $(\theta^*, -\theta^{*\top} K\theta^*)$. But under cost uncertainty, an intercept chosen using a nominal estimate $\widehat{K}$ leaves the residual shift $\theta^\top(K - \widehat{K})\theta$. The following simple identity formalizes the worst-case error of a general fixed model (θ, b), and then specializes it to intercept correction.

Proposition H.1 (Worst-case error in general and for intercept correction)**.** *Let $\mathcal{K}$ be a nonempty compact set of symmetric positive definite inverse-cost matrices and fix (θ, b). Assume the baseline distribution of (X, Y) is held fixed as K varies. Define $\underline{q}(\theta) := \inf_{K\in\mathcal{K}} \theta^\top K\theta$ and $\overline{q}(\theta) := \sup_{K\in\mathcal{K}} \theta^\top K\theta$. Then*

$$\sup_{K\in\mathcal{K}} \{\mathrm{MSE}_K(\theta, b) - \mathrm{OPT}\} = (\theta - \theta^*)^\top \Sigma(\theta - \theta^*) + \max\{(b + \underline{q}(\theta))^2,\ (b + \overline{q}(\theta))^2\}. \tag{28}$$

In particular, for intercept correction, where $\theta = \theta^*$ and $b = -\theta^{*\top}\widehat{K}\theta^*$, we have

$$\sup_{K\in\mathcal{K}} \{\mathrm{MSE}_K(\theta, b) - \mathrm{OPT}\} = \sup_{K\in\mathcal{K}} \left(\theta^{*\top}(K - \widehat{K})\theta^*\right)^2.$$

Proof. Fix (θ, b). The prediction term $(\theta - \theta^*)^\top \Sigma(\theta - \theta^*)$ does not depend on K. Thus

$$\sup_{K\in\mathcal{K}} \left\{\mathrm{MSE}_K(\theta, b) - \sigma^2\right\} = (\theta - \theta^*)^\top \Sigma(\theta - \theta^*) + \sup_{K\in\mathcal{K}} (b + \theta^\top K\theta)^2.$$

Let $\underline{q}(\theta) := \inf_{K\in\mathcal{K}} \theta^\top K\theta$ and $\overline{q}(\theta) := \sup_{K\in\mathcal{K}} \theta^\top K\theta$. Since $K \mapsto \theta^\top K\theta$ is continuous and $\mathcal{K}$ is compact, the extrema are attained. Moreover, for any $q \in [\underline{q}(\theta), \overline{q}(\theta)]$, the function $q \mapsto (b + q)^2$ is convex, so its maximum over the range is attained at an endpoint. Therefore

$$\sup_{K\in\mathcal{K}} (b + \theta^\top K\theta)^2 = \max\{(b + \underline{q}(\theta))^2,\ (b + \overline{q}(\theta))^2\}.$$

This proves the first display. If $b = -\theta^\top \widehat{K}\theta$, then $b + \theta^\top K\theta = \theta^\top(K - \widehat{K})\theta$, which gives the plug-in expression. □

Calculation of the worst-case error in general (28) shows that a model is inherently more robust when its coefficient vector has small exposure to the uncertain manipulation directions. Regularization reduces this exposure by shrinking coefficients, while feature selection can remove highly uncertain or highly manipulable coordinates entirely. We now give two examples showing how regularization and feature selection can improve inherent robustness compared to intercept correction.

Example 1 (Regularization alone)**.** Let $d = 1$, $\Sigma = 1$, and $\theta^* = 1$. Suppose the true inverse cost lies in $\mathcal{K} = [1/2, 5/2]$, and the nominal estimate is the midpoint $\widehat{K} = 3/2$. The plug-in intercept correction deploys $(\theta, b) = (1, -3/2)$, whose worst-case excess strategic MSE is

$$\sup_{K\in[1/2,5/2]} (K - 3/2)^2 = 1.$$

By contrast, for ordinary ridge with $\lambda = 1$ the coefficient is

$$\theta^{\mathrm{OR}}(\lambda) = \frac{1}{1 + \lambda} = \frac{1}{2}.$$

and $b = 0$. Its worst-case excess strategic MSE therefore is

$$\left(\frac{1}{2} - 1\right)^2 + \sup_{K\in[1/2,5/2]} \left(K\left(\frac{1}{2}\right)^2\right)^2 = \frac{1}{4} + \left(\frac{5}{2} \cdot \frac{1}{4}\right)^2 = \frac{1}{4} + \frac{25}{64} = \frac{41}{64} < 1.$$

Thus, shrinkage alone can be inherently more robust than a plug-in intercept correction.

Note that in Example 1 we examine a fixed intercept $b = -\widehat{K} = -3/2$ rather than proposing a robust intercept. Nevertheless, Proposition H.1 implies that if one were allowed to choose the best fixed intercept for a fixed coefficient model θ, the minimizer would place $-b$ at the midpoint of the interval $[\underline{q}(\theta), \overline{q}(\theta)]$, yielding residual squared shift $(\overline{q}(\theta) - \underline{q}(\theta))^2/4$. So in this example, we are in fact already comparing against the best fixed intercept model with unknown K.

Example 2 (Feature selection and regularization together)**.** Now, consider a two-feature example in which the first feature is directly predictive but has uncertain manipulation cost, while the second feature is a less manipulable correlated proxy for the first feature:

$$\theta^* = (1, 0), \qquad \Sigma = \begin{pmatrix} 1 & 49/50 \\ 49/50 & 1 \end{pmatrix}, \quad \text{and} \quad K(k) = \begin{pmatrix} k & 0 \\ 0 & 1/2 \end{pmatrix}, \qquad k \in [1/2, 5/2].$$

Again, take the nominal estimate to be the midpoint $\widehat{K} = K(3/2)$. For the plug-in intercept correction with slope θ^*, the deployed rule is $(\theta, b) = ((1, 0), -3/2)$. Since $\theta^{*\top} K(k)\theta^* = k$, the worst-case excess strategic MSE is

$$\sup_{k \in [1/2, 5/2]} (k - 3/2)^2 = 1,$$

as in Example 1. Full-support ridge reduces this error by shrinking the manipulable coordinate. For example, consider full-support ridge with $\lambda = 1$. Since $\Sigma\theta^* = (1, 49/50)$, we have

$$\theta^{\mathrm{OR}}([2], 1) = (\Sigma + I)^{-1}\Sigma\theta^* = \left(\frac{2599}{7599}, \frac{2450}{7599}\right).$$

It has worst-case excess strategic MSE approximately

$$\sup_{k \in [1/2, 5/2]} \mathrm{MSE}_{K(k)}(\theta^{\mathrm{OR}}([2], 1), 0) - \mathrm{OPT} \approx 0.240.$$

However, support-restricted ridge can do even better by dropping the uncertain coordinate and using the less manipulable proxy. Consider support-restricted ridge on $\mathsf{S} = \{2\}$ with $\lambda = 1/4$. The retained coefficient is

$$\beta = \frac{\Sigma_{2y}}{\Sigma_{22} + \lambda} = \frac{49/50}{1 + 1/4} = \frac{98}{125}.$$

Thus

$$\theta^{\mathrm{OR}}(\{2\}, 1/4) = \left(0, \frac{98}{125}\right).$$

The prediction term is

$$1 - 2\left(\frac{49}{50}\right)\left(\frac{98}{125}\right) + \left(\frac{98}{125}\right)^2.$$

Since this rule places zero weight on the uncertain coordinate, its strategic shift is independent of k:

$$\left[\left(0, \frac{98}{125}\right)^\top K(k) \left(0, \frac{98}{125}\right)\right]^2 = \left[\frac{1}{2}\left(\frac{98}{125}\right)^2\right]^2.$$

Combining the two terms yields the worst-case excess strategic MSE

$$1 - 2\left(\frac{49}{50}\right)\left(\frac{98}{125}\right) + \left(\frac{98}{125}\right)^2 + \left[\frac{1}{2}\left(\frac{98}{125}\right)^2\right]^2 \approx 0.172 < 0.240.$$

For comparison, support-restricted least squares on $\mathsf{S} = \{2\}$ gives $\theta^{\mathrm{LS}}(\{2\}) = (0, 49/50)$, whose worst-case excess strategic MSE is

$$1 - \left(\frac{49}{50}\right)^2 + \left[\frac{1}{2}\left(\frac{49}{50}\right)^2\right]^2 \approx 0.270.$$

Thus, the improvement comes from combining feature selection with regularization: feature selection removes the uncertain manipulation direction, and regularization controls the residual exposure of the retained proxy.

Examples 1 and 2 illustrate that, under cost uncertainty, intercept correction can be sensitive to misspecification, whereas regularization and feature selection can be inherently more robust by reducing exposure to uncertain manipulation directions. Designing explicitly robust versions of these methods, such as choosing b or (S, λ) to minimize worst-case strategic MSE over $\mathcal{K}$, is a natural direction for future work.

The same analysis applies to organizations with different cost matrices, provided that they share the same baseline distribution of (X, Y) and face the same deployed model (θ, b). More precisely, let $\mathcal{H}$ be a nonempty compact set of positive definite cost matrices and set $\mathcal{K} = \{H^{-1} : H \in \mathcal{H}\}$. The inverse-cost set is then compact, and Proposition H.1 gives the worst-case excess MSE across organizations. (Note that each organization-specific unrestricted oracle has value $\mathrm{OPT} = \sigma^2$ under these common distribution assumptions, although its optimal intercept depends on its cost matrix.)

H.2 Error sensitivity under feature-dependent manipulation

An extension of our model allows the quadratic cost matrix to depend measurably on an individual's *pre-manipulation* features: $C_x(a) = \frac{1}{2} a^\top H(x) a$, where $H(x) \succ 0$ almost surely. The matrix $H(x)$ is held fixed while the organization optimizes over a. The same first-order calculation as in Lemma B.3 therefore gives $a^*(x; \theta) = H(x)^{-1}\theta$. Throughout this subsection, the baseline distribution of (X, Y) and the deployed model (θ, b) are held fixed, and reporting manipulation does not change Y.

Write $K(X) = H(X)^{-1}$ and let $K = H^{-1}$ for a constant reference cost matrix $H \succ 0$. The strategic MSE under feature-dependent costs is

$$\mathrm{MSE}_{\text{X-dep}}(\theta, b) := \mathbb{E}[(\theta^\top X + b + \theta^\top K(X)\theta - Y)^2].$$

The corresponding reference risk is $\mathrm{MSE}_K(\theta, b)$, as defined in Appendix H.1.

Proposition H.2 (Error sensitivity under feature-dependent manipulation)**.** *Fix a deployed model (θ, b) and a constant reference inverse-cost matrix $K \succ 0$. Suppose $\mathrm{MSE}_K(\theta, b) < \infty$. Define the directional perturbation $D_\theta(X) := \theta^\top (K(X) - K)\theta$, and $\nu_\theta := (\mathbb{E}[D_\theta(X)^2])^{1/2}$. If $\nu_\theta < \infty$, then*

$$\left|\sqrt{\mathrm{MSE}_{\text{X-dep}}(\theta, b)} - \sqrt{\mathrm{MSE}_K(\theta, b)}\right| \le \nu_\theta, \tag{29}$$

and

$$\left|\mathrm{MSE}_{\text{X-dep}}(\theta, b) - \mathrm{MSE}_K(\theta, b)\right| \le 2\sqrt{\mathrm{MSE}_K(\theta, b)}\,\nu_\theta + \nu_\theta^2. \tag{30}$$

In particular, if $\|K(X) - K\|_{\mathrm{op}} \le \delta$ almost surely, then $\nu_\theta \le \delta\|\theta\|_2^2$, and hence

$$\left|\mathrm{MSE}_{\text{X-dep}}(\theta, b) - \mathrm{MSE}_K(\theta, b)\right| \le 2\delta\|\theta\|_2^2\sqrt{\mathrm{MSE}_K(\theta, b)} + \delta^2\|\theta\|_2^4.$$

Proof. Let $e_K := \theta^\top X + b + \theta^\top K\theta - Y$. The reference prediction residual is e_K, while the residual under feature-dependent costs is $e_K + D_\theta(X)$. The reverse triangle inequality in L_2 gives

$$\left|\|e_K + D_\theta(X)\|_{L_2} - \|e_K\|_{L_2}\right| \le \|D_\theta(X)\|_{L_2},$$

which proves (29). Expanding the difference of the squared residuals and applying Cauchy–Schwarz gives

$$\begin{aligned}\left|\mathrm{MSE}_{\text{X-dep}}(\theta, b) - \mathrm{MSE}_K(\theta, b)\right| &= \left|2\mathbb{E}[e_K D_\theta(X)] + \mathbb{E}[D_\theta(X)^2]\right| \\ &\le 2\|e_K\|_{L_2}\|D_\theta(X)\|_{L_2} + \|D_\theta(X)\|_{L_2}^2,\end{aligned}$$

which proves (30). Finally, under the operator-norm condition, almost surely,

$$|D_\theta(X)| \le \|K(X) - K\|_{\mathrm{op}}\|\theta\|_2^2 \le \delta\|\theta\|_2^2.$$

Thus, $\nu_\theta \le \delta\|\theta\|_2^2$, giving the final bound. □

These bounds allow arbitrary dependence between $H(X)$, X, and the prediction residual. They can help quantify the deployment sensitivity of a fixed model around the inverse-cost matrix K.

The similar role of our feature selection and regularization levers therefore carries over in this setting: feature selection can remove coordinates along which manipulation costs vary the most with X, while regularization can reduce sensitivity to this variation by shrinking the coefficients. More precisely,

the sensitivity is controlled by the directional perturbation $\theta^\top(H(X)^{-1} - H^{-1})\theta$, which vanishes for the perturbations confined to coordinates omitted by the selected support.

The upper- and lower-bounds (Theorems 1 and 2) do not transfer verbatim, as the strategic shift is now random and may be correlated with X and the prediction residual. Nonetheless, the displayed bound above shows that the performance of any fixed support-restricted ridge estimator is well-behaved and changes continuously with the magnitude of the X-dependent cost perturbation.

H.3 Error sensitivity under cost misspecification

Suppose a support $\widehat{\mathsf{S}}$ and regularization parameter $\widehat{\lambda}$ are selected using a nominal cost matrix $H \succ 0$. Fix the resulting support-restricted ridge estimator $\widehat{\theta} := \theta^{\mathrm{OR}}(\widehat{\mathsf{S}}, \widehat{\lambda})$. Let $H' \succ 0$ be the true cost matrix at deployment, and write $K = H^{-1}$ and $K' = (H')^{-1}$. We compare the strategic MSE of the same deployed model $(\widehat{\theta}, 0)$ under these two inverse-cost matrices, holding the baseline distribution of (X, Y) fixed.

Proposition H.3 (Error sensitivity under cost misspecification)**.** *Let $K, K' \succ 0$ be symmetric inverse-cost matrices and fix a coefficient vector $\widehat{\theta} \in \mathbb{R}^d$ supported on $\widehat{\mathsf{S}} \subseteq [d]$. Suppose $\|K' - K\|_{\mathrm{op}} \le \delta$ and $\max\{\|K\|_{\mathrm{op}}, \|K'\|_{\mathrm{op}}\} \le M$. Then, under the same baseline distribution of (X, Y),*

$$\left|\mathrm{MSE}_{K'}(\widehat{\theta}, 0) - \mathrm{MSE}_K(\widehat{\theta}, 0)\right| \le 2M\delta\|\widehat{\theta}\|_2^4. \tag{31}$$

More precisely, if $\widehat{\mathsf{S}} \neq \varnothing$, define $\delta_{\widehat{\mathsf{S}}} := \|(K' - K)_{\widehat{\mathsf{S}}\widehat{\mathsf{S}}}\|_{\mathrm{op}}$, and $M_{\widehat{\mathsf{S}}} := \max\{\|K_{\widehat{\mathsf{S}}\widehat{\mathsf{S}}}\|_{\mathrm{op}}, \|K'_{\widehat{\mathsf{S}}\widehat{\mathsf{S}}}\|_{\mathrm{op}}\}$. Then

$$\left|\mathrm{MSE}_{K'}(\widehat{\theta}, 0) - \mathrm{MSE}_K(\widehat{\theta}, 0)\right| \le 2M_{\widehat{\mathsf{S}}}\delta_{\widehat{\mathsf{S}}}\|\widehat{\theta}\|_2^4. \tag{32}$$

If $\widehat{\mathsf{S}} = \varnothing$, then $\widehat{\theta} = 0$ and the risk is independent of the cost matrix.

Proof. By Lemma B.4, the prediction and noise terms are the same under K and K'. Therefore,

$$\left|\mathrm{MSE}_{K'}(\widehat{\theta}, 0) - \mathrm{MSE}_K(\widehat{\theta}, 0)\right| = \left|(\widehat{\theta}^\top K'\widehat{\theta})^2 - (\widehat{\theta}^\top K\widehat{\theta})^2\right|.$$

Write $q = \widehat{\theta}^\top K\widehat{\theta}$ and $q' = \widehat{\theta}^\top K'\widehat{\theta}$. Since $K, K' \succ 0$, we have $q, q' \ge 0$. Moreover,

$$|q' - q| = |\widehat{\theta}^\top(K' - K)\widehat{\theta}| \le \delta\|\widehat{\theta}\|_2^2, \quad \text{and} \quad q' + q \le 2M\|\widehat{\theta}\|_2^2.$$

Factoring the difference of squares gives

$$|q'^2 - q^2| = |q' - q|(q' + q) \le 2M\delta\|\widehat{\theta}\|_2^4,$$

which proves (31).

For the support-specific bound, we use $q = \widehat{\theta}_{\widehat{\mathsf{S}}}^\top K_{\widehat{\mathsf{S}}\widehat{\mathsf{S}}}\widehat{\theta}_{\widehat{\mathsf{S}}}$, and $q' = \widehat{\theta}_{\widehat{\mathsf{S}}}^\top K'_{\widehat{\mathsf{S}}\widehat{\mathsf{S}}}\widehat{\theta}_{\widehat{\mathsf{S}}}$. Applying the same argument to these retained blocks and noting that $\|\widehat{\theta}_{\widehat{\mathsf{S}}}\|_2 = \|\widehat{\theta}\|_2$ gives (32). For the empty support, both quadratic forms are zero. □

Proposition H.3 shows that only the retained block of the inverse-cost perturbation affects a fixed support-restricted predictor's strategic MSE. In particular, if $(K' - K)_{\widehat{\mathsf{S}}\widehat{\mathsf{S}}} = 0$, then the two risks are equal, even if other blocks differ. Feature selection can therefore remove exposure to uncertain manipulation directions.

On a fixed support, increasing the ordinary ridge parameter makes the coefficient norm nonincreasing and hence reduces the coefficient-norm factor in the sensitivity bound. This does not imply that the actual strategic MSE decreases monotonically with regularization for arbitrary cost geometry: shrinkage also changes baseline prediction error.

The same bound applies uniformly across organizations when the inverse-cost discrepancy and norm bounds hold uniformly, the baseline distribution is common, and the deployed model is held fixed.

We note that this is a "post selection" performance bound analysis for fixed selection and regularization. It does not say that the selected support or the regularization remain unchanged when the nominal cost matrix is perturbed.

We also note that the perturbation is measured in *inverse costs*. A small perturbation in H itself implies a comparable perturbation in H^{-1} only under additional conditioning assumptions on H and H'.

I Discussion and Limitations

We develop a framework for studying strategic classification through the coarse policy levers of feature selection and regularization. Our main message is that features should be selected jointly based on their relative manipulability and predictability, rather than based on the manipulability of individual features alone. Through theoretical analysis and a real-world case study, we show that the interaction between feature selection and regularization yields nuanced implications for policy design that broaden the prevailing narrative. Furthermore, the coarse levers offer greater inherent robustness under cost uncertainty compared to alternative strategies. More broadly, given the ubiquity of prediction systems in practice, we envision a growing need for principled and practical frameworks that bridge theoretical foundations with the policy and institutional contexts in which these systems are deployed.

Several interesting questions remain. We study a one-shot interaction in this work, and it would be valuable to analyze long-term dynamics in this setting. An interesting statistical perspective to explore further is to view strategic tuning as tuning under an endogenous deployment shift. Related work studies optimal regularization under exogenous covariate and regression shifts (see, e.g., [33] and references therein), whereas here the shift is induced by the deployed coefficients themselves and is coupled to support restriction. Another extension is to study more expressive coarse levers, such as grouped ridge or per-feature regularization, as intermediate designs between ordinary ridge and the strategically optimal solution. Our framework can also be extended to incorporate settings where certain variables *must* be retained for policy or institutional reasons, by modifying the support-selection objective.

At the same time, our model makes some simplifying assumptions that are common in the strategic learning literature, such as known and quadratic manipulation costs, which are homogeneous across organizations. In this paper, we take a first step toward understanding misspecification by studying the inherent robustness of a fixed deployed model when the true inverse-cost matrix lies in a plausible uncertainty set. Our examples illustrate that intercept corrections can be sensitive to misspecification, while regularization and feature selection can reduce exposure to uncertain manipulation directions. However, these examples barely scratch the surface of the broader misspecification question. We do not design robust versions of any method in this work; developing worst-case or distributionally robust choices of the intercept, support, and regularization level, examining richer cost structures, and studying the impact of cost heterogeneity across organizations are important directions for future work.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We provide theoretical results in Section 3, our algorithm in Section 5, and experimental results in Section 6 that match our main claims.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We discuss the limitations in Section 7 and Appendix I.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide the full set of assumptions in our results wherever stated (Section 3, Section 5) and provide the proofs in the appendix for the respective sections (Appendix C,E).

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all the details required to reproduce the illustrations in Appendix 3 and the experimental results in Section G.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the code with the submission that can be used to reproduce the main experimental results in the paper.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify all the experimental details and description of the setup in Section 6 and Section G.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Most of our experiments are deterministic. We report algorithmic randomness wherever it applies.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide information on the experiments compute resources in Appendix G.7.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the motivation and societal impacts of our work throughout the introduction of the paper (Section 1) and in the discussion and limitations section as well (Section 7).

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: Our paper does not release data or models that have a high risk for misuse.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We use the upcoding package (`https://github.com/StanfordHPDS/upcoding`) to generate data for experiments and we cite the authors in the paper as well as include the URL.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not release new assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.