Algorithmic Feature Highlighting for Human–AI Decision-Making

Yifan Guo Jann Spiess
Stanford University

September 2026

Abstract

Human decision-makers often face choices about complex cases with many potentially relevant features, but have limited bandwidth to inspect and integrate all available information. In such settings, we study algorithms that *highlight* a small subset of case-specific features for human consideration. Because the choice of which features are highlighted carries information, how that choice is interpreted matters: a *sophisticated* decision-maker correctly conditions on why features are highlighted, while a *naive* decision-maker updates only on revealed feature values and treats the selection as exogenous. In a two-feature example, we characterize the tradeoff between the two types and show that optimizing for one can be very costly to the other. Sophisticated design is hard to approximate even when only one feature can be shown, whereas naive design is tractable for fixed bandwidth and NP-hard when bandwidth is part of the input. We develop approximation algorithms for naive design and bound the excess risk of a naive-optimal policy for any mix of the two types. We illustrate our framework in a simulation calibrated to housing sales data. Overall, our results establish the value of contextual feature highlighting as a practical and computationally feasible tool for achieving human–algorithm complementarity.

1 Introduction

In high-stakes decision-making, algorithms frequently assist human decision-makers rather than replacing them. For example, algorithms may screen resumes, but final hiring decisions are still made by managers. In such settings, having algorithms recommend a final action may not be sufficient for decision-makers to make high-quality decisions that reflect their preferences and private information. Instead, effective algorithmic support tools should surface interpretable and verifiable details that are most relevant to the final decision.

In this article, we propose a model of joint human–algorithm decision-making in which the algorithm *highlights* key features that are relevant to the downstream decision of the human decision-maker. We model highlighting as a constrained information policy that selects which of a large number of features to reveal. We show that simple and computationally tractable algorithms are able to improve markedly over policies that instead always reveal the same features, and that this gain is robust to different ways in which decision-makers may interpret the selected features. Overall, we argue that highlighting presents a practically appealing and computationally feasible tool

Yifan Guo (yifanguo@stanford.edu) and Jann Spiess (jspiess@stanford.edu), Graduate School of Business, Stanford University. We thank Haozhan Gao, Sophie Greenwood, Ziyang Guo, Bryce McLaughlin, Manish Raghavan, Nick Shu, Renyuan Xu, Ruqing Xu, and audiences at ESIF 2026 and Stanford University for helpful conversations and comments. The authors used AI to help revise the writing, set up the simulation study, and develop the proofs. The authors take full responsibility for the content.

for achieving human–algorithm complementarity that raises rich theoretical questions and warrants further investigation.

We model human–algorithm interaction as a sender–receiver game between an algorithm designer (sender) who decides which features to highlight and a human decision-maker (receiver) who observes the highlighted features. The designer's goal is to help the human decision-maker recover a potentially high-dimensional feature vector as accurately as possible. The algorithm designer specifies an algorithm that inspects the full feature vector and selects a subset of its components to report to the bandwidth-constrained decision-maker. The human decision-maker observes only those features highlighted by the algorithm. After receiving the highlighted features, the human decision-maker infers missing information and uses it to inform subsequent decisions. For example, when a candidate applies for a job, the algorithm may observe all information contained in the application and then highlight a subset of features for the human HR manager, who makes the final decision. Throughout, we assume that the preferences of the algorithm designer and the decision-maker are aligned.

We believe that *highlighting* features has two important advantages in practice. First, presenting more than a simple prediction or recommended action allows the human decision-maker to learn richer information that provides important context and can be combined with heterogeneous preferences and private information. Second, the constraint that the algorithm can only pick existing features means that the provided information is interpretable, trustworthy, and verifiable. For example, simply providing a summary generated by large language models (LLMs) could be problematic since we cannot generally guarantee that it contains only factual information about a given case. By contrast, even if we use an LLM to decide *which* features to highlight, the resulting information remains correct by construction.[1]

We consider different modes of human decision-making to highlight challenges that arise in practice when designing highlighting policies. We focus on two benchmark models of human decision-makers: a sophisticated receiver and a naive receiver. Both types of decision-makers observe the features highlighted by the algorithm and update their posteriors accordingly. The *sophisticated* receiver also takes into account the information contained in the algorithm's decision about which features to highlight, in addition to the values of the highlighted features. By contrast, the *naive* receiver ignores why the algorithm highlights specific features and processes only the highlighted features at face value. For example, a hiring manager could be sophisticated if they respond not just to the information contained in a summary, but also to what is left out. Or they could be naive if they simply fill in the blanks based on their best guess, without taking into account why some information may be included or left out. In practice, a real decision-maker may behave as a mixture of these two types, exhibiting both sophisticated and naive behaviors.

We show that these different modes of updating have different implications for optimal highlighting policies, their robustness, and their computation. We first document gaps between the optimal policies for naive and sophisticated receivers, and the cost of optimizing for one when the audience is the other. While the sophisticated receiver does at least as well as the naive receiver under the same policy, a naive-optimal policy can have an arbitrarily large additive gap from the sophisticated optimum, which we call the "price of simplicity." Conversely, even the sophisticated optimum most favorable to naive receivers can impose an arbitrarily large "price of complexity" on them. We characterize this tradeoff in an example with two independent binary features of unequal scale. The naive solution always reveals the feature with the larger scale, whereas a sophisticated receiver can recover both features in optimum. We also study policy design for a population of

[1]This would still leave space for deception by a misaligned sender since the choice of *which* features are selected could be driven by a different objective (akin to Bayesian persuasion, Kamenica and Gentzkow, 2011), but the deception would then mainly be driven by failing to highlight the most relevant information.

naive and sophisticated receivers, and consider optimizing for one type while bounding the loss incurred by the other.

We then turn to the computation of optimal highlighting policies, and ask which types of solutions can be computed efficiently. We begin with fixed highlighting, a special case of highlighting in which the same subset of features is revealed in every case. In this benchmark case, the two receiver types form the same posterior. Under a Gaussian prior and quadratic loss, fixed highlighting is NP-hard when bandwidth is part of the input. Nevertheless, a greedy algorithm offers a spectral approximation guarantee for the reduction in expected loss.

Having considered the benchmark case of fixed highlighting, we next focus on the computational complexity of contextual highlighting, where the highlighted features depend on the specific instance. We show that in this case designing an optimal policy for a *sophisticated* decision-maker remains computationally difficult, even when the number of revealed features is limited. Since the sophisticated receiver considers not only the revealed features but also why they were revealed, the designer's optimization problem should account for this additional information channel. For a prior supported on finitely many states that are given as input, we show that obtaining a 1.001 approximation to the minimum sophisticated risk is NP-hard even with squared loss, binary features, and bandwidth one. For *naive* receivers, optimal contextual highlighting can be computed efficiently when bandwidth is fixed. The algorithm evaluates every permitted highlight separately at each state in the support. When bandwidth is part of the input, however, the problem becomes NP-hard.

Having discussed solutions for sophisticated and naive decision-makers, we also study algorithms for mixed audiences. First, we ask when the solution that is optimal for naive receivers still works well for the overall, mixed audience. For independent binary features and squared loss, the naive-optimal policy has a simple interpretation. It highlights the k most surprising features, where surprise is measured by the squared difference between a realized value and its prior mean. We define its *total excess risk* as its expected total squared error minus the minimum expected total squared error achievable by a policy designed for the same audience mixture. This excess risk is at most $\lambda \log J_{d,k}/2$, where λ is the sophisticated population share and $J_{d,k}$ is the number of permitted selected sets. We identify a regime in which this policy achieves vanishing total excess risk uniformly over audience mixtures. In the accompanying example, naive risk remains bounded while the risk of revealing nothing grows without bound. As a second approach to mixed audiences, we ask whether we can modify a perfect sophisticated solution that achieves zero loss to yield better naive performance, leading to a "profile matching" algorithm. This profile matching algorithm minimizes naive risk subject to perfect sophisticated recovery through a minimum-cost matching formulation, which can also be expressed as a minimum-cost flow problem. Together, these approaches address two endpoints of mixed-audience design, evaluating a naive-optimal policy for a mixed population and minimizing naive risk subject to perfect sophisticated recovery.

We illustrate our framework in a simulation calibrated to data on housing sales in Ames, Iowa. An algorithm chooses which of ten property attributes to show a buyer who wants to recover the property's features and predict its sale price. We compare the best fixed subset with exact selection, actual-loss greedy, projection greedy, and profile matching, evaluating each policy for both receiver types. We also compare an algorithm that has access only to the attributes with one that also sees the price, but cannot highlight it.

In the empirical exercise, we show that the algorithms we consider can substantially outperform a benchmark in which a fixed set of features is highlighted, for both sophisticated and naive receivers. Giving the designer access to price further improves performance under exact contextual selection. Both greedy methods also improve on fixed highlighting across many budgets, with their relative performance depending on the designer's information and the budget. When the algorithm

uses features alone, profile matching lets a sophisticated receiver reconstruct all ten housing attributes from four highlighted fields while achieving near-optimal naive risk. We also show that these algorithms can have some surprising properties. For example, in our setting, it is not generally optimal to always highlight the maximal allowable number of features, since withholding some strategically can improve downstream predictions. In our simulation, we observe this behavior when the algorithm can use price. In this case, allowing a policy to highlight fewer fields can outperform showing all ten, because the choice of which features are highlighted can convey additional price information.

We also discuss extensions to more realistic updating rules and additional frictions between sender and receiver. Beyond sophisticated and naive receivers, we consider specific behavioral biases in which receivers update their beliefs based on stereotypes. We also discuss private information of the receiver, wrong beliefs, and misaligned preferences. In addition to these extensions, we discuss more complex feature distributions. For independent uniform features, we solve both receiver problems exactly and show that selection can reduce sophisticated risk without revealing the full state. We leave the exact sophisticated optimum under a Gaussian prior as an open problem.

The remainder of the article is structured as follows. Section 2 reviews the related literature in human–AI interaction, information design, and feature revelation. Section 3 introduces the model of algorithmic feature highlighting with naive and sophisticated updating. Section 5 studies the cost of targeting the wrong updating type and the tradeoff in designing for a mixed audience. Section 6 studies computational complexity, Gaussian greedy approximation, and guarantees for mixed audiences. Section 7 presents our simulation study based on Ames housing data. Section 8 discusses extensions, and Section 9 concludes.

2 Related Literature

We contribute to an interdisciplinary literature across operations, economics, and computer science that aims to model and develop solutions for human–algorithm interaction. We also build upon approaches from information design and the literature on subset selection under constraints.

2.1 Human–AI Decision-Making and Behavioral Foundations

Our work relates to a growing literature that studies how humans and algorithms jointly make decisions. Vaccaro et al. (2024) conduct a meta-analysis, finding that human–AI combinations on average perform worse than the best of humans or AI alone. Athey et al. (2020) study the allocation of decision authority between humans and AI. Agarwal et al. (2023) study radiologists' belief updating when combining AI predictions and clinical information. Relative to this work, we focus on designing optimal highlighting algorithms that consider human belief updating, including whether humans account for the algorithm's selection of highlighted features.

More closely related to our work are the studies by Noti et al. (2025) and Greenwood et al. (2025). The former study formalizes AI-assisted decision-making with a learning human and shows that the optimal feature-revelation schedule balances short-run accuracy against long-run human learning, yielding a tractable combinatorial characterization. We differ from that study by allowing feature selection to depend on realized values and by distinguishing two types of human receivers. We see a connection between their and our approach, with naive behavior loosely corresponding to an initial interaction and sophisticated behavior to a long-run response shaped by experience. The latter study analyzes instance-contingent algorithmic delegation and shows that the optimal delegate can be strictly better than the best standalone AI. Relative to that article, our focus is

on selecting information for a human decision-maker, whereas that study focuses on designing an algorithmic delegate.

We also connect to the literature modeling human behavior in human–AI decision-making. Boyacı et al. (2024) study accuracy and effort under cognitive limitations. Green and Chen (2019) discuss practical algorithm-in-the-loop design. Bastani et al. (2026) study how to provide interpretable tips for sequential decisions. Balakrishnan et al. (2025) show that feature transparency reduces naive advice-weighting bias in collaboration between humans and algorithms with private information. More broadly, work on designing human–AI collaboration under behavioral responses (e.g., Agarwal et al., 2025; McLaughlin and Spiess, 2022, 2024; Hoong and Dreyfuss, 2025) studies how optimal disclosure, recommendation, and automation designs change when humans exhibit rational or biased responses. Relatedly, we study how accounting for or neglecting endogenous feature selection changes optimal highlighting policies, their computational complexity, and the tradeoff between receiver types.

A closely related strand studies human–AI complementarity, including conditions for joint decisions to improve on standalone performance (Donahue et al., 2022) and the value of human expertise on algorithmically indistinguishable inputs (Alur et al., 2024). Guo et al. (2026b) formalize human–AI complementarity through the expected improvement in best-attainable payoff from an additional signal, and Guo et al. (2026a) operationalize this idea by fine-tuning an LLM to extract signals from a supervisor's context that complement an existing agent's recommendation. Our focus is on a specific communication constraint. The algorithm reports a subset of the original feature values, and we compare receivers who do and do not account for how that subset was selected.

2.2 Information Design and Verifiable Disclosure

Our highlighting policy is a constrained signaling scheme, connecting our model to verifiable disclosure and Bayesian persuasion (Rayo and Segal, 2010; Kamenica and Gentzkow, 2011; Hoffmann et al., 2020; Gradwohl et al., 2022; Doval and Skreta, 2024). Related work incorporates behavioral responses into information design (Yu et al., 2023) and restricts persuasive messages to verifiable evidence (Hoefer et al., 2024). We study aligned objectives with messages restricted to a small subset of observed feature values. Our focus is on how optimal policies and their computation depend on whether the receiver accounts for selection.

2.3 Subset Selection and Instance-Wise Feature Revelation

Fixed highlighting chooses the same subset for every state, providing a benchmark related to classical feature selection. Das and Kempe (2018) analyze greedy subset selection under spectral conditions. We instead allow the highlighted set to depend on the realized state. Vasconcelos and Mitra (2020) study Gaussian quadratic reconstruction from a selected coordinate and its index, where selection itself conveys information. We compare receivers who account for or neglect selection and study the resulting implications for highlighting design. Nahum et al. (2024) select a common subset of item attributes to reduce marketplace congestion while preserving allocation value. Their objective reflects competition for scarce goods, whereas ours is aligned assistance for an individual decision through contextual feature highlighting.

3 A Model of Algorithmic Feature Highlighting

We model the interaction between an algorithm that highlights features and a bandwidth-constrained human decision-maker. Features are given by a random vector $X = (X_i)_{i \in M}$ that is distributed according to a (known) distribution $\mathbb{P}$, where M is a finite index set of size $|M| = d$. We call a realization of X in $\mathcal{X}$ a *state*. The algorithm selects a small number of these features based on a highlighting policy. The decision-maker then observes the highlighted features, updates their belief about the full state X, and incurs a loss that depends on how well they recover X.

Highlighting policy. We model feature *highlighting* as follows. We assume that the features $X = (X_i)_{i \in M} \in \mathcal{X}$ are accessible to an algorithm[2] that selects (or "highlights") a subset $I \subseteq M$ with $|I| \leq k$ and reveals the set I as well as the associated features $X_I = (X_i)_{i \in I}$ to the human decision-maker. Here, k is an exogenously given bandwidth constraint that limits the number of features that can be highlighted. In the main model the algorithm observes every coordinate of X and may reveal any of them; Section 7 adds a price that the algorithm may observe but cannot reveal. A (deterministic) highlighting policy is a map

$$\sigma : \mathcal{X} \to \{I \subseteq M : |I| \leq k\}.$$

A policy is *fixed* if it selects the same set at every state, and *contextual* otherwise. To facilitate our theoretical results, we also allow for σ to be a random policy.[3] We assume that this highlighting policy is chosen by an *algorithm designer*, who acts as the sender in our game, while the human decision-maker acts as the receiver.

The highlighting policy can be interpreted as follows. In hiring, a company may use an algorithm to assist HR in the recruiting process; the algorithm highlights several key features in each applicant's resume. In a hospital, the algorithm may highlight key components of a patient's medical file to assist doctors in their final assessment. In school admissions, the algorithm may highlight key aspects of each candidate's background for the admissions committee.

Human belief formation. Upon observing (I, X_I), the human decision-maker forms a posterior belief $\widehat{\mathbb{P}}$ over the full vector X, including its unknown components. We assume that the human decision-maker knows the distribution $\mathbb{P}$ (which acts as their prior), and we consider two main ways in which the decision-maker updates to form a posterior:

- *Sophisticated updating.* The sophisticated receiver is aware of the policy σ, and updates their belief by conditioning on the selected set I and the values X_I. The sophisticated receiver accounts for both the highlighted values and *why* they were highlighted. We model this update by the distribution
$$\widehat{\mathbb{P}}^S_\sigma = \mathbb{P}(\cdot \mid X_I, \sigma(X) = I).$$
- *Naive updating.* The naive receiver updates on the highlighted values X_I, but ignores the informational content of *which* set I was selected. We model this by the distribution
$$\widehat{\mathbb{P}}^N = \mathbb{P}(\cdot \mid X_I),$$

[2] In Section 8 we discuss extensions to human private information, in which case features are given by $X = (X_M, X_H)$ with machine features X_M and human features X_H.

[3] Formally, write $\mathcal{M}_k = \{I \subseteq M : |I| \leq k\}$ for the menu of feasible sets, where $0 \leq k \leq d$, and let Π_k contain all policies with measurable selection probabilities $K_\sigma(I \mid x)$ on this menu. A deterministic policy puts probability one on a single set at each state. The human observes the highlight $Z = (I, X_I)$, but not the policy's random seed U, which is independent of X. We write $\mathbb{P}_\sigma$ for the joint law induced by the prior and the selection rule.

where I is treated as a fixed set when conditioning on its observed values. Thus the naive receiver treats selection as exogenous.[4] For a fixed set I, we write $\widehat{\mathbb{P}}_I^N = \mathbb{P}(\cdot \mid X_I)$.

Here, we assume for simplicity that updating is based on the *true* distribution $\mathbb{P}$. Our model extends to the case where updating is performed according to some prior $\widetilde{\mathbb{P}}$ that may or may not be equal to $\mathbb{P}$. In that case, we can still write $\widehat{\mathbb{P}}_\sigma^S$ and $\widehat{\mathbb{P}}^N$ for the corresponding posteriors.

Throughout this article, we consider two types of priors $\mathbb{P}$: finitely supported priors mainly for results on computational complexity and Gaussian priors for closed-form posterior means. We think of both types of priors as approximations of real-world distributions that facilitate different qualitative insights.

Decision-maker's recovery loss. We assume that the human decision-maker's loss from having a (potentially imprecise) belief $\widehat{\mathbb{P}}$ reflects how well the belief recovers the unknown vector X. In general, we express this loss by $L(\widehat{\mathbb{P}}, X)$ for a loss function L that is smaller the more accurately $\widehat{\mathbb{P}}$ captures the realized X. In our article, we specifically consider $X \in \mathbb{R}^d$ and express this loss by the quadratic loss function

$$L_Q(\widehat{\mathbb{P}}, X) = \|\widehat{X}(\widehat{\mathbb{P}}) - X\|_Q^2 = (\widehat{X}(\widehat{\mathbb{P}}) - X)^\top Q(\widehat{X}(\widehat{\mathbb{P}}) - X) \tag{1}$$

for the best guess $\widehat{X}(\widehat{\mathbb{P}}) = \mathbb{E}_{\widehat{\mathbb{P}}}[X]$ and a fixed symmetric positive semidefinite weight matrix $Q \in \mathbb{R}^{d\times d}$. This loss function measures how far the belief $\widehat{\mathbb{P}}$ is from the realized feature vector X by the distance between the posterior mean $\widehat{X}(\widehat{\mathbb{P}})$ and X in the seminorm $\|\cdot\|_Q$. We call the case $Q = I_d$ *squared loss*; it sums, rather than averages, the squared errors across the d coordinates. We motivate this specific form of loss function below in terms of downstream tasks of the decision-maker. Writing $\widehat{X}_\sigma^S = \widehat{X}(\widehat{\mathbb{P}}_\sigma^S)$ and $\widehat{X}^N = \widehat{X}(\widehat{\mathbb{P}}^N)$ for the two reconstructions after the highlight (I, X_I), the losses of sophisticated and naive decision-makers are

$$L_Q(\widehat{\mathbb{P}}_\sigma^S, X) = \|\mathbb{E}[X \mid X_I, \sigma(X) = I] - X\|_Q^2 \qquad \text{and} \qquad L_Q(\widehat{\mathbb{P}}^N, X) = \|\mathbb{E}[X \mid X_I] - X\|_Q^2,$$

respectively.

Designer objective. We assume that the algorithm designer has the aligned objective of minimizing expected loss, or risk,

$$R_Q^S(\sigma) = \mathbb{E}_\sigma[L_Q(\widehat{\mathbb{P}}_\sigma^S, X)], \qquad\qquad R_Q^N(\sigma) = \mathbb{E}_\sigma[L_Q(\widehat{\mathbb{P}}^N, X)],$$

respectively, where $\mathbb{E}_\sigma$ denotes the expectation for the distribution over (X, I, X_I) induced by $\mathbb{P}$ and σ. We use *loss* for the realized value L_Q and *risk* for its expectation, and we omit the subscript Q when the weight matrix is fixed. The design problem is then

$$\min_{\sigma\in\Pi_k} R^\tau(\sigma)$$

for $\tau \in \{S, N\}$, with value $R^{\tau,\star} = \inf_{\sigma\in\Pi_k} R^\tau(\sigma)$. If the designer faces uncertainty over human types, a natural Bayesian objective is the population mixture $R^\lambda(\sigma) = \lambda R^S(\sigma) + (1-\lambda)R^N(\sigma)$, where the sophisticated share $\lambda \in [0,1]$ is independent of the state and the policy seed; we write $R^{\lambda,\star}$ for its infimum over Π_k. A robust alternative is to minimize the worst-case risk across types, $\min_{\sigma\in\Pi_k} \max\{R^S(\sigma), R^N(\sigma)\}$.

[4] Neglect of selection of this kind is studied in the literature (e.g., Eyster and Rabin, 2005; Enke, 2020; Di Tillio et al., 2021). Related departures from Bayesian updating appear in the correlation-neglect model of Amin et al. (2026) and the radiology evidence of Agarwal et al. (2023). Our naive receiver instead neglects the information conveyed by feature selection.

Robust microfoundation of recovery loss. We now provide a microfoundation of the decision-maker's recovery loss (1) in terms of downstream tasks of the decision-maker. Specifically, we assume that the decision-maker observes the highlighted set and features (I, X_I), updates their belief to $\widehat{\mathbb{P}}$, and then chooses an action

$$\widehat{A}_\ell(\widehat{\mathbb{P}}) = \operatorname*{argmin}_{a \in \mathcal{A}} \mathbb{E}_{\widehat{\mathbb{P}}}[\ell(a, X)].$$

For example, a hiring manager may update their belief about a candidate's profile and then make a hiring decision based on this belief.

While the decision-maker knows the loss function ℓ corresponding to the specific task, the algorithm designer instead only knows that the receiver chooses according to some loss function from a set $\mathcal{L}$, leading the designer to consider the robust objective

$$L(\widehat{\mathbb{P}}, X) = \sup_{\ell \in \mathcal{L}} \ell(\widehat{A}_\ell(\widehat{\mathbb{P}}), X)$$

that considers the worst-case outcomes across loss functions $\ell \in \mathcal{L}$. For example, an algorithm used to highlight features in applicant resumes may be used for many downstream tasks and not just for the *specific* hiring task of a given manager. We then assume that the algorithm is chosen to work well across cases, in a minimax sense.

Suppose now that the human decision-maker's task is to predict an outcome $Y = \beta^\top X + \varepsilon$ with $\mathbb{E}[\varepsilon \mid X] = 0$ and $\beta, X \in \mathbb{R}^d$, under the standard mean-squared error loss

$$\ell_\beta(a, X) = \mathbb{E}[(Y - a)^2 \mid X] = (a - \beta^\top X)^2 + \mathbb{V}\mathrm{ar}(\varepsilon \mid X).$$

Then, if there is uncertainty about *which* prediction task the receiver faces, for

$$\mathcal{L} = \{\ell_\beta; \beta \in \mathbb{R}^d, \|\beta\|_W \leq 1\}$$

with $\|\beta\|_W^2 = \beta^\top W \beta$ for a positive definite matrix W, we obtain the recovery loss

$$L(\widehat{\mathbb{P}}, X) = \sup_{\|\beta\|_W \leq 1} \left(\beta^\top (\widehat{X} - X)\right)^2 + \mathbb{V}\mathrm{ar}(\varepsilon \mid X) = (\widehat{X} - X)^\top W^{-1} (\widehat{X} - X) + \text{const.},$$

which is (1) with $Q = W^{-1}$, up to an additive term that does not depend on the belief $\widehat{\mathbb{P}}$ and hence not on the highlighting policy. In the hiring example, this prediction task could correspond to an evaluation of a candidate in terms of a score, and the varying vectors β would represent uncertainty about how different features are evaluated for different job positions.

Special case: feature and outcome recovery. A specific case of the quadratic loss function is when the decision-maker aims to recover the full vector X along with an outcome Y of interest, where for simplicity we assume that $Y = \gamma^\top X + \varepsilon$ with $\mathbb{E}[\varepsilon \mid X] = 0$. In this case, we can consider the loss function that combines full recovery of X with targeting of Y by setting

$$L(\widehat{\mathbb{P}}, X) = (1 - \alpha) \|X - \widehat{X}\|^2 + \alpha \, \mathbb{E}[(Y - \gamma^\top \widehat{X})^2 \mid X] = \|X - \widehat{X}\|_Q^2 + \text{const.}$$

where $Q = (1 - \alpha) \, I_d + \alpha \, \gamma\gamma^\top$. Squared loss ($\alpha = 0$, so $Q = I_d$) and simple prediction with a known prediction target ($\alpha = 1$) are special cases. For squared loss, Section 4.1 interprets the loss as the expected loss of humans with private linear prediction tasks, and compares highlighting to two alternative ways in which decisions could be taken: directly by the human or based on an algorithmic recommendation. Specifically, there, we provide conditions under which highlighting is preferable to simple algorithmic recommendations.

4 Highlighting and Recommendations

The recovery objective in Section 3 allows highlighted features to support different downstream tasks. We now ask when this flexibility is valuable relative to human search and algorithmic recommendations. We first compare five modes of collaboration in a benchmark where private tasks give equal average weight to every direction. We then allow the algorithm to optimize a collection of linear scores and show how the preferred interface changes as uncertainty about the human's task spreads across directions. Finally, we allow the human to observe additional state coordinates that are unavailable to the algorithm.

Let $X \sim \mathcal{N}(0, \Sigma)$ with $\Sigma \succ 0$. The human privately observes a task vector $\beta \in \mathbb{R}^d$, independent of X, and predicts $\beta^\top X$ under squared loss. The algorithm observes X and knows the distributions of X and β, but not the realized task. We assume $\mathbb{E}\|\beta\|^2 < \infty$, and any policy randomization is independent of the state and the task. For either updating rule, averaging the prediction loss of the action $a = \beta^\top \widehat{X}^\tau$ over tasks gives the recovery risk $R^\tau_Q(\sigma)$ with $Q = \mathbb{E}[\beta\beta^\top]$.

4.1 Highlighting, Human Search, and Recommendations

For the initial comparison, specialize to $X \sim \mathcal{N}(0, I_d)$ and assume

$$\mathbb{E}[\beta\beta^\top] = I_d, \qquad \overline{\beta} = \mathbb{E}[\beta].$$

We call tasks satisfying this second-moment condition *isotropic tasks*. For example, $\beta \sim \mathcal{N}(0, I_d)$ satisfies this condition. This isotropic second-moment assumption gives equal average weight to every direction in the feature space. We compare five modes of collaboration with a feature budget $1 \le k \le d$. Search and highlighting use this feature budget, and each recommendation consists of one score.

- *Random search.* The human inspects k features chosen uniformly at random.[5]
- *Fixed highlighting.* The algorithm always reveals (I, X_I) for the same set I with $|I| = k$.
- *Contextual highlighting.* The algorithm reveals (I, X_I) with I chosen by a policy $\sigma \in \Pi_k$.
- *Mandatory recommendation.* The algorithm reports $r = \overline{\beta}^\top X$, and the human takes $a = r$.
- *Advisory recommendation.* The algorithm reports the same r. The human knows how r is computed and takes $a = \mathbb{E}[\beta^\top X \mid r, \beta]$.

Under random search and fixed highlighting, the human takes the posterior mean of $\beta^\top X$ given the revealed features and β. Under contextual highlighting, a receiver of type τ takes $\beta^\top \widehat{X}^\tau$. We write R^{rand}_k, $R^{\text{fix}}(I)$, $R^\tau(\sigma)$, R^{mand}, and R^{adv} for the corresponding risks. The next proposition ranks these modes of collaboration, with σ^N denoting the policy that reveals the k coordinates with the largest $|X_j|$.

Proposition 1 (Ranking with isotropic tasks)**.** *Under the setup above, for every set I with $|I| = k$,*

$$R^{\text{mand}} \ge R^{\text{adv}} \ge R^{\text{fix}}(I) = R^{\text{rand}}_k \ge R^N(\sigma^N) = R^S(\sigma^N).$$

Moreover, $R^{\text{adv}} = R^{\text{fix}}(I)$ if and only if $k = 1$ and $\overline{\beta} \neq 0$.

[5] The inspection budget connects this benchmark to the limited information-processing capacity studied in the rational inattention literature (e.g., Sims, 2003; Lipnowski et al., 2020).

Proof. For any fixed $e \in \mathbb{R}^d$, the condition $\mathbb{E}[\beta\beta^\top] = I_d$ gives $\mathbb{E}_\beta[(\beta^\top e)^2] = \|e\|^2$. Thus fixed highlighting and random search both have risk $d-k$, and σ^N is naive-optimal. Let S_k be the sum of the k largest squared coordinates of X. Under σ^N, each hidden coordinate has a conditional distribution symmetric about zero, so naive and sophisticated posterior means coincide and

$$R^N(\sigma^N) = R^S(\sigma^N) = d - \mathbb{E}[S_k] \le d - k,$$

where the inequality follows from $S_k \ge (k/d)\|X\|^2$. For recommendations, direct calculation and Gaussian conditioning give

$$R^{\text{mand}} = d - \|\overline{\beta}\|^2 \ge d-1, \qquad R^{\text{adv}} = \begin{cases} d-1, & \overline{\beta} \neq 0, \\ d, & \overline{\beta} = 0. \end{cases}$$

Since $\mathbb{V}\text{ar}(\beta) = I_d - \overline{\beta}\overline{\beta}^\top \succeq 0$, we have $\|\overline{\beta}\|^2 \le 1$. When $\overline{\beta} \neq 0$, the advisory action is $r\,\beta^\top\overline{\beta}/\|\overline{\beta}\|^2$. Gaussian conditioning gives the risk for a given task vector,

$$\mathbb{E}[(a - \beta^\top X)^2 \mid \beta] = \|\beta\|^2 - \frac{(\beta^\top\overline{\beta})^2}{\|\overline{\beta}\|^2}.$$

Averaging over tasks and using $\mathbb{E}[\beta\beta^\top] = I_d$ therefore yields

$$R^{\text{adv}} = d - \frac{\overline{\beta}^\top\mathbb{E}[\beta\beta^\top]\overline{\beta}}{\|\overline{\beta}\|^2} = d-1.$$

These formulas give the ranking and the stated equality condition. □

Proposition 1 shows how selecting features for the realized case can lower risk relative to random inspection and the mean-task recommendation. The recommendation in this benchmark uses the single score $\overline{\beta}^\top X$. We next allow the algorithm to choose up to k score directions and examine how the comparison depends on the distribution of tasks.

4.2 Optimal Linear Advice

We now allow general $\Sigma \succ 0$ and $Q = \mathbb{E}[\beta\beta^\top]$, and compare highlighting with optimized linear advice under the same budget k. Highlighting reports (I, X_I) with $|I| \le k$, including silence. Linear advice reports $AX \in \mathbb{R}^k$, where A is fixed before the state is realized and is known to the human. The human then takes $a = \mathbb{E}[\beta^\top X \mid AX, \beta]$. The budget counts displayed features, together with their identities, or displayed linear scores. Write $R^{\text{adv},\star}_{Q,k}$ for the infimum of advice risk over $A \in \mathbb{R}^{k\times d}$ and $R^{\tau,\star}_{Q,k} = \inf_{\sigma\in\Pi_k} R^\tau_Q(\sigma)$ for optimal highlighting risk.

Proposition 2 (Optimal linear advice)**.** *Let $C = \Sigma^{1/2}Q\Sigma^{1/2}$ have eigenvalues $\lambda_1(C) \ge \cdots \ge \lambda_d(C)$. Then*

$$R^{\text{adv},\star}_{Q,k} = \sum_{j=k+1}^{d} \lambda_j(C).$$

An optimal report uses $A = U_k^\top\Sigma^{-1/2}$, where the columns of U_k are orthonormal eigenvectors for the k largest eigenvalues of C.

Thus a small number of scores suffices when task-weighted variation is concentrated in a few directions. In particular, advice has zero risk when $\text{rank}(Q) \le k$. The calculation follows by whitening the state and choosing the k directions that remove the most expected prediction error. Detailed proofs of this proposition and the results below are in Appendix C.

4.3 A Reversal with Task Uncertainty

To isolate uncertainty about the task, specialize to $X \sim \mathcal{N}(0, I_d)$ and consider

$$\beta_\eta = \sqrt{1-\eta}\,\mathbf{1}_d + \sqrt{\eta}\,Z, \qquad Z \sim \mathcal{N}(0, I_d), \quad Z \perp X, \qquad \eta \in [0,1]. \tag{2}$$

Here $\mathbf{1}_d$ is the vector of ones, and $Q_\eta = (1-\eta)\mathbf{1}_d\mathbf{1}_d^\top + \eta I_d$. The risk of revealing nothing is $\mathbb{E}[(\beta_\eta^\top X)^2] = d$ throughout this family. As η increases, the task changes from a known common direction to private weights dispersed across coordinates. At $\eta = 1$, the task second moment is I_d, as in Section 4.1. This is the feature and outcome recovery family in Section 3 with $\alpha = 1-\eta$ and $\gamma = \mathbf{1}_d$.

Write $A_k(\eta) = R^{\mathrm{adv},\star}_{Q_\eta,k}$ and $H^\tau_k(\eta) = R^{\tau,\star}_{Q_\eta,k}$. Write σ^{mag} for the magnitude policy used in Proposition 1, which reveals the k largest $|X_j|$ with fixed tie-breaking. Define $h_{d,k}$ as the expected sum of the $d-k$ smallest values among $G_1^2, \ldots, G_d^2$, where $G \sim \mathcal{N}(0, I_d)$.

Theorem 1 (A reversal with task uncertainty)**.** *For $1 \le k < d$ in (2),*

$$A_k(\eta) = (d-k)\eta, \qquad R^N_{Q_\eta}(\sigma^{\mathrm{mag}}) = R^S_{Q_\eta}(\sigma^{\mathrm{mag}}) = h_{d,k}.$$

For each $\tau \in \{S, N\}$, there is a unique $\eta_\tau \in (0,1)$ at which $A_k(\eta_\tau) = H^\tau_k(\eta_\tau)$. Optimal linear advice has lower risk for $\eta < \eta_\tau$, and optimal highlighting has lower risk for $\eta > \eta_\tau$. The thresholds satisfy

$$0 < \eta_S \le \eta_N \le \frac{h_{d,k}}{d-k} < 1.$$

Proof sketch. The advice value follows from the eigenvalues of Q_η. Under magnitude highlighting, symmetry gives zero posterior means for omitted coordinates and eliminates expected cross terms, yielding risk $h_{d,k} < d-k$ for both receivers. Each fixed policy's risk is affine in η, so optimal highlighting risk is concave. At $\eta = 0$, a linear score reveals the target exactly, while every highlighting policy has risk bounded away from zero. At $\eta = 1$, magnitude highlighting improves on linear advice. Concavity and continuity give a unique crossing, and sophisticated dominance under each policy orders the thresholds. △

The comparison turns on the directions in which the human may need information. A score can summarize a common task, while contextual highlighting adapts the displayed inputs to the realized case and leaves the human to apply private task weights. Accounting for selection weakly lowers the level of task uncertainty at which highlighting becomes competitive. The magnitude policy supplies the upper bound on the thresholds, while optimization over all feasible highlights can also use silence when omitted contributions cancel.

4.4 Human Private State Information

We finally allow the human to have private information about both the task and the state. Let $(X_M, X_H) \in \mathbb{R}^d \times \mathbb{R}^p$ be centered jointly Gaussian with positive definite covariance. The algorithm observes X_M, while the human observes X_H and a task vector $\beta = (\beta_M, \beta_H)$ independent of the joint state and with finite second moment. The common prediction target is

$$Y = \beta_M^\top X_M + \beta_H^\top X_H. \tag{3}$$

The algorithm knows the joint distribution and can highlight at most k machine features or report k fixed linear scores AX_M. Its random seed is independent of the state and task. Human private coordinates are observed before the message and do not use the disclosure budget.

After a highlight, the human takes $a^\tau = \beta_H^\top X_H + \beta_M^\top \widehat{X}_M^\tau$, where

$$\widehat{X}_M^S = \mathbb{E}_\sigma[X_M \mid X_H, I, X_I], \qquad \widehat{X}_M^N = \mathbb{E}[X_M \mid X_H, X_I],$$

and the naive conditional mean treats I as fixed. Here X_I contains coordinates of X_M. The known contribution $\beta_H^\top X_H$ cancels from the prediction error, leaving recovery risk for X_M with weight $Q_M = \mathbb{E}[\beta_M \beta_M^\top]$.

Proposition 3 (Linear advice with human private information)**.** *Let $\Omega = \mathbb{V}\mathrm{ar}(X_M \mid X_H)$. The minimum risk from k fixed linear scores is*

$$R^{\mathrm{adv},\star}_{Q_M,k}(X_H) = \sum_{j=k+1}^{d} \lambda_j(\Omega^{1/2} Q_M \Omega^{1/2}),$$

where the eigenvalues are ordered from largest to smallest.

The human can subtract $A\mathbb{E}[X_M \mid X_H]$ from the reported scores and then estimate the remaining uncertainty. Because its highlighting choice can depend only on X_M, the algorithm evaluates risk by averaging over X_H conditional on X_M.

For a concrete comparison, take $p = d$, $\mathbb{V}\mathrm{ar}(X_M) = \mathbb{V}\mathrm{ar}(X_H) = I_d$, and $\mathbb{C}\mathrm{ov}(X_M, X_H) = \rho I_d$ with $0 \le \rho < 1$. Let the machine task weights β_M follow (2) independently of the joint state, and put $r = 1 - \rho^2$. Write $A_k(\eta, \rho)$ and $H_k^\tau(\eta, \rho)$ for the optimal advice and highlighting risks in this model, and $R^\tau_{Q_\eta}(\sigma; \rho)$ for a policy's highlighting risk.

Corollary 1 (A reversal with private state information)**.** *For $1 \le k < d$, optimal advice and magnitude highlighting on X_M satisfy*

$$A_k(\eta, \rho) = r(d-k)\eta, \qquad R^N_{Q_\eta}(\sigma^{\mathrm{mag}}; \rho) = r(1-r)(d-k) + r^2 h_{d,k}.$$

Magnitude highlighting is naive-optimal at $\eta = 1$. For each receiver type there is a unique threshold $\eta_\tau(\rho)$, below which optimal advice has lower risk and above which optimal highlighting has lower risk, with equality at the threshold. Moreover,

$$0 < \eta_S(\rho) \le \eta_N(\rho) \le \rho^2 + (1-\rho^2)\frac{h_{d,k}}{d-k} < 1.$$

Private state information preserves the reversal while making the human's remaining uncertainty partly unobservable to the algorithm. It also makes the treatment of selection consequential even under magnitude highlighting, as the following example illustrates.

Example 1 (Private information and selection neglect)**.** *Let $d = 2$, $k = 1$, and $\eta = 1$ in the correlated model above. Magnitude highlighting is naive-optimal, and $h_{2,1} = 1 - 2/\pi$. When $\rho = 0$, optimal naive risk is approximately* 0.36338*. When $\rho^2 = 0.1$, it is*

$$H_1^N(1, \rho) = 0.09 + 0.81\left(1 - \frac{2}{\pi}\right) \approx 0.38434.$$

Thus additional informative private observations increase optimal naive risk in this example. The naive receiver uses those observations while ignoring the information conveyed by omission. A sophisticated receiver conditions on both, and additional private information weakly reduces risk under each fixed highlighting policy.

5 Naive vs. Sophisticated Updating

Above, we introduced a model of feature highlighting in which an algorithm sends a subset of features to a human decision-maker. We considered two ways in which the decision-maker may update their beliefs, either by accounting for the selection rule (sophisticated updating) or by ignoring it (naive updating). How the receiver interprets selection changes which highlighting policies are optimal. In this section, we first show that deploying a policy optimized for one type to the other can impose an arbitrarily large cost, even with only two independent binary features. We then consider a common policy for a mixed audience. Optimizing for naive receivers minimizes worst-case risk across receiver types, while information about audience composition motivates minimizing average risk. Under quadratic loss, a suitably reweighted population mixture also captures an individual who combines the two posterior beliefs. Finally, we formulate design problems that minimize one type's risk subject to a bound on the other's and characterize the resulting Pareto frontier in the two-feature example.

5.1 The Cost of Misoptimization

Optimizing for one receiver type can impose a large cost on the other. We quantify this cost in both directions and illustrate it with two independent binary features of unequal scale. In order to compare the performance of the two solutions, let $\sigma^N \in \operatorname{argmin}_{\sigma\in\Pi_k} R^N(\sigma)$ and $\sigma^S \in \operatorname{argmin}_{\sigma\in\Pi_k} R^S(\sigma)$ denote optimal policies for naive and sophisticated receivers, respectively. To quantify the mismatch when a policy optimized for one type is deployed to the other, define

$$\text{price of complexity} \qquad \Delta^N := R^N(\sigma^S) - R^N(\sigma^N) \geq 0, \tag{4}$$

$$\text{price of simplicity} \qquad \Delta^S := R^S(\sigma^N) - R^S(\sigma^S) \geq 0. \tag{5}$$

The *price of complexity* is the naive receiver's excess risk from using a sophisticated-optimal policy instead of a naive-optimal policy. Conversely, the *price of simplicity* is the sophisticated receiver's excess risk from using a naive-optimal policy instead of a sophisticated-optimal policy. The following example shows that both prices can be arbitrarily large, even when the naive optimum is unique and we choose the sophisticated optimum most favorable to naive receivers.

Example 2 (Features of unequal scale)**.** *Let $X_1 \sim \text{Unif}\{0, a\}$ and $X_2 \sim \text{Unif}\{0, b\}$ independently, where $0 < a < b$. The algorithm observes both features and may reveal only one of them. Take squared loss, $Q = I_2$.*

Revealing feature 2 gives naive loss $\frac{a^2}{4}$ in every state, while revealing feature 1 gives $\frac{b^2}{4}$. Thus the unique naive-optimal policy always reveals feature 2, and

$$R^{N,\star} = R^N(\sigma^N) = R^S(\sigma^N) = \frac{a^2}{4}.$$

This is also the best fixed policy.

Now consider the policy that reveals feature 1 when $\frac{X_1}{a} = \frac{X_2}{b}$ and feature 2 otherwise. A sophisticated receiver can infer the missing coordinate from the selected index and its value. This policy therefore has zero sophisticated risk, so $R^{S,\star} = 0$, while its naive risk is $\frac{a^2+b^2}{8}$.

We say that a policy achieves *perfect sophisticated recovery* if $R^S(\sigma) = 0$. Under squared loss, the sophisticated receiver then recovers X with probability one. In this example, we choose σ^S to minimize naive risk among sophisticated-optimal policies. The frontier below shows that the policy

just described attains this minimum. With this choice, the price of simplicity is $\Delta^S = \frac{a^2}{4}$, and

$$\Delta^N = \min_{\sigma \in \Pi_1 :\, R^S(\sigma)=0} \left(R^N(\sigma) - R^{N,\star}\right) = \frac{b^2 - a^2}{8}. \tag{6}$$

Holding a fixed and increasing b makes this minimum price of complexity (Δ^N) arbitrarily large, both in absolute terms and relative to $R^{N,\star} = \frac{a^2}{4}$. Conversely, taking $b = \sqrt{a^2+1}$ and increasing a makes the price of simplicity (Δ^S) arbitrarily large while the minimum price of complexity remains $\frac{1}{8}$ and the sophisticated optimum remains zero. Figure 1 illustrates these comparisons through the exact Pareto frontier of achievable risk pairs $(R^N(\sigma), R^S(\sigma))$.

The example illustrates a basic design tradeoff. Contextual selection lets sophisticated receivers infer the omitted feature, but raises the loss of receivers who ignore selection. The frontier in Figure 1 shows how the feature scales determine the cost of moving between the two optima. A designer serving both types must therefore weigh their risks when choosing a common policy.

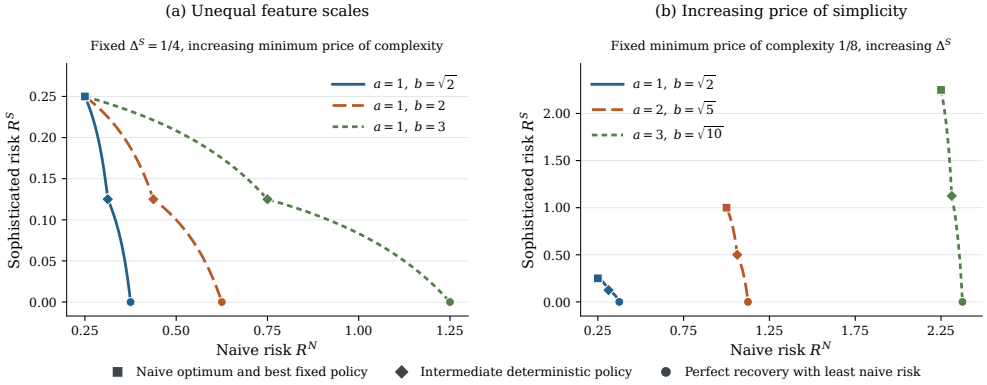


Figure 1: Exact Pareto frontiers in Example 2. The left panel fixes $a = 1$ and increases b, holding the naive optimum at $R^{N,\star} = \frac{1}{4}$ while the price of complexity grows. The right panel increases a with $b = \sqrt{a^2+1}$, holding the minimum price of complexity at $\frac{1}{8}$ while the price of simplicity grows. The unique naive-optimal policy coincides with the best fixed policy and always reveals feature 2.

5.2 Designing for a Mixture of Naive and Sophisticated Receivers

We now consider a designer who serves both naive and sophisticated receivers. The designer can minimize worst-case risk, average risk, or one type's risk subject to a bound on the other's. We also establish the connection between population mixtures and mixtures of posterior beliefs.

Robust design. When audience composition is unknown, the designer can minimize the largest risk across receiver types. The following result shows that this worst-case risk is the naive risk.

Proposition 4 (Robust design reduces to naive optimization)**.** *For every $Q \succeq 0$, every highlighting policy $\sigma \in \Pi_k$ satisfies*

$$R^N_Q(\sigma) - R^S_Q(\sigma) = \mathbb{E}_\sigma\left[\|\widehat{X}^N - \widehat{X}^S_\sigma\|^2_Q\right] \geq 0.$$

Consequently, a naive-optimal policy also minimizes worst-case risk across receiver types.

Population and belief mixtures. Information about audience composition allows the designer to optimize average performance. Suppose a fraction $\lambda \in [0,1]$ of receivers is sophisticated. We call such a population a *mixed audience*. The designer minimizes the population-mixture risk R^λ of Section 3,

$$\min_{\sigma \in \Pi_k} R^\lambda(\sigma). \tag{7}$$

When $\lambda = 1$, the designer optimizes for sophisticated receivers, and when $\lambda = 0$, the designer optimizes for naive receivers. Under quadratic loss, the same design objective also arises when an individual receiver combines the two posterior beliefs, after adjusting the mixture weight.

Proposition 5 (Belief mixtures and population mixtures)**.** *Take the quadratic loss L_Q with fixed $Q \succeq 0$. For a fixed $\theta \in [0,1]$, suppose the receiver uses the posterior*

$$\widehat{\mathbb{P}}^\theta_\sigma(\cdot \mid I, X_I) = \theta \widehat{\mathbb{P}}^S_\sigma(\cdot \mid I, X_I) + (1-\theta)\widehat{\mathbb{P}}^N(\cdot \mid I, X_I).$$

Its posterior-mean reconstruction is

$$\widehat{X}^\theta_\sigma = \widehat{X}(\widehat{\mathbb{P}}^\theta_\sigma) = \theta \widehat{X}^S_\sigma + (1-\theta)\widehat{X}^N.$$

Writing $R^\theta_{Q,\text{within}}(\sigma) = \mathbb{E}_\sigma[L_Q(\widehat{\mathbb{P}}^\theta_\sigma, X)]$ for its true expected loss gives

$$\begin{aligned} R^\theta_{Q,\text{within}}(\sigma) &= (2\theta - \theta^2) R^S_Q(\sigma) + (1-\theta)^2 R^N_Q(\sigma) \\ &= R^{\lambda(\theta)}_Q(\sigma), \qquad \lambda(\theta) = 2\theta - \theta^2. \end{aligned} \tag{8}$$

The two objectives therefore have the same optimal highlighting policies.

The posterior mean is a Bayes action under quadratic loss, and the risk identity follows from the same orthogonality used in Proposition 4. The next corollary allows the belief weight to vary across receivers.

Corollary 2 (Heterogeneous belief weights)**.** *Suppose receivers share the prior and quadratic loss L_Q, and a common policy does not condition on their belief weight $\Theta \in [0,1]$. Assume that Θ is independent of the state X. Then aggregate expected loss equals $R^{\lambda_{\text{eff}}}_Q(\sigma)$, the risk of a population mixture with*

$$\lambda_{\text{eff}} = 1 - \mathbb{E}[(1-\Theta)^2] = 2\mathbb{E}[\Theta] - \mathbb{E}[\Theta^2].$$

Thus a common policy can be designed for heterogeneous belief mixtures using a single effective population share.

Risk constraints and the Pareto frontier. Alternatively, the designer can minimize one type's risk subject to an upper bound on the other type's risk. For feasible risk ceilings $\overline{r}^N$ and $\overline{r}^S$, the two constrained design problems are

$$\begin{aligned} \min_{\sigma \in \Pi_k} \quad & R^S(\sigma) \\ \text{subject to} \quad & R^N(\sigma) \le \overline{r}^N, \end{aligned} \qquad \begin{aligned} \min_{\sigma \in \Pi_k} \quad & R^N(\sigma) \\ \text{subject to} \quad & R^S(\sigma) \le \overline{r}^S. \end{aligned} \tag{9}$$

For the two-feature example, solving the first problem as the naive-risk ceiling varies gives the exact Pareto frontier. The next proposition expresses this ceiling as an allowance above the naive optimum.

Proposition 6 (Exact two-phase Pareto frontier)**.** *In Example 2, define*

$$V(\eta) = \min_{\substack{\sigma\in\Pi_1:\\ R^N(\sigma)\le R^{N,\star}+\eta}} R^S(\sigma), \qquad \eta \ge 0.$$

For $0 \le \eta \le (b^2-a^2)/8$, set $t = 8\eta/(b^2-a^2)$. Then

$$V(\eta) = \frac{a^2}{4}\begin{cases} \dfrac{2-3t}{2(1-t)}, & 0 \le t \le \frac{1}{2}, \\ \dfrac{2(1-t)}{3-2t}, & \frac{1}{2} \le t \le 1. \end{cases} \tag{10}$$

For $\eta \ge (b^2-a^2)/8$, $V(\eta) = 0$. In particular, the smallest naive-risk increase compatible with perfect sophisticated recovery is the minimum price of complexity in (6).

The frontier is attained by revealing feature 1 at $(0,0)$ with probability $\min\{2t,1\}$ and at (a,b) with probability $\max\{2t-1,0\}$, and feature 2 otherwise. We call $t \in [0,1/2]$ and $t \in [1/2,1]$ the two *phases* of the frontier. At $t = 1/2$, the frontier policy reveals feature 1 exactly at $(0,0)$ and is therefore deterministic; we call it the *intermediate deterministic policy*. Figure 1 shows how the two feature scales change the frontier. Each phase is strictly concave, and the slope changes at $t = 1/2$.

The frontier also determines the optimal policy in (7) for each population share λ. Comparing its endpoints gives the threshold

$$\lambda_c = \frac{b^2-a^2}{a^2+b^2}. \tag{11}$$

For $\lambda < \lambda_c$, the naive optimum is optimal. For $\lambda_c < \lambda < 1$, the optimal risk pair is that of perfect sophisticated recovery with minimum naive risk. At $\lambda = \lambda_c$, both endpoints are optimal.

Putting these pieces together, the Pareto frontier in the two-feature example has two concave phases, demonstrating the tradeoff between naive and sophisticated risks. As the sophisticated population share crosses λ_c, the optimal policy switches from minimizing naive risk to achieving perfect sophisticated recovery with the smallest increase in naive risk.

6 Computational Complexity and Algorithms

	Time Complexity	
	Fixed k	k part of the input
Fixed highlighting (Gaussian prior)	Feasible Enumeration using (24)	NP-hard (Theorem 2)
Sophisticated contextual (finite prior)	NP-hard even for $k = 1$ (Theorem 4)	NP-hard (Theorem 4)
Naive contextual (finite prior)	Feasible (Proposition 7)	NP-hard (Theorem 5)

Table 1: Computational complexity of optimal highlighting under quadratic recovery loss.

Notes. Feasible means polynomial-time exact optimization. Finite priors are given by explicitly listed support states and probabilities. For fixed k, fixed highlighting enumerates $O(d^k)$ subsets and evaluates each risk using (24). Sophisticated hardness holds even for a 1.001-approximation. The naive problem with k as part of the input has an NP-complete decision version.

We have shown that optimal highlighting policies can differ substantially depending on whether they are designed for naive or sophisticated decision-makers. We now examine how this distinction affects the computation of optimal policies and consider design for a mixture of the two types. Table 1 summarizes the complexity of exact optimization.

We begin with fixed highlighting, for which the same subset of features is revealed for every state. Under fixed highlighting, the sophisticated and naive receivers have the same posterior, so the two design problems coincide. This problem is NP-hard when bandwidth is part of the input, even under a Gaussian prior. We then study contextual feature highlighting, in which the highlighted set can depend on the realized state. In the finite-prior model, sophisticated design is hard to approximate even with bandwidth one, whereas naive design is tractable for fixed bandwidth and NP-hard when bandwidth is part of the input. Under Gaussian priors, we develop fixed greedy and contextual projection greedy algorithms with spectral approximation guarantees.

Finally, we study design for a mixed audience. We bound the additional expected loss from deploying a naive-optimal policy to a mixed population and give a minimum-cost matching formulation for minimizing naive risk while preserving perfect sophisticated recovery. These are complementary approaches to mixed-audience design: tractable naive optimization at fixed bandwidth, and minimum naive risk when perfect sophisticated recovery is feasible. We compare the performance of the policies in the case study in Section 7.

6.1 Fixed Highlighting

Fixed highlighting is closely related to subset selection, since the designer chooses one set before observing the state. We show that this problem is NP-hard even under a Gaussian prior and quadratic loss, assuming that the bandwidth k increases. Nevertheless, a greedy algorithm, which we call fixed greedy, achieves a spectral approximation guarantee for the improvement over revealing nothing.

Let $X \sim \mathcal{N}(\mu, \Sigma)$ with $\mu \in \mathbb{R}^d$ and $\Sigma \succ 0$, and suppose every coordinate is revealable. Take the quadratic loss L_Q for $Q \succeq 0$. For a fixed set I, both receivers use the posterior $\widehat{\mathbb{P}}_I^N$ and its mean $\mathbb{E}[X \mid X_I]$. Writing σ_I for the policy that always reveals I, their common risk is

$$R_Q^{\mathrm{fix}}(I) = R_Q^N(\sigma_I) = R_Q^S(\sigma_I) = \mathbb{E}\Big[L_Q(\widehat{\mathbb{P}}_I^N, X)\Big].$$

Theorem 2 (Hardness of fixed highlighting)**.** *Given $\Sigma \succ 0$ and $Q \succeq 0$, finding a set $I \subseteq [d]$ with $|I| \le k$ that minimizes $R_Q^{\mathrm{fix}}(I)$ is NP-hard when k is part of the input.*

The proof gives a reduction from exact cover by three-sets through a full-rank sparse-regression construction, following the approach of Natarajan (1995). Although finding an optimal set is NP-hard in general, the greedy algorithm below provides a spectral approximation guarantee for the gain over revealing nothing.

Fixed greedy. Starting from the empty set, fixed greedy adds the feature that most reduces expected loss and repeats for k steps. Write I^{greedy} for its output and $f_Q(I) = R_Q^{\mathrm{fix}}(\emptyset) - R_Q^{\mathrm{fix}}(I)$ for the gain over revealing nothing. Each step accounts for how the selected features change the value of the remaining ones.

Theorem 3 (Approximation for fixed highlighting)**.** *Suppose $\Sigma \succ 0$, $Q \succeq 0$, and $1 \le k \le d$. Let $\Psi = (\Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}})_{i,j=1}^d$ be the correlation matrix of X, and define*

$$m_{2k} = \min_{\substack{T \subseteq [d] \\ 1 \le |T| \le \min(2k, d)}} \lambda_{\min}(\Psi_{TT}),$$

where $\lambda_{\min}$ *denotes the smallest eigenvalue. For* $I^* \in \operatorname{argmin}_{|I|\le k} R_Q^{\mathrm{fix}}(I)$*, fixed greedy satisfies*

$$f_Q(I^{\mathrm{greedy}}) \ge (1 - e^{-m_{2k}}) f_Q(I^*). \tag{12}$$

Proof sketch. The gain f_Q is a sum of regression gains in standardized coordinates, so, as in Das and Kempe (2018), the sparse eigenvalues of Ψ bound its submodularity ratio below by m_{2k}. Hence at each step some unselected feature closes at least a fraction m_{2k}/k of the remaining gap to $f_Q(I^*)$, and greedy does at least as well. After k steps, the remaining gap is at most $(1 - m_{2k}/k)^k \le e^{-m_{2k}}$ times its initial value. △

Appendix A.2 gives the marginal-gain formula, implementation, and proofs.

6.2 Contextual Feature Highlighting

Unlike fixed highlighting, contextual highlighting allows the algorithm to select a different subset of features for each state. For a naive receiver, the designer can choose the highlighted set separately for each state because the posterior depends only on the selected features and their values. For a sophisticated receiver, the posterior also depends on which other states the policy assigns to the same selected set and values. We first examine this computational distinction in the *finite-prior model*. We then return to the Gaussian model and give an approximation for naive contextual design under general quadratic recovery loss.

In the finite-prior model, the prior is given as input by an explicit list of distinct *support states* $X^1, \dots, X^n \in \mathbb{R}^d$ and positive probabilities $p_1, \dots, p_n$ summing to one. Features may be arbitrarily dependent, and running times are measured in n and d. With *equally likely binary support states*, every X^s lies in $\{0,1\}^d$ and $p_s = 1/n$. In this section, we focus on the quadratic recovery loss with $Q \succeq 0$.

Sophisticated design. We first show that deterministic policies suffice for the optimization problem, then formulate sophisticated design as a mixed-integer nonlinear program. Furthermore, we show that the problem is NP-hard even for equally likely binary support states and $Q = I_d$.

Lemma 1 (Deterministic policies suffice for sophisticated risk)**.** *In the finite-prior model with* $Q \succeq 0$*, every randomized highlighting policy has a deterministic counterpart with weakly lower sophisticated risk. In particular, the minimum sophisticated risk is attained by a deterministic policy.*

Sophisticated risk is concave in the policy because it minimizes a linear objective over message-contingent reconstructions. A randomized policy is a convex combination of deterministic policies, so one component has weakly lower risk.

We now formulate the sophisticated design problem over deterministic policies. Recall the menu $\mathcal{M}_k$ and write $J_{d,k} = |\mathcal{M}_k|$ for its size. Let $\mathcal{Z} = \{(I, X_I^s) : I \in \mathcal{M}_k,\ s = 1, \dots, n\}$ be the set of possible messages. For $z = (I, v)$, let $\mathcal{S}_z = \{s : X_I^s = v\}$ index the states compatible with that message. The binary variable u_{sz} assigns state $s \in \mathcal{S}_z$ to message z, and $\widehat{X}_z$ is the reconstruction after that message. The design problem is

$$\begin{aligned} \min_{u, \widehat{X}} \quad & \sum_{z \in \mathcal{Z}} \sum_{s \in \mathcal{S}_z} p_s u_{sz} \|\widehat{X}_z - X^s\|_Q^2 & \\ \text{subject to} \quad & \sum_{z:\, s \in \mathcal{S}_z} u_{sz} = 1 & s = 1, \dots, n, \\ & u_{sz} \in \{0,1\} & z \in \mathcal{Z},\ s \in \mathcal{S}_z, \\ & \widehat{X}_z \in \mathbb{R}^d & z \in \mathcal{Z}. \end{aligned} \tag{13}$$

For fixed assignments, an optimal reconstruction at each used message is the posterior mean of its assigned states,

$$\widehat{X}_z = \frac{\sum_{s\in\mathcal{S}_z} p_s u_{sz} X^s}{\sum_{s\in\mathcal{S}_z} p_s u_{sz}}.$$

Thus minimizing over reconstructions gives exactly the sophisticated risk of the assigned policy, and Lemma 1 shows that the program also attains the optimum over randomized policies. The formulation has $nJ_{d,k}$ binary variables and requires enumerating the menu, which can be exponentially large when k varies.

Theorem 4 (Inapproximability of sophisticated highlighting)**.** *It is NP-hard to compute a* 1.001*-approximation to the minimum sophisticated expected loss, even in the finite-prior model with bandwidth $k = 1$, equally likely binary support states, and $Q = I_d$. Thus the problem has no polynomial-time approximation scheme unless* P = NP*.*

The reduction starts from the binary Euclidean clustering instances of Lee et al. (2017). Most messages are reserved for individually identified auxiliary states. The remaining messages represent clusters whose centroids are posterior means, as detailed in Appendix A.3.

Theorem 4 shows that sophisticated design is hard to approximate even with bandwidth $k = 1$. This difficulty arises because the receiver accounts for the highlighting rule, which links the policy's choices across states. We next turn to naive design, where the designer can optimize separately at each state.

Naive design. In contrast to sophisticated updating, a naive receiver's posterior depends only on the highlighted values and treats the selected set as fixed. The designer can therefore choose the highlighted set separately for each state, making the problem polynomial-time solvable for each fixed k in the finite-prior model. The problem becomes NP-hard when k is part of the input. Under a Gaussian prior, revealing an additional feature can increase realized naive loss, so actual-loss greedy can stop prematurely. This motivates the projection greedy method below, for which we establish an approximation guarantee.

Proposition 7 (Optimality for naive contextual highlighting with constant k)**.** *In the finite-prior model, define $L_Q^N(I;X) = \|\widehat{X}(\widehat{\mathbb{P}}_I^N) - X\|_Q^2$. An optimal naive policy is obtained by choosing*

$$\sigma^N(X) \in \operatorname*{argmin}_{I\subseteq[d],\,|I|\le k} L_Q^N(I;X)$$

separately at every support state. The problem is polynomial-time solvable for each fixed k.

When k is part of the input, the number of feasible sets grows exponentially, and the problem is NP-hard even for equally likely binary support states and $Q = I_d$.

Theorem 5 (NP-hardness of naive contextual highlighting with general k)**.** *In the finite-prior model with $Q \succeq 0$, deciding whether there exists a contextual policy with $R_Q^N(\sigma) \le t$ for a threshold t is NP-complete when k is part of the input. This remains true for equally likely binary support states, $Q = I_d$, and $t = 0$.*

The zero-risk reduction in Appendix A.3 also rules out a finite multiplicative approximation of naive risk on this class unless P = NP. We therefore turn to a Gaussian prior and seek an efficient approximation with a guarantee relative to the risk of revealing nothing.

Greedy approximation under a Gaussian prior. We return to the Gaussian model of Section 6.1, with $X \sim \mathcal{N}(\mu, \Sigma)$, $\Sigma \succ 0$, $Q \succ 0$, and every coordinate observed and revealable. The realized naive loss $L_Q^N(I; X)$ is defined as in Proposition 7, and $L_{Q,0}(X) = \|X - \mu\|_Q^2$ is the loss of the empty highlight.

Contextual actual-loss greedy starts from the empty set and adds up to k coordinates, each time choosing the largest positive reduction in realized naive loss and stopping when none exists. It can stop prematurely because individual disclosures may raise loss even when their combination permits exact recovery, as Example 3 shows.

Example 3 (Correlated features and premature stopping)**.** *Let $X \sim \mathcal{N}(0, \Sigma)$, with*

$$\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \qquad Q = I_2, \qquad k = 2.$$

At the realized state $X = (1, -1)^\top$, the naive losses are

$$L_Q^N(\emptyset; X) = 2, \qquad L_Q^N(\{1\}; X) = L_Q^N(\{2\}; X) = 2.25, \qquad L_Q^N(\{1, 2\}; X) = 0.$$

Actual-loss greedy stops without revealing anything, although revealing both features permits exact recovery.

Contextual projection greedy projects $Q^{1/2}(X - \mu)$ onto the span of the selected covariance directions $Q^{1/2}\Sigma_{:j}$. Write $\widetilde{L}_Q(I; X)$ for the squared projection residual. Each step adds the coordinate that most reduces this residual, even if actual loss rises. The search runs for k steps or until a prefix has zero actual loss, and returns the prefix $\sigma_Q^{\text{proj}}(X)$ with the lowest actual loss, including the empty set. Appendix A.3 gives the formulas and Algorithm 1.

Define m_{2k}^Q analogously to m_{2k} from the Gram matrix $\Gamma_Q = \Sigma Q \Sigma$ after normalizing its diagonal to one, and put $\zeta_Q = 1 - e^{-m_{2k}^Q}$. Let $\omega_{Q,k}$ be the smallest constant such that $L_Q^N(I; X) \le \omega_{Q,k} \widetilde{L}_Q(I; X)$ for every X and $|I| \le k$. This finite constant measures the discrepancy between actual and projected loss.

Theorem 6 (Approximation for contextual projection greedy)**.** *Under this Gaussian model, let $1 \le k \le d$ and $L_{Q,k}^{N,\star}(X) = \min_{|I| \le k} L_Q^N(I; X)$. For every X, contextual projection greedy satisfies*

$$L_Q^N(\sigma_Q^{\text{proj}}(X); X) \le \min\left\{ L_{Q,0}(X),\, \omega_{Q,k}\left[(1 - \zeta_Q) L_{Q,0}(X) + \zeta_Q L_{Q,k}^{N,\star}(X)\right] \right\}. \tag{14}$$

Proof sketch. The projection gain is a regression gain, so the fixed-greedy argument gives the factor ζ_Q. The naive reconstruction supplies feasible projection coefficients, which bounds the best projected loss by the best actual loss. Multiplying by $\omega_{Q,k}$ controls actual loss at the search endpoint, and best-prefix selection also bounds loss by the empty highlight. △

Corollary 3 (Mahalanobis weights)**.** *Under the assumptions of Theorem 6, if $Q = \Sigma^{-1}$, then $\omega_{Q,k} = 1$ and projection gain equals actual naive gain $F_Q(I; X) = L_{Q,0}(X) - L_Q^N(I; X)$. Let $\sigma(X)$ be the set returned by either contextual actual-loss greedy or contextual projection greedy. Then, for every X,*

$$F_Q(\sigma(X); X) \ge (1 - e^{-m_{2k}}) \max_{|I| \le k} F_Q(I; X),$$

where m_{2k} is the constant of Theorem 3.

Here the projection coefficients coincide with those of the naive conditional mean, so the two rules have the same marginal gains. Appendix A.3 gives the detailed proofs and expected-loss implications.

6.3 Mixed Design for Sophisticated and Naive Receivers

In practice, a designer may face a mixture of naive and sophisticated receivers. An optimal policy for one type need not be optimal for the other. We first bound the cost of deploying a naive-optimal policy to a mixed audience using the information conveyed by the selected set. For independent binary features, a sharper bound yields vanishing total excess risk when the expected number of atypical features grows and bandwidth exceeds it by a fixed proportion. We then consider the opposite endpoint, achieving perfect sophisticated recovery with the smallest loss for naive receivers.

Naive-oriented design. We call a policy *naive-oriented* if it minimizes, or nearly minimizes, naive risk. There are three reasons to optimize for naive receivers. First, a naive-optimal policy is robust to unknown sophistication, since it is optimal for the naive receiver and the sophisticated receiver cannot do worse under the same policy. Second, a naive-optimal policy is often easier to compute than a sophisticated-optimal policy, as in Section 6.2. Third, a naive-optimal policy is often easier to interpret and deploy in practice. We therefore ask what performance guarantee optimizing for naive receivers provides for a mixed audience. We then illustrate the result with independent binary features, where naive optimization has a simple form.

Theorem 7 (Excess risk of naive-oriented design)**.** *Suppose $X_i \sim \text{Bernoulli}(p_i)$ independently, with $p_i \in (0,1)$, and take $Q = I_d$. A naive-optimal policy σ^N reveals the k coordinates with the largest values of $(X_i - p_i)^2$, with any fixed tie-breaking rule. For every $\lambda \in [0,1]$,*

$$0 \le R_Q^{\lambda}(\sigma^N) - R_Q^{\lambda,\star} \le \frac{\lambda}{2} \log J_{d,k}. \tag{15}$$

The proof bounds the difference between naive and sophisticated risk under any policy by one half of the entropy of its selected set. This controls the possible value of selection information beyond the highlighted values. We defer the proof to Appendix A.4.

Corollary 4 (Bounded-block dependence)**.** *Suppose the binary coordinates form mutually independent blocks $B_1, \dots, B_G$ with $\max_g |B_g| \le b_{\max}$, allowing arbitrary dependence within each block. Retain $Q = I_d$ and the menu $\mathcal{M}_k$. A naive-optimal policy under the joint block prior satisfies*

$$0 \le R_Q^{\lambda}(\sigma^N) - R_Q^{\lambda,\star} \le \frac{\lambda b_{\max}}{2} \log J_{d,k}.$$

For independent coordinates, we can sharpen the guarantee when bandwidth accommodates most deviations from the modal feature values.

Proposition 8 (Sparse deviations)**.** *Under the assumptions of Theorem 7, suppose $p_i \neq 1/2$ and $1 \le k \le d$. For the naive-optimal policy σ^N defined there, let*

$$b_i = \mathbb{1}\{p_i > 1/2\}, \qquad q_i = \min\{p_i, 1-p_i\}, \qquad N_{\text{dev}} = \sum_{i=1}^{d} \mathbb{1}\{X_i \neq b_i\}, \qquad s = \mathbb{E}[N_{\text{dev}}] = \sum_{i=1}^{d} q_i.$$

For every $\lambda \in [0,1]$,

$$0 \le R_Q^{\lambda}(\sigma^N) - R_Q^{\lambda,\star} \le \lambda R_Q^{S}(\sigma^N) \le \lambda \mathbb{E}[(N_{\text{dev}} - k)_+], \tag{16}$$

where $(u)_+ = \max\{u, 0\}$. If $k \ge (1+\delta)s$ for some $\delta > 0$, then

$$R_Q^{\lambda}(\sigma^N) - R_Q^{\lambda,\star} \le \frac{\lambda}{\delta} e^{-sh(\delta)}, \qquad h(\delta) = (1+\delta)\log(1+\delta) - \delta > 0. \tag{17}$$

The policy reveals atypical features first, so imputing modal values for omitted coordinates incurs loss $(N_{\text{dev}} - k)_+$. This bounds sophisticated risk, and naive optimality bounds the excess risk by λ times that risk. A concentration bound gives the exponential rate. Appendix A.4 provides the proof and the entropy bound's per-feature implication.

Consider a sequence of problems satisfying the proposition's assumptions, and write $R^{\lambda,\star}_{Q,d}$ for the mixed optimum in dimension d. If

$$s_d \longrightarrow \infty, \qquad s_d/d \longrightarrow 0, \qquad k_d = \lceil (1+\delta)s_d \rceil$$

for fixed $\delta > 0$, then $k_d/d \to 0$ and

$$\sup_{\lambda \in [0,1]} \left[R^{\lambda}_{Q}(\sigma^N_d) - R^{\lambda,\star}_{Q,d} \right] \longrightarrow 0. \tag{18}$$

Thus a vanishing highlighted fraction suffices for vanishing total excess risk uniformly over audience composition in this regime.

For example, let $p_i = d^{-1/2}$ for every coordinate and $k_d = \lceil 2\sqrt{d} \rceil$, with integers $d \geq 5$. Then $s_d = \sqrt{d}$, and the total excess risk is at most $e^{-c\sqrt{d}}$ for every λ, where $c = 2\log 2 - 1 > 0$. The no-disclosure risk is $R_{Q,0} = \sqrt{d} - 1$, while

$$R^N_Q(\sigma^N_d) \leq 1 + e^{-c\sqrt{d}}, \qquad R^S_Q(\sigma^N_d) \leq e^{-c\sqrt{d}}.$$

Highlighting therefore keeps naive risk bounded and makes sophisticated risk vanish even as the risk of no disclosure diverges.

Perfect sophisticated recovery. We next ask how to achieve the sophisticated optimum when perfect recovery is feasible, while minimizing the resulting naive risk.

Use the finite-prior model of Section 6.2, with a nonempty menu $\mathcal{M}$ of sets satisfying $|I| \leq k$ given as part of the input. Let $\Pi(\mathcal{M})$ contain all policies selecting from this menu. A possible message is $z = (I, v)$ with $I \in \mathcal{M}$ and $v = X^s_I$ for some state s. Let $\mathcal{Z}$ be the set of these messages and $\mathcal{S}_z = \{s : X^s_I = v\}$ its compatibility class. Write $\widehat{\mathbb{P}}^N_z = \mathbb{P}(\cdot \mid X_I = v)$ for the naive posterior. The naive reconstruction and the probability-weighted cost of assigning state $s \in \mathcal{S}_z$ to message z are

$$\widehat{X}^N_z = \frac{\sum_{t \in \mathcal{S}_z} p_t X^t}{\sum_{t \in \mathcal{S}_z} p_t}, \qquad c_{sz} = p_s L_Q(\widehat{\mathbb{P}}^N_z, X^s) = p_s \|\widehat{X}^N_z - X^s\|^2_Q.$$

Theorem 8 (Perfect sophisticated recovery with minimum naive risk)**.** *Suppose $Q \succeq 0$ and any two distinct states s, t satisfy $\|X^s - X^t\|^2_Q > 0$. The* compatibility graph *has a vertex for each support state and each message $z \in \mathcal{Z}$, and an edge between s and z whenever $s \in \mathcal{S}_z$; a matching that covers every state vertex is a* covering matching*. Then*

$$\min_{\sigma \in \Pi(\mathcal{M}):\, R^S_Q(\sigma) = 0} R^N_Q(\sigma) \tag{19}$$

is a minimum-cost bipartite matching problem on this graph. An integral optimum of the following linear program gives an optimal deterministic policy, assigning state s to its unique message with $u_{sz} = 1$,

$$\begin{aligned} \min_{u \geq 0} \quad & \sum_{z \in \mathcal{Z}} \sum_{s \in \mathcal{S}_z} c_{sz} u_{sz} \\ \textit{subject to} \quad & \sum_{z:\, s \in \mathcal{S}_z} u_{sz} = 1 \quad \textit{for every state } s, \\ & \sum_{s \in \mathcal{S}_z} u_{sz} \leq 1 \quad \textit{for every message } z. \end{aligned} \tag{20}$$

Perfect sophisticated recovery is feasible if and only if this linear program is feasible. The problem can be solved in time polynomial in n, $|\mathcal{M}|$, and d.

The result permits dependent features and unequal state probabilities. Since the support states are distinct, $Q \succ 0$ guarantees the required separation. With the unrestricted at-most-k menu, the runtime is polynomial in the input length and $J_{d,k}$. In Example 2, the least naive risk compatible with zero sophisticated risk is $(a^2 + b^2)/8$, giving the endpoint of the frontier in Proposition 6.

Remark 1 (Extension beyond quadratic loss)**.** *The matching formulation extends to nonnegative losses $\ell(a, X)$ with attained posterior Bayes minima if each state admits a zero-loss action and no action has zero loss at distinct states. Zero sophisticated risk then requires state-identifying messages. Use $c_{sz} = p_s \ell(a_z^N, X^s)$, where a_z^N is a Bayes action under $\widehat{\mathbb{P}}_z^N$ with fixed tie-breaking. Matching and integrality are unchanged.*

7 Simulation Case Study Based on Ames Housing

We now illustrate feature highlighting in a calibrated numerical exercise using Ames housing data. Home valuation involves many potentially relevant attributes, yet a buyer or appraiser may have limited bandwidth to inspect the details of a listing. An algorithm can help by choosing which attributes to highlight. After observing the highlighted information, the receiver updates beliefs about the property's features and price. The receiver may care about both feature recovery and price prediction.

Throughout, we consider two designer information settings. In the first, the algorithm has access only to housing attributes. In the second, it also conditions on price, which cannot itself be highlighted. We evaluate the policies introduced in Section 6 under these two settings.

The simulation study yields four main findings. First, contextual highlighting generally outperforms fixed highlighting. Second, how receivers interpret information matters for policy design. Third, actual-loss greedy is near optimal at large budgets, while the relative performance of the greedy methods depends on the setting. Finally, highlighting fewer features can reduce the receiver's risk.

7.1 Setup and Data

The Ames data describe residential property sales in Ames, Iowa, from 2006 to 2010 (De Cock, 2011). We calibrate the model on 1,756 properties, which we call the calibration sample. The data include ten features that can be highlighted and a sale price which we assume cannot be highlighted. The features, which we call fields, cover construction, size, quality, and amenities, with one binary field and nine three-category fields. Their categories appear in Table 3 in Appendix B. Location, identifiers, and sale-calendar variables are excluded.

7.2 Belief Model and Risk

We fit a joint Gaussian to the ten numerical measurements underlying the fields, in their recorded units (for example, year built or living area in square feet), and to natural-log sale price, using their empirical mean and covariance. We then draw 100,000 observations, bin each numerical measurement into its categories, and exponentiate log price to obtain dollar price. The resulting categorical profile X^r and unbinned price V^r define the empirical prior

$$\mathbb{P}_n = \frac{1}{n} \sum_{r=1}^{n} \delta_{(X^r, V^r)}, \qquad n = 100{,}000. \tag{21}$$

We call (21) the primary prior. It assigns positive probability to 10,808 distinct categorical profiles, which we call *support profiles*. Both receivers compute posteriors from this same empirical prior. Appendix B gives the calibration details.

Write $D = X$ when the designer observes only the categorical profile and $D = (X, V)$ when it also observes price. In the first setting, identical profiles receive identical highlights and selection uses the conditional price mean $m_x = \mathbb{E}[V \mid X = x]$. In either setting, the receiver sees only the final unordered message (I, X_I), with $I = \sigma(D)$. The naive receiver predicts price by $\mathbb{E}[V \mid X_I = x_I]$ as if I were fixed, whereas the sophisticated receiver also conditions on $\sigma(D) = I$. Each predicts categorical features using the corresponding posterior probabilities. Thus sophisticated updating accounts for both the highlighted values and the selection rule, including the information in a shorter or empty highlight. All expectations below refer to (21).

Loss and risk. For price prediction $\widehat{v}$ and feature-probability predictions $\widehat{p} = (\widehat{p}_j)_{j=1}^{10}$, we use

$$\ell_\alpha(\widehat{v}, \widehat{p}; x, v) = \alpha \frac{(\widehat{v} - v)^2}{\mathbb{V}\mathrm{ar}(V)} + (1 - \alpha) \sum_{j=1}^{10} w_j \frac{\|\widehat{p}_j - e_{x_j}\|_2^2}{b_j}, \qquad \sum_{j=1}^{10} w_j = 1. \tag{22}$$

Here e_{x_j} is the realized one-hot vector and $b_j = 1 - \|\pi_j\|_2^2$ is the no-disclosure Brier risk for prior category probabilities π_j. The loss ℓ_α combines price-prediction and feature-recovery errors. The fixed feature weights w_j come from grouped ridge coefficients fitted on the calibration sample. We set $\alpha = 1/2$, weighting price prediction and feature recovery equally.

We assume that the designer knows this loss and that preferences are aligned. We write R^N and R^S for the naive and sophisticated risks, each defined as the expected loss $\mathbb{E}_\sigma[\ell_\alpha]$ under (21). We compute each risk by averaging the loss over the n draws. This leads to the quadratic recovery risk of Section 3 with X replaced by the stacked target of feature indicators and price and Q by the diagonal weight matrix Q_α of Appendix B.3.

When the designer observes the full feature vector but not price, *full disclosure*, which highlights all ten fields, attains the common lower bound on both receivers' risks

$$\kappa_\alpha = \alpha \frac{\mathbb{E}[\mathbb{V}\mathrm{ar}(V \mid X)]}{\mathbb{V}\mathrm{ar}(V)}. \tag{23}$$

If the designer also observes price, selective highlighting can communicate information beyond X and achieve lower risk than full disclosure.

7.3 Highlighting Policies

We compare five methods at bandwidths $k = 1, \ldots, 10$, which we call disclosure budgets in this section. The suffix $(=k)$ marks a method that highlights exactly k fields, and $(\leq k)$ marks one that may highlight fewer, including none. The selection score is the naive receiver's conditional risk given the designer's information D. It integrates over price uncertainty when $D = X$ and uses realized price error when $D = (X, V)$. The methods implement the fixed and contextual design problems in Section 6 under the empirical prior.

- *Best fixed subset* $(=k)$. Enumerate all size-k subsets and choose the one with the lowest risk, following the fixed-design formulation in Section 6.1.
- *Exact selection* $(\leq k)$. Following the enumeration principle in Proposition 7, minimize the selection score separately for each value of D over all subsets of at most k fields. The variant exact selection $(=k)$ restricts the enumeration to subsets of exactly k fields.

- *Actual-loss greedy* ($\leq k$). Starting from the empty highlight, add the field with the smallest resulting selection score for k steps, including steps that raise the score. Return the lowest-score prefix among steps $0, \ldots, k$, including the empty highlight. This extends the rule in Section 6.2 with continued search and best-prefix selection.

- *Projection greedy* ($\leq k$). Group each field's one-hot indicators and add fields by greatest projection gain for k steps. Return the prefix with the lowest selection score, including the empty set, as in Algorithm 1. We describe the details in Appendix B.3.

- *Profile matching* ($=k$). Use the covering-matching formulation in Theorem 8 to require recovery of the profile X. Minimize naive risk over these assignments using exactly k fields, and write σ_{match} for the resulting policy. Its messages identify X but convey no additional price information within a profile because selection depends only on X.

7.4 Main Findings

We evaluate budgets $k = 1, \ldots, 10$ for the five main methods and also examine the exact selection ($=k$) variant. Figure 2 illustrates selected budgets, while Table 2 reports the complete five-method comparison.

Contextual highlighting outperforms fixed highlighting. Contextual policies generally reduce both receivers' risks relative to the best fixed subset at the same budget. At $k = 5$, the best fixed subset attains $(R^N, R^S) = (0.30, 0.30)$. Exact selection ($\leq k$) reduces these risks to $(0.20, 0.13)$ when the designer observes features alone and to $(0.14, 0.05)$ when it also observes price. Adapting the highlight to the state improves naive predictions, while sophisticated updating extracts additional information from the choice of highlighted features.

Policy choice depends on the receiver. A policy optimized for naive receivers need not minimize sophisticated risk. For example, at $k = 4$, when the designer observes features alone, exact selection ($\leq k$) yields $(R^N, R^S) = (0.25, 0.16)$, while profile matching yields $(0.26, 0.11)$. Profile matching achieves perfect recovery of X for sophisticated receivers and attains their lower bound $\kappa_{1/2} = 0.11$. In this example, it also achieves near-optimal naive risk.

Actual-loss greedy versus projection greedy. With features alone, projection greedy achieves lower naive risk at budgets $k = 4, \ldots, 9$, whereas actual-loss greedy does so at $k = 1, 2, 3$. When price is observed, actual-loss greedy achieves lower naive risk at every budget and lower sophisticated risk except at $k = 4$. At $k = 10$, its risks are $(0.07, 0.01)$, compared with $(0.09, 0.03)$ for projection greedy. Both methods select their best prefix, but actual-loss greedy searches using posterior loss and projection greedy uses covariance-based projection gains. Actual-loss greedy is near optimal for naive receivers at the largest budget. With price observed and $k = 10$, its naive risk is 0.0712, compared with 0.0683 under exact selection ($\leq k$), a gap of 4.2%.

Less can be more. Figure 3 shows that, when the designer observes price, the naive risk of exact selection ($=k$) reaches its minimum of 0.08 at $k = 8$ and rises to 0.11 at $k = 10$. Exact selection ($\leq k$) with $k = 10$ instead attains $(R^N, R^S) = (0.07, 0.01)$ while highlighting 8.51 fields on average. Figure 4 explains the tradeoff for the naive receiver. As k increases, highlighting more features may worsen price prediction while improving feature recovery. In this comparison, the price-prediction gain from allowing fewer disclosures outweighs the feature-recovery cost under equal weights. These

gains rely on the designer's price information and the tradeoff between the prediction and recovery tasks.

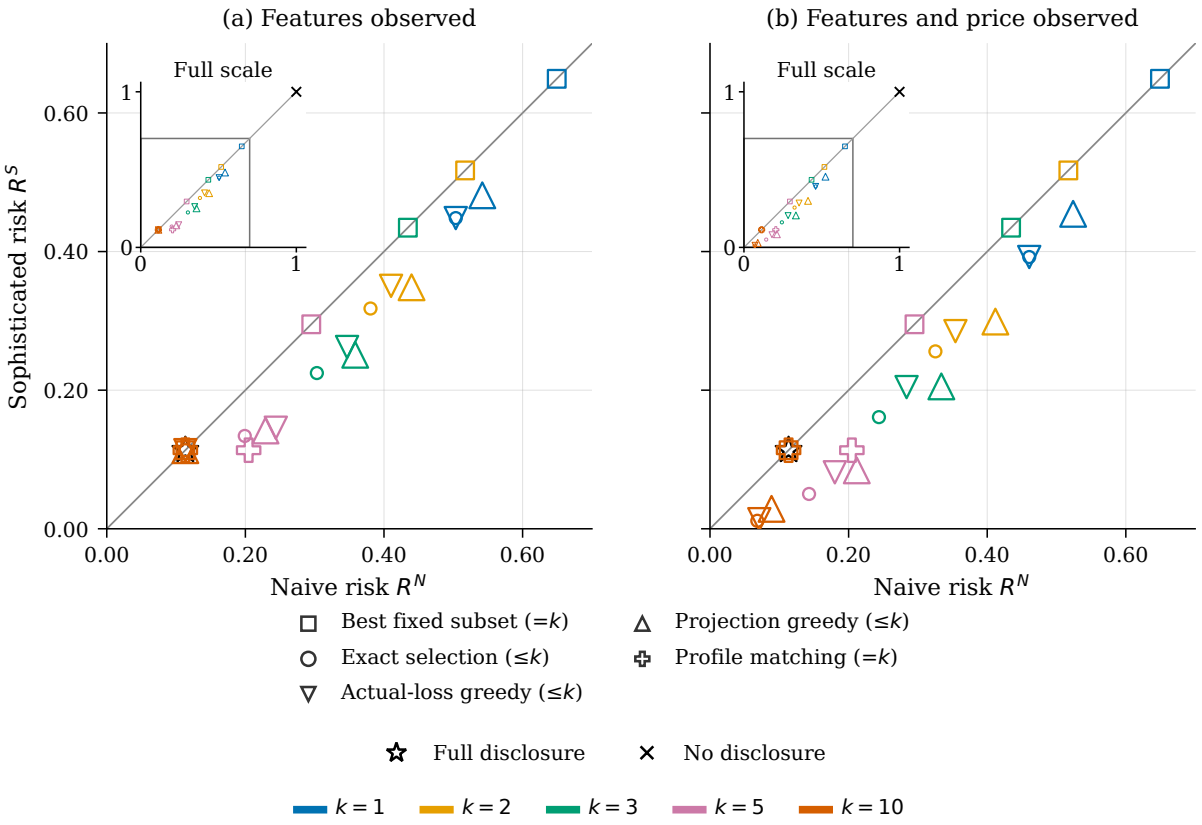


Figure 2: Naive and sophisticated performance under two designer information settings (with price observed or unobserved). Risk combines the price-prediction and feature-recovery components with $\alpha = 1/2$. We compare best fixed subset ($=k$), exact selection ($\leq k$), actual-loss greedy ($\leq k$), projection greedy ($\leq k$), and profile matching ($=k$) at budgets $k \in \{1, 2, 3, 5, 10\}$. No disclosure yields normalized risks $(1, 1)$.

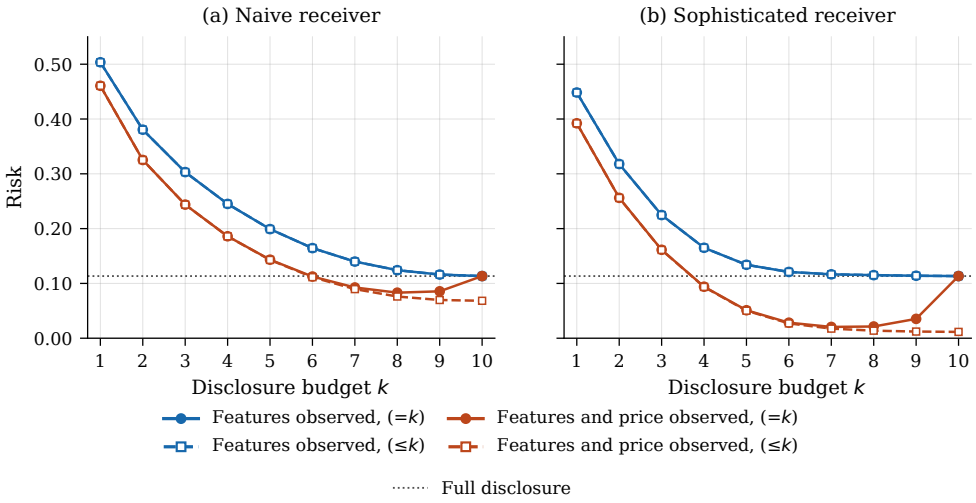


Figure 3: Exact selection ($=k$) and ($\leq k$) for $k = 1, 2, \ldots, 10$. The panels report the naive and sophisticated risks of the naive-optimal policies under the two information settings, with $\alpha = 1/2$. Policies marked ($=k$) use the entire budget, whereas policies marked ($\leq k$) may leave capacity unused.

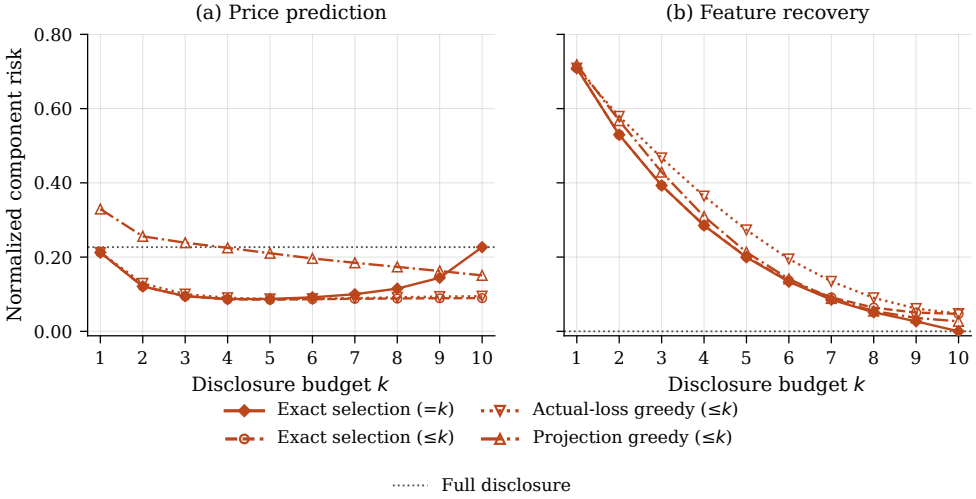


Figure 4: Price prediction and feature recovery when the designer observes price. We consider naive risk with $\alpha = 1/2$. We compare exact selection ($=k$), exact selection ($\leq k$), actual-loss greedy ($\leq k$), and projection greedy ($\leq k$) at budgets $k = 1, 2, \ldots, 10$. The panels report the normalized price-prediction and feature-recovery risk components. Effective feature highlighting should balance both components.

Table 2: Naive and sophisticated risk for all budgets in the five-method comparison. Figure 2 shows selected budgets. Each entry is (R^N, R^S) under the primary empirical prior with equal price and feature weights.

(a) Features observed

k	Best fixed subset ($=k$)	Exact selection ($\leq k$)	Actual-loss greedy ($\leq k$)	Projection greedy ($\leq k$)	Profile matching ($=k$)
1	(0.65, 0.65)	(0.50, 0.45)	(0.50, 0.45)	(0.54, 0.48)	Infeasible
2	(0.52, 0.52)	(0.38, 0.32)	(0.41, 0.35)	(0.44, 0.35)	Infeasible
3	(0.43, 0.43)	(0.30, 0.22)	(0.35, 0.26)	(0.36, 0.25)	Infeasible
4	(0.36, 0.36)	(0.25, 0.16)	(0.29, 0.19)	(0.29, 0.18)	(0.26, 0.11)
5	(0.30, 0.30)	(0.20, 0.13)	(0.24, 0.15)	(0.23, 0.14)	(0.20, 0.11)
6	(0.25, 0.25)	(0.16, 0.12)	(0.20, 0.13)	(0.19, 0.13)	(0.17, 0.11)
7	(0.20, 0.20)	(0.14, 0.12)	(0.17, 0.12)	(0.15, 0.12)	(0.14, 0.11)
8	(0.16, 0.16)	(0.12, 0.12)	(0.14, 0.12)	(0.13, 0.12)	(0.12, 0.11)
9	(0.13, 0.13)	(0.12, 0.11)	(0.12, 0.11)	(0.12, 0.12)	(0.12, 0.11)
10	(0.11, 0.11)	(0.11, 0.11)	(0.11, 0.11)	(0.11, 0.11)	(0.11, 0.11)

(b) Features and price observed

k	Best fixed subset ($=k$)	Exact selection ($\leq k$)	Actual-loss greedy ($\leq k$)	Projection greedy ($\leq k$)	Profile matching ($=k$)
1	(0.65, 0.65)	(0.46, 0.39)	(0.46, 0.39)	(0.52, 0.45)	Infeasible
2	(0.52, 0.52)	(0.33, 0.26)	(0.35, 0.29)	(0.41, 0.30)	Infeasible
3	(0.43, 0.43)	(0.24, 0.16)	(0.28, 0.20)	(0.33, 0.21)	Infeasible
4	(0.36, 0.36)	(0.19, 0.09)	(0.23, 0.14)	(0.27, 0.13)	(0.26, 0.11)
5	(0.30, 0.30)	(0.14, 0.05)	(0.18, 0.08)	(0.21, 0.08)	(0.20, 0.11)
6	(0.25, 0.25)	(0.11, 0.03)	(0.14, 0.05)	(0.17, 0.06)	(0.17, 0.11)
7	(0.20, 0.20)	(0.09, 0.02)	(0.11, 0.03)	(0.14, 0.04)	(0.14, 0.11)
8	(0.16, 0.16)	(0.08, 0.01)	(0.09, 0.02)	(0.11, 0.03)	(0.12, 0.11)
9	(0.13, 0.13)	(0.07, 0.01)	(0.08, 0.02)	(0.10, 0.03)	(0.12, 0.11)
10	(0.11, 0.11)	(0.07, 0.01)	(0.07, 0.01)	(0.09, 0.03)	(0.11, 0.11)

Notes. Values are rounded to two decimals. Methods marked ($=k$) highlight exactly k fields, and methods marked ($\leq k$) highlight at most k. Actual-loss greedy ($\leq k$) searches through k additions, including non-improving steps, and returns the lowest-score prefix, including silence. Infeasible denotes the absence of a covering matching on the support profiles. Both panels use the same profile-matching policies; Appendix B.4 describes their computation and a lower bound on their naive risk. No disclosure has risks $(1.00, 1.00)$, and full disclosure has risks $(0.11, 0.11)$ in both settings.

8 Extensions

We discuss several extensions of our framework. First, we comment on continuous state spaces, where the set of possible states is infinite. Second, we discuss stereotypical updating as an alternative behavioral rule. Third, we incorporate private human information, which affects belief formation. Finally, we discuss incorrect beliefs and misaligned preferences as important extensions for future research.

8.1 Continuous Features and Uniform Highlighting

Our main results focus on discrete distributions, where a highlighting policy maps each state in a finite support to a subset of coordinates. In many applications, however, X is continuous, and the posteriors induced by a policy can be even more complex. In such settings, the designer's problem for a sophisticated receiver becomes a continuous-state information design problem. The following example with uniform features shows that the optimal policies for the two receivers can differ fundamentally.

Example 4 (Continuous uniform features)**.** *Let X_1 and X_2 be independent uniform variables on $[-1, 1]$. Take $Q = I_2$ and $k = 1$. We restrict the algorithm to reveal exactly one feature in every state. For a naive receiver, revealing coordinate j sets $\widehat{X}_j = X_j$ and leaves the other coordinate at its prior mean 0. The optimal naive policy therefore reveals the coordinate with larger magnitude, with arbitrary tie-breaking. Writing this policy as σ^N, we obtain*

$$R^{N,\star} = R^N(\sigma^N) = \mathbb{E}\big[\min\{X_1^2, X_2^2\}\big] = \frac{1}{6}.$$

Under this policy $R^S(\sigma^N) = 1/6$ as well.

For a sophisticated receiver, an optimal policy instead reveals feature 1 when the two coordinates have the same sign and feature 2 otherwise,

$$\sigma^S(x_1, x_2) = \begin{cases} \{1\}, & \text{if } x_1 x_2 \geq 0, \\ \{2\}, & \text{otherwise.} \end{cases}$$

The sophisticated receiver imputes the omitted coordinate as $\operatorname{sgn}(X_1)/2$ after feature 1 is revealed and as $-\operatorname{sgn}(X_2)/2$ after feature 2 is revealed. This yields the sophisticated risk

$$R^{S,\star} = R^S(\sigma^S) = \frac{1}{12}, \qquad R^N(\sigma^S) = \frac{1}{3}.$$

We prove it is the optimal policy in Appendix A.5.

The example shows that the optimal policies for the two receivers can differ. The naive-optimal policy reveals the more extreme feature, whereas the sophisticated-optimal policy uses the choice of feature to encode additional information. The problem becomes much harder when the underlying distribution is Gaussian. For $X \sim \mathcal{N}(0, I_2)$ under the same exactly-one restriction, naive-optimal highlighting still reveals the larger magnitude and has risk $1 - 2/\pi$. Sophisticated design requires optimizing the selection regions jointly with their induced conditional means. We leave the exact globally optimal sophisticated policy and its loss as an open problem, even for this two-dimensional Gaussian benchmark.

8.2 Alternative Behavioral Updating and Stereotypes

Our main model considers sophisticated updating, which conditions on the selection event, and naive updating, which ignores it, as two benchmark rules for belief formation. Both updating rules provide natural but extreme benchmarks. Decision-makers' belief updating may also be influenced by anchoring, availability, and representativeness. As a specific behavioral deviation, we consider *stereotypical updating* (e.g., Bordalo et al., 2016), where the receiver replaces Bayesian conditioning by finding a representative state. In one stylized version, after observing (I, X_I), the receiver selects a single representative state

$$\widehat{X} \in \operatorname*{argmax}_{x \in \mathrm{supp}(\mathbb{P}):\ \ x_I = X_I} \mathbb{P}(X = x \mid \ X_I),$$

and then plays the action that is optimal for $\widehat{X}$. Under squared loss, this corresponds to imputing all missing coordinates by the most likely completion. A detailed study of optimal highlighting against such behavioral rules is an interesting direction for future work.

8.3 Private Human Information and a Minimax Property

Above, we have assumed that a decision state is fully characterized by the vector X that is available to the algorithm. In practice, the human decision-maker may have additional private information. We now allow the human to observe such private information X_H in addition to the algorithm's message (I, X_I). The posterior definitions in Section 3 extend directly to

$$\widehat{\mathbb{P}}^S_\sigma(\cdot \mid X_H, I, X_I) = \mathbb{P}(\cdot \mid X_H, X_I, \sigma(X_M) = I),$$
$$\widehat{\mathbb{P}}^N(\cdot \mid X_H, I, X_I) = \mathbb{P}(\cdot \mid X_H, X_I).$$

The design objective remains to minimize expected loss, except that the Bayes action is now computed under posteriors conditional on X_H.

Private information also provides an additional motivation for robust design. If the algorithm is uncertain about what X_H the human has access to, a natural robust objective is to choose a policy that performs well uniformly over possible private information structures. Because additional private information can only reduce a Bayesian receiver's expected squared error, the worst case for the designer is the case in which the human observes nothing beyond the highlighted features. Consequently, finding the policy that minimizes expected loss under the assumption that $X_H = \emptyset$ may itself be viewed as a robust solution to the problem with private information.

Formally, fix a jointly distributed model for (X_M, X_H) with $X_M \in \mathbb{R}^d$ the machine features available to the designer and X_H an additional square-integrable signal that the human may or may not observe. Write $\mathcal{H}$ for the collection of such private information structures. The degenerate element $\emptyset \in \mathcal{H}$ corresponds to a human with no private information. For a highlighting policy σ with $|\sigma(x)| \le k$ and a receiver type $\tau \in \{S, N\}$, let $\widehat{X}^\tau_M(\sigma; X_I, X_H)$ denote the induced Bayes action for reconstructing X_M, and define the risk with private information H by

$$R^\tau(\sigma; H) \ := \ \mathbb{E}\Big[\big\|X_M - \widehat{X}^\tau_M(\sigma; X_I, X_H)\big\|_2^2\Big], \qquad I = \sigma(X_M),$$

so that $R^\tau(\sigma; \emptyset) = R^\tau_{I_d}(\sigma)$ is the risk of Section 3 under squared loss. Throughout, the revealed vector X_I carries its index tag I, so conditioning on X_I means conditioning on (I, X_I).

Proposition 9 (Policies optimized without private information are minimax for the sophisticated receiver)**.** *For any highlighting policy σ and any private information structure $H \in \mathcal{H}$,*

$$R^S(\sigma; H) \ \le \ R^S(\sigma; \emptyset).$$

Consequently, any policy that is optimal in the problem without private information,

$$\sigma^\star \in \operatorname*{argmin}_{\sigma:\,|\sigma(x)|\le k} R^S(\sigma;\emptyset),$$

is also minimax-optimal over private information structures,

$$\sigma^\star \in \operatorname*{argmin}_{\sigma:\,|\sigma(x)|\le k} \sup_{H\in\mathcal{H}} R^S(\sigma;H).$$

The above minimax proposition implies that optimizing the highlighting policy for a sophisticated human with no private information delivers a robust solution against arbitrary private information channels the designer may not observe or control. For the sophisticated receiver, the posterior mean is the orthogonal projection of X_M onto the full information σ-field generated by the private signal and the highlighting rule, so adding information can only reduce squared loss.

8.4 Wrong Beliefs and Misaligned Preferences

The above model assumes that both sophisticated and naive receivers have a prior that coincides with the actual distribution. Furthermore, the preferences of the sender (designer) and the receiver (decision-maker) are fully aligned. An important natural extension is to a model where both priors and preferences can differ between sender and receiver. Specifically, the receiver may start with a prior $\widetilde{\mathbb{P}}$ and minimize expected loss for a loss function $\widetilde{\ell}$. The sender may then optimize a highlighting policy under their prior $\mathbb{P}$ and loss function ℓ, taking both frictions into account and connecting the model further to Bayesian persuasion (Kamenica and Gentzkow, 2011). Highlighting may be an attractive form of collaboration in this case since revealing features may still lead to reasonable solutions when the sender and receiver have partially misaligned preferences or different beliefs and may not fully trust each other.

9 Conclusion

We study feature highlighting as a way for algorithms to support human decision-makers who can inspect only a few features. Our main takeaway is that good highlighting depends on how people interpret both the highlighted values and the choice of what to show. Sophisticated receivers can learn from selection, but policies that exploit this information can perform poorly for naive receivers. We characterize this tradeoff in a two-feature example and study how to design a common policy for audiences containing both types.

The differences between naive and sophisticated solutions also shape which policies are practical to compute. In the finite-prior model, naive design is tractable at fixed bandwidth, while sophisticated design is hard to approximate even when only one feature can be shown. Our approximation algorithms and mixed-audience guarantees identify settings in which simple policies perform well. A calibrated exercise using Ames housing data illustrates the gains from contextual highlighting and the importance of the designer's information.

More broadly, our results motivate further work on how people interpret selected information and on tractable highlighting policies that account for their private information, beliefs, and preferences. Such policies offer a promising approach to achieving complementarity between human judgment and algorithmic support.

References

Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. Combining human expertise with artificial intelligence: Experimental evidence from radiology. Technical report, National Bureau of Economic Research, 2023. (Cited on pages 4 and 7.)

Nikhil Agarwal, Alex Moehring, and Alexander Wolitzky. Designing human-ai collaboration: A sufficient-statistic approach. Technical report, National Bureau of Economic Research, 2025. (Cited on page 5.)

Rohan Alur, Manish Raghavan, and Devavrat Shah. Human expertise in algorithmic prediction. *Advances in Neural Information Processing Systems*, 37:138088–138129, 2024. (Cited on page 5.)

Saurabh Amin, Amine Bennouna, Daniel Huttenlocher, Dingwen Kong, Liang Lyu, and Asuman Ozdaglar. A bayesian framework for human-ai collaboration: Complementarity and correlation neglect. *arXiv preprint arXiv:2602.14331*, 2026. (Cited on page 7.)

Susan C. Athey, Kevin A. Bryan, and Joshua S. Gans. The allocation of decision authority to human and artificial intelligence. *AEA Papers and Proceedings*, 110:80–84, 2020. (Cited on page 4.)

Maya Balakrishnan, Kris Johnson Ferreira, and Jordan Tong. Human-algorithm collaboration with private information: Naïve advice-weighting behavior and mitigation. *Management Science*, 2025. (Cited on page 5.)

Hamsa Bastani, Osbert Bastani, and Wichinpong Park Sinchaisri. Improving human sequential decision making with reinforcement learning. *Management Science*, 72(1):733–755, 2026. (Cited on page 5.)

Pedro Bordalo, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. Stereotypes. *The Quarterly Journal of Economics*, 131(4):1753–1794, 2016. (Cited on page 30.)

Tamer Boyacı, Caner Canyakmaz, and Francis De Véricourt. Human and machine: The impact of machine input on decision making under cognitive limitations. *Management Science*, 70(2): 1258–1275, 2024. (Cited on page 5.)

Abhimanyu Das and David Kempe. Approximate submodularity and its applications: Subset selection, sparse approximation and dictionary selection. *Journal of Machine Learning Research*, 19(3):1–34, 2018. (Cited on pages 5 and 18.)

Dean De Cock. Ames, iowa: Alternative to the boston housing data as an end of semester regression project. *Journal of Statistics Education*, 19(3), 2011. (Cited on page 23.)

Alfredo Di Tillio, Marco Ottaviani, and Peter Norman Sørensen. Strategic sample selection. *Econometrica*, 89(2):911–953, 2021. (Cited on page 7.)

Kate Donahue, Alexandra Chouldechova, and Krishnaram Kenthapadi. Human-algorithm collaboration: Achieving complementarity and avoiding unfairness. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1639–1656, 2022. (Cited on page 5.)

Laura Doval and Vasiliki Skreta. Constrained information design. *Mathematics of Operations Research*, 49(1):78–106, 2024. (Cited on page 5.)

Benjamin Enke. What you see is all there is. *The Quarterly Journal of Economics*, 135(3):1363–1398, 2020. (Cited on page 7.)

Erik Eyster and Matthew Rabin. Cursed equilibrium. *Econometrica*, 73(5):1623–1672, 2005. (Cited on page 7.)

Ronen Gradwohl, Niklas Hahn, Martin Hoefer, and Rann Smorodinsky. Algorithms for persuasion with limited communication. *Mathematics of Operations Research*, 47(3):2520–2545, 2022. (Cited on page 5.)

Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–24, 2019. (Cited on page 5.)

Sophie Greenwood, Karen Levy, Solon Barocas, Hoda Heidari, and Jon Kleinberg. Designing algorithmic delegates: The role of indistinguishability in human-ai handoff. In *Proceedings of the 26th ACM Conference on Economics and Computation*, pages 306–336, 2025. (Cited on page 4.)

Ziyang Guo, Yifan Wu, Jason Hartline, Kenneth Holstein, and Jessica Hullman. Complllm: Fine-tuning llms to discover complementary signals for decision-making. *arXiv preprint arXiv:2602.19458*, 2026a. (Cited on page 5.)

Ziyang Guo, Yifan Wu, Jason Hartline, and Jessica Hullman. The value of information in human-ai decision-making. In *International Conference on Learning Representations*, volume 2026, pages 41826–41857, 2026b. (Cited on page 5.)

Martin Hoefer, Pasin Manurangsi, and Alexandros Psomas. Algorithmic persuasion with evidence. *ACM Transactions on Economics and Computation*, 12(4):1–34, 2024. (Cited on page 5.)

Florian Hoffmann, Roman Inderst, and Marco Ottaviani. Persuasion through selective disclosure: Implications for marketing, campaigning, and privacy regulation. *Management Science*, 66(11): 4958–4979, 2020. (Cited on page 5.)

Ruru Hoong and Bnaya Dreyfuss. Improving AI-assisted decision-making through calibrated coarsening, 2025. (Cited on page 5.)

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101 (6):2590–2615, 2011. (Cited on pages 2, 5, and 31.)

Euiwoong Lee, Melanie Schmidt, and John Wright. Improved and simplified inapproximability for k-means. *Information Processing Letters*, 120:40–43, 2017. (Cited on pages 19 and 39.)

Elliot Lipnowski, Laurent Mathevet, and Dong Wei. Attention management. *American Economic Review: Insights*, 2(1):17–32, 2020. (Cited on page 9.)

Bryce McLaughlin and Jann Spiess. Algorithmic assistance with recommendation-dependent preferences, 2022. arXiv:2208.07626. (Cited on page 5.)

Bryce McLaughlin and Jann Spiess. Designing algorithmic recommendations to achieve human-ai complementarity. *arXiv preprint arXiv:2405.01484*, 2024. (Cited on page 5.)

Omer Nahum, Gali Noti, David Parkes, and Nir Rosenfeld. Decongestion by representation: Learning to improve economic welfare in marketplaces. In *International Conference on Learning Representations*, volume 2024, pages 30047–30071, 2024. (Cited on page 5.)

Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995. (Cited on page 17.)

Gali Noti, Kate Donahue, Jon Kleinberg, and Sigal Oren. Ai-assisted decision making with human learning. In *Proceedings of the 26th ACM Conference on Economics and Computation*, pages 38–39, 2025. (Cited on page 4.)

Luis Rayo and Ilya Segal. Optimal information disclosure. *Journal of Political Economy*, 118(5): 949–987, 2010. (Cited on page 5.)

Daniel Russo and James Zou. How much does your data exploration overfit? controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1):302–323, 2019. (Cited on page 45.)

Christopher A Sims. Implications of rational inattention. *Journal of Monetary Economics*, 50(3): 665–690, 2003. (Cited on page 9.)

Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12), 2024. (Cited on page 4.)

Marcos M. Vasconcelos and Urbashi Mitra. Observation-driven scheduling for remote estimation of two Gaussian random variables. *IEEE Transactions on Control of Network Systems*, 7(1): 232–244, 2020. doi: 10.1109/TCNS.2019.2900864. (Cited on page 5.)

Guanghui Yu, Wei Tang, Saumik Narayanan, and Chien-Ju Ho. Encoding human behavior in information design through deep learning. *Advances in neural information processing systems*, 36:7506–7528, 2023. (Cited on page 5.)

A Proofs

A.1 Omitted proofs for Section 5

Proof of Proposition 4. Under the common true prior, the sophisticated reconstruction is the conditional mean given the highlight (I, X_I). Write

$$D(I, X_I) = \widehat{X}^N(I, X_I) - \widehat{X}^S_\sigma(I, X_I).$$

Since $\widehat{X}^S_\sigma(I, X_I) = \mathbb{E}_\sigma[X \mid I, X_I]$, the conditional residual has mean zero,

$$\mathbb{E}_\sigma[\widehat{X}^S_\sigma(I, X_I) - X \mid I, X_I] = 0.$$

The vector $D(I, X_I)$ is determined by the highlight, and Q is fixed. The law of iterated expectations therefore gives

$$\begin{aligned}\mathbb{E}_\sigma[(\widehat{X}^S_\sigma - X)^\top Q D] &= \mathbb{E}_\sigma\Big[\mathbb{E}_\sigma[\widehat{X}^S_\sigma - X \mid I, X_I]^\top Q D\Big] \\ &= 0.\end{aligned}$$

Expanding $\widehat{X}^N - X = (\widehat{X}^S_\sigma - X) + D$ in the quadratic loss and taking expectations yields

$$R^N_Q(\sigma) - R^S_Q(\sigma) = \mathbb{E}_\sigma\|D\|^2_Q = \mathbb{E}_\sigma\big[\|\widehat{X}^N - \widehat{X}^S_\sigma\|^2_Q\big] \geq 0,$$

where the inequality follows from $Q \succeq 0$. □

Proof of Proposition 5 and Corollary 2. Linearity of the posterior mean gives $\widehat{X}^{\theta}_{\sigma} = \theta\widehat{X}^{S}_{\sigma} + (1-\theta)\widehat{X}^{N}$, which minimizes posterior expected quadratic loss. With $D = \widehat{X}^{N} - \widehat{X}^{S}_{\sigma}$ as in the proof of Proposition 4, $\widehat{X}^{\theta}_{\sigma} - X = (\widehat{X}^{S}_{\sigma} - X) + (1-\theta)D$, and the same orthogonality gives

$$\begin{aligned} R^{N}_{Q}(\sigma) - R^{S}_{Q}(\sigma) &= \mathbb{E}_{\sigma}\|D\|^{2}_{Q}, \\ R^{\theta}_{Q,\text{within}}(\sigma) &= R^{S}_{Q}(\sigma) + (1-\theta)^{2}\mathbb{E}_{\sigma}\|D\|^{2}_{Q}. \end{aligned}$$

Substituting the first identity into the second gives

$$\begin{aligned} R^{\theta}_{Q,\text{within}}(\sigma) &= R^{S}_{Q}(\sigma) + (1-\theta)^{2}\big[R^{N}_{Q}(\sigma) - R^{S}_{Q}(\sigma)\big] \\ &= (2\theta - \theta^{2})R^{S}_{Q}(\sigma) + (1-\theta)^{2}R^{N}_{Q}(\sigma) \\ &= R^{\lambda(\theta)}_{Q}(\sigma), \qquad \lambda(\theta) = 2\theta - \theta^{2}. \end{aligned}$$

Thus the receiver's true expected loss under the belief mixture equals the population-mixture risk with sophisticated share $\lambda(\theta)$. This proves (8) for every feasible policy, so the two objectives have the same optimal highlighting policies.

For the corollary, the common policy does not condition on Θ, and Θ is independent of X. Thus the joint law of the state and highlight does not vary with Θ. Averaging the risk identity over Θ gives aggregate expected loss $R^{\lambda_{\text{eff}}}_{Q}(\sigma)$, where

$$\lambda_{\text{eff}} = \mathbb{E}[\lambda(\Theta)] = 2\mathbb{E}[\Theta] - \mathbb{E}[\Theta^{2}] = 1 - \mathbb{E}[(1-\Theta)^{2}].$$

This is the population-mixture representation in Corollary 2. □

Proof of Proposition 6. In Example 2, revealing feature 2 is uniquely naive-optimal in every state. For an arbitrary randomized policy, let $r = \mathbb{P}_{\sigma}(I \neq \{2\})$. Relative to revealing feature 2, revealing feature 1 increases naive loss by $(b^{2} - a^{2})/4$, while the empty highlight increases it by $b^{2}/4$. Consequently, any policy satisfying the naive-risk constraint obeys

$$r \leq \frac{4\eta}{b^{2} - a^{2}} = \frac{t}{2}.$$

Consider either message that reveals feature 2. Let $u, v \in [0, 1/4]$ be the joint probabilities of this message and its two compatible states. Its sophisticated risk contribution is $a^{2}uv/(u+v)$, with value zero when $u + v = 0$. For fixed total mass $s = u + v$, minimizing $u(s-u)$ over the feasible interval gives

$$h(s) = \min_{\substack{u+v=s \\ 0\leq u,v\leq 1/4}} \frac{uv}{u+v} = \begin{cases} 0, & 0 \leq s \leq 1/4, \\ \dfrac{s - 1/4}{4s}, & 1/4 \leq s \leq 1/2. \end{cases}$$

Let s_{0}, s_{1} be the total masses of the two feature-2 messages. Then $s_{0} + s_{1} = 1 - r$ and $s_{0}, s_{1} \leq 1/2$. The function h is increasing and concave on $[1/4, 1/2]$. For $0 \leq r \leq 1/4$, both masses belong to this interval, and concavity gives a minimum at $(s_{0}, s_{1}) = (1/2, 1/2 - r)$, up to exchange. For $1/4 \leq r \leq 1/2$, a minimum occurs at $(1/4, 3/4 - r)$, up to exchange. Indeed, when both masses are at least $1/4$, concavity again places the minimum at an endpoint. When one mass is less than $1/4$, increasing it to $1/4$ leaves its cost zero and weakly reduces the cost of the other message. All remaining messages have nonnegative risk, so

$$R^{S}(\sigma) \geq a^{2} \begin{cases} \dfrac{1}{8} + \dfrac{1/4 - r}{4(1/2 - r)}, & 0 \leq r \leq 1/4, \\ \dfrac{1/2 - r}{4(3/4 - r)}, & 1/4 \leq r \leq 1/2. \end{cases}$$

This lower bound decreases in r. Substituting $r \le t/2$ gives (10).

To attain the bound, reveal feature 1 at $(0,0)$ with probability $\min\{2t,1\}$ and at (a,b) with probability $\max\{2t-1,0\}$. Reveal feature 2 in all other cases. The total probability of revealing feature 1 is $t/2$, so the naive-risk increase is exactly η. Each feature-1 message identifies a single state, and the masses in the feature-2 messages attain the lower bound above. At $t=1$, all messages identify a single state and sophisticated risk is zero. For every $t<1$, the lower bound is strictly positive. □

Population threshold. On each of $0 \le t \le 1/2$ and $1/2 \le t \le 1$, the frontier in (10) is concave as a function of the naive-risk allowance. Adding the linear naive-risk term preserves concavity, so a population optimum occurs at one of the three segment endpoints. Their risk pairs are

$$\left(\frac{a^2}{4}, \frac{a^2}{4}\right), \qquad \left(\frac{3a^2+b^2}{16}, \frac{a^2}{8}\right), \qquad \left(\frac{a^2+b^2}{8}, 0\right).$$

They are collinear, and the difference between the population objective at the last and first pairs is

$$\frac{(1-\lambda)(b^2-a^2) - 2\lambda a^2}{8}.$$

Its sign changes at the threshold in (11). Below that threshold the naive endpoint is optimal, and above it the minimum-naive-risk perfect-recovery endpoint is optimal. At equality the intermediate deterministic policy also attains the same value.

A.2 Omitted proofs for Section 6.1

Fixed greedy and conditional covariance. For a fixed set I, Gaussian conditioning gives the residual covariance

$$\Sigma_{\cdot|I} = \Sigma - \Sigma_{:I}\Sigma_{II}^{-1}\Sigma_{I:}, \qquad \Sigma_{\cdot|\emptyset} = \Sigma.$$

Here $\Sigma_{:I}$ contains the columns indexed by I, and Σ_{II} is the corresponding principal submatrix. The fixed risk and the gain over revealing nothing are

$$R_Q^{\mathrm{fix}}(I) = \mathrm{tr}(Q\Sigma_{\cdot|I}), \tag{24}$$

$$f_Q(I) = R_Q^{\mathrm{fix}}(\emptyset) - R_Q^{\mathrm{fix}}(I) = \mathrm{tr}(Q\Sigma_{:I}\Sigma_{II}^{-1}\Sigma_{I:}), \tag{25}$$

with $f_Q(\emptyset) = 0$. For $j \notin I$, write $c_{j|I} = (\Sigma_{\cdot|I})_{:j}$ and $v_{j|I} = (\Sigma_{\cdot|I})_{jj} > 0$. Gaussian conditioning gives the marginal gain

$$\Delta_Q(j \mid I) = f_Q(I \cup \{j\}) - f_Q(I) = \frac{c_{j|I}^\top Q c_{j|I}}{v_{j|I}} \ge 0. \tag{26}$$

Thus expected loss cannot increase when a fixed highlight reveals another feature.

Fixed greedy begins with $I_0 = \emptyset$ and, for $t = 0, \dots, k-1$, selects

$$j_t \in \operatorname*{argmax}_{j \notin I_t} \Delta_Q(j \mid I_t), \qquad I_{t+1} = I_t \cup \{j_t\}.$$

It returns $I^{\mathrm{greedy}} = I_k$, with ties broken arbitrarily.

The posterior reconstruction is $\widehat{X}(\widehat{\mathbb{P}}_I^N) = \mu + \Sigma_{:I}\Sigma_{II}^{-1}(X_I - \mu_I)$. Taking its expected quadratic loss gives (24), and subtracting from $\mathrm{tr}(Q\Sigma)$ gives (25). For $j \notin I$, the Schur-complement update is

$$\Sigma_{\cdot|I\cup\{j\}} = \Sigma_{\cdot|I} - \frac{c_{j|I}c_{j|I}^\top}{v_{j|I}}.$$

Taking the trace against Q proves (26). Fixed greedy runs in $O(d^3)$ operations.

Proof of Theorem 2. We reduce exact cover by three-sets to the fixed highlighting problem. Take an instance with $3k$ elements, $d \geq k$ candidate triples, and incidence matrix $A_0 \in \{0,1\}^{3k \times d}$, where $k \geq 1$. Define

$$\overline{A} = \begin{pmatrix} A_0 \\ (2k)^{-1} I_d \end{pmatrix}, \qquad \overline{b} = \begin{pmatrix} \mathbf{1}_{3k} \\ 0_d \end{pmatrix}.$$

The matrix $\overline{A}$ has full column rank. If an exact cover exists, its indicator vector has k nonzero entries and squared residual $\|\overline{b} - \overline{A}\beta\|_2^2 = 1/(4k) \leq 1/4$. If no exact cover exists, every choice of at most k triples leaves an element uncovered. That coordinate contributes one to the squared residual, regardless of the coefficients assigned to the selected columns. Thus a residual threshold of $1/2$ separates the two cases.

Set

$$\Sigma = \overline{A}^\top \overline{A}, \qquad g = \overline{A}^\top \overline{b}, \qquad c = \Sigma^{-1} g, \qquad Q = cc^\top + \Sigma^{-1}.$$

Both Σ and Q are positive definite. For $X \sim \mathcal{N}(0, \Sigma)$, substitution in (25) gives

$$f_Q(I) = |I| + g_I^\top \Sigma_{II}^{-1} g_I.$$

The second term is the value explained by least squares on the selected columns and cannot decrease when a column is added. The first term strictly increases, so an optimum uses k coordinates. For each such set,

$$f_Q(I) = k + \|\overline{b}\|_2^2 - \min_{\mathrm{supp}(\beta) \subseteq I} \|\overline{b} - \overline{A}\beta\|_2^2.$$

Since $\|\overline{b}\|_2^2 = 3k$, the exact-cover instance has a solution if and only if

$$\max_{|I| \leq k} f_Q(I) \geq 4k - \frac{1}{2}.$$

This is equivalently a decision about whether the minimum fixed risk is at most $\mathrm{tr}(Q\Sigma) - 4k + 1/2$.

Since $4k^2\Sigma = 4k^2 A_0^\top A_0 + I_d$ is an integer matrix with entries at most $12k^2 + 1$, the matrices Σ^{-1}, c, and Q have polynomial encoding length, and the risk of any set can be evaluated exactly in polynomial time. An algorithm for either the optimal value or an optimal set would therefore decide exact cover by three-sets in polynomial time. □

Proof of Theorem 3. Let $D = \mathrm{diag}(\Sigma_{11}, \ldots, \Sigma_{dd})$, so that $\Psi = D^{-1/2}\Sigma D^{-1/2}$. Write $Q = \sum_r \varphi_r \varphi_r^\top$ for vectors $\varphi_r \in \mathbb{R}^d$, for example the columns of a Cholesky factor. Put $b^{(r)} = D^{-1/2}\Sigma\varphi_r$. Equation (25) can be written as

$$\begin{aligned} f_Q(I) &= \sum_r b_I^{(r)\top} \Psi_{II}^{-1} b_I^{(r)} \\ &= \sum_r \max_{\mathrm{supp}(\beta) \subseteq I} \{2b^{(r)\top}\beta - \beta^\top \Psi \beta\}. \end{aligned}$$

Take disjoint sets U, V with $|U \cup V| \leq 2k$. For response r, let $\beta_U^{(r)}$ maximize its quadratic over coefficients supported on U, and let $g^{(r)} = 2b^{(r)} - 2\Psi\beta_U^{(r)}$ be the gradient there. Its coordinates on U are zero. For any $j \in V$, allowing a displacement only in coordinate j gives improvement $\big(g_j^{(r)}\big)^2/(4\Psi_{jj}) = \big(g_j^{(r)}\big)^2/4$. Reoptimizing over $U \cup \{j\}$ can only increase this improvement. Consequently,

$$\sum_{j \in V} \Delta_Q(j \mid U) \geq \sum_r \frac{\|g_V^{(r)}\|_2^2}{4}.$$

For a displacement h supported on $U \cup V$, the change in response r's objective is

$$g_V^{(r)\top} h_V - h^\top \Psi h \le g_V^{(r)\top} h_V - m_{2k}\|h\|_2^2 \le \frac{\|g_V^{(r)}\|_2^2}{4m_{2k}}.$$

Maximizing the left side and summing over responses therefore yields

$$\sum_{j \in V} \Delta_Q(j \mid U) \ge m_{2k}\big(f_Q(U \cup V) - f_Q(U)\big). \tag{27}$$

Let I_t be the greedy set after $t < k$ steps and apply (27) with $U = I_t$ and $V = I^* \setminus I_t$. By monotonicity, $f_Q(I_t \cup I^*) \ge f_Q(I^*)$. Since $|V| \le k$, the largest available marginal is at least $m_{2k}\big(f_Q(I^*) - f_Q(I_t)\big)/k$. Starting from $f_Q(I_0) = 0$, the greedy recurrence gives

$$f_Q(I_k) \ge \big[1 - (1 - m_{2k}/k)^k\big] f_Q(I^*) \ge (1 - e^{-m_{2k}}) f_Q(I^*).$$

Subtracting this inequality from the no-disclosure risk gives the equivalent risk bound

$$R_Q^{\mathrm{fix}}(I^{\mathrm{greedy}}) \le e^{-m_{2k}} R_Q^{\mathrm{fix}}(\emptyset) + (1 - e^{-m_{2k}}) R_Q^{\mathrm{fix}}(I^*).$$

□

A.3 Omitted proofs for Section 6.2

Proof of Lemma 1. Let $\mathcal{Z}$ be the finite set of feasible index-value messages, and let $\pi_{sz} = K_\sigma(I \mid X^s)$ be the probability that state X^s sends message $z = (I, X_I^s)$. Incompatible pairs have probability zero, and $\sum_z \pi_{sz} = 1$ for every state. Thus the feasible policy kernels form a product of finite simplices. The contribution of message z to sophisticated risk is

$$\mathrm{risk}_z(\pi) = \min_{\widehat{X} \in \mathbb{R}^d} \sum_s p_s \pi_{sz} \|\widehat{X} - X^s\|_Q^2.$$

For every fixed reconstruction $\widehat{X}$, the expression is linear in π. Its minimum over $\widehat{X}$ is therefore concave in π. Summing over messages shows that R_Q^S is concave in π.

Every kernel is a convex combination $\pi = \sum_r \alpha_r \pi^r$ of deterministic kernels. Concavity gives

$$R_Q^S(\pi) \ge \sum_r \alpha_r R_Q^S(\pi^r) \ge \min_r R_Q^S(\pi^r).$$

There are finitely many deterministic policies, so one of them attains the minimum over all kernels. □

For (13), fix a binary assignment, put $q_z = \sum_{s \in \mathcal{S}_z} p_s u_{sz}$, and let $\overline{X}_z$ be the posterior mean of the states assigned to z when $q_z > 0$. Then

$$\sum_{s \in \mathcal{S}_z} p_s u_{sz} \|\widehat{X}_z - X^s\|_Q^2 = \sum_{s \in \mathcal{S}_z} p_s u_{sz} \|\overline{X}_z - X^s\|_Q^2 + q_z \|\widehat{X}_z - \overline{X}_z\|_Q^2,$$

so the inner minimum is attained at $\widehat{X}_z = \overline{X}_z$ and equals the assigned policy's sophisticated risk.

Proof of Theorem 4. We reduce Euclidean r-means to the sophisticated design problem. Lee et al. (2017) show that Euclidean r-means is NP-hard to approximate within 1.0013 on distinct binary points of the form $\mathbf{e}_u+\mathbf{e}_v$, with r part of the input. Take such a source instance $y_1,\dots,y_h\in\{0,1\}^{d_0}$ with $1\le r<h$. Its objective is the sum of squared distances to cluster means. Set

$$U_0=\sum_{i=1}^{h}\|y_i-y_1\|_2^2,\qquad T=4U_0+1.$$

The integer U_0 is positive and at most hd_0. A partition into at most r clusters has cost at most U_0, since one cluster already does. If exactly r nonempty clusters are required, split clusters without increasing cost.

Choose the smallest integer $q\ge 2$ satisfying

$$2^{q-1}-1\ge r+2d_0+2qT+1,$$

and put $d=r+d_0+qT$ and $g=1+2d-r$. Create a data state

$$X^{i,\mathrm{data}}=(0_r,y_i,0_{qT})\in\{0,1\}^d$$

for every source point. Reserve the r messages $(\{j\},0)$, $j=1,\dots,r$, for source clusters. At bandwidth one there are $2d+1$ possible messages, including the empty message. Let $\mathcal{Z}_0$ be the other g messages.

For each $z\in\mathcal{Z}_0$, create an auxiliary state

$$X^{z,\mathrm{aux}}=(1_r,b_z,\mathrm{rep}_T(c_z)),\qquad b_z\in\{0,1\}^{d_0},\quad c_z\in\{0,1\}^q\setminus\{0_q\},$$

where rep_T repeats each bit T times. Choose distinct codes c_z so that $X^{z,\mathrm{aux}}$ can send its designated message z. If z specifies a middle-block coordinate, set that coordinate of b_z to the required value and set the remaining entries to zero. If z specifies a repeated-code coordinate, constrain its corresponding code bit to the required value. The messages $(\{j\},1)$ with $j\le r$ and the empty message impose no code constraint. A prescribed zero code bit leaves $2^{q-1}-1\ge g$ nonzero codes, and a prescribed one bit leaves $2^{q-1}\ge g$ codes. Greedy assignment therefore finds a distinct compatible code at every step.

Give all $n=h+g$ states probability $1/n$ and take $Q=I_d$. The construction has polynomial size: $T\le 4hd_0+1$, and minimality of q gives $2^q=O(r+d_0+qT)$, so all candidate codes and the state table can be enumerated in polynomial time.

We first consider deterministic policies. Every auxiliary state differs from every other state in at least T code coordinates. For a message used by a deterministic policy, call the set C of states that send it its cell. For a nonempty cell C, the pairwise variance identity gives

$$\mathrm{risk}(C)=\frac{1}{n}\sum_{X\in C}\|X-\overline{X}_C\|_2^2=\frac{1}{n|C|}\sum_{\{X,X'\}\subseteq C}\|X-X'\|_2^2.$$

If C contains an auxiliary state and at least one other state, the pairs involving that auxiliary state imply

$$\mathrm{risk}(C)\ge\frac{(|C|-1)T}{n|C|}\ge\frac{T}{2n}>\frac{2U_0}{n}.$$

On the other hand, implement any source partition of cost at most U_0 using the reserved messages, and assign each auxiliary state its designated message. This gives a feasible policy of risk at most

U_0/n. Consequently every optimal deterministic policy isolates each auxiliary state in a singleton cell.

An auxiliary state cannot send a reserved message because its first r coordinates are all one. Its g singleton cells therefore occupy all g nonreserved messages. Every data state must send a reserved message. The first and code blocks are constant across data states, so the resulting loss is exactly the source clustering cost divided by n. Conversely, every source partition can be implemented in this way. Lemma 1 then yields

$$\mathrm{OPT}_{\text{highlighting}} = \frac{\mathrm{OPT}_{r\text{-means}}}{n}$$

even when the highlighting optimum allows randomization. Since n is computed in polynomial time, a polynomial-time 1.001-approximation to the minimum sophisticated risk would approximate the r-means objective within $1.001 < 1.0013$, which is NP-hard. □

Proof of Proposition 7. The naive posterior is fixed by the prior and the highlighted values, so $L_Q^N(I; X^s)$ does not depend on the policy. For selection probabilities $\pi_{sI} = K_\sigma(I \mid X^s)$,

$$R_Q^N(\pi) = \sum_s p_s \sum_{I:|I|\le k} \pi_{sI} L_Q^N(I; X^s).$$

Each inner sum is minimized by assigning probability one to a loss-minimizing set. Grouping support states by their highlighted values computes the losses for each set in polynomial time. Since $J_{d,k} = O(d^k)$ for fixed k, exhaustive search gives the claimed algorithm. □

Proof of Theorem 5. We reduce vertex cover to the naive contextual design problem. Take a simple graph $G = (V, E)$ with $d = |V|$ and $|E| \ge 1$. We may restrict its cover size to $2 \le k < d$.

Give equal positive probability to the distinct binary states

$$X^0 = 0_d, \qquad X^e = \mathbf{e}_u + \mathbf{e}_v \quad \text{for } e = \{u, v\} \in E.$$

Set $Q = I_d$. At the zero state, revealing coordinates in I leaves exactly the zero state and the edge states whose edges avoid I compatible with the highlight. Writing $E_I = \{e \in E : e \cap I = \emptyset\}$, the naive mean is

$$\mathbb{E}[X \mid X_I = 0] = \frac{\sum_{e \in E_I} X^e}{1 + |E_I|}.$$

The numerator is coordinatewise nonnegative and is zero if and only if E_I is empty. Thus a zero-loss highlight at X^0 exists exactly when G has a vertex cover of size at most k.

At an edge state, highlighting its two endpoints identifies the state, since the graph is simple. If a cover of size at most k exists, reveal it at the zero state and reveal each edge's endpoints at its edge state. The resulting policy has zero population risk. Conversely, zero population risk implies zero loss at every positive-probability state, including X^0, and hence yields a vertex cover. This proves the population reduction.

A policy certificate specifies one set per support state. Its conditional means and quadratic risk can be verified in polynomial time, so the decision problem is in NP, including for general $Q \succeq 0$. By Proposition 7, randomization cannot improve the optimum.

A finite multiplicative approximation of nonnegative risk would have to return zero whenever the optimum is zero. It would therefore decide these zero-risk instances, proving the approximation limitation stated after the theorem unless P = NP. □

Contextual actual-loss greedy. For a realized state, let $\delta = X - \mu$. The naive reconstruction and its realized loss are

$$\widehat{X}(\widehat{\mathbb{P}}_I^N) = \mu + \Sigma_{:I}\Sigma_{II}^{-1}\delta_I, \tag{28}$$

$$L_Q^N(I;X) = \|e_I\|_Q^2, \qquad e_I = \delta - \Sigma_{:I}\Sigma_{II}^{-1}\delta_I. \tag{29}$$

At $I = \emptyset$, take $\widehat{X}(\widehat{\mathbb{P}}_\emptyset^N) = \mu$ and write $L_{Q,0}(X) = \|\delta\|_Q^2$.

Projection greedy and marginal gains. Define $\Gamma_Q = \Sigma Q \Sigma$ and $\xi_Q = \Sigma Q \delta$, and let

$$\widetilde{L}_Q(I;X) = \min_{\mathrm{supp}(\beta)\subseteq I} \|\delta - \Sigma\beta\|_Q^2, \tag{30}$$

$$\widetilde{F}_Q(I;X) = L_{Q,0}(X) - \widetilde{L}_Q(I;X) = \xi_{Q,I}^\top \Gamma_{Q,II}^{-1}\xi_{Q,I}. \tag{31}$$

The surrogate projects $r = Q^{1/2}(X-\mu)$ onto the span of the selected directions $Q^{1/2}\Sigma_{:j}$, $j \in I$. The projection gain is zero at the empty set and is monotone in I. Its marginal gain is

$$\widetilde{\Delta}_Q(j \mid I;X) = \frac{(\xi_{Q,j} - \Gamma_{Q,jI}\Gamma_{Q,II}^{-1}\xi_{Q,I})^2}{\Gamma_{Q,jj} - \Gamma_{Q,jI}\Gamma_{Q,II}^{-1}\Gamma_{Q,Ij}} \geq 0. \tag{32}$$

The search uses these nonnegative gains and evaluates each prefix by actual naive loss. Algorithm 1 returns the prefix with the lowest loss, including the empty set. Projection greedy runs in $O(d^3)$ operations per state.

ALGORITHM 1: Projection greedy with best-prefix selection

Input: μ, $\Sigma \succ 0$, $Q \succ 0$, precomputed Γ_Q, state X, bandwidth $0 \leq k \leq d$
Output: Highlighted set $\sigma_Q^{\mathrm{proj}}(X)$
$\delta \leftarrow X - \mu$, $\xi_Q \leftarrow \Sigma Q \delta$
$I_0 \leftarrow \emptyset$, $I^{\mathrm{best}} \leftarrow \emptyset$, $L^{\mathrm{best}} \leftarrow \|\delta\|_Q^2$
for $t \leftarrow 0$ **to** $k-1$ **do**
  if $L^{\mathrm{best}} = 0$ **then**
    break
  end
  $j_t \in \mathrm{argmax}_{j \notin I_t} \widetilde{\Delta}_Q(j \mid I_t; X)$
  $I_{t+1} \leftarrow I_t \cup \{j_t\}$
  Compute $L_Q^N(I_{t+1};X)$ using (29)
  if $L_Q^N(I_{t+1};X) < L^{\mathrm{best}}$ **then**
    $I^{\mathrm{best}} \leftarrow I_{t+1}$, $L^{\mathrm{best}} \leftarrow L_Q^N(I_{t+1};X)$
  end
end
return $\sigma_Q^{\mathrm{proj}}(X) \leftarrow I^{\mathrm{best}}$

For the approximation constant, normalize Γ_Q to unit diagonal by setting

$$D_Q = \mathrm{diag}(\Gamma_{Q,11}, \ldots, \Gamma_{Q,dd}), \qquad \Psi_Q = D_Q^{-1/2}\Gamma_Q D_Q^{-1/2}.$$

Then

$$m_{2k}^Q = \min_{\substack{T\subseteq[d] \\ 1\leq|T|\leq\min(2k,d)}} \lambda_{\min}(\Psi_{Q,TT}), \qquad \zeta_Q = 1 - \exp(-m_{2k}^Q).$$

Projection identities. For any selected set, expanding the quadratic objective in the supported coefficients gives

$$(\delta - \Sigma_{:I}\beta_I)^\top Q(\delta - \Sigma_{:I}\beta_I) = L_{Q,0} - 2\xi_{Q,I}^\top \beta_I + \beta_I^\top \Gamma_{Q,II}\beta_I.$$

Its minimizer is $\widetilde{\beta}_I = \Gamma_{Q,II}^{-1}\xi_{Q,I}$. Completing the square proves (31). The Schur complement gives (32). The actual naive coefficients $\beta_I^N = \Sigma_{II}^{-1}\delta_I$ are feasible, but generally not optimal, in (30).

Bounding the discrepancy. For a proper nonempty set I, define

$$\omega_Q(I) = \sup_{X:\, \widetilde{L}_Q(I;X)>0} \frac{L_Q^N(I;X)}{\widetilde{L}_Q(I;X)}, \qquad \omega_{Q,k} = \max_{|I|\le k} \omega_Q(I),$$

with $\omega_Q(\emptyset) = \omega_Q([d]) = 1$. Write

$$\Lambda_Q = \Sigma^{1/2} Q \Sigma^{1/2}, \qquad \chi_Q = \frac{\lambda_{\max}(\Lambda_Q)}{\lambda_{\min}(\Lambda_Q)}.$$

We show that $1 \le \omega_{Q,k} \le (\chi_Q + 1)^2/(4\chi_Q)$. Thus weight matrices close to a positive multiple of Σ^{-1} limit the discrepancy between the two criteria.

Let $u = \Sigma^{-1/2}\delta$, let $\mathcal{V}_I = \operatorname{span}\{\Sigma^{1/2}\mathbf{e}_j : j \in I\}$, and let P_I be Euclidean projection onto this subspace. The naive coefficients solve the Euclidean projection normal equations. Consequently,

$$L_Q^N(I;X) = \|(I_d - P_I)u\|_{\Lambda_Q}^2, \qquad \widetilde{L}_Q(I;X) = \min_{v\in\mathcal{V}_I} \|u - v\|_{\Lambda_Q}^2,$$

where $\|w\|_{\Lambda_Q}^2 = w^\top \Lambda_Q w$. For a proper nonempty I, use orthonormal coordinates for $\mathcal{V}_I \oplus \mathcal{V}_I^\perp$ and write

$$\Lambda_Q = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}, \qquad S_I = \Lambda_{22} - \Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12}.$$

The two losses are respectively $u_2^\top \Lambda_{22} u_2$ and $u_2^\top S_I u_2$. Both matrices are positive definite. It follows that

$$\omega_Q(I) = \frac{1}{1 - \rho_Q(I)^2}, \qquad \rho_Q(I) = \|\Lambda_{11}^{-1/2}\Lambda_{12}\Lambda_{22}^{-1/2}\|_2 < 1.$$

If projected loss is zero, then δ belongs to the span of $\Sigma_{:I}$, and the naive loss is also zero. Thus $L_Q^N(I;X) \le \omega_Q(I)\widetilde{L}_Q(I;X)$ holds for all X.

To obtain a bound that does not depend on I, put $a_0 = \lambda_{\min}(\Lambda_Q)$ and $b_0 = \lambda_{\max}(\Lambda_Q)$. Spectral calculus gives

$$\Lambda_Q + a_0 b_0 \Lambda_Q^{-1} \preceq (a_0 + b_0) I_d.$$

If the columns of U form an orthonormal basis for $\mathcal{V}_I^\perp$, put $B = U^\top \Lambda_Q U$ and $C = U^\top \Lambda_Q^{-1} U$. Compress the preceding inequality and multiply on both sides by $B^{1/2}$ to get

$$B^{1/2} C B^{1/2} \preceq \frac{(a_0 + b_0)B - B^2}{a_0 b_0} \preceq \frac{(a_0 + b_0)^2}{4 a_0 b_0} I_{d-|I|}.$$

The block inverse formula gives $C = S_I^{-1}$. The largest eigenvalues of $B^{1/2}CB^{1/2}$ and $C^{1/2}BC^{1/2}$ coincide, and the latter equals $\omega_Q(I)$. Hence

$$1 \le \omega_{Q,k} \le \frac{(\chi_Q + 1)^2}{4\chi_Q}.$$

Proof of Theorem 6. Normalize Γ_Q by D_Q and put $b = D_Q^{-1/2}\xi_Q$. Then

$$\widetilde{F}_Q(I;X) = \max_{\mathrm{supp}(\beta)\subseteq I} \{2b^\top\beta - \beta^\top\Psi_Q\beta\}.$$

The sparse-eigenvalue argument in the proof of Theorem 3 applies to this single quadratic. It gives a submodularity ratio of at least m_{2k}^Q through (27). Let I_T be the final search set, where $T \le k$ is the number of additions.

If the algorithm stops before k additions, some evaluated prefix $S \subseteq I_T$ has zero actual loss. Its naive coefficients provide an exact fit in the projection problem on I_T. Hence $\widetilde{L}_Q(I_T;X) = 0$, and the terminal search set attains the full projection gain $L_{Q,0}(X)$.

If the search reaches k additions, monotonicity and the greedy recurrence, starting from zero gain, give

$$\widetilde{F}_Q(I_k;X) \ge \left[1 - \left(1 - \frac{m_{2k}^Q}{k}\right)^k\right] \max_{|I|\le k} \widetilde{F}_Q(I;X) \ge \zeta_Q \max_{|I|\le k} \widetilde{F}_Q(I;X).$$

Thus the projection-gain guarantee holds for I_T in either case. Let $I^*(X)$ minimize actual naive loss among the feasible sets. Its conditional-mean coefficients are feasible in the projection problem, so $\widetilde{L}_Q(I^*(X);X) \le L_{Q,k}^{N,\star}(X)$. Subtracting the gain guarantee from $L_{Q,0}(X)$ gives

$$\widetilde{L}_Q(I_T;X) \le (1-\zeta_Q)L_{Q,0}(X) + \zeta_Q L_{Q,k}^{N,\star}(X).$$

Multiply by $\omega_{Q,k}$ to bound actual loss at I_T. The returned prefix has no greater actual loss than either I_T or the empty set. This proves (14).

For expected risk, let $R_{Q,0} = R_Q^{\mathrm{fix}}(\emptyset) = \mathrm{tr}(Q\Sigma)$. Choosing the first minimizing set in a fixed ordering defines a measurable naive-optimal policy, and randomization cannot improve on the pointwise minimum, so $R_Q^{N,\star} = \mathbb{E}[L_{Q,k}^{N,\star}(X)]$. Taking expectations in the two pointwise bounds gives

$$R_Q^N(\sigma_Q^{\mathrm{proj}}) \le \min\left\{R_{Q,0},\, \omega_{Q,k}\left[(1-\zeta_Q)R_{Q,0} + \zeta_Q R_Q^{N,\star}\right]\right\}. \tag{33}$$

□

Proof of Corollary 3. For $Q = \Sigma^{-1}$, $\Gamma_Q = \Sigma$ and $\xi_Q = \delta$. The fitted coefficients and the naive coefficients therefore coincide for every set. The discrepancy factor is one, and (31) gives the actual naive gain

$$F_Q(I;X) = L_{Q,0}(X) - L_Q^N(I;X) = \delta_I^\top \Sigma_{II}^{-1}\delta_I.$$

By (32), both rules select by the squared conditional standardized residual

$$\frac{(\delta_j - \Sigma_{jI}\Sigma_{II}^{-1}\delta_I)^2}{\Sigma_{jj} - \Sigma_{jI}\Sigma_{II}^{-1}\Sigma_{Ij}}.$$

Also $\Psi_Q = \Psi$, so the general theorem yields the stated factor.

If actual-loss greedy stops before k steps because all remaining marginals vanish, then $\delta_j - \Sigma_{jI}\Sigma_{II}^{-1}\delta_I = 0$ for every unselected j. The same residual is zero on the selected coordinates. Thus $e_I = 0$, and the current highlight already has zero loss. Continuing the chain cannot improve its gain. □

Positive definite weights. Theorem 6 requires $Q \succ 0$, since otherwise the discrepancy factor can be infinite. For example, take $\Sigma = I_2$, let Q be the matrix of ones, and set $\delta = (0,1)^\top$. For $I = \{1\}$, the projected loss is zero but the actual naive loss is one.

A.4 Omitted proofs for Section 6.3

Lemma 2 (Selection information and posterior means)**.** *Under the assumptions of Theorem 7, every randomized policy satisfies*

$$0 \le R_Q^N(\sigma) - R_Q^S(\sigma) \le \frac{1}{2}\,\mathrm{H}_\sigma(I) \le \frac{1}{2}\log J_{d,k},$$

where $\mathrm{H}_\sigma(I)$ *is the entropy of the selected set under the policy. Under the independent-block assumptions of Corollary 4, the factor* $1/2$ *is replaced by* $b_{\max}/2$.

Proof. Let $Z = (I, X_I)$ denote the highlight, $\mathbb{P}_\sigma$ its joint law with the state, and $K_\sigma(I \mid X)$ the policy's selection probabilities. For a positive-probability message $z = (A, v)$, write

$$\mu_z = \mathbb{P}_\sigma(X \in \cdot \mid Z = z), \qquad \nu_z = \mathbb{P}(X \in \cdot \mid X_A = v).$$

The index set $A \subseteq M$ is fixed in the second conditional. Under coordinate independence, ν_z is a product distribution over the unrevealed coordinates. Writing $q_{z,i} = \mathbb{E}_{\mu_z} X_i$, decomposition of relative entropy against a product distribution gives

$$\begin{aligned} D_{\mathrm{KL}}(\mu_z \| \nu_z) &\ge \sum_{i \notin A} D_{\mathrm{KL}}\big(\mathrm{Bernoulli}(q_{z,i}) \| \,\mathrm{Bernoulli}(p_i)\big) \\ &\ge 2 \sum_{i \notin A} (q_{z,i} - p_i)^2 = 2\|\widehat{X}_\sigma^S(z) - \widehat{X}^N(z)\|_2^2. \end{aligned}$$

For the first inequality, the difference is the nonnegative relative entropy from μ_z to the product of its marginals. For the second, the Bernoulli relative entropy has value and derivative zero at $q = p$ and second derivative $1/[q(1-q)] \ge 4$ on $(0,1)$. The endpoint cases follow by continuity.

For any finite prior, Bayes' rule gives the likelihood ratio

$$\frac{d\mu_{(A,v)}}{d\nu_{(A,v)}}(x) = \frac{K_\sigma(A \mid x)\mathbb{P}(X_A = v)}{\mathbb{P}_\sigma(I = A)\mathbb{P}_\sigma(X_A = v \mid I = A)}.$$

Taking logarithms and averaging under the actual joint distribution yields

$$\begin{aligned} \mathbb{E}_Z D_{\mathrm{KL}}(\mu_Z \| \nu_Z) &= \mathrm{I}_\sigma(X; I) \\ &\quad - \sum_{A \in \mathcal{M}_k} \mathbb{P}_\sigma(I = A) D_{\mathrm{KL}}\big(\mathbb{P}_\sigma(X_A \in \cdot \mid I = A) \| \mathbb{P}(X_A \in \cdot)\big) \\ &\le \mathrm{H}_\sigma(I) \le \log J_{d,k}, \end{aligned} \tag{34}$$

where $\mathrm{I}_\sigma(X; I)$ is mutual information. Orthogonality of the sophisticated conditional mean gives

$$R_Q^N(\sigma) - R_Q^S(\sigma) = \mathbb{E}_\sigma \|\widehat{X}^N(Z) - \widehat{X}_\sigma^S(Z)\|_2^2.$$

Combining these relations proves the independent-coordinate bound.

For independent blocks, fixed-value conditioning preserves independence across blocks, so write $\nu_z = \bigotimes_g \nu_{z,g}$. Let $\mu_{z,g}$ be the marginal of μ_z on block B_g. Data processing and the Bernoulli inequality above give, for each unrevealed $i \in B_g$,

$$|\mathbb{E}_{\mu_z} X_i - \mathbb{E}_{\nu_z} X_i|^2 \le \frac{1}{2} D_{\mathrm{KL}}(\mu_{z,g} \| \nu_{z,g}).$$

Summing over coordinates and blocks gives

$$\|\widehat{X}^S_\sigma(z) - \widehat{X}^N(z)\|_2^2 \le \frac{b_{\max}}{2} \sum_g D_{\mathrm{KL}}(\mu_{z,g} \| \nu_{z,g}) \le \frac{b_{\max}}{2} D_{\mathrm{KL}}(\mu_z \| \nu_z).$$

Using (34) and the risk identity proves the block bound. □

This proof uses an entropy method related to the analysis of adaptive selection by Russo and Zou (2019).

Proof of Theorem 7 and Corollary 4. Set $C = \frac{1}{2} \log J_{d,k}$ under coordinate independence and $C = \frac{b_{\max}}{2} \log J_{d,k}$ under block independence. Lemma 2 gives $0 \le R^N_Q(\sigma) - R^S_Q(\sigma) \le C$ for every feasible policy. Hence

$$R^\lambda_Q(\sigma) = R^N_Q(\sigma) - \lambda\big(R^N_Q(\sigma) - R^S_Q(\sigma)\big) \ge R^{N,\star}_Q - \lambda C,$$

and taking the infimum over policies gives $R^{\lambda,\star}_Q \ge R^{N,\star}_Q - \lambda C$. For a naive-optimal policy, $R^\lambda_Q(\sigma^N) \le R^N_Q(\sigma^N) = R^{N,\star}_Q$, and subtracting proves both upper bounds. Under coordinate independence, the naive loss after revealing I at state X is

$$L^N_Q(I; X) = \sum_{i \notin I} (X_i - p_i)^2.$$

Every summand is positive because $0 < p_i < 1$, so revealing the k largest squared deviations is optimal, with any fixed tie-breaking rule. □

Per-feature implication of the entropy bound. Dividing the bound in Corollary 4 by d gives

$$\frac{R^\lambda_Q(\sigma^N) - R^{\lambda,\star}_Q}{d} \le \frac{\lambda b_{\max}}{2} \frac{k}{d} \log \frac{ed}{k}, \qquad 1 \le k \le d. \tag{35}$$

To see this, let $t = k/d \in (0, 1]$. Since $t^j \ge t^k$ for $0 \le j \le k$,

$$(1+t)^d = \sum_{j=0}^{d} \binom{d}{j} t^j \ge t^k \sum_{j=0}^{k} \binom{d}{j} = t^k J_{d,k}.$$

Taking logarithms and using $\log(1+t) \le t$ gives

$$\log J_{d,k} \le d \log(1 + k/d) + k \log(d/k) \le k \log(ed/k).$$

Substituting this inequality into the block bound and dividing by d proves the displayed claim. For bounded $b_{\max}$, the bound vanishes as $k/d \to 0$ because $t \log(e/t) \to 0$.

Proof of Proposition 8. Independence gives $L^N_Q(I; x) = \sum_{i \notin I} (x_i - p_i)^2$. Revealing the k largest summands therefore attains $R^{N,\star}_Q$ over Π_k, including randomized policies. Moreover,

$$(X_i - p_i)^2 = \begin{cases} q_i^2 < 1/4, & X_i = b_i, \\ (1 - q_i)^2 > 1/4, & X_i \ne b_i. \end{cases}$$

Every atypical coordinate is selected before every typical coordinate, so exactly $(N_{\mathrm{dev}}-k)_+$ atypical coordinates remain unhighlighted. Consider the action determined by the highlight,

$$\widetilde{a}_i(I, x_I) = \begin{cases} x_i, & i \in I, \\ b_i, & i \notin I. \end{cases}$$

Its realized squared loss under σ^N is $(N_{\mathrm{dev}} - k)_+$. The sophisticated posterior mean minimizes expected squared loss given the actual highlight, hence

$$R_Q^S(\sigma^N) \le \mathbb{E}[(N_{\mathrm{dev}} - k)_+].$$

Every feasible policy satisfies

$$R_Q^\lambda(\sigma) \ge (1-\lambda) R_Q^N(\sigma) \ge (1-\lambda) R_Q^{N,\star}.$$

Taking the infimum and using naive optimality gives

$$R_Q^\lambda(\sigma^N) - R_Q^{\lambda,\star} \le (1-\lambda) R_Q^{N,\star} + \lambda R_Q^S(\sigma^N) - (1-\lambda) R_Q^{N,\star} = \lambda R_Q^S(\sigma^N).$$

The lower bound holds by definition of $R_Q^{\lambda,\star}$, which proves (16).

The deviation indicators are independent Bernoulli variables with means q_i. For every $t > 0$,

$$\mathbb{E}[e^{tN_{\mathrm{dev}}}] = \prod_{i=1}^{d} [1 + q_i(e^t - 1)] \le \exp\big(s(e^t - 1)\big).$$

The tail-sum identity and exponential Markov inequality give

$$\begin{aligned} \mathbb{E}[(N_{\mathrm{dev}} - k)_+] &= \sum_{r=1}^{\infty} \mathbb{P}(N_{\mathrm{dev}} \ge k + r) \\ &\le \sum_{r=1}^{\infty} \exp\big(s(e^t - 1) - t(k + r)\big) = \frac{\exp\big(s(e^t - 1) - tk\big)}{e^t - 1}. \end{aligned}$$

Set $t = \log(1 + \delta)$ and use $k \ge (1 + \delta)s$ to obtain

$$\mathbb{E}[(N_{\mathrm{dev}} - k)_+] \le \delta^{-1} e^{-s h(\delta)}.$$

Since $h(0) = 0$ and $h'(\delta) = \log(1 + \delta) > 0$ for $\delta > 0$, we have $h(\delta) > 0$. Substituting into (16) proves (17). □

Asymptotic implication and example. For the sequence preceding (18),

$$k_d/d \le (1 + \delta) s_d/d + 1/d \longrightarrow 0.$$

The budgets are therefore feasible for all sufficiently large d. Equation (17) bounds the total excess risk uniformly over λ by $\delta^{-1} e^{-s_d h(\delta)}$, which tends to zero as $s_d \to \infty$.

In the example with $p_i = d^{-1/2}$ and $d \ge 5$, each modal value is zero, $s_d = \sqrt{d}$, and $k_d = \lceil 2\sqrt{d} \rceil \le d$. Taking $\delta = 1$ gives $h(1) = c = 2 \log 2 - 1$ and

$$\mathbb{E}[(N_{\mathrm{dev}} - k_d)_+] \le e^{-c\sqrt{d}}.$$

An omitted typical coordinate contributes $p_i^2 = 1/d$ to naive loss, while an omitted atypical coordinate contributes at most one. Thus, at every state,

$$L_Q^N(\sigma_d^N(X); X) \le 1 + (N_{\mathrm{dev}} - k_d)_+.$$

Taking expectations yields the stated naive-risk bound, and the modal-imputation argument yields the sophisticated-risk bound. With no disclosure, both receivers use the prior mean and incur

$$R_{Q,0} = \sum_{i=1}^{d} p_i(1 - p_i) = \sqrt{d} - 1.$$

Proof of Theorem 8. For a randomized policy, let $\pi_{sz} = K_\sigma(I \mid X^s)$ be the probability of sending compatible message $z = (I, X_I^s)$ at state s. Put $w_{sz} = p_s \pi_{sz}$ and $q_z = \sum_{s \in \mathcal{S}_z} w_{sz}$. When $q_z > 0$, the sophisticated conditional mean is

$$\widehat{X}_z^S = \frac{1}{q_z} \sum_{s \in \mathcal{S}_z} w_{sz} X^s.$$

The contribution of this message to sophisticated risk satisfies the pairwise variance identity

$$\sum_{s \in \mathcal{S}_z} w_{sz} \|X^s - \widehat{X}_z^S\|_Q^2 = \frac{1}{q_z} \sum_{\substack{s<t \\ s,t \in \mathcal{S}_z}} w_{sz} w_{tz} \|X^s - X^t\|_Q^2. \tag{36}$$

Both sides equal $\sum_s w_{sz} \|X^s\|_Q^2 - q_z \|\widehat{X}_z^S\|_Q^2$. By the separation assumption, total sophisticated risk is zero if and only if each used message receives positive probability from at most one state.

For a deterministic policy, every state sends exactly one compatible message. Thus perfect recovery is equivalent to a matching covering the state side of the compatibility graph. The naive reconstruction $\widehat{X}_z^N$ is fixed by the original prior, so a matching has naive risk equal to the sum of its edge costs c_{sz}. Its incidence vector is an integral feasible solution of (20), and every integral feasible solution defines such a matching.

The constraint matrix of (20) is the vertex–edge incidence matrix of the bipartite compatibility graph and is therefore totally unimodular. Since the right-hand sides are integral and the feasible region is bounded, every feasible instance has an integral optimal solution. This proves the equivalence of LP feasibility and deterministic perfect recovery, and identifies the minimum naive risk among deterministic perfect-recovery policies.

Now consider any randomized policy with zero sophisticated risk. The first part of the proof implies that the sets of messages used with positive probability by different states are disjoint. For each state s, choose a message z_s of minimum c_{sz} among those it uses. The resulting messages remain distinct across states, so they define a perfect-recovery matching. Moreover,

$$\sum_s c_{sz_s} \le \sum_s \sum_{z:\, s \in \mathcal{S}_z} \pi_{sz} c_{sz} = R_Q^N(\pi).$$

Finally, put $J = |\mathcal{M}|$. There are at most nJ distinct messages and exactly nJ state-message edges, since each state has one compatible message for each selected set in the menu. A minimum-cost bipartite matching algorithm therefore computes the optimum, or reports infeasibility, in polynomial time in the stated input size. □

A.5 Omitted proofs for Section 8.1

Proof of Example 4. The naive receiver imputes zero for the omitted coordinate, so minimizing realized loss selects the larger absolute value. Since $|X_1|$ and $|X_2|$ are independent uniform variables on $[0,1]$,

$$R^{N,\star} = 2\int_0^1 u^2(1-u)\,du = \frac{1}{6}.$$

Under this policy, observing $X_j = x_j$ restricts the omitted coordinate to $[-|x_j|, |x_j|]$. Its mean remains zero, proving $R^S(\sigma^N) = 1/6$.

For the sophisticated lower bound, first consider a measurable function $w\colon \mathbb{R} \to [0,1]$ with mass $m = \int w(v)\,dv < \infty$. At most $2t$ of this mass can lie within distance t of any $c \in \mathbb{R}$. Integrating the mass outside this interval gives

$$\int (v-c)^2 w(v)\,dv \geq \int_0^{m/2} 2t(m-2t)\,dt = \frac{m^3}{12}.$$

Let $q(x_1, x_2) \in [0,1]$ be any measurable probability of revealing coordinate 1. Define the section masses

$$m_1(x_1) = \int_{-1}^1 q(x_1, x_2)\,dx_2, \qquad m_2(x_2) = \int_{-1}^1 [1 - q(x_1, x_2)]\,dx_1.$$

Let $h(x_1)$ and $g(x_2)$ be the actual sophisticated conditional means of the omitted coordinate after revealing features 1 and 2, respectively. Since the joint density is $1/4$, the risk is

$$R^S(q) = \frac{1}{4}\int_{-1}^1\int_{-1}^1 \big[(x_2 - h(x_1))^2 q(x_1, x_2) + (x_1 - g(x_2))^2(1 - q(x_1, x_2))\big]\,dx_2\,dx_1.$$

Applying the scalar bound to each section yields

$$R^S(q) \geq \frac{1}{48}\left(\int_{-1}^1 m_1(x_1)^3\,dx_1 + \int_{-1}^1 m_2(x_2)^3\,dx_2\right) \geq \frac{1}{12}.$$

The last inequality is Jensen's inequality on the disjoint union of two intervals of total length four, using $\int m_1 + \int m_2 = 4$.

Under σ^S, each nonempty section is a half-interval of length one. The omitted-coordinate posterior therefore has variance $1/12$, attaining the lower bound. The omitted coordinate has second moment $1/3$, so the naive receiver, which imputes zero, incurs $R^N(\sigma^S) = 1/3$. □

A.6 Omitted proofs for Section 8.3

Proof of Proposition 9. Fix a policy and let $Z = (I, X_I)$ be its message, with $I = \sigma(X_M)$. Write $\mathcal{G}_0$ for the σ-algebra generated by Z and $\mathcal{G}_H$ for that generated by (Z, X_H). Then $\mathcal{G}_0 \subseteq \mathcal{G}_H$, and the sophisticated reconstructions are the corresponding conditional means of X_M. Conditional-mean orthogonality gives

$$R^S(\sigma;\emptyset) - R^S(\sigma;H) = \mathbb{E}\big\|\mathbb{E}[X_M \mid \mathcal{G}_H] - \mathbb{E}[X_M \mid \mathcal{G}_0]\big\|_2^2 \geq 0.$$

Because $\emptyset \in \mathcal{H}$, the supremum over private information structures equals $R^S(\sigma;\emptyset)$ for every policy. Minimizing over policies proves the minimax claim. □

B Ames Calibration and Implementation

This appendix describes the construction of the empirical prior (21) and the feature weights in (22). It then gives the selection scores used by the contextual methods, the projection greedy search, and the computation of profile matching in Section 7.

B.1 Housing Attributes and Calibration

Table 3 lists the ten fields and the categories into which each numerical measurement is binned.

Table 3: Housing attributes and their categories.

Field	Categories in recorded units
Construction year	Below 1960, 1960 to below 2000, 2000 or later
Living area	Below 1,200, 1,200 to below 1,800, at least 1,800 sq ft
Garage capacity	Below 2, 2 to below 3, at least 3 vehicles
Fireplaces	Below 1, at least 1
Remodel or addition year	Below 1980, 1980 to below 2000, 2000 or later
Overall quality	Below 6, 6 to below 7, at least 7
Basement area	Below 800, 800 to below 1,300, at least 1,300 sq ft
Full bathrooms	Below 2, 2 to below 3, at least 3
Lot area	Below 7,500, 7,500 to below 11,500, at least 11,500 sq ft
Bedrooms	Below 3, 3 to below 4, at least 4

Notes. Full bathrooms and bedrooms count above-grade rooms. Overall quality is the recorded material and finish rating. The cutpoints apply directly to the continuous Gaussian draws in recorded units. Gaussian tails enter the endpoint categories.

Let $H \in \mathbb{R}^{10}$ denote the ten numerical measurements and T natural-log sale price. All of them are observed for every property in the calibration sample, and we fit a joint Gaussian to (H, T) using the maximum-likelihood mean and covariance, without shrinkage, truncation, or reweighting. We then draw 100,000 joint observations (H^r, T^r) and set $V^r = \exp(T^r)$ and $X^r = g(H^r)$, where g bins each measurement at the cutpoints in Table 3. Calibration is therefore performed in log-price units, whereas policy evaluation uses unadjusted dollar transaction prices. The primary prior assigns positive probability to 10,808 of the $2 \cdot 3^9 = 39{,}366$ possible categorical profiles. The Gaussian approximation and thresholding change some marginal category frequencies substantially, most for the four count measurements (garage capacity, fireplaces, full bathrooms, and bedrooms) and for lot area.

B.2 Feature Weights

The feature weights are calibrated on the ten categorical fields of the calibration sample. A ridge regression predicts natural-log sale price from all-level one-hot encodings, fitted on a random subset of 1,317 of its 1,756 properties. Five-fold cross-validation within this subset selects 10 from the penalty grid

$$\{0.01,\ 0.03162,\ 0.1,\ 0.3162,\ 1,\ 3.162,\ 10,\ 31.62,\ 100\}.$$

If β_{jc} is the fitted coefficient for category c of field j, the weights are

$$\widetilde{w}_j = \left(\sum_c \beta_{jc}^2 \mathbb{V}\mathrm{ar}(\mathbb{1}\{X_j = c\}) \right)^{1/2}, \qquad w_j = \frac{\widetilde{w}_j}{\sum_{h=1}^{10} \widetilde{w}_h},$$

with the indicator variances computed on that subset. The weights range from 0.01 for bedrooms to 0.16 for living area; they are fitted on log price and held fixed for the dollar-price loss.

B.3 Selection Scores and Projection Greedy

Exact selection, actual-loss greedy, and projection greedy choose a highlight separately for each value of the designer's information D. Their score is the conditional risk of the naive receiver. Naive posteriors depend only on the highlighted values through (21), and not on the policy, so the scores can be computed before any policy is chosen. For a profile x and a set I, let $\widehat{p}_j^N(x_I)$ be the naive posterior probabilities of field j and $\widehat{v}_I^N(x_I)$ the naive posterior price mean. The resulting feature-recovery loss is

$$s_{xI} = \sum_{j=1}^{10} w_j \frac{\|\widehat{p}_j^N(x_I) - e_{x_j}\|_2^2}{b_j}.$$

Write $\varsigma_x^2 = \mathbb{V}\mathrm{ar}(V \mid X = x)$ for the conditional price variance of profile x under (21); profiles drawn once have $\varsigma_x^2 = 0$. The scores in the two information settings are

$$c_x(I) = (1-\alpha)s_{xI} + \frac{\alpha}{\mathbb{V}\mathrm{ar}(V)}\big[(\widehat{v}_I^N(x_I) - m_x)^2 + \varsigma_x^2\big], \tag{37}$$

$$c_{x,v}(I) = (1-\alpha)s_{xI} + \frac{\alpha}{\mathbb{V}\mathrm{ar}(V)}(\widehat{v}_I^N(x_I) - v)^2. \tag{38}$$

Averaging the term in ς_x^2 over profiles gives the residual κ_α in (23), which is common to all features-only policies. Price-observed policies are evaluated directly against the sampled prices V^r. Exact selection ($\leq k$) minimizes the applicable score over every subset with at most k fields, and exact selection ($=k$) minimizes it over subsets with exactly k fields. Actual-loss greedy ($\leq k$) starts from the empty highlight and adds the field with the smallest resulting score for k steps, including steps that raise the score. At each step $s = 0, \ldots, k$, it records the applicable score in (37) or (38) and returns the lowest-score prefix, including the empty highlight. Field choices within 10^{-12} of the smallest resulting score follow the field order in Table 3. A new prefix replaces the incumbent only if its score is smaller by more than 10^{-12}, so numerical ties favor the shorter prefix.

Projection greedy uses a surrogate to choose its search path and the score above to choose its final highlight. Stack the one-hot features and dollar price into the target vector

$$Y = (e_{X_1}^\top, \ldots, e_{X_{10}}^\top, V)^\top,$$

and let Q_α be the diagonal matrix representing (22) on this target. Its entries are $(1-\alpha)w_j/b_j$ for each indicator of field j and $\alpha/\mathbb{V}\mathrm{ar}(V)$ for price. For designer information D, define

$$r_D = Q_\alpha^{1/2}\big(\mathbb{E}[Y \mid D] - \mathbb{E}[Y]\big), \qquad B_j = Q_\alpha^{1/2}\mathbb{C}\mathrm{ov}(Y, e_{X_j}).$$

With features alone, the price coordinate of $\mathbb{E}[Y \mid D]$ is m_x. With price observed, it is v, so the price coordinate of r_D is

$$(r_D)_V = \sqrt{\frac{\alpha}{\mathbb{V}\mathrm{ar}(V)}}\big(v - \mathbb{E}[V]\big).$$

Observed price therefore enters both the projection search and, through (38), the comparison of prefixes, each with the price weight in (22). Let P_I project orthogonally onto the combined column spaces of the blocks B_j for $j \in I$. The projection gain for adding field j is

$$\Delta^{\mathrm{proj}}(j \mid I, D) = \|P_{I\cup\{j\}} r_D\|_2^2 - \|P_I r_D\|_2^2. \tag{39}$$

Each field counts once toward the budget, regardless of its number of indicators, and using whole column spaces makes the gain invariant to category relabeling.

Starting from the empty highlight, projection greedy adds a field maximizing (39) at each step. For budget k, it returns the prefix with the smallest score among the empty prefix and the first k additions. Because the search path does not depend on k, the policies for different budgets select prefixes of a common path. Ties in the gain follow the field order in Table 3, and ties in the score favor the shorter prefix. Since full disclosure is a candidate prefix at $k = 10$, projection greedy attains the lower bound κ_α at that budget when the designer observes features alone. When the designer observes price, its best prefix can improve on full disclosure.

B.4 Profile Matching

Profile matching applies Theorem 8 to the support profiles, with the menu of all sets of exactly k fields. Under the primary prior, a covering matching exists for $k = 4, \ldots, 10$ and does not exist for $k \leq 3$. Both information settings use the same policy $\sigma_{\text{match}}(X)$. Each message identifies X, so the sophisticated receiver has zero feature-recovery loss and predicts price by $m_X = \mathbb{E}[V \mid X]$. Its risk is therefore κ_α in (23) in both information settings.

For $4 \leq k \leq 9$, we solve the resulting minimum-cost bipartite matching problem. Dropping the distinct-message requirement lets each profile take its cheapest compatible highlight, so the features-only naive risk of exact selection ($=k$) is a lower bound on the naive risk of profile matching. At $k = 10$, the full highlight is the only compatible message, so profile matching coincides with full disclosure.

C Proofs for Highlighting and Recommendations

Proof of Proposition 2. Independence of β and X gives

$$\mathbb{E}[\beta^\top X \mid AX, \beta] = \beta^\top \mathbb{E}[X \mid AX].$$

Averaging the prediction error over β therefore gives the weighted recovery risk with $Q = \mathbb{E}[\beta\beta^\top]$. Write $G = \Sigma^{-1/2}X$ and $B = A\Sigma^{1/2}$. Let P_B be the orthogonal projector onto the row space of B. Gaussian conditioning gives $\mathbb{E}[G \mid BG] = P_B G$, including when A has deficient row rank. With $C = \Sigma^{1/2}Q\Sigma^{1/2}$, the resulting risk is

$$R^{\mathrm{adv}}(A) = \operatorname{tr}\big(C(I_d - P_B)\big). \tag{40}$$

Every subspace of dimension at most k can be the row space of B. To maximize $\operatorname{tr}(CP_B)$, let $u_1, \dots, u_d$ be an orthonormal eigenbasis of C, with decreasing eigenvalues $\lambda_j(C)$. Then

$$\operatorname{tr}(CP_B) = \sum_{j=1}^{d} \lambda_j(C)\|P_B u_j\|^2 \le \sum_{j=1}^{k} \lambda_j(C),$$

because $0 \le \|P_B u_j\|^2 \le 1$ and their sum is $\operatorname{rank}(P_B) \le k$. Equality is attained by the projector onto the first k eigenvectors. Thus $A^* = U_k^\top \Sigma^{-1/2}$ attains risk $\sum_{j=k+1}^{d} \lambda_j(C)$. □

The following bound controls the information conveyed by the choice of a highlighted set. Its formulation also permits private state information.

Lemma 3 (Residual uncertainty after highlighting)**.** *Let $M \in \mathbb{R}^d$ and a human signal H be jointly Gaussian, with $\Omega = \operatorname{Var}(M \mid H) \succ 0$. The signal H may be absent, in which case $\Omega = \operatorname{Var}(M)$. For $1 \le k < d$, define*

$$\mathcal{M}_k = \{I \subseteq [d] : |I| \le k\}, \qquad J_{d,k} = \sum_{\ell=0}^{k} \binom{d}{\ell}, \qquad \delta_{b,k} = \min_{I \in \mathcal{M}_k} \operatorname{Var}(b^\top M \mid M_I, H)$$

for a fixed $b \in \mathbb{R}^d$. Every possibly randomized selection of $I \in \mathcal{M}_k$ and every measurable decoder $g_I(M_I, H)$ satisfy

$$\mathbb{E}\big[(b^\top M - g_I(M_I, H))^2\big] \ge \frac{\pi\delta_{b,k}}{6J_{d,k}^2}. \tag{41}$$

In particular, this bound applies to both posterior receivers' weighted recovery risks whenever the task matrix dominates $bb^\top$.

Proof. The result is immediate if $\delta_{b,k} = 0$. Otherwise, for each fixed I, the conditional distribution of $T = b^\top M$ given (M_I, H) is Gaussian with constant variance at least $\delta_{b,k}$. Its density is bounded by $(2\pi\delta_{b,k})^{-1/2}$. For every $t \ge 0$, conditioning on (M_I, H) and then averaging yields

$$\mathbb{P}\big(|T - g_I(M_I, H)| \le t\big) \le \sqrt{\frac{2}{\pi\delta_{b,k}}}\, t.$$

The event that the selected decoder has error at most t is contained in the union of these events over all $J_{d,k}$ sets. This containment also holds for randomized selection. Writing $c = J_{d,k}\sqrt{2/(\pi\delta_{b,k})}$, we obtain

$$\mathbb{P}\big(|T - g_I(M_I, H)| \le t\big) \le ct.$$

Integrating the tail of the squared error gives

$$\mathbb{E}\big[(T - g_I(M_I, H))^2\big] \geq \int_0^{1/c} 2t(1 - ct)\,\mathrm{d}t = \frac{1}{3c^2} = \frac{\pi\delta_{b,k}}{6J_{d,k}^2}.$$

If a task matrix Q satisfies $Q \succeq bb^\top$, then $\|M - \widehat{M}\|_Q^2 \geq (b^\top M - b^\top \widehat{M})^2$ for either reconstruction. Taking $g_I = b^\top \widehat{M}$ proves the last assertion. □

Proof of Theorem 1. The eigenvalues of Q_η are $d - (d-1)\eta$ and η with multiplicity $d-1$. Proposition 2 therefore gives $A_k(\eta) = (d-k)\eta$. Under magnitude highlighting, flipping the sign of any omitted coordinate preserves both the selected set and the revealed values. Each omitted coordinate has sophisticated posterior mean zero, which is also its naive posterior mean. The same sign symmetry makes every expected cross product of distinct omitted coordinates zero. Every diagonal entry of Q_η equals one, so both receivers' risks under this policy equal $h_{d,k}$ for every η.

We have $0 < h_{d,k} < d - k$. Positivity follows because all Gaussian squared coordinates are positive almost surely. For strictness of the upper bound, compare magnitude highlighting with any fixed set of k coordinates. The realized sum of omitted squares is weakly smaller under magnitude highlighting and is strictly smaller with positive probability. The fixed set has expected risk $d-k$.

Fix a highlighting policy σ and a receiver type τ. Its reconstruction is independent of η, and its risk has the form

$$R_{Q_\eta}^\tau(\sigma) = (1-\eta)u_\sigma^\tau + \eta v_\sigma^\tau, \qquad \begin{aligned} u_\sigma^\tau &= \mathbb{E}\big[(\mathbf{1}_d^\top (X - \widehat{X}^\tau))^2\big], \\ v_\sigma^\tau &= \mathbb{E}\|X - \widehat{X}^\tau\|^2. \end{aligned} \tag{42}$$

For naive reconstruction, $v_\sigma^N \leq \mathbb{E}\|X\|^2 = d$ and $u_\sigma^N \leq d v_\sigma^N \leq d^2$. Conditional-mean optimality gives $u_\sigma^S \leq u_\sigma^N$ and $v_\sigma^S \leq v_\sigma^N$ for the same policy. Consequently, all the affine functions in (42) have uniformly bounded slopes. Their infimum $H_k^\tau(\eta)$ is concave and Lipschitz continuous on $[0,1]$.

Apply Lemma 3 with $M = X$, no signal H, and $b = \mathbf{1}_d$. For a fixed set I, the residual variance of $\mathbf{1}_d^\top X$ is $d - |I|$, so $\delta_{b,k} = d - k$. It follows that

$$(1-\eta)\frac{\pi(d-k)}{6J_{d,k}^2} \leq H_k^\tau(\eta) \leq h_{d,k}. \tag{43}$$

Thus the continuous concave function $F_\tau(\eta) = H_k^\tau(\eta) - (d-k)\eta$ is positive at zero and negative at one. It has a zero by continuity. For any zero $a \in (0,1)$ and $b > a$, concavity implies

$$\frac{F_\tau(b) - F_\tau(a)}{b - a} \leq \frac{F_\tau(a) - F_\tau(0)}{a} < 0.$$

It follows that the zero is unique, with the stated strict rankings on either side. Same-policy sophisticated dominance implies $H_k^S \leq H_k^N$, hence $\eta_S \leq \eta_N$. The upper bound in (43) gives $\eta_N \leq h_{d,k}/(d-k) < 1$. The lower bound also yields

$$\eta_S \geq \frac{\pi}{6J_{d,k}^2 + \pi}.$$

□

Proof of Proposition 3. Write $\mu(H) = \mathbb{E}[X_M \mid X_H = H]$ and $E = X_M - \mu(X_H)$. Joint Gaussianity implies that $E \sim \mathcal{N}(0, \Omega)$ is independent of X_H. From a report AX_M and the private observation

X_H, the human can calculate $AE = AX_M - A\mu(X_H)$. The target's contribution $\beta_H^\top X_H + \beta_M^\top \mu(X_H)$ is already known to the human. The remaining prediction problem is therefore to estimate $\beta_M^\top E$ from (AE, β_M). Independence of the task vector and the joint state allows us to apply Proposition 2 with state covariance Ω and task matrix Q_M. The optimal risk is the eigenvalue tail of $\Omega^{1/2} Q_M \Omega^{1/2}$. An optimal report uses $A = U_k^\top \Omega^{-1/2}$, where U_k contains its first k orthonormal eigenvectors. □

Proof of Corollary 1. Write $M = X_M$, $H = X_H$, and $r = 1 - \rho^2 > 0$. Then $\mathrm{Var}(M \mid H) = rI_d$, so Proposition 3 gives $A_k(\eta, \rho) = r(d-k)\eta$. For a fixed highlighted set I, the naive receiver reconstructs M_j exactly for $j \in I$ and uses ρH_j otherwise. Conditional on $M = m$, write $H = \rho m + \sqrt{r}\,\xi$, where $\xi \sim \mathcal{N}(0, I_d)$. The error on an omitted coordinate is therefore $r m_j - \rho\sqrt{r}\,\xi_j$. Define

$$f_\eta(I, m) = \eta \sum_{j \notin I} m_j^2 + (1 - \eta) \left(\sum_{j \notin I} m_j \right)^2 .$$

The conditional expected naive loss for selecting I is

$$r^2 f_\eta(I, m) + r(1 - r)(d - |I|). \tag{44}$$

This calculation uses the conditional independence of the ξ_j and the unit diagonal of Q_η. The designer can minimize (44) using its observation m alone. Randomizing among highlighted sets cannot improve this conditional minimum.

Under magnitude highlighting, sign symmetry makes the expected cross terms in f_η vanish, and $|I| = k$ almost surely. Its naive risk is therefore

$$B_{d,k}(\rho) = r(1 - r)(d - k) + r^2 h_{d,k}. \tag{45}$$

At $\eta = 1$, revealing coordinate j reduces the score by $r^2 m_j^2 + r(1 - r)$. Selecting the k largest $|m_j|$ minimizes the conditional score, so magnitude highlighting is naive-optimal at this endpoint.

For any fixed policy, either receiver's reconstruction depends on ρ and the policy but is independent of η. Its risk is again affine in η, with coefficients u_σ^τ and v_σ^τ defined as in (42) using M in place of X and conditioning on the human's observation H. For naive reconstruction, the conditional error calculation gives

$$v_\sigma^N \le r^2 \mathbb{E}\|M\|^2 + r(1 - r)d = rd,$$
$$u_\sigma^N \le r^2 d\, \mathbb{E}\|M\|^2 + r(1 - r)d \le rd^2.$$

The same bounds hold for the sophisticated receiver by conditional-mean optimality for each policy. Thus both optimized highlighting risks are concave and Lipschitz continuous in η.

For $b = \mathbf{1}_d$, conditioning on (M_I, H) leaves variance $r(d - |I|)$ in $b^\top M$. Lemma 3 yields

$$H_k^\tau(0, \rho) \ge \frac{\pi r(d - k)}{6 J_{d,k}^2} > 0.$$

At the other endpoint, same-policy sophisticated dominance and (45) give

$$H_k^\tau(1, \rho) \le B_{d,k}(\rho) = r(d - k) - r^2(d - k - h_{d,k}) < r(d - k).$$

The concavity argument in Theorem 1 therefore gives a unique crossing $\eta_\tau(\rho) \in (0, 1)$ for each receiver type. Sophisticated dominance orders the crossings. The feasible magnitude policy provides the uniform bound $H_k^\tau(\eta, \rho) \le B_{d,k}(\rho)$, so

$$0 < \eta_S(\rho) \le \eta_N(\rho) \le \frac{B_{d,k}(\rho)}{r(d - k)} = \rho^2 + (1 - \rho^2)\frac{h_{d,k}}{d - k} < 1.$$

□

Calculation for Example 1. For two independent standard Gaussian coordinates G_1, G_2,

$$h_{2,1} = \mathbb{E}\min(G_1^2, G_2^2) = 1 - \frac{1}{2}\mathbb{E}\big|(G_1 - G_2)(G_1 + G_2)\big| = 1 - \frac{2}{\pi}.$$

The final equality uses independence of $G_1 - G_2$ and $G_1 + G_2$, each with distribution $\mathcal{N}(0, 2)$ and mean absolute value $2/\sqrt{\pi}$. At $\eta = 1$, Corollary 1 gives the exact optimal naive risk

$$H_1^N(1, \rho) = r(1 - r) + r^2\left(1 - \frac{2}{\pi}\right).$$

At $\rho = 0$, this is $1 - 2/\pi \approx 0.363380228$. At $\rho^2 = 0.1$, it is $0.09 + 0.81(1 - 2/\pi) \approx 0.384337984$. □