

Strategic Feature Selection

Jivat Neet Kaur¹, Pratik Patil², Divya Shanmugam³,
Emma Pierson¹, Michael I. Jordan^{1,7}, Nika Haghtalab¹,
Meena Jagadeesan^{4,5*}, Ahmed Alaa^{1*}, Serena Wang^{6*}

¹University of California, Berkeley, ²University of Texas, Austin, ³Cornell Tech, ⁴Stanford University,
⁵University of Pennsylvania, ⁶Harvard University, ⁷Inria, Paris

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 (Geruso and Layton, 2020). 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 (U.S. Department of Justice, 2026). 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 (Medicare Payment Advisory Commission, 2025).

These types of incentives are captured by the framework of strategic classification (Hardt et al., 2016; Podimata, 2025), 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.

*Equal contribution

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 tend to be 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 responses. This raises a different question than the one typically studied in strategic learning: not how to redesign the entire predictor, but how to address strategic behavior using only the coarse 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 (Björkegren et al., 2025; Kronick et al., 2025a). 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 (Centers for Medicare & Medicaid Services, 2013, 2023). 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 (Centers for Medicare & Medicaid Services, 2023). 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 (Centers for Medicare & Medicaid Services, 2023). 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 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 3). 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 (Centers for Medicare & Medicaid Services, 2024; Kronick et al., 2025a).

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 (Kronick et al., 2025a) (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 does 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 (Section 7), 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. These works aim to learn optimal prediction rules when decision subjects can manipulate their features at a cost (e.g., (Hardt et al., 2016; Dong et al., 2018; Chen et al., 2020; Shavit et al., 2020; Levanon and Rosenfeld, 2021; Björkegren et al., 2025)). Closely related to our work, Björkegren et al. (2025) 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 (2025) for a survey), which also studies strategic behavior from an improvement and causality perspective (Miller et al., 2020; Haghtalab et al., 2020; Kleinberg and Raghavan, 2020; Shavit et al., 2020), 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.

Closely related in spirit to our work, Handina and Mazumdar (2024) 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.

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 (U.S. Department of Justice, 2026). The health policy literature has put forth several proposals to combat such behavior. Rose (2016) 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 (Meyers and Trivedi, 2021), incorporating patient survey data (Bellerose et al., 2025; McWilliams et al., 2025), or excluding diagnoses coded most differentially by insurers in order to reduce overpayments (Kronick et al., 2025a). Closer to the levers we study, McGuire et al. (2021) show how constrained regression and variable selection can improve the performance of risk adjustment systems. In a similar vein, Arias (2025) 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 (Martinelli and Parker, 2009; Camacho and Conover, 2011). Work on proxy means tests emphasizes that the variables used to determine eligibility should be both predictive of welfare and verifiable in practice (Grosh and Baker, 1995). Niehaus et al. (2013) 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.

Adversarial robustness. Our work is also related to adversarially robust feature selection and poisoning-robust feature selection (Zhang et al., 2016; Xiao et al., 2015), 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.

2 Model and Preliminaries

We study the strategic interaction between a decision maker who specifies a predictor and organizations that can modify the reported features of the individuals they serve. Building on the common linear specification in strategic learning (Shavit et al., 2020; Kleinberg and Raghavan, 2020; Harris et al., 2022), 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 a predicted outcome.

We model the population as a distribution $\mathcal{P}_{X,Y}$ over (X, Y) , where $X \in \mathbb{R}^d$ denotes the feature vector and $Y \in \mathbb{R}$ denotes the outcome. We assume that X and Y have finite second moments, and define the feature covariance matrix $\Sigma := \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top] \succ 0$. For simplicity, we assume $\mathbb{E}[X] = 0$ and $\mathbb{E}[Y] = 0$. This centering is without loss of generality, as nonzero means can be absorbed into the intercept.

Because our focus is on linear predictors, we define θ^* to be the coefficient of the best linear projection of Y onto X :

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

Since $\Sigma \succ 0$, this minimizer is unique and is given by $\theta^* = \Sigma^{-1} \mathbb{E}[XY]$. We can therefore write $Y = \theta^{*\top} X + \varepsilon$ where $\varepsilon := Y - \theta^{*\top} X$ is the projection residual. By the normal equations, the residual is orthogonal to the features: $\mathbb{E}[X\varepsilon] = 0$. Under our centering conventions, $\mathbb{E}[\varepsilon] = 0$ as well. Let $\sigma^2 := \mathbb{E}[\varepsilon^2]$.

Following Hardt et al. (2016); Björkegren et al. (2025), 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 $C(a) > 0$. Adopting the quadratic manipulation costs in earlier work (Shavit et al., 2020; Björkegren et al., 2025), 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

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

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 $\text{MSE}(\theta, b) := \mathbb{E}[(\theta^\top X + b - Y)^2]$, but the mean squared error after the organization best responds:

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

As a benchmark, we use the notion of the *strategic optimum* (Hardt et al., 2016), allowing the decision maker to optimize over all coefficients in \mathbb{R}^d and intercept in \mathbb{R} . The unrestricted strategic optimum is

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

2.1 Policy levers: Feature selection and regularization

Our analysis focuses on two *coarse* policy levers that are leveraged in practice to combat strategic behavior within existing prediction pipelines. Throughout this section, the estimators are defined with respect to the unmanipulated distribution $\mathcal{P}_{X,Y}$.

Feature selection. Intuitively, feature selection provides a direct way to limit strategic manipulation by excluding features whose vulnerability to gaming outweighs their predictive value. The decision maker chooses a subset of features $S \subseteq [d]$ and fits a linear predictor using only X_S . Let $R = [d] \setminus S$ denote the omitted coordinates, and define the coordinate subspace $\Theta(S) := \{\theta \in \mathbb{R}^d : \text{supp}(\theta) \subseteq S\}$. The support-restricted least squares estimator is:

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

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 with scalar regularization parameter $\lambda \geq 0$. The ordinary ridge estimator is:

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

These two levers are often deployed jointly, with some features dropped entirely and the remaining coefficients regularized. To capture this combination, we consider the *support-restricted ridge* estimator. For a subset $S \subseteq [d]$ and regularization level $\lambda \geq 0$, define:

$$(\theta^{\text{OR}}(S, \lambda), b^{\text{OR}}(S, \lambda)) \in \arg \min_{\theta \in \Theta(S), b \in \mathbb{R}} \mathbb{E}[(Y - \theta^\top X - b)^2] + \lambda \|\theta\|_2^2. \quad (7)$$

This estimator nests the previous two cases: (i) When $\lambda = 0$, it reduces to the support-restricted least squares estimator: $\theta^{\text{OR}}(S, 0) = \theta^{\text{LS}}(S)$. (ii) When $S = [d]$, it reduces to ordinary ridge: $\theta^{\text{OR}}([d], \lambda) = \theta^{\text{OR}}(\lambda)$. In our theoretical and empirical results, we will show that combining feature selection with regularization can yield larger gains than using either lever alone.

All three estimators (5), (6), and (7) admit closed-form expressions, as detailed in Lemma A.5.

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 $\widetilde{\text{OPT}} := \min_{\theta \in \mathbb{R}^d} \text{MSE}_{\text{strat}}(\theta, 0)$ and $\widetilde{\text{OPT}}(\mathbf{S}) := \min_{\theta \in \Theta(\mathbf{S})} \text{MSE}_{\text{strat}}(\theta, 0)$.

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

$$\text{MSE}_{\text{strat}}(\theta^{\text{OR}}(\mathbf{S}, \lambda), b^{\text{OR}}(\mathbf{S}, \lambda)) - \text{OPT} \geq \widetilde{\text{OPT}} - \text{OPT}. \quad (8)$$

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., $b^{\text{OR}}(\mathbf{S}, \lambda) = 0$ (see Lemma A.5). For a fixed support \mathbf{S} , the estimator $\theta^{\text{OR}}(\mathbf{S}, \lambda)$ belongs to the class $\Theta(\mathbf{S})$, and therefore cannot outperform the support-restricted zero-intercept benchmark $\widetilde{\text{OPT}}(\mathbf{S})$. Moreover, $\widetilde{\text{OPT}}(\mathbf{S})$ is monotonically decreasing as the support \mathbf{S} grows. \square

Theorem 1 shows that the limits of the coarse levers can be characterized by a gap $\widetilde{\text{OPT}} - \text{OPT}$. 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^* : $0 \leq \widetilde{\text{OPT}} - \text{OPT} \leq (\theta^{*\top} H^{-1} \theta^*)^2$ (stated formally in Proposition B.1).

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

3.2 Upper bound on the excess strategic MSE

We now turn to upper bounds on the $\text{MSE}_{\text{strat}}(\theta^{\text{OR}}(\mathbf{S}, \lambda))$. Since $\widetilde{\text{OPT}}$ serves as a lower bound per Theorem 1, we bound the gap between $\text{MSE}_{\text{strat}}(\theta^{\text{OR}}(\mathbf{S}, \lambda))$ and $\widetilde{\text{OPT}}$. 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 $\mathbf{S} \subseteq [d]$,*

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

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

1. $\Gamma(\mathbf{S}) := \min_{\eta \in \mathbb{R}^{|\mathbf{S}|}} \{(\eta - \theta_{\mathbf{S}}^{\text{LS}}(\mathbf{S}))^\top \Sigma_{\mathbf{S}\mathbf{S}} (\eta - \theta_{\mathbf{S}}^{\text{LS}}(\mathbf{S})) + (\eta^\top H_{\mathbf{S}\mathbf{S}}^{-1} \eta)^2\}$,
2. $L_{\text{pred}}(\mathbf{S}) := (\theta_{\mathbf{R}}^*)^\top \Sigma_{\mathbf{R}|\mathbf{S}} \theta_{\mathbf{R}}^*$ and $\Sigma_{\mathbf{R}|\mathbf{S}} := \Sigma_{\mathbf{R}\mathbf{R}} - \Sigma_{\mathbf{R}\mathbf{S}} \Sigma_{\mathbf{S}\mathbf{S}}^{-1} \Sigma_{\mathbf{S}\mathbf{R}}$ for $\mathbf{R} = [d] \setminus \mathbf{S}$,
3. $\delta_H(\mathbf{S}) := \inf_{\gamma \geq 0} \|\Sigma_{\mathbf{S}\mathbf{S}}^{-1/2} (H_{\mathbf{S}\mathbf{S}}^{-1} - \gamma I) \Sigma_{\mathbf{S}\mathbf{S}}^{-1/2}\|_{\text{op}}$,

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

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

The first term, $-(\Gamma([d]) - \Gamma(\mathbf{S}))$, is the *manipulability gain*. For a given feature set, $\Gamma(\mathbf{S})$ captures how much performance is lost because the predictor must move away from the best-fitting coefficients on \mathbf{S} in order to account for manipulability: the first term $(\eta - \theta_{\mathbf{S}}^{\text{LS}}(\mathbf{S}))^\top \Sigma_{\mathbf{S}\mathbf{S}} (\eta - \theta_{\mathbf{S}}^{\text{LS}}(\mathbf{S}))$ measures the fit lost by deviating from the least squares fit on \mathbf{S} , and the second term $(\eta^\top H_{\mathbf{S}\mathbf{S}}^{-1} \eta)^2$ measures the loss due to manipulability. Thus, manipulability gain is the *strategic burden* relieved by restricting attention to \mathbf{S} instead of the full feature set

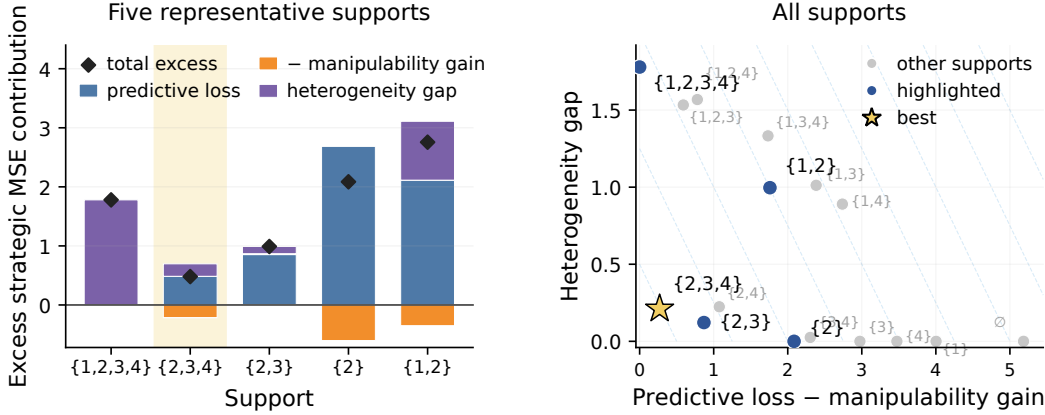


Figure 1: **The best subset depends on the joint predictability-manipulability structure.** In a $d = 4$ example, feature 1 is highly predictive but also highly manipulable. Feature 2 is strongly correlated with feature 1 and therefore acts as a less manipulable proxy. Features 3, 4 are less manipulable 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 heterogeneity gap; dashed lines indicate equal excess strategic MSE. 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 C.1 for details.

[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, $L_{\text{pred}}(\mathbf{S})$, 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 \mathbf{R} that cannot be linearly recovered from \mathbf{S} .

The third term, $C_H(\mathbf{S}) \delta_H(\mathbf{S})^2$, is the *heterogeneity gap* on feature set \mathbf{S} . The quantity $\delta_H(\mathbf{S})$ is small when the features in \mathbf{S} have sufficiently homogeneous manipulation costs (i.e., $H_{\mathbf{S}\mathbf{S}}^{-1}$ 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.

The complete proof for Theorem 2 is deferred to Appendix B.5. Here, we provide the main intuition through a proof sketch.

Proof sketch. There are two steps. First, for any fixed λ , we decompose the excess strategic MSE by adding and subtracting the benchmark $\widetilde{\text{OPT}}(\mathbf{S})$:

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

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

When are the coarse levers near-optimal? Theorem 2 gives an immediate ϵ -optimality condition. If there exists a support \mathbf{S} such that $L_{\text{pred}}(\mathbf{S}) - (\Gamma([d]) - \Gamma(\mathbf{S})) + C_H(\mathbf{S}) \delta_H(\mathbf{S})^2 \leq \epsilon$, then optimally tuned support-

restricted ridge satisfies $\min_{\mathbb{T} \subseteq [d], \lambda \geq 0} \text{MSE}_{\text{strat}}(\theta^{\text{OR}}(\mathbb{T}, \lambda)) \leq \widetilde{\text{OPT}} + \epsilon$. Thus, relative to the unrestricted strategic optimum, the coarse levers are near-optimal up to the irreducible gap $\widetilde{\text{OPT}} - \text{OPT}$ 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, then scalar ridge exactly attains the support-restricted zero-intercept oracle.

Example 1 (Isotropic cost model). Consider an isotropic cost model in which all features have the same manipulability. Formally, for any support $S \subseteq [d]$, $H_{SS}^{-1} = \gamma I$ for some $\gamma \geq 0$. Therefore, on any support contained in $[d]$, the retained manipulation costs are homogeneous, so scalar ridge is exactly expressive:

$$\min_{\lambda \geq 0} \text{MSE}_{\text{strat}}(\theta^{\text{OR}}(S, \lambda)) = \widetilde{\text{OPT}}(S), \quad S \subseteq [d].$$

Thus a highly manipulable group need not be dropped if it is predictive and internally homogeneous; it can instead be retained and regularized. A formal statement for the isotropic cost model is given in Appendix B.7.

Remark 1. Example 1 shows that the heterogeneity gap is exactly zero under an isotropic cost model on the retained support. However, isotropy is not necessary for exact optimality: a strengthened version of the upper bound (derived in Appendix B.8) yields a necessary-and-sufficient characterization, showing that the heterogeneity gap is zero if and only if the support-restricted oracle lies in an eigenspace of the inverse-cost matrix on the retained support.

The isotropic case in Example 1 is useful but idealized. A more realistic idealization is a two-level inverse-cost model in which one group of features is much easier to manipulate than the rest.

Example 2 (Two-level cost model). Partition the features as $[d] = S_m \cup S_b$, where S_m denotes a highly manipulable group and S_b denotes a baseline homogeneous group. For simplicity, suppose that in this example we work in whitened coordinates, so that $\Sigma = I$, and assume

$$H^{-1} = \begin{pmatrix} \kappa I_{S_m} & 0 \\ 0 & I_{S_b} \end{pmatrix}, \quad \kappa > 1.$$

The limiting regime $\kappa \rightarrow \infty$ corresponds to a manipulable group with zero manipulation cost.

Example 2 separates two effects. On any support contained entirely in S_m or entirely in S_b , the retained manipulation costs are fully homogeneous, so scalar ridge is exactly expressive:

$$\min_{\lambda \geq 0} \text{MSE}_{\text{strat}}(\theta^{\text{OR}}(S, \lambda)) = \widetilde{\text{OPT}}(S), \quad S \subseteq S_m \text{ or } S \subseteq S_b.$$

By contrast, any support that contains variables from both groups has two distinct inverse-cost levels. In the whitened coordinates above:

$$\delta_H(S) = \begin{cases} 0, & S \subseteq S_m \text{ or } S \subseteq S_b, \\ (\kappa - 1)/2, & S \cap S_m \neq \emptyset \text{ and } S \cap S_b \neq \emptyset. \end{cases}$$

Thus, as κ grows, the gap between the inverse-cost levels on S_m and S_b grows. On a mixed support, the oracle would shrink the highly manipulable features more aggressively than the baseline features, but scalar ridge has only one shrinkage parameter and cannot match both levels well. Hence, the heterogeneity gap increases with κ . A formal statement for the two-level inverse-cost family is given in Appendix B.9.

Remark 2. Example 2 illustrates why feature selection can become the more important lever when heterogeneity increases. If the highly manipulable group S_m is sufficiently predictive and internally homogeneous, the decision maker may retain it and regularize it aggressively. If its predictive value is small, or if its signal can be substituted by the baseline group S_b , the decision maker may instead exclude S_m .

The main message is that highly manipulable features should not always be dropped; rather, the optimal response depends jointly on their predictiveness, their substitutability by less manipulable features, and the homogeneity of the retained manipulation costs.

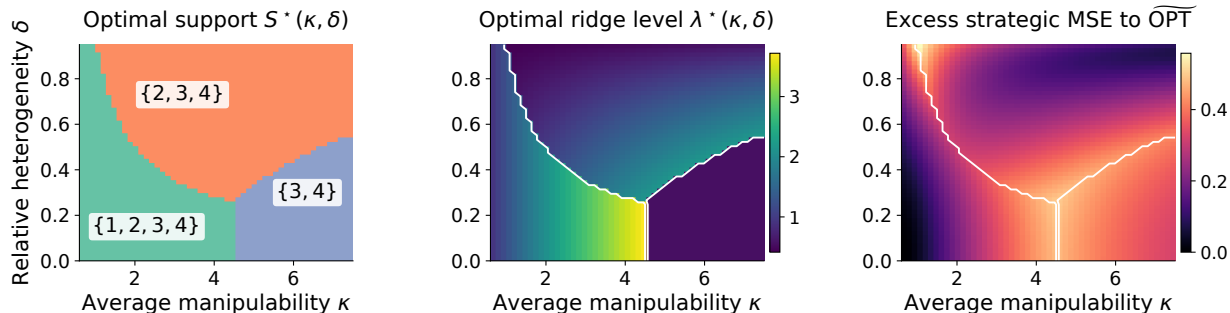


Figure 2: **Manipulable feature groups with homogeneous costs can be retained and regularized.** We compute the optimal support and regularization level exactly for a $d = 4$ example with diagonal costs. Features $\{1, 2\}$ are a manipulable pair with average manipulability κ and relative cost heterogeneity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

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 $\delta_H(S)$ 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 $L_{\text{pred}}(S)$ 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

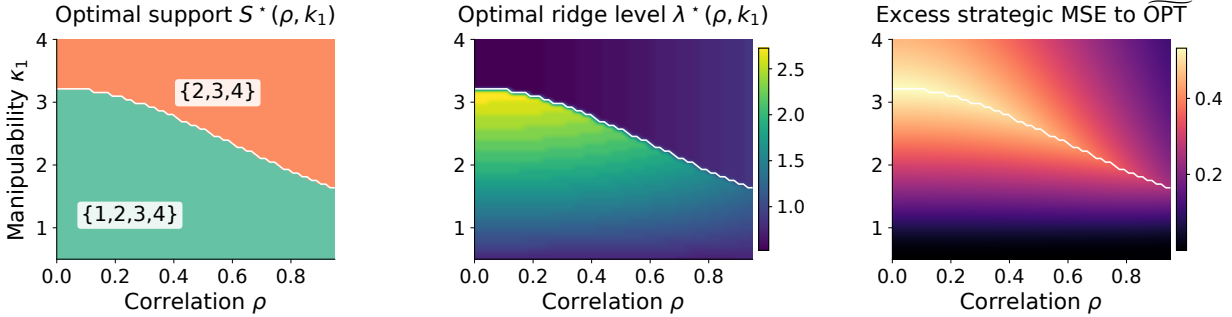


Figure 3: **Less manipulable correlated proxies can replace manipulable features.** We compute the optimal support and regularization level exactly for a $d = 4$ example with correlated features. Feature 1 is predictive but manipulable, feature 2 is a correlated less manipulable proxy for feature 1 with 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 $S^*(\rho, k_1)$ 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 $\widetilde{\text{OPT}}$. 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 $\widetilde{\text{OPT}}$ over a broad region of the parameter space. See Appendix C.3 for details.

manipulable feature can dominate the full support. Figure 3 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 B.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

$$(S^*, \lambda^*) \in \arg \min_{\substack{S \subseteq [d] \\ |S| \leq s}} \min_{\lambda \geq 0} \Phi(S, \lambda), \quad \Phi(S, \lambda) := \text{MSE}_{\text{strat}}(\theta^{\text{OR}}(S, \lambda), 0).$$

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 $\mathbf{1}^\top w \leq s$. For a fixed $\lambda > 0$, define the weighted ridge estimator:

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

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

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

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

Output: $(\hat{\lambda}, \hat{\mathbf{S}}_{\hat{\lambda}}, \hat{\theta}_{\hat{\lambda}})$.

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 $W = \text{diag}(w)$ gives $\theta_\lambda(w) = (\Sigma + \lambda W^{-1})^{-1} \Sigma \theta^*$ (see Lemma D.2 for more details). At fixed λ , we minimize the exact strategic objective $\text{MSE}_{\text{strat}}(\theta_\lambda(w), 0)$ over the weight polytope $\{w \in [0, 1]^d : \mathbf{1}^\top w \leq s\}$.

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

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

$$\Phi_{\lambda, s}^* := \min_{\substack{\mathbf{S} \subseteq [d] \\ |\mathbf{S}| \leq s}} \Phi(\mathbf{S}, \lambda), \quad L_{\lambda, s}^{\text{wt}} := \min_{w \in \mathcal{W}_s} \text{MSE}_{\text{strat}}(\theta_\lambda(w), 0),$$

then $L_{\lambda, s}^{\text{wt}} \leq \Phi_{\lambda, s}^*$ (see Proposition D.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 $\text{Top}_s(w)$ denote the indices of the s largest entries of w . For a current support \mathbf{S} , let $\mathcal{N}_s(\mathbf{S})$ denote the add/drop/swap neighborhood of \mathbf{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.

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 of the form $\theta^{\text{OR}}(\hat{\mathbf{S}}_{\hat{\lambda}}, \hat{\lambda})$, 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 $\Phi(\mathbf{S}, \lambda)$, so the refinement step terminates after finitely many moves and never worsens the rounded support (see Proposition D.2). We provide implementation details and a convex reformulation for the diagonal-covariance case in Appendix D and synthetic benchmarks in Appendix E.

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 (Kronick et al., 2025a). 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>.

6.1 Experimental setup

To simulate realistic upcoding, we build on prior work to first generate diagnosis codes in the absence of coding incentives, and then modify them to match the shifts in diagnosis prevalence associated with more intense coding. We use the `upcoding` package¹, which simulates realistic co-occurring HCCs for the Medicare population using self-reported health conditions from a large national survey that are unaffected by coding incentives (Enache and Rose, 2026). Starting from this baseline data, we upcode selected HCCs among the top-ten diagnosis groups identified by Kronick et al. (2025a) to match their reported prevalence shifts. The baseline HCC data provide a realistic empirical covariance structure for diagnoses, while the induced prevalence shifts provide policy-grounded information about coding vulnerability that we use to construct the manipulation-ease matrix H^{-1} . The resulting baseline dataset contains $n = 5000$ beneficiaries and 115 V28 HCC indicators. Of these, 30 HCCs have nonzero baseline prevalence.² Support sizes in the results that follow are therefore reported over the 30 HCCs with nonzero baseline prevalence.

We use the published CMS-HCC V28 coefficients³ as our signal θ^* and generate outcomes from the baseline HCC covariates. All methods are fit using only baseline, pre-manipulation data, following the MA setting. Evaluation follows the strategic response model studied in this paper: for a fitted coefficient vector $\hat{\theta}$, features at deployment respond as $X \mapsto X^{(0)} + \alpha H^{-1} \hat{\theta}$, where $\alpha \geq 0$ controls manipulation intensity. We report the resulting post-manipulation MSE.

We compare our support-restricted ridge method to the following baselines: full ridge, which tunes ordinary ridge on all HCCs with nonzero baseline prevalence; top-ten exclusion, which removes HCCs with nonzero baseline prevalence that belong to the top-ten diagnosis groups identified by Kronick et al. (2025a) and then tunes ridge on the remaining features; prediction-only, which selects features using predictive value alone and then tunes ridge; cost-only, which selects features using manipulation-ease scores alone and then tunes ridge; and subset-only, which performs support-restricted least squares without ridge shrinkage. We also include the full-support zero-intercept oracle $\widetilde{\text{OPT}}$ (introduced in Section 3.1) as a benchmark. For ease of visualization and comparison across methods, we report post-manipulation MSE normalized by the full-support ridge value at the same manipulation intensity, so values below one indicate an improvement over full-support ridge.

We refer the reader to Appendix F for further details on the data, semi-synthetic outcome construction, cost matrix, baselines, and evaluation metrics.

6.2 Results

Our empirical results match the qualitative findings of our theoretical analysis: low post-manipulation error need not require blanket exclusion, and the best policy designs come from combining feature selection with regularization.

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 $\widetilde{\text{OPT}}$. The left panel of Figure 4 shows that the support-restricted ridge design with $k = 25$ retained features reduces post-manipulation MSE by about 40% relative to full ridge and outperforms the

¹<https://github.com/StanfordHPDS/upcoding>

²This is partly due to export restrictions in the underlying All of Us data, which exclude co-occurring HCC sets reported by only a small number of survey respondents.

³<https://www.cms.gov/files/document/2024-announcement-pdf.pdf>

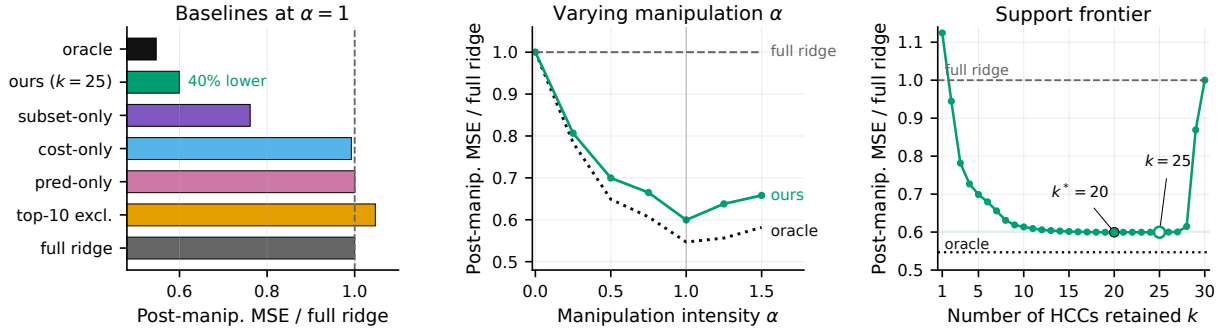


Figure 4: **Feature selection under optimal regularization retains many features while keeping post-manipulation error low.** (Left) Post-manipulation MSE at manipulation intensity $\alpha = 1$ for all methods. (Middle) Post-manipulation MSE under varying α for the oracle $\widetilde{\text{OPT}}$ and 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.

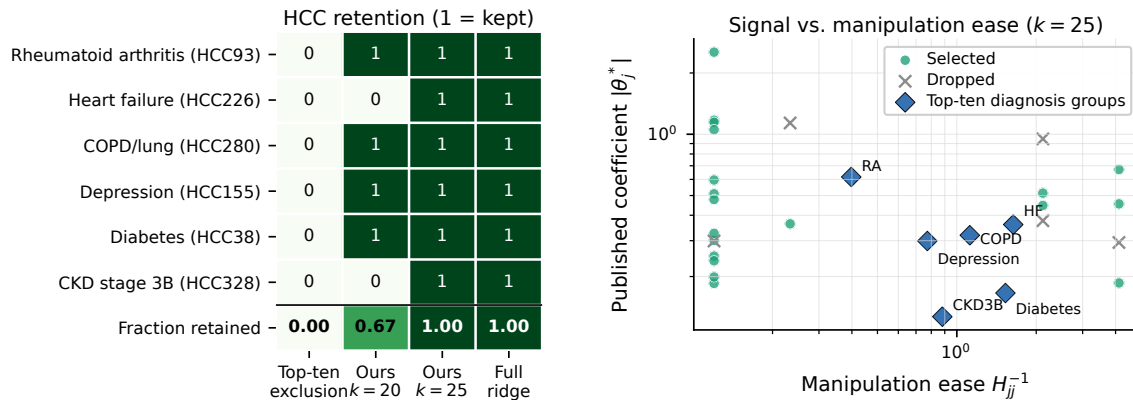


Figure 5: **The selected model reflects the joint predictability-manipulability structure, rather than dropping all intensely coded features.** (Left) HCC groups with nonzero prevalence in the data among the top-ten diagnosis groups identified by [Kronick et al. \(2025a\)](#). The top-ten exclusion policy 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.

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 4 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 4 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 5, 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. (2025a), while excluding other HCCs (see the right panel of Figure 5). 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 5 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 Inherent Robustness under Cost Uncertainty

Our main results assume that the cost matrix H is known. In practice, however, manipulation costs are difficult to estimate precisely. We therefore briefly examine the case where cost matrix H is *unknown*.

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 $b^{\text{OPT}} = -\theta^{*\top} K \theta^*$, achieving the irreducible error σ^2 (see Lemma A.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

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

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 2 (Worst-case error in general and for intercept correction). *Let \mathcal{K} be a compact set of possible inverse-cost matrices and fix (θ, b) . Define $\underline{q}(\theta) := \inf_{K \in \mathcal{K}} \theta^\top K \theta$ and $\bar{q}(\theta) := \sup_{K \in \mathcal{K}} \theta^\top K \theta$. Then*

$$\sup_{K \in \mathcal{K}} \{\text{MSE}_K(\theta, b) - \text{OPT}\} = (\theta - \theta^*)^\top \Sigma (\theta - \theta^*) + \max\{(b + \underline{q}(\theta))^2, (b + \bar{q}(\theta))^2\}. \quad (11)$$

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

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

Calculation of the worst-case error in general (11) 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 3 (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 $\hat{K} = 3/2$. 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, ordinary ridge with $\lambda = 1$ gives $\theta^{\text{OR}}(\lambda) = 1/2$ and $b = 0$. Its worst-case excess strategic MSE is

$$\left(\frac{1}{2} - 1\right)^2 + \sup_{K \in [1/2, 5/2]} \left(K \left(\frac{1}{2}\right)\right)^2 = \frac{1}{4} + \left(\frac{5}{8}\right)^2 = \frac{41}{64} < 1.$$

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

Note that in Example 3 we examine a fixed intercept $b = -\hat{K} = -3/2$ rather than proposing a robust intercept. Nevertheless, Proposition 2 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 $[\underline{q}(\theta), \bar{q}(\theta)]$, yielding residual squared shift $(\bar{q}(\theta) - \underline{q}(\theta))^2/4$. So in this example, we are in fact already comparing against the best fixed intercept model with unknown K .

Example 4 (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:

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

Again, take the nominal estimate to be the midpoint $\hat{K} = K(3/2)$. The plug-in intercept correction with slope θ^* has worst-case excess strategic MSE equal to 1, as in Example 3.

Full-support ridge reduces this error by shrinking the manipulable coordinate. For example, full-support ridge with $\lambda = 1$ has worst-case excess strategic MSE approximately

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

However, support-restricted ridge can do even better by dropping the uncertain coordinate and using the less manipulable proxy. On the support $S = \{2\}$, ridge with $\lambda = 1/4$ gives

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

Since this rule places zero weight on the uncertain coordinate, its strategic shift is independent of k . Its worst-case excess strategic MSE is

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

For comparison, support-restricted least squares on $S = \{2\}$ gives $\theta^{\text{LS}}(\{2\}) = (0, 49/50)$, whose worst-case excess strategic MSE is approximately 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 3 and 4 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.

8 Discussion

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., Patil et al. (2024) 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.

Acknowledgments

We would like to thank Oana M. Enach, 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

- All of Us Research Program Investigators. The “All of Us” research program. *New England Journal of Medicine*, 381(7):668–676, 2019.
- Camilo Arias. Selecting risk adjusters with penalized regression and expert judgment: evidence from Colombia. *Health Economics Review*, 16(1):6, 2025.
- 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.
- Daniel Björkegren, Joshua E. Blumenstock, and Samsun Knight. Manipulation-robust prediction. *American Economic Review*, 2025. Forthcoming.
- Adriana Camacho and Emily Conover. Manipulation of social program eligibility. *American Economic Journal: Economic Policy*, 3(2):41–65, 2011.
- 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/medicareadvtspecratestats/downloads/advance2014.pdf>.
- 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>.
- 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>.
- Yiling Chen, Yang Liu, and Chara Podimata. Learning strategy-aware linear classifiers. In *Advances in Neural Information Processing Systems*, 2020.
- 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.
- Oana M. Enache and Sherri Rose. Time-to-event estimation with unreliably reported events in medicare health plan payment. arXiv: 2602.04092, 2026.
- Michael Geruso and Timothy Layton. Upcoding: Evidence from Medicare on squishy risk adjustment. *Journal of Political Economy*, 128(3):984–1026, 2020.
- 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>.
- Laszlo Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.
- 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.
- Tinashe Handina and Eric Mazumdar. Understanding model selection for learning in strategic environments. In *Advances in Neural Information Processing Systems*, 2024.
- 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.
- Keegan Harris, Dung Daniel T. Ngo, Logan Stapleton, Hoda Heidari, and Zhiwei Steven Wu. Strategic instrumental variable regression: Recovering causal relationships from strategic responses. In *Proceedings of the International Conference on Machine Learning*, 2022.
- 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.

- 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, 2025a.
- 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, 2025b.
- Sagi Levanon and Nir Rosenfeld. Strategic classification made practical. In *Proceedings of the International Conference on Machine Learning*, 2021.
- César Martinelli and Susan Wendy Parker. Deception and misreporting in a social program. *Journal of the European Economic Association*, 7(4):886–908, 2009.
- 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.
- 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.
- 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/>.
- 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.
- John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *Proceedings of the International Conference on Machine Learning*, 2020.
- Paul Niehaus, Antonia Atanassova, Marianne Bertrand, and Sendhil Mullainathan. Targeting with agents. *American Economic Journal: Economic Policy*, 5(1):206–38, 2013.
- 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.
- 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.
- 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.
- Chara Podimata. Incentive-aware machine learning: Robustness, fairness, improvement & causality. arXiv: 2505.05211, 2025.
- Sherri Rose. A machine learning framework for plan payment risk adjustment. *Health Services Research*, 51(6):2358–2374, 2016.
- Yonadav Shavit, Benjamin L. Edelman, and Brian Axelrod. Causal strategic linear regression. In *Proceedings of the International Conference on Machine Learning*, 2020.
- 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>.

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.

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.

A Theoretical Details and Proofs for Section 2

A.1 Strategic optimum characterization

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

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

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

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

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

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

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

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

A.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., Györfi et al. (2002)), whose derivation we collect below for completeness.

Lemma A.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 $\mathbb{E}[X\varepsilon] = 0$. □

A.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 A.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:

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

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

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

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 . \square

Under this best response per Lemma A.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 A.4 (Strategic MSE identity). *For any $\theta \in \mathbb{R}^d$ and $b \in \mathbb{R}$,*

$$\text{MSE}_{\text{strat}}(\theta, b) = (\theta - \theta^*)^\top \Sigma (\theta - \theta^*) + (b + \theta^\top H^{-1}\theta)^2 + \sigma^2. \quad (12)$$

Proof. By Lemma A.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:

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

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

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

as claimed. \square

We can interpret the three terms in the strategic MSE decomposition (12) 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.

A.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 $S \subseteq [d]$, we let $R = [d] \setminus S$. We write $\Sigma_{S_y} := \mathbb{E}[X_S Y]$. Note that since $Y = \theta^{*\top} X + \varepsilon$ and $\mathbb{E}[X\varepsilon] = 0$, we have

$$\Sigma_{S_y} = \Sigma_{SS}\theta_S^* + \Sigma_{SR}\theta_R^*. \quad (13)$$

Lemma A.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 $S \subseteq [d]$, the support-restricted least squares estimator is:

$$\theta^{\text{LS}}(S) = (\theta_S^{\text{LS}}(S), 0_R),$$

where

$$\theta_S^{\text{LS}}(S) = \Sigma_{SS}^{-1}\Sigma_{S_y} = \theta_S^* + \Sigma_{SS}^{-1}\Sigma_{SR}\theta_R^*.$$

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

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

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

$$\theta^{\text{OR}}(S, \lambda) = (\theta_S^{\text{OR}}(S, \lambda), 0_R),$$

where

$$\theta_S^{\text{OR}}(S, \lambda) = (\Sigma_{SS} + \lambda I_{|S|})^{-1}\Sigma_{S_y}.$$

Equivalently,

$$\theta_S^{\text{OR}}(S, \lambda) = (\Sigma_{SS} + \lambda I_{|S|})^{-1}(\Sigma_{SS}\theta_S^* + \Sigma_{SR}\theta_R^*).$$

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

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

The objective is strictly convex because $\Sigma_{SS} \succ 0$. Hence, the solution is unique given by $\theta_S^{\text{LS}}(S) = \Sigma_{SS}^{-1}\Sigma_{S_y}$, and using (13) gives

$$\theta_S^{\text{LS}}(S) = \Sigma_{SS}^{-1}(\Sigma_{SS}\theta_S^* + \Sigma_{SR}\theta_R^*) = \theta_S^* + \Sigma_{SS}^{-1}\Sigma_{SR}\theta_R^*.$$

Ridge. By Lemma A.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

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

Support-restricted ridge. Again write $\theta = (\eta, 0_R)$. The support-restricted ridge problem reduces to

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

Since $\Sigma_{\mathcal{S}\mathcal{S}} + \lambda I_{|\mathcal{S}|} \succ 0$, the solution is unique again and after substituting (13) equals

$$\theta_{\mathcal{S}}^{\text{OR}}(\mathcal{S}, \lambda) = (\Sigma_{\mathcal{S}\mathcal{S}} + \lambda I_{|\mathcal{S}|})^{-1} \Sigma_{\mathcal{S}y} = (\Sigma_{\mathcal{S}\mathcal{S}} + \lambda I_{|\mathcal{S}|})^{-1} (\Sigma_{\mathcal{S}\mathcal{S}} \theta_{\mathcal{S}}^* + \Sigma_{\mathcal{S}R} \theta_R^*),$$

as desired. \square

The formula $\theta_{\mathcal{S}}^{\text{LS}}(\mathcal{S}) = \theta_{\mathcal{S}}^* + \Sigma_{\mathcal{S}\mathcal{S}}^{-1} \Sigma_{\mathcal{S}R} \theta_R^*$ 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 A.1 (Nesting of the policy levers). *The support-restricted ordinary ridge estimator satisfies $\theta^{\text{OR}}(\mathcal{S}, 0) = \theta^{\text{LS}}(\mathcal{S})$ and, for $\mathcal{S} = [d]$, $\theta^{\text{OR}}([d], \lambda) = \theta^{\text{OR}}(\lambda)$.*

Proof. Both identities follow immediately from the closed forms in Lemma A.5. \square

B Theoretical Details and Proofs for Section 3

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

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

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

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

and

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

B.1 Proof of Theorem 1

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

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

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

$$\widetilde{\text{OPT}}(\mathcal{S}) \geq \widetilde{\text{OPT}}.$$

Combining the two inequalities gives

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

Subtracting $\widetilde{\text{OPT}}$ from both sides and minimizing over (\mathcal{S}, λ) yields

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

This proves the claimed lower bound.

B.2 Proof of Proposition B.1

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

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

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

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

B.3 Proof of Proposition B.2

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

$$L_{\text{pred}}(S) := (\theta_R^*)^\top \Sigma_{R|S} \theta_R^*, \quad \Sigma_{R|S} := \Sigma_{RR} - \Sigma_{RS} \Sigma_{SS}^{-1} \Sigma_{SR},$$

and the support-restricted strategic burden

$$\Gamma(S) := \min_{\eta \in \mathbb{R}^{|S|}} \{(\eta - \theta_S^{\text{LS}}(S))^\top \Sigma_{SS} (\eta - \theta_S^{\text{LS}}(S)) + (\eta^\top H_{SS}^{-1} \eta)^2\}. \quad (14)$$

Then

$$\widetilde{\text{OPT}}(S) - \widetilde{\text{OPT}} = L_{\text{pred}}(S) - (\Gamma([d]) - \Gamma(S)). \quad (15)$$

We first record a useful support-restricted decomposition.

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

$$\theta_S^{\text{LS}}(S) = \Sigma_{SS}^{-1} \Sigma_{Sy} = \theta_S^* + \Sigma_{SS}^{-1} \Sigma_{SR} \theta_R^*.$$

Then

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

where

$$L_{\text{pred}}(S) = (\theta_R^*)^\top \Sigma_{R|S} \theta_R^*, \quad \Sigma_{R|S} = \Sigma_{RR} - \Sigma_{RS} \Sigma_{SS}^{-1} \Sigma_{SR}.$$

Proof. Let $z := \eta - \theta_S^*$. Since $\theta = (\eta, 0_R)$, we have $\theta - \theta^* = (z, -\theta_R^*)$. Expanding the quadratic form blockwise gives

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

Now define $z_0 := \eta - \theta_S^{\text{LS}}(S) = z - \Sigma_{SS}^{-1} \Sigma_{SR} \theta_R^*$. Completing the square yields

$$z^\top \Sigma_{SS} z - 2z^\top \Sigma_{SR} \theta_R^* = z_0^\top \Sigma_{SS} z_0 - (\theta_R^*)^\top \Sigma_{RS} \Sigma_{SS}^{-1} \Sigma_{SR} \theta_R^*.$$

Substituting this into the previous display gives

$$(\theta - \theta^*)^\top \Sigma (\theta - \theta^*) = z_0^\top \Sigma_{SS} z_0 + (\theta_R^*)^\top (\Sigma_{RR} - \Sigma_{RS} \Sigma_{SS}^{-1} \Sigma_{SR}) \theta_R^*,$$

which is the claimed identity. \square

Proof of Proposition B.2. Fix $S \subseteq [d]$ and write $R = [d] \setminus S$. For any $\theta = (\eta, 0_R) \in \Theta(S)$, Lemma B.1 gives

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

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

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

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

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

where

$$\Gamma(S) = \min_{\eta \in \mathbb{R}^{|S|}} \{(\eta - \theta_S^{\text{LS}}(S))^\top \Sigma_{SS}(\eta - \theta_S^{\text{LS}}(S)) + (\eta^\top H_{SS}^{-1}\eta)^2\}.$$

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

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

Subtracting gives

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

as claimed. \square

B.4 Proof of Theorem B.1

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

$$0 \leq \min_{\lambda \geq 0} \left\{ \text{MSE}_{\text{strat}}(\theta^{\text{OR}}(S, \lambda)) - \widetilde{\text{OPT}}(S) \right\} \leq C_H(S) \delta_H(S)^2,$$

where $C_H(S) := 4 L_H(S) \|\Sigma_{SS}^{-1/2} H_{SS}^{-1} \Sigma_{SS}^{-1/2}\|_{\text{op}}^2 \|\Sigma_{SS}^{1/2} \theta_S^{\text{LS}}(S)\|_2^6$, with $L_H(S) := 1 + 6 \|\Sigma_{SS}^{-1/2} H_{SS}^{-1} \Sigma_{SS}^{-1/2}\|_{\text{op}}^2 \|\Sigma_{SS}^{1/2} \theta_S^{\text{LS}}(S)\|_2^2$.

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

$$\eta_S^* \in \arg \min_{\eta \in \mathbb{R}^{|S|}} \{(\eta - \theta_S^{\text{LS}}(S))^\top \Sigma_{SS}(\eta - \theta_S^{\text{LS}}(S)) + (\eta^\top H_{SS}^{-1}\eta)^2\}.$$

For $\lambda \geq 0$, let $\theta_S^{\text{OR}}(\lambda)$ denote the retained-coordinate vector of $\theta^{\text{OR}}(S, \lambda)$:

$$\theta_S^{\text{OR}}(\lambda) = (\Sigma_{SS} + \lambda I)^{-1} \Sigma_{SS} \theta_S^{\text{LS}}(S).$$

Lemma B.2 (Support-restricted oracle is generalized ridge). *The minimizer η_S^* is unique. Moreover, defining $\alpha_S := 2 \eta_S^{*\top} H_{SS}^{-1} \eta_S^*$, we have*

$$(\Sigma_{SS} + \alpha_S H_{SS}^{-1}) \eta_S^* = \Sigma_{SS} \theta_S^{\text{LS}}(S).$$

Equivalently,

$$\eta_S^* = (\Sigma_{SS} + \alpha_S H_{SS}^{-1})^{-1} \Sigma_{SS} \theta_S^{\text{LS}}(S).$$

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

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

Thus the objective is strongly convex and has a unique minimizer. The first-order condition at η_S^* is

$$2\Sigma_{SS}(\eta_S^* - \theta_S^{\text{LS}}(S)) + 4(\eta_S^{*\top} H_{SS}^{-1} \eta_S^*) H_{SS}^{-1} \eta_S^* = 0.$$

Dividing by 2 and using the definition of α_S gives

$$\Sigma_{SS}(\eta_S^* - \theta_S^{\text{LS}}(S)) + \alpha_S H_{SS}^{-1} \eta_S^* = 0.$$

Equivalently,

$$(\Sigma_{SS} + \alpha_S H_{SS}^{-1}) \eta_S^* = \Sigma_{SS} \theta_S^{\text{LS}}(S).$$

□

We use whitened coordinates on the support. Define

$$u_S^* := \Sigma_{SS}^{1/2} \eta_S^*, \quad u_S^{\text{LS}} := \Sigma_{SS}^{1/2} \theta_S^{\text{LS}}(S), \quad u_{\lambda,S} := \Sigma_{SS}^{1/2} \theta_S^{\text{OR}}(\lambda).$$

The support-restricted oracle satisfies

$$\left(I + \alpha_S \Sigma_{SS}^{-1/2} H_{SS}^{-1} \Sigma_{SS}^{-1/2} \right) u_S^* = u_S^{\text{LS}},$$

while the ordinary ridge path satisfies

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

Lemma B.3 (Distance from the oracle to the ordinary ridge path). *For every $\gamma \geq 0$, if $\lambda = \alpha_S \gamma$, then*

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

Consequently,

$$\inf_{\lambda \geq 0} \|u_{\lambda,S} - u_S^*\|_2 \leq \alpha_S \delta_H(S) \|u_S^*\|_2.$$

Proof. Subtracting the whitened normal equations gives

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

Therefore

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

Choosing $\lambda = \alpha_S \gamma$ gives

$$u_{\alpha_S \gamma, S} - u_S^* = \alpha_S (I + \alpha_S \gamma \Sigma_{SS}^{-1})^{-1} \Sigma_{SS}^{-1/2} (H_{SS}^{-1} - \gamma I) \Sigma_{SS}^{-1/2} u_S^*.$$

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

$$\|(I + \alpha_S \gamma \Sigma_{SS}^{-1})^{-1}\|_{\text{op}} \leq 1.$$

Thus

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

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

$$\inf_{\lambda \geq 0} \|u_{\lambda,S} - u_S^*\|_2 \leq \alpha_S \delta_H(S) \|u_S^*\|_2.$$

□

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

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

Then

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

for every $\lambda \geq 0$.

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

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

The Hessian of this function is

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

Hence its operator norm is at most

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

Both $u_{\mathbf{S}}^*$ and $u_{\lambda, \mathbf{S}}$ lie in the ball $\|u\|_2 \leq \|u_{\mathbf{S}}^{\text{LS}}\|_2$. Indeed,

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

and

$$u_{\lambda, \mathbf{S}} = \left(I + \lambda \Sigma_{\text{SS}}^{-1} \right)^{-1} u_{\mathbf{S}}^{\text{LS}},$$

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

$$\text{MSE}_{\text{strat}}(\theta^{\text{OR}}(\mathbf{S}, \lambda)) - \widetilde{\text{OPT}}(\mathbf{S}) \leq L_H(\mathbf{S}) \|u_{\lambda, \mathbf{S}} - u_{\mathbf{S}}^*\|_2^2.$$

□

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

$$\min_{\lambda \geq 0} \left\{ \text{MSE}_{\text{strat}}(\theta^{\text{OR}}(\mathbf{S}, \lambda)) - \widetilde{\text{OPT}}(\mathbf{S}) \right\} \leq \alpha_{\mathbf{S}}^2 L_H(\mathbf{S}) \delta_H(\mathbf{S})^2 \|u_{\mathbf{S}}^*\|_2^2.$$

Now

$$\|u_{\mathbf{S}}^*\|_2 \leq \|u_{\mathbf{S}}^{\text{LS}}\|_2 = \left\| \Sigma_{\text{SS}}^{1/2} \theta_{\mathbf{S}}^{\text{LS}}(\mathbf{S}) \right\|_2.$$

Also,

$$\alpha_{\mathbf{S}} = 2\eta_{\mathbf{S}}^{*\top} H_{\text{SS}}^{-1} \eta_{\mathbf{S}}^* = 2u_{\mathbf{S}}^{*\top} \Sigma_{\text{SS}}^{-1/2} H_{\text{SS}}^{-1} \Sigma_{\text{SS}}^{-1/2} u_{\mathbf{S}}^*,$$

so

$$\alpha_{\mathbf{S}} \leq 2 \left\| \Sigma_{\text{SS}}^{-1/2} H_{\text{SS}}^{-1} \Sigma_{\text{SS}}^{-1/2} \right\|_{\text{op}} \|u_{\mathbf{S}}^*\|_2^2 \leq 2 \left\| \Sigma_{\text{SS}}^{-1/2} H_{\text{SS}}^{-1} \Sigma_{\text{SS}}^{-1/2} \right\|_{\text{op}} \|u_{\mathbf{S}}^{\text{LS}}\|_2^2.$$

Therefore

$$\alpha_{\mathbf{S}}^2 \|u_{\mathbf{S}}^*\|_2^2 \leq 4 \left\| \Sigma_{\text{SS}}^{-1/2} H_{\text{SS}}^{-1} \Sigma_{\text{SS}}^{-1/2} \right\|_{\text{op}}^2 \|u_{\mathbf{S}}^{\text{LS}}\|_2^6.$$

Combining the displays gives

$$\min_{\lambda \geq 0} \left\{ \text{MSE}_{\text{strat}}(\theta^{\text{OR}}(\mathbf{S}, \lambda)) - \widetilde{\text{OPT}}(\mathbf{S}) \right\} \leq C_H(\mathbf{S}) \delta_H(\mathbf{S})^2,$$

with the valid choice

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

□

B.5 Proof of Theorem 2

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

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

By Proposition B.2,

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

By Theorem B.1,

$$\min_{\lambda \geq 0} \left\{ \text{MSE}_{\text{strat}}(\theta^{\text{OR}}(S, \lambda)) - \widetilde{\text{OPT}}(S) \right\} \leq C_H(S) \delta_H(S)^2.$$

Therefore, for every S ,

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

B.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 B.2.

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

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

Equivalently, defining

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

we have

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

and hence

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

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

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

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

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

Its gradient is

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

and its Hessian is

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

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

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

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

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

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

Dividing by 2 and defining

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

yields

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

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

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

□

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

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

Proof of Corollary B.1. Let η_S^* denote the support-restricted zero-intercept oracle on S . By Lemma B.2, there exists $\alpha_S := 2\eta_S^{*\top} H_{SS}^{-1} \eta_S^*$ such that

$$(\Sigma_{SS} + \alpha_S H_{SS}^{-1}) \eta_S^* = \Sigma_{SS} \theta_S^{\text{LS}}(S).$$

If $H_{SS}^{-1} = \gamma I$, then

$$(\Sigma_{SS} + \alpha_S \gamma I) \eta_S^* = \Sigma_{SS} \theta_S^{\text{LS}}(S).$$

Thus η_S^* is exactly the ordinary ridge estimator on support S with ridge parameter $\lambda = \alpha_S \gamma$. Therefore the ordinary ridge path contains the support-restricted zero-intercept oracle, and so

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

□

Corollary B.1 shows that scalar ridge is exact when the retained manipulation costs are isotropic. This condition is sufficient but not necessary. Proposition B.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.

B.8 Necessary and sufficient condition for ridge optimality

The isotropy condition in Corollary B.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 B.4 (Necessary and sufficient condition for exact fixed-support ridge optimality). *Fix a support $S \subseteq [d]$. Let η_S^* denote the support-restricted zero-intercept oracle on S . Then*

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

if and only if there exists $\gamma \geq 0$ such that $(H_{SS}^{-1} - \gamma I) \eta_S^ = 0$.*

Proof. The support-restricted oracle η_S^* satisfies

$$(\Sigma_{SS} + \alpha_S H_{SS}^{-1})\eta_S^* = \Sigma_{SS}\theta_S^{\text{LS}}(\mathbf{S}),$$

where $\alpha_S = 2\eta_S^{*\top} H_{SS}^{-1} \eta_S^*$. The ordinary ridge estimator at level λ satisfies

$$(\Sigma_{SS} + \lambda I)\theta_S^{\text{OR}}(\lambda) = \Sigma_{SS}\theta_S^{\text{LS}}(\mathbf{S}).$$

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 $\theta_S^{\text{OR}}(\lambda) = \eta_S^*$. Subtracting the two normal equations gives $(\alpha_S H_{SS}^{-1} - \lambda I)\eta_S^* = 0$. If $\alpha_S > 0$, set $\gamma = \lambda/\alpha_S$. Then $(H_{SS}^{-1} - \gamma I)\eta_S^* = 0$. If $\alpha_S = 0$, then $\eta_S^* = 0$ because $H_{SS}^{-1} \succ 0$. In this case, $(H_{SS}^{-1} - \gamma I)\eta_S^* = 0$ holds for every $\gamma \geq 0$.

Conversely, suppose there exists $\gamma \geq 0$ such that $(H_{SS}^{-1} - \gamma I)\eta_S^* = 0$. If $\eta_S^* = 0$, then the first-order condition implies $\Sigma_{SS}\theta_S^{\text{LS}}(\mathbf{S}) = 0$, and hence $\theta_S^{\text{LS}}(\mathbf{S}) = 0$. Ordinary ridge also returns 0 for every $\lambda \geq 0$, so the ridge gap is zero.

If $\eta_S^* \neq 0$, then $\alpha_S > 0$. Set $\lambda = \alpha_S \gamma$. Then

$$(\Sigma_{SS} + \lambda I)\eta_S^* = (\Sigma_{SS} + \alpha_S \gamma I)\eta_S^* = (\Sigma_{SS} + \alpha_S H_{SS}^{-1})\eta_S^* = \Sigma_{SS}\theta_S^{\text{LS}}(\mathbf{S}).$$

Therefore η_S^* is the ordinary ridge estimator at level λ , so ordinary ridge attains $\widetilde{\text{OPT}}(\mathbf{S})$. \square

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

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

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

Proposition B.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 $\mathbf{T} \subseteq [d]$,*

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

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

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

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

$$\delta_H(\mathbf{T}) = \inf_{\gamma \geq 0} \|H_{\mathbf{T}\mathbf{T}}^{-1} - \gamma I\|_{\text{op}}.$$

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

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

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

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 B.1. It also follows directly from Proposition B.5, since $H_{\mathsf{T}\mathsf{T}}^{-1}$ is a scalar multiple of the identity on such supports. \square

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

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

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

then

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

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

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

Therefore

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

Using the support restriction identity from Proposition B.2,

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

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. \square

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.

B.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

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

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

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

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

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

Proof. The derivative of the ridge path is

$$\frac{d}{d\lambda} \theta_S^{\text{OR}}(\lambda) = -(\Sigma_{\text{SS}} + \lambda I)^{-1} \theta_S^{\text{OR}}(\lambda).$$

Thus

$$\frac{d}{d\lambda} \theta_S^{\text{OR}}(0) = -\Sigma_{\text{SS}}^{-1} \theta_S^{\text{LS}}.$$

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

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

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

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

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

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

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

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

Proof. Let

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

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

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

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

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

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

$$(\Sigma_{\text{SS}} + \lambda I) \theta_S^{\text{OR}}(\lambda) = \Sigma_{\text{SS}} \theta_S^{\text{LS}},$$

we have

$$\theta_S^{\text{OR}}(\lambda) - \theta_S^{\text{LS}} = -\lambda \Sigma_{\text{SS}}^{-1} \theta_S^{\text{OR}}(\lambda).$$

Also,

$$\frac{d}{d\lambda}\theta_S^{\text{OR}}(\lambda) = -(\Sigma_{\text{SS}} + \lambda I)^{-1}\theta_S^{\text{OR}}(\lambda).$$

Hence

$$G'(\lambda) = 2(\theta_S^{\text{OR}}(\lambda) - \theta_S^{\text{LS}})^\top \Sigma_{\text{SS}} \frac{d}{d\lambda}\theta_S^{\text{OR}}(\lambda) = 2\lambda \theta_S^{\text{OR}}(\lambda)^\top (\Sigma_{\text{SS}} + \lambda I)^{-1}\theta_S^{\text{OR}}(\lambda).$$

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

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

Therefore

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

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

C Simulation Details and Illustrations for Section 4

C.1 Details for Figure 1

We consider the following $d = 4$ example:

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

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 $S \subseteq [4]$. For each support, we compute the support-restricted zero-intercept oracle value

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

and the best support-restricted ridge value

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

by one-dimensional optimization over λ . The plotted quantities are predictive loss $L_{\text{pred}}(S)$, negative manipulability gain $-(\Gamma([d]) - \Gamma(S))$, the heterogeneity gap $C_H(S)\delta_H(S)^2$, and total excess, given by the sum of these three quantities. The support restriction identity in Proposition B.2 gives the support restriction cost $\widetilde{\text{OPT}}(S) - \widetilde{\text{OPT}} = L_{\text{pred}}(S) - (\Gamma([d]) - \Gamma(S))$.

The left panel illustrates five representative supports. The full support $\{1, 2, 3, 4\}$ has zero support restriction cost but a large heterogeneity gap 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 heterogeneity gap, 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 heterogeneity gap 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 heterogeneity gap.

The right panel plots every support by support restriction cost and heterogeneity gap. 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 heterogeneity gap is not monotone in support size: adding features can either improve or worsen the heterogeneity gap, depending on whether the added features make the retained inverse-cost geometry more heterogeneous.

C.2 Details for Figure 2

We consider the following $d = 4$ example:

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

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

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

and the exact support-restricted zero-intercept oracle

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

We also compute the full-support zero-intercept oracle

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

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

$$\theta(q) = (\Sigma + 2qH^{-1})^{-1}\Sigma\theta^*, \quad q = \theta(q)^\top H^{-1}\theta(q). \quad (16)$$

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

$$(\mathsf{S}^*, \lambda^*) \in \arg \min_{\mathsf{S} \subseteq [4], \lambda \geq 0} \text{MSE}_{\text{strat}}(\theta^{\text{OR}}(\mathsf{S}, \lambda), 0).$$

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

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.

C.3 Details for Figure 3

We consider the following $d = 4$ example:

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

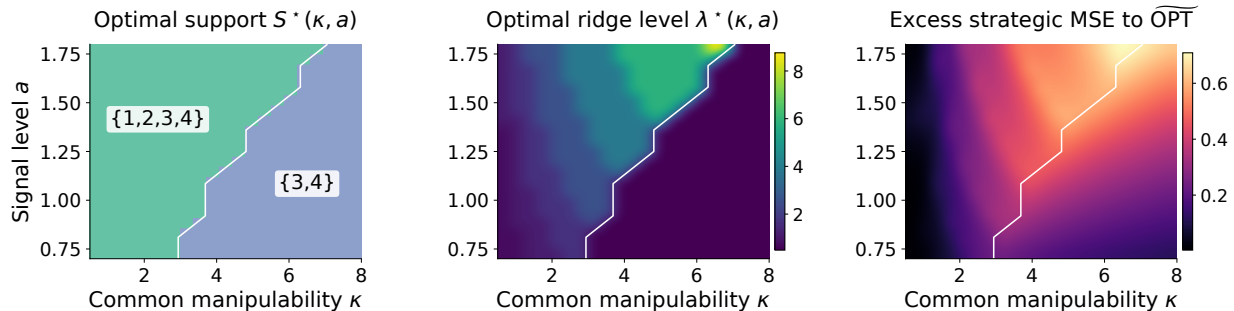


Figure 6: **A homogeneous manipulable pair can be retained and regularized when sufficiently predictive.** We compute the optimal support and regularization level for $d = 4$ features with diagonal costs. Features $\{1, 2\}$ are a homogeneous manipulable pair with common signal level a and common manipulability κ ; features 3 and 4 have low manipulability. (Left) The optimal support $S^*(\kappa, a)$ 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 $\widetilde{\text{OPT}}$. 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 $\widetilde{\text{OPT}}$ over a broad range of the parameter space.

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 $S \subseteq [4]$. For each support, we compute the exact support-restricted ridge optimum

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

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

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

is computed using the generalized-ridge fixed-point equation (16) described above.

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

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.

C.4 Predictiveness versus manipulability with homogeneous costs

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

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

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 $(\mathbf{S}^*, \lambda^*)$ and evaluate the resulting strategic MSE. To assess near-optimality, we compute the full-support zero-intercept oracle

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

Because Σ and H^{-1} are diagonal here, the oracle can be computed from the scalar fixed-point equation (16):

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

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

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

D Theoretical Details and Proofs for Section 5

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

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

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

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

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

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

D.1 Proof of Proposition 1

Let $w \in \{0, 1\}^d$ and define $\mathbf{S}(w) = \{i \in [d] : w_i = 1\}$. 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 $\text{supp}(\theta) \subseteq \mathbf{S}(w)$. On coordinates $i \in \mathbf{S}(w)$, we have $w_i = 1$, so $\varphi(\theta_i, w_i) = \theta_i^2$. Therefore the weighted ridge problem becomes:

$$\arg \min_{\theta \in \mathbb{R}^d : \text{supp}(\theta) \subseteq \mathbf{S}(w)} \{ \mathbb{E}[(Y - \theta^\top X)^2] + \lambda \|\theta\|_2^2 \}.$$

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

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

D.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 $L_{\lambda, s}^{\text{wt}} := \min_{w \in \mathcal{W}_s} \text{MSE}_{\text{strat}}(\theta_\lambda(w))$. 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 D.1 (Weighted relaxation lower bound). *For every $\lambda > 0$ and support budget s , $L_{\lambda, s}^{\text{wt}} \leq \Phi_{\lambda, s}^*$.*

Proof. By Proposition 1, $\Phi_{\lambda, s}^* = \min_{\substack{w \in \{0, 1\}^d \\ \mathbf{1}^\top w \leq s}} \text{MSE}_{\text{strat}}(\theta_\lambda(w))$. 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:

$$L_{\lambda, s}^{\text{wt}} = \min_{w \in \mathcal{W}_s} \text{MSE}_{\text{strat}}(\theta_\lambda(w)) \leq \Phi_{\lambda, s}^*.$$

□

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

$$0 \leq \Phi(\widehat{\mathcal{S}}_\lambda, \lambda) - \Phi_{\lambda, s}^* \leq \Phi(\widehat{\mathcal{S}}_\lambda, \lambda) - L_{\lambda, s}^{\text{wt}}.$$

More generally, the same bound holds if $L_{\lambda, s}^{\text{wt}}$ is replaced by any certified lower bound on $\Phi_{\lambda, s}^$.*

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

$$0 \leq \Phi(\widehat{\mathcal{S}}_\lambda, \lambda) - \Phi_{\lambda, s}^*.$$

By Proposition D.1, $L_{\lambda, s}^{\text{wt}} \leq \Phi_{\lambda, s}^*$. Therefore

$$\Phi(\widehat{\mathcal{S}}_\lambda, \lambda) - \Phi_{\lambda, s}^* \leq \Phi(\widehat{\mathcal{S}}_\lambda, \lambda) - L_{\lambda, s}^{\text{wt}}.$$

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

D.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 $\text{Top}_s(w)$ 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 $\Phi(\mathcal{S}, \lambda)$.

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

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

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

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

Proof. Each accepted move replaces the current support \mathcal{S} by a support \mathcal{T} satisfying $\Phi(\mathcal{T}, \lambda) < \Phi(\mathcal{S}, \lambda)$. Thus the exact fixed- λ 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: $\Phi(\widehat{\mathcal{S}}_\lambda, \lambda) \leq \Phi(\mathcal{S}^{(0)}, \lambda)$. At termination, there is no neighbor $\mathcal{T} \in \mathcal{N}_s(\widehat{\mathcal{S}}_\lambda)$ with $\Phi(\mathcal{T}, \lambda) < \Phi(\widehat{\mathcal{S}}_\lambda, \lambda)$, which is precisely the stated local optimality condition. □

D.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: $\Sigma = \text{diag}(v_1, \dots, v_d)$ with $v_i > 0$ for all $i \in [d]$. Write $\theta^* = (\theta_1^*, \dots, \theta_d^*)$. Fix $\lambda > 0$ and recall that $\mathcal{W}_s = \{w \in [0, 1]^d : \mathbf{1}^\top w \leq s\}$.

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

$$\theta_{\lambda,i}(w) = \frac{v_i w_i}{v_i w_i + \lambda} \theta_i^*, \quad i = 1, \dots, d. \quad (17)$$

This motivates the reparameterization

$$r_i := \frac{v_i w_i}{v_i w_i + \lambda}, \quad i = 1, \dots, d. \quad (18)$$

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 (17) becomes $\theta_{\lambda,i}(w) = \theta_i^* r_i$.

The change of variables in (18) is one-to-one for $\lambda > 0$, with inverse

$$w_i = \frac{\lambda r_i}{v_i(1 - r_i)}, \quad 0 \leq r_i \leq \frac{v_i}{v_i + \lambda}. \quad (19)$$

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

$$\sum_{i=1}^d \frac{\lambda r_i}{v_i(1 - r_i)} \leq s. \quad (20)$$

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

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

$$\begin{aligned} \min_{r \in \mathbb{R}^d} \quad & \sum_{i=1}^d v_i (\theta_i^*)^2 (1 - r_i)^2 + \left((\theta^* \odot r)^\top H^{-1} (\theta^* \odot r) \right)^2 \\ \text{subject to} \quad & 0 \leq r_i \leq \frac{v_i}{v_i + \lambda}, \quad i = 1, \dots, d, \\ & \sum_{i=1}^d \frac{\lambda r_i}{v_i(1 - r_i)} \leq s, \end{aligned} \quad (21)$$

where \odot denotes coordinatewise multiplication. Moreover, (21) is a convex optimization problem.

Proposition D.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 (20) plays the role of a soft support-size constraint.

When H is also diagonal, say $H = \text{diag}(h_1, \dots, h_d)$ 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 D.4 and D.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 $\Sigma = \text{diag}(v_1, \dots, v_d)$ with $v_i > 0$. For $w \in \mathcal{W}_s$, the weighted ridge problem decouples coordinate-wise:

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

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

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

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 $\mathbf{1}^\top w \leq s$ is equivalent to

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

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

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

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 $r \mapsto ((\theta^* \odot r)^\top H^{-1} (\theta^* \odot r))^2$ 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 $0 \leq r_i < 1$, 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 $r_i \leq v_i/(v_i + \lambda) < 1$, the budget constraint is convex. Thus the entire problem is a convex optimization problem. \square

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

$$\theta_S^{\text{LS}}(S) = \Sigma_{SS}^{-1} \Sigma_{Sy}, \quad \theta_S^{\text{OR}}(\lambda) = (\Sigma_{SS} + \lambda I)^{-1} \Sigma_{Sy}.$$

Then

$$\Phi(S, \lambda) = \sigma^2 + L_{\text{pred}}(S) + \lambda^2 \theta_S^{\text{OR}}(\lambda)^\top \Sigma_{SS}^{-1} \theta_S^{\text{OR}}(\lambda) + (\theta_S^{\text{OR}}(\lambda)^\top H_{SS}^{-1} \theta_S^{\text{OR}}(\lambda))^2.$$

Proof. By Lemma B.1, for every $\theta = (\eta, 0_R) \in \Theta(S)$,

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

Substituting $\eta = \theta_S^{\text{OR}}(\lambda)$ gives

$$\Phi(S, \lambda) = \sigma^2 + L_{\text{pred}}(S) + (\theta_S^{\text{OR}}(\lambda) - \theta_S^{\text{LS}}(S))^\top \Sigma_{SS} (\theta_S^{\text{OR}}(\lambda) - \theta_S^{\text{LS}}(S)) + (\theta_S^{\text{OR}}(\lambda)^\top H_{SS}^{-1} \theta_S^{\text{OR}}(\lambda))^2.$$

The ridge normal equation is

$$(\Sigma_{SS} + \lambda I) \theta_S^{\text{OR}}(\lambda) = \Sigma_{SS} \theta_S^{\text{LS}}(S).$$

Equivalently,

$$\theta_S^{\text{OR}}(\lambda) - \theta_S^{\text{LS}}(S) = -\lambda \Sigma_{SS}^{-1} \theta_S^{\text{OR}}(\lambda).$$

Therefore

$$(\theta_S^{\text{OR}}(\lambda) - \theta_S^{\text{LS}}(S))^\top \Sigma_{SS} (\theta_S^{\text{OR}}(\lambda) - \theta_S^{\text{LS}}(S)) = \lambda^2 \theta_S^{\text{OR}}(\lambda)^\top \Sigma_{SS}^{-1} \theta_S^{\text{OR}}(\lambda).$$

This proves the claim. \square

D.6 Closed form for the weighted ridge

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

D.7 Exact set function in the diagonal case

Assume

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

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

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

Define

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

Then define

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

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

$$\Phi_\lambda(\mathbf{S}) = \Phi_\lambda(\emptyset) - G_\lambda(\mathbf{S}),$$

where

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

Consequently,

$$\arg \min_{\substack{\mathbf{S} \subseteq [d] \\ |\mathbf{S}| \leq s}} \Phi_\lambda(\mathbf{S}) = \arg \max_{\substack{\mathbf{S} \subseteq [d] \\ |\mathbf{S}| \leq s}} G_\lambda(\mathbf{S}).$$

Proof. Let

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

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

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

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

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

Since

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

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

The strategic term is

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

Thus

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

which is equivalent to the stated identity. \square

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

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

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

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

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

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

D.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,

$$r_i = \lambda \frac{\theta_i}{w_i}, \quad i = 1, \dots, d,$$

or

$$\lambda\theta_i = w_i r_i, \quad i = 1, \dots, d.$$

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

$$\begin{aligned} \min_{\theta, r, w} \quad & (\theta - \theta^*)^\top \Sigma (\theta - \theta^*) + (\theta^\top H^{-1} \theta)^2 + \sigma^2 \\ \text{s.t.} \quad & r = \Sigma(\theta^* - \theta), \\ & \lambda\theta_i = w_i r_i, \quad i = 1, \dots, d, \\ & 0 \leq w_i \leq 1, \quad i = 1, \dots, d, \\ & \mathbf{1}^\top w \leq s. \end{aligned} \tag{22}$$

Proof. For any w with strictly positive coordinates, the constraints $r = \Sigma(\theta^* - \theta)$ and $\lambda\theta_i = w_i r_i$ for $i = 1, \dots, d$, are exactly the weighted ridge normal equations. Therefore the feasible θ associated with w is precisely $\theta_\lambda(w)$. The objective is exactly $\text{MSE}_{\text{strat}}(\theta)$. 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. \square

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:

$$\lambda\theta_i \geq -R_i w_i, \tag{23}$$

$$\lambda\theta_i \leq R_i w_i, \tag{24}$$

$$\lambda\theta_i \geq r_i - R_i(1 - w_i), \tag{25}$$

$$\lambda\theta_i \leq r_i + R_i(1 - w_i). \tag{26}$$

E 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.

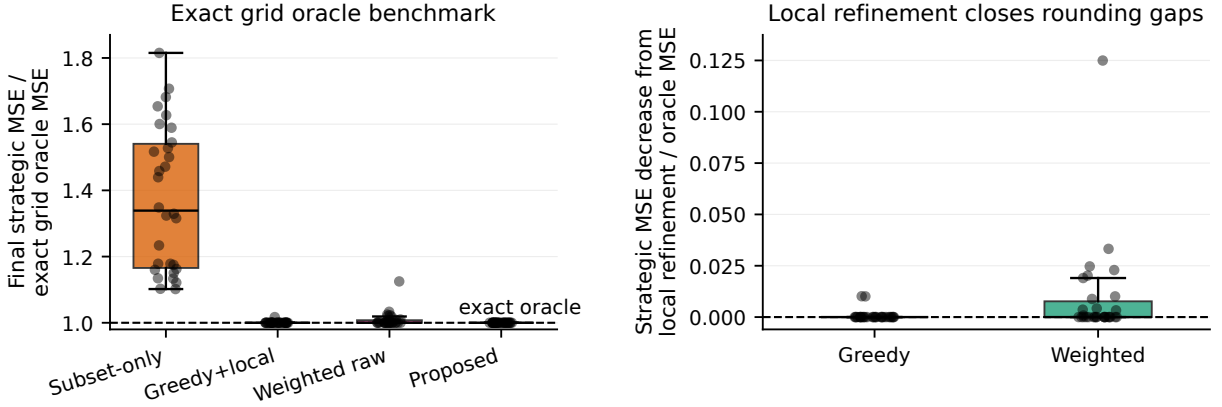


Figure 7: **Low-dimensional synthetic benchmarks validate our proposed weighted screen-and-refit method.** (Left) Final strategic MSE normalized by the exact grid oracle MSE on the same instance, $\text{MSE}_{\text{strat}}(\hat{\theta})/\text{MSE}_{\text{strat}}(\theta^{\text{OR}}(\mathbf{S}^*, \lambda^*))$. 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.

Problem size and oracle. All instances use

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

For each instance, the exact grid oracle is

$$(\mathbf{S}^*, \lambda^*) \in \arg \min_{\substack{S \subseteq [d] \\ |S| \leq s}} \min_{\lambda \in \Lambda} \Phi(\mathbf{S}, \lambda), \quad \Phi(\mathbf{S}, \lambda) = \text{MSE}_{\text{strat}}(\theta^{\text{OR}}(\mathbf{S}, \lambda)).$$

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 $z_j, \varepsilon_j, \eta_j \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1)$ and set $\theta_j^* = |0.95 + 0.60z_j + 0.20\varepsilon_j| + 0.05$, with $H_{jj}^{-1} = \exp(0.95z_j + 0.35\eta_j)$. For each of the 3×2 regime combinations, we generate 5 independent replications, giving 30 total benchmark instances.

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

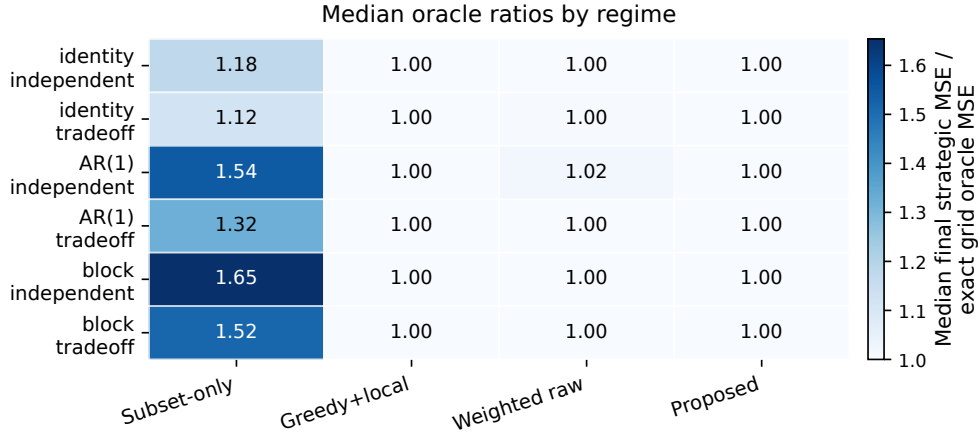


Figure 8: **Median oracle ratios by synthetic regime.** Each entry reports the median final strategic MSE normalized by the exact grid oracle MSE, $\text{MSE}_{\text{strat}}(\hat{\theta})/\text{MSE}_{\text{strat}}(\theta^{\text{OR}}(\mathbf{S}^*, \lambda^*))$, across the five replications in the corresponding 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.

start from the empty support and repeatedly add the coordinate whose inclusion gives the largest decrease in the exact fixed- λ objective $\Phi(\mathbf{S}, \lambda)$, 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 $\mathbf{1}^\top w \leq s$, 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 $\Phi(\mathbf{S}, \lambda)$ until no improving move remains. This step can only improve the rounded support and terminates after finitely many moves by Proposition D.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 D.3.

F Experimental Details and Illustrations for Section 6

F.1 Data description

The `upcoding` package⁴ 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 (All of Us Research Program Investigators, 2019). 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 $\text{HCC1}, \text{HCC2}, \dots$ 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 (Kronick et al., 2025a) 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 (Kronick et al., 2025b) 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 (Kronick et al., 2025a) 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

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

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. (2025a), 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 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 F.3.

⁴<https://github.com/StanfordHPDS/upcoding>

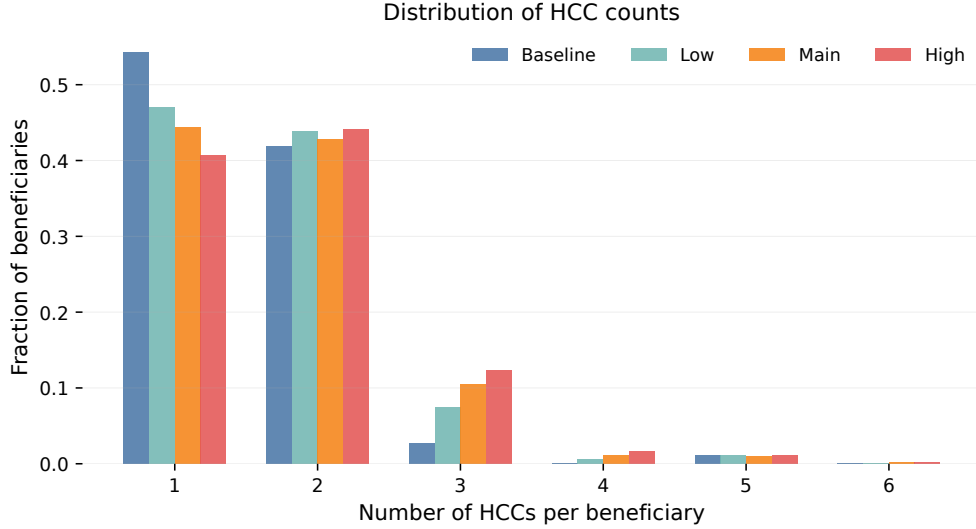


Figure 9: **Distribution of HCC counts per beneficiary.** The figure shows the number of V28 HCCs per beneficiary in the baseline data and in the low, main, and high upcoding scenarios. Only HCCs with nonzero baseline prevalence are included. The baseline is sparse, with mean 1.52 HCCs, median 1, and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

F.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.

F.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 $\tilde{X}_i^{(0)}$, we construct the semi-synthetic outcome as

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

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

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

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

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

$$\text{MSE}_\alpha(\hat{\theta}) = (\hat{\theta} - \theta^*)^\top \hat{\Sigma}(\hat{\theta} - \theta^*) + \alpha^2 (\hat{\theta}^\top H^{-1} \hat{\theta})^2 + \sigma^2. \quad (27)$$

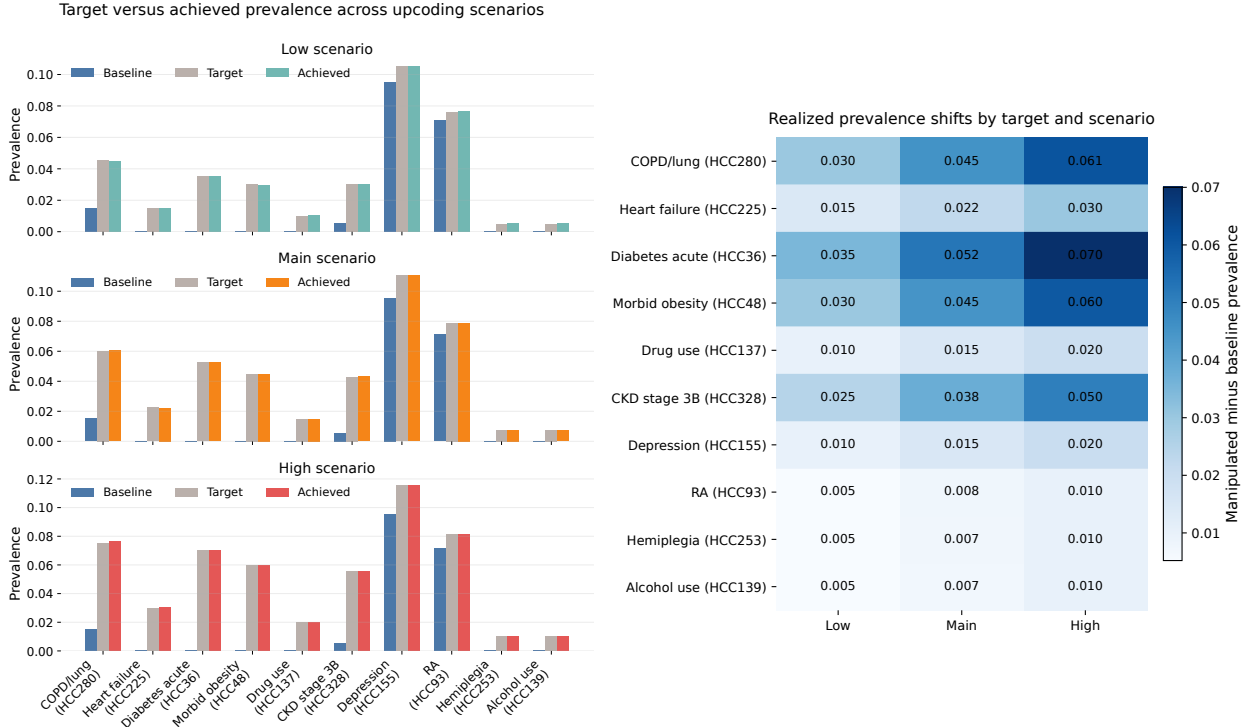


Figure 10: **Calibration of simulated upcoding across target HCCs and scenarios.** (Left) Baseline, target, and achieved prevalence after applying the V28 hierarchy for each target HCC under the low, main, and high scenarios. The achieved prevalences closely track the calibration targets. (Right) Realized 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.

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.

F.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 F.1 that the realized upcoded matrices are used to check that these calibration targets are achieved.)

The final manipulation-ease matrix has the form:

$$H^{-1} = \tau\{D + \rho B\} + \eta I. \quad (28)$$

Here D is diagonal and controls the marginal ease of manipulating each HCC. The matrix B is a block

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

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: $d_j^{\text{aw}} = d_{\text{floor}}$. 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 $d_j^{\text{aw}} \leftarrow \max\{d_j^{\text{aw}}, d_{\text{floor}} + s_g\}$, 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 5). Specifically, for HCC j in this group, we set $d_j = d_{\text{floor}} + \xi(d_j^{\text{aw}} - d_{\text{floor}})$, 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 $d_j = d_j^{\text{aw}}$.

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}_ℓ for $\ell = 1, \dots, L$, 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 (28) should be interpreted as a proxy, not as an estimated structural inverse cost matrix. Its purpose in this paper is to encode three

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

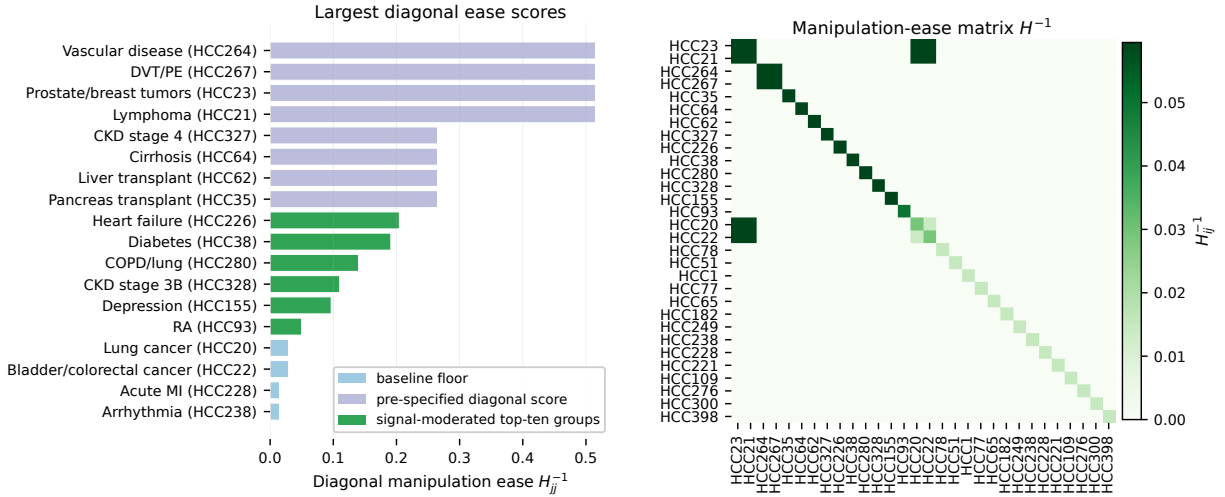


Figure 11: **Visualization of the manipulation-ease matrix.** (Left) Largest HCC diagonal entries of the final H^{-1} , grouped by the source of the diagonal score. (Right) The final inverse-cost matrix over HCCs with nonzero baseline prevalence, including the block component.

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.

F.5 Baselines and tuning

All methods are evaluated using the same plug-in post-manipulation MSE_α objective (27). 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, $|\hat{\Sigma}\theta^*|$, 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.

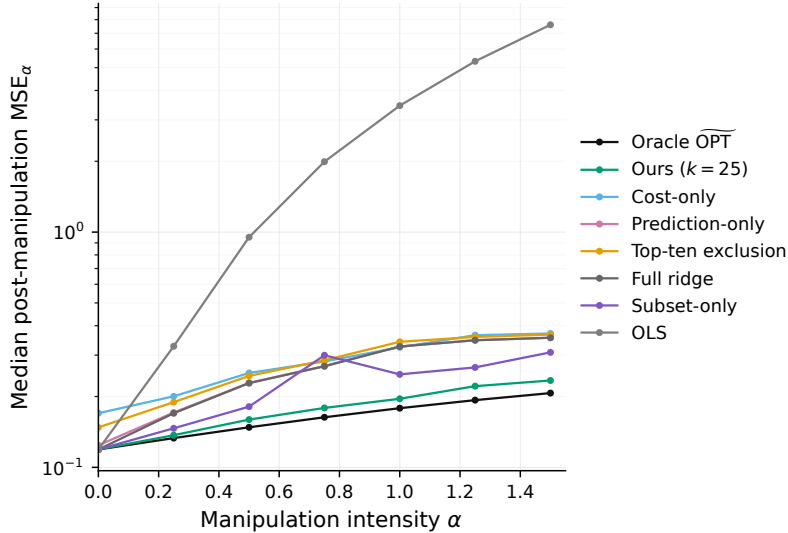


Figure 12: **Post-manipulation MSE under varying manipulation intensity α for all baselines.** Each curve reports median post-manipulation MSE as the manipulation intensity α varies, shown on a log scale. The proposed support-restricted ridge estimator with $k = 25$ remains close to the zero-intercept 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.

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 $\widetilde{\text{OPT}}$: 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).

For the oracle at manipulation intensity α , the first-order condition has the generalized-ridge form (16), 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. (2021, 2022). Extending these guarantees to strategic settings, along with joint support selection, is an interesting direction for future work.

F.6 Supporting illustrations

Figure 12 extends the comparison under varying manipulation intensity in Figure 4 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 5 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

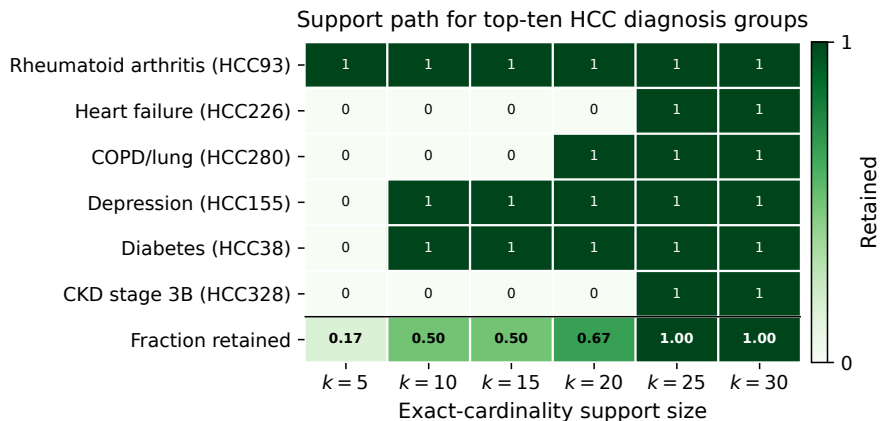


Figure 13: **Support path for top-ten diagnosis groups with nonzero prevalence.** Rows show the HCC groups that belong to the top-ten diagnosis groups identified by [Kronick et al. \(2025a\)](#), and columns show exact-cardinality support-restricted ridge solutions at different support sizes. The final 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.

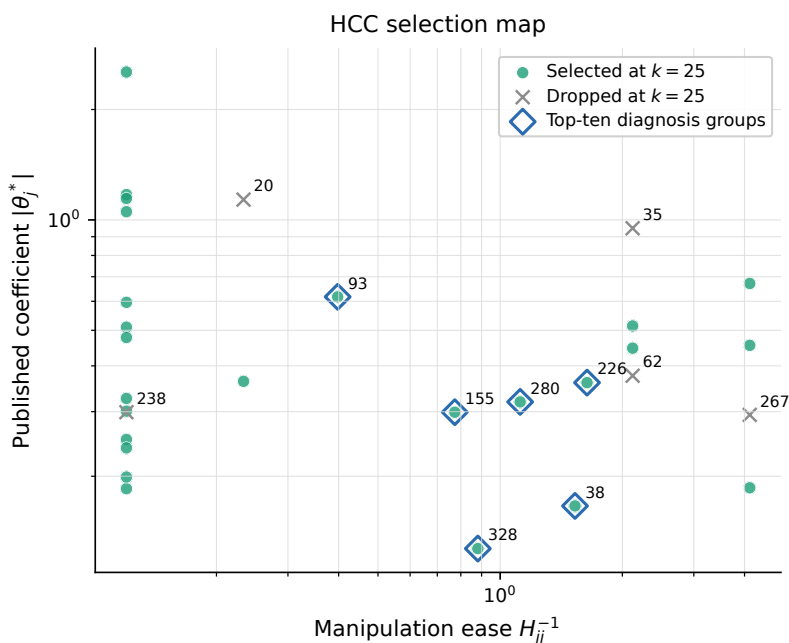


Figure 14: **Signal–manipulability selection map.** Each point is an HCC with nonzero prevalence in the data. Diamonds mark HCCs among the top-ten diagnosis groups identified by [Kronick et al. \(2025a\)](#), and labels are shown for the diagnosis groups that are dropped or belong to the top-ten set. 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.

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 Theoretical Details and Proofs for Section 7

Throughout this section, recall that we write $K := H^{-1}$ and make the dependence of the strategic MSE on K explicit:

$$\text{MSE}_K(\theta, b) = (\theta - \theta^*)^\top \Sigma (\theta - \theta^*) + (b + \theta^\top K \theta)^2 + \sigma^2.$$

G.1 Proof of Proposition 2

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

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

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

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

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.

G.2 Details for Example 3

Let $d = 1$, $\Sigma = 1$, $\theta^* = 1$, and $\mathcal{K} = [1/2, 5/2]$. The nominal estimate is $\widehat{K} = 3/2$. The plug-in intercept correction deploys $(\theta, b) = (1, -3/2)$. Its worst-case excess strategic MSE is

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

For ridge with $\lambda = 1$, the coefficient is

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

Its worst-case excess strategic MSE therefore is

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

G.3 Details for Example 4

Consider

$$\theta^* = (1, 0), \quad \Sigma = \begin{pmatrix} 1 & 49/50 \\ 49/50 & 1 \end{pmatrix},$$

and

$$K(k) = \begin{pmatrix} k & 0 \\ 0 & 1/2 \end{pmatrix}, \quad k \in [1/2, 5/2].$$

The nominal estimate is $\widehat{K} = K(3/2)$.

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

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

Now consider full-support ridge with $\lambda = 1$. Since $\Sigma\theta^* = (1, 49/50)$, we have

$$\theta^{\text{OR}}([2], 1) = (\Sigma + I)^{-1}\Sigma\theta^* = \begin{pmatrix} 2599 & 2450 \\ 7599 & 7599 \end{pmatrix}.$$

The worst case occurs at $k = 5/2$, and direct substitution gives the worst-case excess strategic MSE of approximately 0.240.

Next consider support-restricted ridge on $\mathbf{S} = \{2\}$ 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

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

The prediction term is

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

The strategic term is independent of k , because the first coordinate is dropped:

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

Combining the two terms yields the worst-case excess strategic MSE of approximately $0.172 < 0.240$.

Finally, support-restricted least squares on $\mathbf{S} = \{2\}$ gives $\theta^{\text{LS}}(\{2\}) = (0, 49/50)$. Its 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, support restriction alone is not as effective as combining support restriction with ridge shrinkage in this example.