Reliable Learning Using Misaligned Agents

Nischal Aryal
Oregon State University
Corvallis, OR 97331
aryaln@oregonstate.edu

Arash Termehchy
Oregon State University
Corvallis, OR 97331
termehca@oregonstate.edu

Ali Vakilian
Rutgers University
New Brunswick, NJ 08854
a.vakilian@rutgers.edu

Marianne Winslett
University of Illinois Urbana-Champaign
Urbana, IL 61801
winslett@illinois.edu

Abstract

Financial, social, and political factors often prevent the interests of the owners of ML systems and their users from being perfectly aligned. These systems often produce biased information that can influence users to make decisions that are not in their best interest, and vice versa. Current approaches require ML systems to implement protocols to mitigate their biases, or attempt to compel users to report data honestly. However, system owners and users usually have little incentive to implement these protocols. We believe a successful solution to this problem must recognize the conflict of interest between the systems and their users, and use this information to protect both systems and users against information that adversely influences their decisions. Toward this end, we propose a game-theoretic framework that models the interaction between supervised learning systems and users with conflicts of interest. We present scalable algorithms with theoretical guarantees that maximize the amount of desired information and actions and minimize biased and manipulative actions in interaction with supervised learning systems.

1 Introduction

Many factors prevent the incentives of the organizations that develop and operate popular ML systems and services from being perfectly aligned with those of their users [31]. This conflict is especially acute in systems where a user's own data trains or personalizes a model that will directly affect that user. For example, streaming and shopping services use self-reports of new users' preferences to fit an initial personalization model for recommendations [18], and users often underreport preferences they consider unflattering or unrepresentative. Meanwhile, platforms have an incentive to interpret preferences in ways that favor engagement or higher-margin content over the user's stated tastes. Similarly, institutional traders split a large order into smaller visible pieces [19], because revealing their true intended volume would let other market participants front-run the trade at the institution's expense [6]. Algorithmic traders detect and exploit such partial disclosures [33], treating stealth-trading patterns as a stronger signal of hidden intent than comparable-sized orders from less strategic traders [29]. In these examples, a user with private knowledge of their true data strategically decides what to reveal to a learner whose training objective may not fully align with the user's own interest.

Current proposals to combat the undesirable actions and influences of ML systems require them to implement certain rules and restrictions [31, 32]. However, these restrictions usually conflict with the objectives of the entities behind these systems, so they do not have an incentive to implement them. In addition, some argue that these restrictions limit the freedom to consume, learn from, and

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: Dynamic Alignment on Human-AI Coupled System (BiAlign Workshop).

distribute information or conduct business and may not be ethical or would have long-term adverse impacts on society. Also, they often impose significant computational overhead [31].

In this paper, we propose a game-theoretic framework for supervised learning in settings where the user and the learner have conflicts of interest. In our framework, the user has private training data that, if truthfully conveyed to the ML system (learner), may lead the system to learn a biased model to the benefit of the system. Importantly, our framework also captures the case where the system first learns an unbiased model and then injects bias at inference time. In both cases, the model's predictions undermine the user's interest, e.g., by providing biased recommendations.

The owner of the ML system and the user do share some interests: the ML system should satisfy the user's objective well enough to keep them engaged. If the system learns an overly biased model, the model's inferences may be too biased and discourage the users from using the system, e.g., recommending a set of products that only remotely reflect the training data collected from the user. We can leverage this partial common interest and design algorithms that take the user's true data as input and compute the optimal training data to share with the learner to minimize the bias of the learned model. But the ML system may be aware that users know about its undesirable biases in learning, e.g., based on their previous experiences. Thus, it may try to interpret the input training data strategically to offset the user's modifications, e.g., by using its prior information about the user.

We propose novel algorithms that find the equilibria of this reasoning and detect the equilibria in which the user and the learner have enough common interest to make learning an accurate model possible. We also provide efficient algorithms that find strategies for the user to modify the training data so that the resulting data reflects the true training data, to learn an accurate model, and convince the learner to deploy the least biased model possible in the setting.

2 Framework

Private Training Dataset. Let $\mathcal{X} \subseteq \mathbb{R}^p$ be the feature space and $\mathcal{Y}$ the label space, with $\mathcal{Y} = \mathbb{R}$ for regression and $\mathcal{Y} = \{1, \dots, K_y\}$ for K_y-class classification. We denote the set of possible training datasets of size N as $\mathcal{D} = (\mathcal{X} \times \mathcal{Y})^N$. A private training dataset is $D = ((\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_N, y_N)) \in \mathcal{D}$. The user picks their private dataset from a prior distribution π.

Example 2.1. *Consider an institutional order-execution problem with feature space $\mathcal{X} = [0,1]^2$, where $\boldsymbol{x}_i = (x_{i1}, x_{i2})$ records the normalized visible size and submission frequency of trade tranche i within a large institutional order. A simple private dataset is $D = ((\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), (\boldsymbol{x}_3, y_3))$ with $\boldsymbol{x}_1 = (1, 0)$, $\boldsymbol{x}_2 = (0, 1)$, $\boldsymbol{x}_3 = (1, 1)$, and $\boldsymbol{y} = (0.2, 0.2, 0.8)^\top$, where y_i represents the realized price-impact score attributable to tranche i with feature vector $\boldsymbol{x}_i$, as later inferred by a market-observing high-frequency trading (HFT) system.*

User's Strategy. For a private dataset D, the user submits a dataset report $M \in \mathcal{M} \subseteq \mathcal{D}$ to the learner. The dataset report may be different from D to manipulate the learner's decision. For instance, in Example 2.1, a report may keep labels fixed and alter feature values. Reports are *costless* and *unverifiable*: the learner observes M but cannot directly verify whether it is a truthful description of D. A **user strategy** is a mapping from a private dataset to a dataset report, $\sigma_r : \mathcal{D} \to \mathcal{M}$.

Learner's Strategy. Let $\mathcal{W} \subseteq \mathbb{R}^d$ be a feasible parameter set and $\mathcal{H} = \{h_{\boldsymbol{w}} \mid \boldsymbol{w} \in \mathcal{W}\}$ a hypothesis class parameterized by $\boldsymbol{w}$, where each $h_{\boldsymbol{w}} : \mathcal{X} \to \mathbb{R}^{K_y}$ scores a feature vector, e.g., a linear trade characterization scoring model $h_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{w}$. The learner's action is the choice of a hypothesis $h \in \mathcal{H}$, or equivalently model parameters $\boldsymbol{w} \in \mathcal{W}$. A **learner strategy** is a mapping from the dataset report to a hypothesis, $\sigma_\ell : \mathcal{M} \to \mathcal{H}$.

Utility Functions. The utility functions of the user and the learner measure the desirability of the learned hypothesis for the user and the learner, respectively. The learner's and user's utilities are functions from $\mathcal{D} \times \mathcal{H}$ to $\mathbb{R}$. We denote the learner's and user's utilities by $U^\ell(D, h)$ and $U^r(D, h)$, respectively. If there is a conflict of interest between the user and the learner, U^r and U^ℓ take their maximum values for different values of h for the same dataset D.

Example 2.2. *Continuing Example 2.1, let $\mathcal{H} = \{h : \boldsymbol{x} \mapsto \boldsymbol{x}^\top \boldsymbol{w},\ \boldsymbol{w} \in \mathbb{R}^2\}$. Assume the HFT system's (the learner's) utility is the negative empirical squared loss: $U^\ell(D, h_{\boldsymbol{w}}) = -\hat{L}^\ell(D, \boldsymbol{w})$, where $\hat{L}^\ell(D, \boldsymbol{w}) = \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2$. Hence $\boldsymbol{w}_\ell^*(D) \in \arg\max_{\boldsymbol{w}} U^\ell(D, h_{\boldsymbol{w}}) = \arg\min_{\boldsymbol{w}} \hat{L}^\ell(D, \boldsymbol{w}) = \left(\frac{1}{3}, \frac{1}{3}\right)$. Suppose the institutional trader (the user) prefers the HFT system to underestimate the*

magnitude of its total trade, so that the HFT gives smaller weights to the size and frequency coefficients. Assume the user's utility has the same squared-loss form, but with a conflict vector $\boldsymbol{c} = (-0.1, -0.1)^\top$ *and* $U^r(D, h_{\boldsymbol{w}}) = -\|\boldsymbol{X}(\boldsymbol{w} - \boldsymbol{c}) - \boldsymbol{y}\|_2^2$. *Then,* $\boldsymbol{w}_r^*(D) \in \arg\max_{\boldsymbol{w}} U^r(D, h_{\boldsymbol{w}}) = \left(\frac{7}{30}, \frac{7}{30}\right)$. *Thus, for the same dataset* D*, the learner's and the user's utility functions are maximized for different ideal hypotheses.*

Error in learning. Training is stochastic, so the hypothesis the learner wants, $h_{\boldsymbol{w}}$, is not learned exactly. The deployed hypothesis is $\hat{h} = h_{\boldsymbol{w}+\boldsymbol{\xi}}$, where $\boldsymbol{\xi} \sim \nu$ for an error distribution ν on $\mathbb{R}^d$. We assume the error is *data-independent*: ν does not depend on D, M, or $\boldsymbol{w}$, with $\mathbb{E}_{\boldsymbol{\xi}\sim\nu}[\boldsymbol{\xi}] = \mathbf{0}$ and $\mathbb{E}_{\boldsymbol{\xi}\sim\nu}[\|\boldsymbol{\xi}\|_2^2] = \kappa < \infty$. This includes, for example, Gaussian error $\nu = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, Laplace error, and discrete Rademacher error. In Appendix E we investigate data-dependent error.

Common knowledge. We assume the prior π, the report space $\mathcal{M}$, and the utility functions U^ℓ, U^r are common knowledge. This follows the standard game-theoretic convention used in strategic classification, where agents' costs/utilities are framework primitives [8, 10, 14].

Equilibria. As explained in Section 1, the user and the learner reason about each other's strategies to maximize their utilities. We ask whether this reasoning converges to a stable state in which the user and the learner settle on fixed strategies where neither can gain by deviating. Since the learner does not know the user's training dataset, we use the concept of *Bayesian Nash Equilibrium* [9, 16].

We assume the user's strategy is deterministic.[1] A deterministic strategy partitions $\mathcal{D}$ according to the report the user sends, so Bayes' rule reduces to conditioning on an event. Writing $\sigma_r^{-1}(M) = \{D \in \mathcal{D} : \sigma_r(D) = M\}$, the learner's *posterior belief* after a report M with $\pi(\sigma_r^{-1}(M)) > 0$ is $\pi(\cdot \mid M) = \pi(\cdot \mid \sigma_r^{-1}(M)) \in \Delta(\mathcal{D})$.

Given a report M, the learner's strategy maps it to a hypothesis $h = \sigma_\ell(M)$, which is deployed as $\hat{h}$ under learning error $\xi \sim \nu$. To simplify notation, we fold the learning error into the utility function. Formally, for both user and learner $q \in \{\ell, r\}$ we write $U^q(D, h) := \mathbb{E}_{\boldsymbol{\xi}\sim\nu}[U^q(D, \hat{h})]$ as the expected utility of a hypothesis h, and leave the expectation implicit. Hereafter, unless stated otherwise, all results are under the error distribution ν.

For a fixed dataset D, the *user's utility* from report M is $U^r(D, \sigma_\ell(M))$. For a fixed report M, the *learner's expected utility* from a response h is $\mathbb{E}_{D\sim\pi(\cdot|M)}[U^\ell(D, h)]$.

Definition 2.1. An equilibrium is a pair of strategies (σ_r, σ_ℓ) such that for every $D \in \mathcal{D}$, $\sigma_r(D) \in \arg\max_{M\in\mathcal{M}} U^r(D, \sigma_\ell(M))$, and for every M with $\pi(\sigma_r^{-1}(M)) > 0$, $\sigma_\ell(M) \in \arg\max_{h\in\mathcal{H}} \mathbb{E}_{D\sim\pi(\cdot|M)}[U^\ell(D, h)]$.

At equilibrium, the user sends reports that are optimal given the learner's response, and the learner chooses hypotheses that are optimal under the induced beliefs. Equilibria are stable states of mutual best response, *so the same characterizations support either party*: a user can use our algorithms to find the optimal dataset report that steers the learner toward a less conflicted hypothesis, or a learner can use them to choose an optimal learning strategy. We adopt the user-side convention throughout for concreteness.

3 Possibility of Influence

3.1 Influential Information Release

An equilibrium (σ_r, σ_ℓ) is *influential* if there exist two distinct reports $M, M' \in \mathcal{M}$, both sent with positive probability under (σ_r, π), such that the learner responds differently after the two reports: $\sigma_\ell(M) \neq \sigma_\ell(M')$. Otherwise, the equilibrium is *non-influential*. A non-influential equilibrium always exists: the user sends the same report for every private dataset, so the report conveys no information, and the learner best responds to its prior. The central question is therefore not whether communication can collapse, but whether meaningful communication can survive despite the conflict of interest. In other words, can the user still transmit information that changes the learned hypothesis, even when the user strategically chooses what to reveal?

[1] Randomization can help the user only for datasets indifferent between reports. Under a continuous prior these form a π-null set, so determinism is without loss. Under a discrete prior they may carry mass; we then break ties by an arbitrary rule.

We say that a learning setting is *influential* if it admits at least one influential equilibrium, and *non-influential* if every equilibrium is non-influential.

Problem 1. *Given the learning setting (D, U^r, U^ℓ), determine whether the setting is influential.*

Conflict over Hypothesis. At a high level, the user and the learner disagree over which hypothesis to learn. To capture this conflict in a tractable and interpretable way, we adopt a standard reduced-form assumption from the game-theoretic literature on strategic communication [9, 16], in which the preferences of the two parties differ by a fixed vector.

We identify each model with its parameter $\boldsymbol{w} \in \mathcal{W} \subseteq \mathbb{R}^d$, so $\mathcal{H} = \{h_{\boldsymbol{w}} : \boldsymbol{w} \in \mathcal{W}\}$. Each dataset D induces a learner-ideal model $\boldsymbol{w}^*_\ell(D) \in \mathcal{W}$. We assume the user's ideal model is $\boldsymbol{w}^*_\ell(D) + \boldsymbol{c}$, where $\boldsymbol{c} \in \mathbb{R}^d$ encodes the direction and magnitude of the **conflict**.[2] Both utilities depend on a hypothesis only through its parameters, so we write $U^\ell(D, \boldsymbol{w})$ and $U^r(D, \boldsymbol{w})$ for $U^\ell(D, h_{\boldsymbol{w}})$ and $U^r(D, h_{\boldsymbol{w}})$, respectively. They take the form $U^\ell(D, \boldsymbol{w}) = -\|\boldsymbol{w} - \boldsymbol{w}^*_\ell(D)\|_2^2$, and $U^r(D, \boldsymbol{w}) = -\|\boldsymbol{w} - (\boldsymbol{w}^*_\ell(D) + \boldsymbol{c})\|_2^2$, evaluated in expectation over the learning error $\boldsymbol{\xi}$. We identify the learner's response with its model vector $\sigma_\ell(M) \in \mathcal{W}$. Since each D induces the model $\boldsymbol{w}^*_\ell(D)$, the prior π over datasets induces a distribution over models which we denote by $\pi_{\boldsymbol{w}}$. Concretely, $\pi_{\boldsymbol{w}}$ is the distribution of $\boldsymbol{w}^*_\ell(D)$ when $D \sim \pi$. We assume $\pi_{\boldsymbol{w}}$ is common knowledge, and that it has finite second moment, which is automatic whenever $\mathcal{W}$ is bounded.

Knowledge about learning algorithm. The user does not know the learner's algorithm. Instead, the user has a training algorithm $\mathcal{A} : \mathcal{D} \to \mathcal{W}$ that produces $\hat{\boldsymbol{w}}(D) = \mathcal{A}(D)$, an approximation of $\boldsymbol{w}^*_\ell(D)$ that need not equal it; the user selects its dataset report based on $\hat{\boldsymbol{w}}(D)$. We assume the learner knows the user's training algorithm. We provide additional details in Appendix B

Reports, regions, and partitions. A **dataset report** $M \in \mathcal{M}$ is the information the user sends. Since the user's strategy depends on D only through $\boldsymbol{w}^*_\ell(D)$, each report has a **region** $R_M := \{\boldsymbol{w}^*_\ell(D) : D \in \mathcal{D},\ \sigma_r(D) = M\} \subseteq \mathcal{W}$. Let $\mathcal{M}_+ := \{M \in \mathcal{M} : \pi_{\boldsymbol{w}}(R_M) > 0\}$ be the reports with positive probability. The collection $\mathcal{R} := \{R_M : M \in \mathcal{M}_+\}$ is the resulting **partition** of the parameter space, up to prior-null sets and tie boundaries. For indexed dataset reports we write $R_i := R_{M_i}$. The learner's posterior after M_i is $\pi_{\boldsymbol{w}}$ conditioned on R_i. In practice, the user implements M by submitting any dataset $\tilde{D}$ with $\boldsymbol{w}^*_\ell(\tilde{D}) \in R_M$. The choice is utility-irrelevant, since the learner's response depends on $\tilde{D}$ only through the region $\boldsymbol{w}^*_\ell(\tilde{D})$ falls in.

3.2 Single Dimension

As in prior work on strategic classification [8, 14], we adopt a screening setup for the user's information release problem. We are motivated by problems such as school admissions and lending, where an individual's qualification is captured by a real-valued score. In a strategic setting, the user can "game" the learner by being intentionally vague about that score, for instance reporting a pass/fail category. All missing proofs are in Appendix C.

Setup. We specialize the framework to $d = 1$, so $U^\ell(D, w) = -|w - w^*_\ell(D)|^2$ and $U^r(D, w) = -|w - (w^*_\ell(D) + c)|^2$, where $w^*_\ell(D) \in \mathcal{W} \subseteq \mathbb{R}$ is the learned model on D and w the learner's response. We take $c > 0$ ($c < 0$ is symmetric, $c = 0$ removes the conflict). Utility functions carry the expectation over learning error ξ implicitly.

Statistics oracle. Since we work with an arbitrary prior π_w, we assume access to an oracle that, given a measurable region $R \subseteq \mathcal{W}$ in the model space, returns its prior mass, posterior mean, and posterior second moment (Algorithm 1). For $d > 1$ we use the same oracle with the bold notation $\pi_{\boldsymbol{w}}$ and $\boldsymbol{w}^*_\ell(D)$. For structured priors and simple regions, these statistics have closed form, and the oracle abstraction is unnecessary. For instance, a uniform prior on a box gives polynomial integrals, and a Gaussian on an interval or halfspace gives the mean and second moment through the standard truncated-moment formulas. For general priors and regions, the statistics are computed numerically, e.g. by Monte Carlo integration [27].

[2] We place the conflict on the user's side, but equilibrium depends only on the gap between the two ideal models. Since the utilities are common knowledge, the framework applies equally when the conflict is shifted to the learner.

Algorithm 1 $\mathrm{Oracle}(\pi_w, R)$

Input: prior π_w, region $R \subseteq \mathcal{W}$ with positive prior mass
Output: prior mass $p(R)$, posterior mean $\Phi(R)$, posterior second moment $\Psi(R)$
$p(R) \leftarrow \Pr_{\pi_w}\big(w_\ell^*(D) \in R\big)$
$\Phi(R) \leftarrow \mathbb{E}_{\pi_w}\big[w_\ell^*(D) \mid w_\ell^*(D) \in R\big]$
$\Psi(R) \leftarrow \mathbb{E}_{\pi_w}\big[w_\ell^*(D) w_\ell^*(D)^\top \mid w_\ell^*(D) \in R\big]$
return $(p(R), \Phi(R), \Psi(R))$

Theorem 3.1. *Let $w_\ell^*(D) \in [0,1]$ have a continuous prior π_w with full support. For $t \in (0,1)$, define $\Phi_L(t) = \mathbb{E}[w_\ell^*(D) \mid w_\ell^*(D) \in [0,t]]$ and $\Phi_H(t) = \mathbb{E}[w_\ell^*(D) \mid w_\ell^*(D) \in (t,1]]$. Then learning is influential if and only if there exists $t \in (0,1)$ with $t + c = \frac{1}{2}(\Phi_L(t) + \Phi_H(t))$.*

Algorithm. Theorem 3.1 reduces influence detection to finding a root of $g(c,t) = \frac{1}{2}\big(\Phi(R(t)) + \Phi(R'(t))\big) - t - c$, where $R(t) = [0,t]$ and $R'(t) = (t,1]$. Each evaluation of g uses two oracle calls, for $R(t)$ and $R'(t)$, so a solver evaluating g at K points makes $2K$ calls. Given a bracketing interval, standard root solvers find an ϵ-accurate root using $O(\log(1/\epsilon))$ evaluations [15]. For some priors, the cutoff condition can be solved analytically, yielding a constant-time algorithm as follows.

Corollary 3.1.1. *If the prior π_w is uniform on $[0,1]$, learning is influential if and only if $c < \frac{1}{4}$.*

3.3 Multiple Dimensions

We now extend the setup to higher dimensions $d > 1$. As in Section 2, each dataset $D \in \mathcal{D}$ trains to a model $\boldsymbol{w}_\ell^*(D) \in \mathbb{R}^d$ with prior $\pi_{\boldsymbol{w}}$. We give a sufficient condition for learning to be influential.

Theorem 3.2. *Suppose $\pi_{\boldsymbol{w}}$ is continuous, non-degenerate with finite first moment, and is spherically symmetric around its mean $\bar{\boldsymbol{w}} := \mathbb{E}_{\pi_{\boldsymbol{w}}}[\boldsymbol{w}_\ell^*(D)]$. Then learning is influential for every $\boldsymbol{c} \in \mathbb{R}^d$.*

This condition covers the rotation-invariant priors standard in Bayesian deep learning, where an isotropic Gaussian $\mathcal{N}(\bar{\boldsymbol{w}}, \sigma^2 I_d)$ is the de facto weight prior [4, 13], along with uniform distributions on balls or spheres. For general priors, there is no scalar cutoff as in one dimension; the source of the hardness is discussed in Section 4.2.

4 User's Optimal Information Release Problem

The user seeks to choose a dataset report $M \in \mathcal{M}$ that maximizes their utility U^r. Among multiple equilibria, a user-optimal equilibrium maximizes the user's expected utility over $D \sim \pi$. A naive algorithm would search over all feasible equilibria. Such an approach is computationally inefficient in general. Our goal is therefore to identify structures that yield efficient algorithms for computing a user-optimal equilibrium and its prescribed dataset report.

Problem 2. *Given the learning setting (D, U^r, U^ℓ), find the dataset report maximizing user's utility.*

4.1 Single Dimension

We begin with the scalar case $d = 1$, where the equilibrium structure reduces the search for the user-optimal report to a one-dimensional partition problem.

Setup. We consider the same utility functions as in Theorem 3.1. Let $w_\ell^*(D) \in \mathcal{W} = [\underline{w}, \bar{w}] \subseteq \mathbb{R}$ and let π_w denote the prior over $w_\ell^*(D)$.

Lemma 4.1. *When $d = 1$, after identifying dataset reports that result in the same learner hypothesis, the equilibrium dataset reports are ordered by $w_\ell^*(D)$.*

Lemma 4.2. *Under a continuous full-support prior π_w on $[\underline{w}, \bar{w}]$ and $c > 0$, every equilibrium has at most $N_{\max} = \big\lfloor 1 + \frac{\bar{w} - \underline{w}}{2c} \big\rfloor$ distinct dataset reports.*

The two lemmas reduce the problem of finding the user's optimal dataset report to a finite search over ordered partitions. Lemma 4.1 says that any one-dimensional equilibrium can be represented by ordered report regions. Thus, if an equilibrium has n distinct reports, its regions form the partition

$\mathcal{R} = \{R_i = [t_{i-1}, t_i] : i \in [n]\}$ for cutpoints $\underline{w} = t_0 < t_1 < \cdots < t_n = \bar{w}$. Lemma 4.2 bounds the number of reports that need to be considered.

Theorem 4.3. *Assume π_w is continuous with full support on $[\underline{w}, \bar{w}]$, and let $c > 0$. A feasible equilibrium is a partition $\underline{w} = t_0 < \cdots < t_n = \bar{w}$, $n \le N_{\max}$, satisfying*

$$t_i = \tfrac{1}{2}\big(\Phi(t_{i-1}, t_i) + \Phi(t_i, t_{i+1})\big) - c, \quad i = 1, \ldots, n-1. \tag{1}$$

The user-optimal equilibrium minimizes the loss $J = \sum_{i=1}^{n} p(t_{i-1}, t_i)\big(\Psi(t_{i-1}, t_i) - \Phi(t_{i-1}, t_i)^2\big) + \kappa$ over feasible partitions subject to (1)*, where $\kappa = \mathbb{E}_{\xi\sim\nu}\|\xi\|_2^2$ is the learning-error second moment (equal to its variance since $\mathbb{E}_{\xi\sim\nu}[\xi] = 0$). If $\boldsymbol{w}_\ell^*(D) \in [t_{i-1}, t_i]$, the user-optimal report is M_i whose region is $R_i = R_{M_i} = [t_{i-1}, t_i]$.*

Informally, every equilibrium is a monotone partition of $[\underline{w}, \bar{w}]$, and the learner responds on each region with the posterior mean of $\boldsymbol{w}_\ell^*(D)$. Taking the expectation over the learning error, the user's expected loss from a region is the within-region posterior variance, plus c^2, plus the constant κ. Because κ is the same for every region, it cancels from the indifference condition (1) and does not affect the optimal partition: maximizing user utility is equivalent to minimizing total within-region posterior variance. The user then sends the report whose region contains $\boldsymbol{w}_\ell^*(D)$.

Corollary 4.3.1. *Suppose π_w has finite support $w^{(1)} < \cdots < w^{(m)}$. A feasible finite-support equilibrium is a partition of the ordered support into contiguous blocks $R_i = \{w^{(t_{i-1}+1)}, \ldots, w^{(t_i)}\}$, $0 = t_0 < t_1 < \cdots < t_k = m$. For adjacent blocks R_i and R_{i+1}, Eq.* (1) *is replaced by*

$$w^{(t_i)} \le \frac{\Phi(R_i) + \Phi(R_{i+1})}{2} - c \le w^{(t_i+1)}. \tag{2}$$

Algorithm. Algorithm 2 appears in Appendix D. For a finite prior with m sorted support points, it computes statistics for all contiguous blocks, keeps adjacent pairs satisfying (2), and solves the resulting shortest-path problem with block cost $p(R)\big(\Psi(R) - \Phi(R)^2\big)$. The runtime is $O(m^3)$. The learning-error variance adds a constant κ, so it does not change the shortest path objective.

For a continuous prior, Lemma 4.2 restricts the search to $n \le N_{\max}$. If the solver uses I_n residual evaluations and returns S_n feasible partitions for a given n, the total posterior-oracle complexity is $O\Big(\sum_{n=1}^{N_{\max}} n(I_n + S_n)\Big)$. Once an optimal partition is stored, report lookup takes $O(\log n)$ time by binary search.

4.2 Multiple Dimensions

We now consider the same learning setting as Section 4.1 for $d > 1$. In multiple dimensions, equilibria are again represented by partitions of $\mathcal{W}$, but unlike the single-dimensional case, there is no canonical ordering of partitions.

Problem. Let $\mathcal{R} = \{R_1, \ldots, R_k\}$ be a measurable finite partition of $\mathcal{W}$, and let $\Phi_i = \Phi(R_i) = \mathbb{E}_{\pi_{\boldsymbol{w}}}[\boldsymbol{w} \mid \boldsymbol{w} \in R_i]$ be the posterior mean of R_i. Then $\mathcal{R}$ is an equilibrium partition if and only if, up to tie boundaries,

$$R_i = \big\{\boldsymbol{w} \in \mathcal{W} : \|\Phi_i - \boldsymbol{w} - \boldsymbol{c}\|_2^2 \le \|\Phi_j - \boldsymbol{w} - \boldsymbol{c}\|_2^2 \ \forall j\big\}. \tag{3}$$

Expected over the learning error, each squared distance gains the same constant κ, which cancels from the comparison, so (3) is unchanged. Among all partitions satisfying (3), the user-optimal partition solves

$$\max_{\mathcal{R}} \ -\sum_{i=1}^{k} \int_{R_i} \|\Phi_i - \boldsymbol{w} - \boldsymbol{c}\|_2^2 \, d\pi_{\boldsymbol{w}}(\boldsymbol{w}). \tag{4}$$

The learning error shifts (4) by the constant $-\kappa$, leaving the user-optimal partition unchanged. If $\mathcal{R}^* = \{R_1^*, \ldots, R_{k^*}^*\}$ is user-optimal, then $M^*(D) = M_i^*$ whenever $w_\ell^*(D) \in R_i^*$.

Algorithm. We show that the user's optimal information-release problem is NP-hard for arbitrary priors via a reduction from the NP-hard *planar k-means* problem. Our proof establishes NP-hardness for finding a user-optimal equilibrium with at most k reports, where k is part of the input. In

Model	Dataset	# Params.	Acc. (%)	Time (s)	Non-strategic	Strategic
Linear SVM	Credit Approval	47	85.4 ± 2.3	0.012 ± 0.007	-49.3 ± 26.1	$\mathbf{-10.2 \pm 12.0}$
	Census	104–106	85.0 ± 0.3	0.11 ± 0.01	-31.6 ± 10.1	$\mathbf{-11.5 \pm 13.0}$
	School Admission	52–57	95.0 ± 0.1	0.064 ± 0.005	-4.9 ± 4.2	$\mathbf{-4.2 \pm 4.7}$
	Prosper Loan	2090	36.7 ± 0.3	110 ± 12	-10.3 ± 4.5	$\mathbf{-6.4 \pm 6.6}$
MLP	Credit Approval	1570	83.8 ± 1.5	0.43 ± 0.39	-24.8 ± 7.5	$\mathbf{-10.3 \pm 10.8}$
	Census	3458	85.7 ± 0.1	5.6 ± 0.8	-60.4 ± 4.5	$\mathbf{-14.3 \pm 16.1}$
	School Admission	1762–1826	94.8 ± 0.1	3.3 ± 1.5	-30.8 ± 13.3	$\mathbf{-10.7 \pm 12.2}$
	Prosper Loan	6443	49.4 ± 0.3	52 ± 6	-397.3 ± 39.8	$\mathbf{-15.9 \pm 23.1}$

Table 1: Algorithm runtime, model accuracy, and average user utility ($\times 10^3$). "# Params." denotes the number of model parameters. Truthful reporting is an upper bound rather than an equilibrium and yields -3.2 ± 4.2 in all settings.

equilibrium, the learner deploys one model per report, so k bounds the number of distinct models it deploys. This budget is natural in practice, since deploying many models is costly: each must be validated and monitored. Learners therefore serve a few models and route each user to one of them. For example, a lender splits applicants into k geographic segments and serves one credit approval model per segment. We also show that exact evaluation of the statistics oracle under general priors is #P-hard via a uniform $[0, 1]^d$ prior. The formal hardness statements and their proofs are in the Appendix D.

4.2.1 Efficient Algorithm for Multiple Dimensions.

Given the hardness of the problem for arbitrary prior distributions, we identify a structured class of priors for which the strategic part of the problem becomes tractable. The key observation is that the conflict vector $\boldsymbol{c}$ determines the direction of disagreement between the user and the learner. Orthogonal to this direction, their objectives are aligned. Proofs appear in Appendix D.2.

Factorized priors. Assume $\boldsymbol{c} \neq 0$ and let $\widehat{\boldsymbol{c}} = \boldsymbol{c}/\|\boldsymbol{c}\|_2$. We decompose $\boldsymbol{w}^*_\ell(D) = s\widehat{\boldsymbol{c}} + \boldsymbol{z}$, where $s = \widehat{\boldsymbol{c}}^\top \boldsymbol{w}^*_\ell(D) \in [\underline{s}, \overline{s}]$ is the conflict projection and $\boldsymbol{z}$ the agreement component, and write $\pi_s, \pi_{\boldsymbol{z}}$ for their marginals. We assume $\pi_{\boldsymbol{w}} = \pi_s \otimes \pi_{\boldsymbol{z}}$, so the conflict projection is independent of the agreement component, with π_s of bounded support. Factorized structure over learned parameters is the standard tractability assumption in Bayesian deep learning, where mean-field variational inference approximates the distribution over weights by a product of independent per-coordinate factors [4, 13].

Theorem 4.4. *Under the factorization assumption, any one-dimensional equilibrium partition of s with conflict $\beta = \|\boldsymbol{c}\|_2$ lifts to a multidimensional equilibrium that reveals $\boldsymbol{z}$ exactly and manipulates only along $\hat{\boldsymbol{c}}$. If the scalar partition is user-optimal, its lift is user-optimal among lifted equilibria.*

Algorithm. Given a dataset D, we compute $\boldsymbol{w}^*_\ell(D)$ and decompose it into the conflict coordinate $s(D) = \hat{\boldsymbol{c}}^\top \boldsymbol{w}^*_\ell(D)$ and the agreement component $\boldsymbol{z}(D) = \boldsymbol{w}^*_\ell(D) - s(D)\hat{\boldsymbol{c}}$. We reuse the algorithm of Section 4.1 (Algorithm 2) on the scalar coordinate s with conflict $\beta = \|\boldsymbol{c}\|_2$, obtaining cutpoints $t_0 < t_1 < \cdots < t_{n^\star}$. Let i be the unique index with $s(D) \in [t_{i-1}, t_i]$. The algorithm returns the report $M^\star(D)$ associated with the region $R_{M^\star(D)} = \big\{\boldsymbol{w} \in \mathcal{W} : \hat{\boldsymbol{c}}^\top \boldsymbol{w} \in [t_{i-1}, t_i],\ \boldsymbol{w} - (\hat{\boldsymbol{c}}^\top \boldsymbol{w})\hat{\boldsymbol{c}} = \boldsymbol{z}(D)\big\}$. Computing $s(D)$ and $\boldsymbol{z}(D)$ takes $O(d)$ time. The scalar cutpoints are computed as in Section 4.1. For the empirical priors used in our experiments, the cost is $O(m^3 + d)$.

5 Experimental Evaluation

We evaluate our algorithms on four real-world datasets using linear SVM and two-layer MLP learners. Additional experiment setup details appear in Appendix F.

Computing the User-Optimal Report. We evaluate the efficient multidimensional algorithm under the factorization assumption (Section 4.2.1) at four conflict magnitudes and randomly sampled conflict directions. For each dataset–learner pair, we construct an empirical prior from 50 independent stratified bootstrap samples. Results are averaged over three runs. Table 1 compares the average

user utility of the *Strategic* report returned by our algorithm with two baselines. *Truthful* reports the user's private dataset without modification. Under the nonzero conflicts considered here, however, it is not an equilibrium because the user can profitably deviate. *Non-strategic* induces a non-influential equilibrium in which the report conveys no information (Section 3). Strategic reporting achieves higher user utility in all eight dataset–learner settings and substantially closes the gap to the truthful upper bound. Runtime is below six seconds in six settings. Appendix F.2 studies additional scalability results.

References

[1] Yakov Babichenko, Inbal Talgam-Cohen, Haifeng Xu, and Konstantin Zabarnyi. Algorithmic cheap talk. In *Proceedings of the 25th ACM Conference on Economics and Computation*, EC '24, page 5–6, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400707049. doi: 10.1145/3670865.3673604. URL https://doi.org/10.1145/3670865.3673604.

[2] Marco Battaglini. Multiple referrals and multidimensional cheap talk. *Econometrica*, 70 (4):1379–1401, 2002. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/3082002.

[3] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[4] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *Proceedings of the 32nd International Conference on Machine Learning - Volume 37*, ICML'15, page 1613–1622. JMLR.org, 2015.

[5] Oliver Board and Andreas Blume. Intentional vagueness. Working Paper 365, Department of Economics, University of Pittsburgh, Aug 2008. URL https://ideas.repec.org/p/pit/wpaper/365.html.

[6] Markus K. Brunnermeier and Lasse H. Pedersen. Predatory trading. *Journal of Finance*, 60(4): 1825—-1863, August 2005.

[7] Archishman Chakraborty and Rick Harbaugh. Comparative cheap talk. *Journal of Economic Theory*, 132(1):70–94, 2007. ISSN 0022-0531. doi: https://doi.org/10.1016/j.jet.2005.07.003. URL https://www.sciencedirect.com/science/article/pii/S0022053105001729.

[8] Lee Cohen, Saeed Sharifi-Malvajerdi, Kevin Stangl, Ali Vakilian, and Juba Ziani. Bayesian strategic classification. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 111649–111678. Curran Associates, Inc., 2024. doi: 10.52202/079017-3546. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ca774047bc3b46cc81e53ead34cd5d5a-Paper-Conference.pdf.

[9] Vincent P. Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, 50 (6):1431–1451, 1982. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1913390.

[10] Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70, 2018.

[11] M. E. Dyer and A. M. Frieze. On the complexity of computing the volume of a polyhedron. *SIAM Journal on Computing*, 17(5):967–974, 1988. doi: 10.1137/0217060. URL https://doi.org/10.1137/0217060.

[12] Michael C. Ferris and Todd S. Munson. Complementarity problems in gams and the path solver1this material is based on research supported by national science foundation grant ccr-9619765 and gams corporation. the paper is an extended version of a talk presented at cefes '98 (computation in economics, finance and engineering: Economic systems) in cambridge, england, in july 1998.1. *Journal of Economic Dynamics and Control*, 24(2):165–188, 2000.

ISSN 0165-1889. doi: https://doi.org/10.1016/S0165-1889(98)00092-X. URL https://www.sciencedirect.com/science/article/pii/S016518899800092X.

[13] Alex Graves. Practical variational inference for neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'11, page 2348–2356, Red Hook, NY, USA, 2011. Curran Associates Inc. ISBN 9781618395993.

[14] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, ITCS '16, page 111–122, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340571. doi: 10.1145/2840728.2840730. URL https://doi.org/10.1145/2840728.2840730.

[15] Alexandros Hollender, Chester Lawrence, and Erel Segal-Halevi. Computing approximate roots of monotone functions. *J. Complex.*, 88(C), June 2025. ISSN 0885-064X. doi: 10.1016/j.jco.2025.101930. URL https://doi.org/10.1016/j.jco.2025.101930.

[16] Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, October 2011. doi: 10.1257/aer.101.6.2590. URL https://www.aeaweb.org/articles?id=10.1257/aer.101.6.2590.

[17] Volodymyr Kuleshov and Okke Schrijvers. Inverse game theory: Learning utilities innbsp;succinct games. In *Web and Internet Economics: 11th International Conference, WINE 2015, Amsterdam, The Netherlands, December 9-12, 2015, Proceedings*, page 413–427, Berlin, Heidelberg, 2015. Springer-Verlag. ISBN 978-3-662-48994-9. doi: 10.1007/978-3-662-48995-6_30. URL https://doi.org/10.1007/978-3-662-48995-6_30.

[18] Wonbin Kweon, Seongku Kang, Junyoung Hwang, and Hwanjo Yu. Deep rating elicitation for new users in collaborative filtering. In *Proceedings of The Web Conference 2020*, WWW '20, page 2810–2816, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370233. doi: 10.1145/3366423.3380042. URL https://doi.org/10.1145/3366423.3380042.

[19] Albert S. Kyle. Continuous auctions and insider trading. *Econometrica*, 53(6):1315–1335, November 1985.

[20] Junyi Liao, Zihan Zhu, Ethan X Fang, Zhuoran Yang, and Vahid Tarokh. Decoding rewards in competitive games: Inverse game theory with entropy regularization. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 37610–37622. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/liao25i.html.

[21] Xinpeng Lv, Yunxin Mao, Renzhe Xu, Chunyuan Zheng, Yikai Chen, Haoxuan Li, Yang Shi, Jinxuan Yang, Zhouchen Lin, Yuanlong Chen, Yuanxing Zhang, Shaowu Yang, Wenjing Yang, and Haotian Wang. Beyond rational illusion: Behaviorally realistic strategic classification, 2026. URL https://arxiv.org/abs/2605.19674.

[22] Meena Mahajan, Prajakta Nimbhorkar, and Kasturi Varadarajan. The planar k-means problem is np-hard. *Theoretical computer science*, 442:13–21, 2012.

[23] Paul R. Milgrom. Good news and bad news: Representation theorems and applications. *Bell Journal of Economics*, 12(2):380–391, Autumn 1981. doi: None. URL https://ideas.repec.org/a/rje/bellje/v12y1981iautumnp380-391.html.

[24] Dan Ofer. School Admission, 1998. https://www.kaggle.com/datasets/danofer/law-school-admissions-bar-passage.

[25] Henry Okam. Prosper Loan, 2021. https://www.kaggle.com/datasets/henryokam/prosper-loan-data.

[26] J. R. Quinlan. Credit Approval. UCI Machine Learning Repository, 1987. DOI: https://doi.org/10.24432/C5FS30.

[27] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 0387212396.

[28] Adrien Saumard and Jon A. Wellner. Log-concavity and strong log-concavity: a review, 2014. URL https://arxiv.org/abs/1404.5886.

[29] Alexis Stenfors and Masayuki Susai. Stealth trading in fx markets, February 2021. University of Portsmouth.

[30] Jibang Wu, Weiran Shen, Fei Fang, and Haifeng Xu. Inverse game theory for stackelberg games: the blessing of bounded rationality. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 32186–32198. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/cfce833814505906445f8df2f65ab548-Paper-Conference.pdf.

[31] Min-Hsuan Yeh, Jeffrey Wang, Xuefeng Du, Seongheon Park, Leitian Tao, Shawn Im, and Yixuan Li. Position: Challenges and future directions of data-centric AI alignment. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=bXfF6Dqe9s.

[32] Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in ranking, part ii: Learning-to-rank and recommender systems. *ACM Comput. Surv.*, 55(6), dec 2022. ISSN 0360-0300. doi: 10.1145/3533380. URL https://doi.org/10.1145/3533380.

[33] D. Zotikov and A. Antonov. CME iceberg order detection and prediction. *Quantitative Finance*, 21(11):1977–1992, 2021. doi: 10.1080/14697688.2020.1813904.

[34] Zhiqun Zuo and Mohammad Mahdi Khalili. Individual fairness in strategic classification, 2026. URL https://arxiv.org/abs/2602.05084.

A Additional Related Work

Strategic Communication. A large literature in economics and game theory studies how an agent with private information can influence another party's decision through costless, unverifiable messages [9]. Extensions to multidimensional states show that influential equilibria can persist even under conflict, but typically under additional structural assumptions, such as comparative or ordered message spaces or multiple informed agents [2, 7]. We instead study a single informed user with general multidimensional messages (dataset reports).

Our setting differs from this literature in several ways. First, they do not address algorithmic problems in the game. Second, the elements of the game in our settings are different from the ones investigated in the aforementioned line of work. In our setting, the private information is not a single point, but rather a training dataset that contains several samples. The messages are not points in some multi-dimensional space; our messages are sets of samples. The action is a trained hypothesis produced by the (learning) algorithm that is only partially known to the other player. These modifications increase the complexity of our problem. Third, the information asymmetry in our setting is two-sided: the learner does not know the dataset, and the user does not observe the learner's training algorithm. Fourth, learning adds stochastic error, which is absent from the classical theory. Fifth, the questions are correspondingly algorithmic: when influence is possible, how the user computes an optimal report, and where these problems become provably hard. Finally, we study the inverse problem of recovering the unknown conflict from a learner's behavior, which, to the best of our knowledge, has no counterpart in this literature.

Researchers have recently investigated algorithmic problems in cheap talk [1]. Our work differs from theirs in three ways. First, they study games with finite state and action sets, whereas our state space is a multidimensional space of learned models described by a distribution. Second, their hardness result reduces from a variant of 3SAT, whereas ours reduces from planar k-means and exploits the geometry of report-induced regions. Third, their tractability results assume either a constant number of states or two actions, whereas ours exploit learning-specific structure—particularly the direction of conflict between the user's and the learner's ideal hypotheses. Moreover, this line of work does not address the added complexity to the original framework in our setting explained in the preceding paragraph.

Inverse Game Theory. A complementary line of work infers agents' objectives from their observed behavior in specific domains. Passive methods rationalize observed equilibrium play, classically in succinct simultaneous-move games [17] and more recently in zero-sum Markov games [20]. Active methods instead query the agent, as in Stackelberg games where a leader commits to strategies and observes responses [30]. These assumptions of purely adversarial play and leader commitment do not hold in our setting. These methods also treat the inverse problem in isolation, whereas our framework couples it with the forward problem and learns the conflict under either access model.

B User's and Learner's Knowledge About Learning Algorithms

The user does not know the learner's algorithm. Instead, the user has a training algorithm $\mathcal{A} : \mathcal{D} \to \mathcal{W}$ that produces $\hat{\boldsymbol{w}}(D) = \mathcal{A}(D)$, an approximation of $\boldsymbol{w}_\ell^*(D)$ that need not equal it; the user selects its dataset report based on $\hat{\boldsymbol{w}}(D)$. We assume the learner knows the user's training algorithm. Formally, we assume there exists an injective Borel-measurable map $\psi : \mathcal{W} \to \mathcal{W}$, monotone when $d = 1$, such that $\mathcal{A}(D) = \psi\big(\boldsymbol{w}_\ell^*(D)\big)$ for all $D \in \mathcal{D}$, and $\mathcal{A}$ (hence ψ) is known to the learner. Strategies in the reported game are identified on the image $\psi(\mathcal{W})$. The user need not know ψ and in general cannot recover $\boldsymbol{w}_\ell^*(D)$ from $\hat{\boldsymbol{w}}(D)$.

Proposition B.1. *Under the above assumption, the game in which the user reports based on $\hat{\mathbf{w}}(D)$ is strategically equivalent to the game of Section 2: (σ_r, σ_ℓ) is an equilibrium of the former if and only if $(\sigma_r \circ \psi, \sigma_\ell)$ is an equilibrium of the latter, and the induced report regions satisfy $\hat{R}_M = \psi(R_M)$. Moreover, corresponding equilibria induce identical outcomes: for every $D \in \mathcal{D}$, $\sigma_r(\hat{\mathbf{w}}(D)) = (\sigma_r \circ \psi)(\mathbf{w}_\ell^*(D))$, so reports, responses, and realized utilities coincide.*

Proof. Since ψ is an injective Borel map, $\psi(R)$ is measurable and the events $\{\hat{\boldsymbol{w}}(D) \in \psi(R)\}$ and $\{\boldsymbol{w}_\ell^*(D) \in R\}$ coincide for every measurable $R \subseteq \mathcal{W}$. Hence any reporting strategy in $\hat{\boldsymbol{w}}$, restricted to $\psi(\mathcal{W})$, induces the same partition of $\mathcal{D}$, the same posteriors $\pi(\cdot \mid M)$, and the same expected

utilities as the corresponding strategy in $\boldsymbol{w}_\ell^*(D)$; best-response sets are therefore identical under the bijection $R \mapsto \psi(R)$, and matched strategies prescribe the same report at every D. □

Remark B.2. *When $d = 1$, monotonicity of ψ maps intervals to intervals, so the ordered partition structure of Section 4.1 is preserved in the reported variable.*

For the remainder of the paper, our standing assumption is that the user does not know the learner's algorithm. Whenever the user is said to "evaluate" or "compute" $\boldsymbol{w}_\ell^*(D)$, it is understood that this is done using the user's own training algorithm. The prior $\pi_{\boldsymbol{w}}$ over the learner's preferred model $\boldsymbol{w}_\ell^*(D)$, supported on the convex parameter space $\mathcal{W}$, remains common knowledge (Section 3.1).

C Missing Proofs of Section 3

We first provide a useful Lemma that we use inside the proof.

Lemma C.1. *Suppose the learning error $\boldsymbol{\xi} \sim \nu$ is independent of the private dataset, satisfies $\mathbb{E}[\boldsymbol{\xi}] = \mathbf{0}$, and has $\kappa := \mathbb{E}[\|\boldsymbol{\xi}\|_2^2] < \infty$. For any measurable report region A with positive probability, let $\Phi(A) := \mathbb{E}_{\pi_{\boldsymbol{w}}}\left[\boldsymbol{w}_\ell^*(D) \mid \boldsymbol{w}_\ell^*(D) \in A\right]$. Then the learner's best response to A is $\Phi(A)$. Moreover, for any private dataset D and learner response $\boldsymbol{w}$,*

$$U^r(D, \boldsymbol{w}) = -\left\|\boldsymbol{w} - \left(\boldsymbol{w}_\ell^*(D) + \boldsymbol{c}\right)\right\|_2^2 - \kappa.$$

Consequently, the learning error does not change the learner's posterior-mean response or the user's comparisons between dataset reports.

Proof. For any candidate learner response $\boldsymbol{w}$,

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{w} + \boldsymbol{\xi} - \boldsymbol{w}_\ell^*(D)\|_2^2 \;\middle|\; \boldsymbol{w}_\ell^*(D) \in A\right] \\
&\quad = \|\boldsymbol{w} - \Phi(A)\|_2^2 + \mathbb{E}\left[\|\boldsymbol{w}_\ell^*(D) - \Phi(A)\|_2^2 \;\middle|\; \boldsymbol{w}_\ell^*(D) \in A\right] + \kappa.
\end{aligned}$$

The cross term containing $\boldsymbol{\xi}$ is zero by data independence and $\mathbb{E}[\boldsymbol{\xi}] = \mathbf{0}$. The only term depending on $\boldsymbol{w}$ is $\|\boldsymbol{w} - \Phi(A)\|_2^2$, so the learner's best response is $\Phi(A)$.

Similarly,

$$\begin{aligned}
U^r(D, \boldsymbol{w}) &= -\mathbb{E}_{\boldsymbol{\xi}\sim\nu}\left[\left\|\boldsymbol{w} + \boldsymbol{\xi} - \left(\boldsymbol{w}_\ell^*(D) + \boldsymbol{c}\right)\right\|_2^2\right] \\
&= -\left\|\boldsymbol{w} - \left(\boldsymbol{w}_\ell^*(D) + \boldsymbol{c}\right)\right\|_2^2 - \kappa.
\end{aligned}$$

Thus, the same term $-\kappa$ appears under every learner response and cancels from all comparisons between dataset reports. □

Proof of Theorem 3.1. For any measurable $A \subseteq [0,1]$ with positive π_w-probability, write $\Phi(A) := \mathbb{E}_{\pi_w}[w_\ell^*(D) \mid w_\ell^*(D) \in A]$, so that $\Phi_L(t) = \Phi([0,t])$ and $\Phi_H(t) = \Phi((t,1])$. By Lemma C.1, the learner responds to every positive-probability report region A with $\Phi(A)$, and learning error does not affect the user's comparisons between reports.

($\Leftarrow$) Fix $t \in (0,1)$ satisfying

$$t + c = \frac{\Phi([0,t]) + \Phi((t,1])}{2}.$$

Choose two distinct reports M_1, M_2, and let their report regions be $R_1 = [0,t]$ and $R_2 = (t,1]$, respectively. Since π_w has full support, both regions have positive probability. Let the learner respond to M_i with $\Phi(R_i)$, and to every other report with $\Phi(R_1)$.

For a private dataset D, the user's gain from selecting M_2 rather than M_1 is

$$\begin{aligned}
&U^r(D, \Phi(R_2)) - U^r(D, \Phi(R_1)) \\
&\quad = \big(\Phi(R_2) - \Phi(R_1)\big)\Big(2(w_\ell^*(D) + c) - \Phi(R_1) - \Phi(R_2)\Big).
\end{aligned}$$

Because $\Phi(R_2) > \Phi(R_1)$, its sign is the sign of $w_\ell^*(D) + c - (\Phi(R_1) + \Phi(R_2))/2$, which by the cutoff condition is the sign of $w_\ell^*(D) - t$. Thus, the user selects M_1 below t, selects M_2 above t,

and is indifferent at t. These strategies form an equilibrium. Since the two reports result in distinct learner responses, learning is influential.

($\Rightarrow$) Suppose learning is influential in some equilibrium. Consider two positive-probability report regions R, R' satisfying $\Phi(R) < \Phi(R')$. The user's gain from selecting the report corresponding to R' rather than that corresponding to R is

$$\big(\Phi(R') - \Phi(R)\big) \times \Big(2(w_\ell^*(D) + c) - \Phi(R) - \Phi(R')\Big),$$

which is strictly increasing in $w_\ell^*(D)$. Therefore, after identifying reports that result in the same learner response and ignoring null boundary sets, the equilibrium report regions are intervals ordered by $w_\ell^*(D)$.

Write the regions in increasing order as $R_i = (t_{i-1}, t_i]$, with $t_0 = 0$, and let $w_i := \Phi(R_i)$. At each interior cutpoint t_i, users immediately below t_i weakly prefer the report corresponding to R_i, while users immediately above t_i weakly prefer the report corresponding to R_{i+1}. Full support and continuity of the utility difference therefore imply

$$t_i + c = \frac{w_i + w_{i+1}}{2}. \tag{5}$$

The partition is finite. Indeed, $w_i \le t_i$ and $w_{i+1} \le t_{i+1}$, so

$$t_i + c = \frac{w_i + w_{i+1}}{2} \le \frac{t_i + t_{i+1}}{2}.$$

Hence $t_{i+1} - t_i \ge 2c$. Since $c > 0$ and the support has length one, only finitely many regions can occur. Write $0 = t_0 < t_1 < \cdots < t_n = 1$, where $n \ge 2$.

Define

$$g(t) := \frac{\Phi([0,t]) + \Phi((t,1])}{2} - t - c.$$

The function g is continuous on $(0,1)$. At t_1, $\Phi([0,t_1]) = w_1$, while $\Phi((t_1,1])$ is a weighted average of $w_2, \ldots, w_n$. Therefore,

$$g(t_1) \ge \frac{w_1 + w_2}{2} - t_1 - c = 0.$$

Similarly, $\Phi((t_{n-1},1]) = w_n$ and $\Phi([0,t_{n-1}]) \le w_{n-1}$, so

$$g(t_{n-1}) \le \frac{w_{n-1} + w_n}{2} - t_{n-1} - c = 0.$$

By the intermediate value theorem, there exists $t^* \in [t_1, t_{n-1}] \subset (0,1)$ such that $g(t^*) = 0$. Therefore,

$$t^* + c = \frac{\Phi([0,t^*]) + \Phi((t^*,1])}{2},$$

as required. □

Remark C.2. *When the cutoff is computed numerically, the correctness of the algorithm depends on the guarantees of the solver and the error in the posterior means returned by the oracle. For specific priors, the cutoff condition can be solved analytically, as shown next.*

Proof of Corollary 3.1.1. Under the uniform prior, $\Phi_L(t) = t/2$ and $\Phi_H(t) = (1+t)/2$. Substituting into the cutoff condition in Theorem 3.1 gives

$$t + c = \frac{1}{2}\left(\frac{t}{2} + \frac{1+t}{2}\right) = \frac{1+2t}{4}.$$

Rearranging yields $t = 1/2 - 2c$. An influential equilibrium exists if and only if $t \in (0,1)$, or equivalently,

$$0 < \frac{1}{2} - 2c < 1.$$

Since $c \ge 0$, the upper inequality holds automatically, while the lower inequality is equivalent to $c < 1/4$. Therefore, learning is influential if and only if $0 \le c < 1/4$. □

Intuition for Corollary 3.1.1. Under the uniform prior, the existence condition depends only on the magnitude of the conflict parameter c, so it can be checked in constant time. The user's private dataset determines the learner's preferred model parameter $w_\ell^*(D)$, while the user prefers the shifted parameter $w_\ell^*(D) + c$. When c is small, datasets associated with lower values of $w_\ell^*(D)$ can be grouped under one report and those associated with higher values under another. The learner responds to each report with the corresponding posterior-mean model, and users in each group prefer their assigned report to switching to the other report. The reported dataset can therefore influence the learned model. When $c \geq 1/4$, no interior cutoff satisfies this equilibrium condition, so learning cannot be influential.

Proof of Theorem 3.2. Fix $\boldsymbol{c} \in \mathbb{R}^d$. Since $d > 1$, choose a unit vector $\boldsymbol{v}$ satisfying $\boldsymbol{v}^\top \boldsymbol{c} = 0$, and define

$$R_+ := \{\boldsymbol{w} : \boldsymbol{v}^\top(\boldsymbol{w} - \bar{\boldsymbol{w}}) \geq 0\}, \quad R_- := \{\boldsymbol{w} : \boldsymbol{v}^\top(\boldsymbol{w} - \bar{\boldsymbol{w}}) < 0\}.$$

Choose two distinct dataset reports M_+ and M_-. Let the user select M_+ when $\boldsymbol{w}_\ell^*(D) \in R_+$, and M_- when $\boldsymbol{w}_\ell^*(D) \in R_-$.

Continuity and spherical symmetry imply that the separating hyperplane has probability zero and that the two regions have equal probability. Moreover, their posterior means are symmetric around $\bar{\boldsymbol{w}}$ and differ from $\bar{\boldsymbol{w}}$ only in the direction $\boldsymbol{v}$. Therefore,

$$\Phi(R_+) = \bar{\boldsymbol{w}} + \alpha \boldsymbol{v}, \qquad \Phi(R_-) = \bar{\boldsymbol{w}} - \alpha \boldsymbol{v}$$

for some $\alpha > 0$. Strict positivity follows from non-degeneracy.

By Lemma C.1, let the learner respond to M_+ and M_- with $\Phi(R_+)$ and $\Phi(R_-)$, respectively, and respond to every other report with $\Phi(R_-)$. Learning error does not affect the user's comparison between these responses.

For a private dataset D, write $\boldsymbol{w} := \boldsymbol{w}_\ell^*(D)$. The user weakly prefers M_+ to M_- if and only if

$$\begin{aligned} &\|\bar{\boldsymbol{w}} + \alpha \boldsymbol{v} - (\boldsymbol{w} + \boldsymbol{c})\|_2^2 \\ &\quad \leq \|\bar{\boldsymbol{w}} - \alpha \boldsymbol{v} - (\boldsymbol{w} + \boldsymbol{c})\|_2^2 . \end{aligned}$$

After cancelling common terms, this is equivalent to

$$\boldsymbol{v}^\top(\boldsymbol{w} - \bar{\boldsymbol{w}} + \boldsymbol{c}) \geq 0.$$

Since $\boldsymbol{v}^\top \boldsymbol{c} = 0$, this reduces to $\boldsymbol{v}^\top(\boldsymbol{w} - \bar{\boldsymbol{w}}) \geq 0$. Thus, when $\boldsymbol{w}_\ell^*(D) \in R_+$, the user weakly prefers M_+, and when $\boldsymbol{w}_\ell^*(D) \in R_-$, the user strictly prefers M_-.

Therefore, the user's strategy is a best response, and the learner responds to both reports with their corresponding posterior means. The strategies form an equilibrium. Since $\Phi(R_+) \neq \Phi(R_-)$, learning is influential. □

D Missing Details of Section 4

D.1 Missing Details of Section 4.1

This section includes the proofs of the user's optimal information release problem in a single dimension setting. We first provide an example that provides intuition of the solution.

Example D.1. *Suppose $w_\ell^*(D)$ is the hypothesis learned from the private dataset D, representing the HFT system's assessed level of execution accommodation for the tranche. Here larger values of $w_\ell^*(D)$ correspond to systematically more favorable (less predatory) trading decisions. The learner prefers to deploy exactly $w_\ell^*(D)$, while the institutional-trader user prefers a more accommodating hypothesis $w_\ell^*(D) + c$. Let the prior on $w_\ell^*(D)$ be uniform on $[0, 1]$, and let $c = 0.1$. Then $n^* = 2$, and the optimal cutpoints are $t_0 = 0$, $t_1 = \frac{1}{2} - 2c = 0.3$, and $t_2 = 1$. Hence, the user-optimal equilibrium partition is $[0, 1] = [0, 0.3] \cup [0.3, 1]$. Thus, the user-optimal dataset report does not reveal the exact private dataset D. Instead, it reveals only whether the hypothesis learned from D lies in the low-accommodation partition $[0, 0.3]$ or the high-accommodation partition $[0.3, 1]$. For instance, if the true private dataset D results in $w_\ell^*(D) = 0.6$, then the user sends the report corresponding to the region $[0.3, 1]$, i.e., a report that reveals only that the dataset belongs to the set of datasets whose resulting hypotheses lie in $[0.3, 1]$.*

Proof of Lemma 4.1. If two dataset reports result in the same learner response, identifying them as a single report does not change the utility of either the user or the learner. Hence, it suffices to compare reports that result in distinct learner responses.

Consider two dataset reports M' and M'' such that the learner responds with $w' < w''$, respectively. By Lemma C.1, the user's gain from selecting M'' rather than M' when the private dataset is D is

$$\begin{aligned}\Delta(D) &:= U^r(D, w'') - U^r(D, w') \\ &= (w'' - w')\big(2(w_\ell^*(D) + c) - w' - w''\big).\end{aligned}$$

Since $w'' - w' > 0$, $\Delta(D)$ is strictly increasing in $w_\ell^*(D)$.

Suppose the report regions are not ordered. Then there exist private datasets D, D' such that $w_\ell^*(D) < w_\ell^*(D')$, but $\sigma_r(D) = M''$ and $\sigma_r(D') = M'$. Because selecting M'' is optimal when the private dataset is D, we have $\Delta(D) \geq 0$. Strict monotonicity then gives $\Delta(D') > \Delta(D) \geq 0$, so the user with private dataset D' strictly prefers M'' to M', contradicting $\sigma_r(D') = M'$.

Therefore, higher values of $w_\ell^*(D)$ cannot be assigned reports that result in lower learner responses. Hence, equilibrium report regions are ordered by $w_\ell^*(D)$. □

Proof of Lemma 4.2. By Lemma 4.1, after identifying reports that result in the same learner response, an equilibrium has an ordered partition $\underline{w} = t_0 < t_1 < \cdots < t_n = \bar{w}$. Let $R_i = (t_{i-1}, t_i]$ and $w_i := \Phi(t_{i-1}, t_i)$.

At an interior cutpoint t_i, users immediately below t_i weakly prefer the report corresponding to R_i, while users immediately above t_i weakly prefer the report corresponding to R_{i+1}. By Lemma C.1, learning error does not affect this comparison. Full support and continuity of the utility difference therefore imply

$$-\big(w_i - (t_i + c)\big)^2 = -\big(w_{i+1} - (t_i + c)\big)^2.$$

Since $w_i < w_{i+1}$, this is equivalent to

$$t_i + c = \frac{w_i + w_{i+1}}{2}.$$

Moreover, $w_i \leq t_i$ and $w_{i+1} \leq t_{i+1}$, so

$$t_i + c = \frac{w_i + w_{i+1}}{2} \leq \frac{t_i + t_{i+1}}{2}.$$

Hence, $t_{i+1} - t_i \geq 2c$ for every $i = 1, \ldots, n-1$. Summing these inequalities gives

$$2c(n-1) \leq \sum_{i=1}^{n-1}(t_{i+1} - t_i) \leq \bar{w} - \underline{w}.$$

Therefore,

$$n \leq 1 + \frac{\bar{w} - \underline{w}}{2c}.$$

□

Proof of Theorem 4.3. By Lemma 4.1, it suffices to consider ordered partitions $\underline{w} = t_0 < t_1 < \cdots < t_n = \bar{w}$. Let $R_i = [t_{i-1}, t_i]$ and $w_i := \Phi(t_{i-1}, t_i)$. Since π_w is continuous, the assignment of boundary points does not affect posterior means or expected utilities. By Lemma C.1, the learner responds to the report corresponding to R_i with w_i.

Consider an interior cutpoint t_i. As in the proof of Lemma 4.2, full support and continuity imply that the user is indifferent between the two adjacent learner responses at t_i:

$$-\big(w_i - (t_i + c)\big)^2 = -\big(w_{i+1} - (t_i + c)\big)^2.$$

Since $w_i < w_{i+1}$, this is equivalent to

$$t_i + c = \frac{w_i + w_{i+1}}{2},$$

or

$$t_i = \frac{1}{2}\left(\Phi(t_{i-1}, t_i) + \Phi(t_i, t_{i+1})\right) - c,$$

which is exactly (1). Thus, every equilibrium partition satisfies the cutpoint system.

Conversely, suppose an ordered partition satisfies (1). For adjacent responses $w_i < w_{i+1}$, the user's gain from selecting the report corresponding to R_{i+1} rather than that corresponding to R_i has the same sign as $w_\ell^*(D) - t_i$. Hence, for $w_\ell^*(D) \in R_i$,

$$\frac{w_{i-1} + w_i}{2} \le w_\ell^*(D) + c \le \frac{w_i + w_{i+1}}{2},$$

with the first or last inequality omitted for the endpoint regions. Therefore, w_i is a closest learner response to $w_\ell^*(D) + c$ among the ordered responses $w_1, \ldots, w_n$. The user consequently prefers the report corresponding to R_i to every other report. Thus, (1) characterizes feasible equilibrium partitions.

By Lemma 4.2, every feasible equilibrium has at most $N_{\max}$ regions.

It remains to select the user-optimal equilibrium. On region R_i, the learner responds with $w_i = \Phi(t_{i-1}, t_i)$. By Lemma C.1, the conditional expected user loss is

$$\begin{aligned}\mathbb{E}_{\pi_w}\left[(w_i - (w_\ell^*(D) + c))^2 \;\middle|\; w_\ell^*(D) \in R_i\right] + \kappa \\ = \Psi(t_{i-1}, t_i) - \Phi(t_{i-1}, t_i)^2 + c^2 + \kappa.\end{aligned}$$

Weighting by $p(t_{i-1}, t_i)$ and summing over all regions gives the user's expected loss

$$\sum_{i=1}^{n} p(t_{i-1}, t_i)\left(\Psi(t_{i-1}, t_i) - \Phi(t_{i-1}, t_i)^2\right) + c^2 + \kappa.$$

Since c^2 is fixed across all feasible partitions, minimizing this loss is equivalent to minimizing

$$J = \sum_{i=1}^{n} p(t_{i-1}, t_i)\left(\Psi(t_{i-1}, t_i) - \Phi(t_{i-1}, t_i)^2\right) + \kappa. \tag{6}$$

over partitions satisfying (1) with $n \le N_{\max}$.

Therefore, any feasible partition minimizing J is user-optimal. If $w_\ell^*(D) \in [t_{i-1}, t_i]$, the corresponding user-optimal report is M_i, whose region is $R_i = R_{M_i} = [t_{i-1}, t_i]$. □

Proof of Corollary 4.3.1. By Lemma 4.1, after identifying reports that result in the same learner response, equilibrium regions are ordered. Because π_w has finite support, these regions are contiguous blocks of $w^{(1)} < \cdots < w^{(m)}$.

By Lemma C.1, the learner responds to the report corresponding to block R_i with its posterior mean $\Phi(R_i)$, and learning error does not affect the user's comparisons between reports. For adjacent blocks R_i and R_{i+1}, define the model value at which the user is indifferent between their learner responses as

$$q_i := \frac{\Phi(R_i) + \Phi(R_{i+1})}{2} - c.$$

All support points in R_i weakly prefer their assigned report if and only if $w^{(t_i)} \le q_i$, while all support points in R_{i+1} weakly prefer their assigned report if and only if $q_i \le w^{(t_i+1)}$. Hence, feasibility of the adjacent blocks is equivalent to

$$w^{(t_i)} \le \frac{\Phi(R_i) + \Phi(R_{i+1})}{2} - c \le w^{(t_i+1)},$$

which is exactly (2).

Now we show that optimality follows using the same selection argument as Theorem 4.3. Because the learner responses are ordered, these adjacent conditions also rule out deviations to non-adjacent reports. Therefore, a contiguous-block partition is feasible if and only if every adjacent pair satisfies (2).

On block R_i, the conditional expected user loss is

$$\mathbb{E}_{\pi_w}\Big[(\Phi(R_i) - (w_\ell^*(D) + c))^2 \,\Big|\, w_\ell^*(D) \in R_i\Big] + \kappa$$
$$= \Psi(R_i) - \Phi(R_i)^2 + c^2 + \kappa.$$

After weighting by $p(R_i)$ and summing over the blocks, the user's expected loss is

$$\sum_{i=1}^{k} p(R_i)\left(\Psi(R_i) - \Phi(R_i)^2\right) + c^2 + \kappa.$$

Since c^2 is fixed across feasible partitions, a user-optimal finite-support equilibrium minimizes

$$\sum_{i=1}^{k} p(R_i)\left(\Psi(R_i) - \Phi(R_i)^2\right) + \kappa$$

over feasible contiguous-block partitions. □

Corollary 4.3.1 Runtime. There are $O(m^2)$ contiguous blocks. After computing prefix sums of the prior probabilities, first moments, and second moments, the quantities $p(R)$, $\Phi(R)$, and $\Psi(R)$ can be evaluated using Algorithm 1 for any block R. There are $O(m^3)$ pairs of adjacent blocks

$$R = \{w^{(r)}, \ldots, w^{(s)}\}, \qquad R' = \{w^{(s+1)}, \ldots, w^{(u)}\},$$

and each pair can be checked using (2) in constant time.

Construct a directed acyclic graph whose nodes are contiguous blocks and whose edges connect feasible adjacent blocks. Assign each block R the cost

$$p(R)\big(\Psi(R) - \Phi(R)^2\big).$$

A shortest path from a block beginning at $w^{(1)}$ to a block ending at $w^{(m)}$ gives a feasible partition minimizing the user-optimal objective. The graph has $O(m^2)$ nodes and $O(m^3)$ edges, so the exact dynamic program runs in $O(m^3)$ time.

Remark D.1. *For finite priors, the dynamic program gives an exact certificate of optimality. For general continuous priors, there is no distribution-free guarantee without specifying a global solver for Eq.* (1)*. If the solver only finds local or approximate solutions, then the returned report is optimal only over the feasible solutions found, up to the solver tolerance.*

Complexity Intuition The optimal information-release problem is to choose an equilibrium partition $\mathcal{R} = \{R_1, \ldots, R_k\}$ that maximizes (4).

For a finite prior supported on $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n\}$, a **brute-force algorithm** enumerates all partitions of the support into nonempty report regions. For each candidate partition, it computes the prior mass and posterior mean $\Phi(R_i)$ of every region, verifies that each support point assigned to R_i weakly prefers $\Phi(R_i)$ to every other learner response, and evaluates (4). It returns a feasible partition attaining the largest objective value. Given a private dataset D, the user selects the report M_i associated with the region R_{M_i} containing $\boldsymbol{w}_\ell^*(D)$. There are B_n partitions of n support points, where B_n is the n-th Bell number. Checking one partition requires at most $O(n^2 d)$ time, so the brute-force algorithm runs in $O(B_n n^2 d)$ time. It is therefore exact but super-exponential in the support size.

For a continuous prior, the same procedure is only conceptual because there are infinitely many possible partitions. Moreover, even for a fixed region, exact computation of its prior mass and posterior mean can itself be intractable. The following result contrasts with the single-dimensional uniform case (Corollary 4.3.1), where the oracle has a closed-form implementation.

Proposition D.2. *Suppose $\pi_{\boldsymbol{w}}$ is uniform on $[0,1]^d$, where d is part of the input, and let $R \subseteq [0,1]^d$ be a positive-volume convex polytope given in halfspace representation. Exact evaluation of* $\mathrm{Oracle}(\pi_{\boldsymbol{w}}, R)$*, which returns the prior mass and posterior mean of R, is #P-hard.*

Proof of Proposition D.2. Under the uniform prior on $[0,1]^d$, the density is 1. Hence the prior mass of a region $R \subseteq [0,1]^d$ is

$$p(R) = \Pr_{\pi_{\boldsymbol{w}}}(\boldsymbol{w} \in R) = \int_R 1 \, d\boldsymbol{w} = \mathrm{vol}(R).$$

Algorithm 2 User-Optimal Report in One Dimension

Input: prior π_w, conflict $c > 0$, private dataset D, statistics oracle, global constrained solver
Output: user-optimal report $M^*(D)$
if π_w has finite support of size m **then**
  Compute an optimal block-partition $\mathcal{R}^*$ using Corollary 4.3.1
else
  $N_{\max} \leftarrow \lfloor 1 + (\bar{w} - \underline{w})/(2c) \rfloor$
  $J^* \leftarrow +\infty$, $\boldsymbol{t}^* \leftarrow \varnothing$
  for $n = 1, \ldots, N_{\max}$ **do**
    Solve the constrained optimization problem

$$\begin{aligned} \min_{t_1,\ldots,t_{n-1}} \quad & \sum_{i=1}^{n} p(t_{i-1}, t_i) \left(\Psi(t_{i-1}, t_i) - \Phi(t_{i-1}, t_i)^2 \right) + \kappa \\ \text{s.t.} \quad & t_i = \frac{\Phi(t_{i-1}, t_i) + \Phi(t_i, t_{i+1})}{2} - c, i = 1, ., n-1, \\ & \underline{w} = t_0 < t_1 < \cdots < t_n = \bar{w}. \end{aligned}$$

    if the problem is feasible and its objective value $J_n < J^*$ **then**
      $J^* \leftarrow J_n$
      $\boldsymbol{t}^* \leftarrow (t_0, \ldots, t_n)$
    end if
  end for
  Let $\mathcal{R}^*$ be the partition defined by $\boldsymbol{t}^*$
end if
Compute $w_\ell^*(D)$ ▷ Proposition B.1
Find the region $R_i^* \in \mathcal{R}^*$ containing $w_\ell^*(D)$
return the report M_i associated with R_i^*

Thus, exact oracle evaluation contains, as a special case, exact computation of the volume of a convex polytope given in halfspace representation. Exact polytope volume computation is $\#P$-hard [11]. Therefore exact evaluation of $\mathrm{Oracle}(\pi_{\boldsymbol{w}}, R)$ is #P-hard. □

Theorem D.3. *For $d \geq 3$, the user's optimal information-release problem under quadratic utilities is NP-hard. Specifically, fix $\boldsymbol{c} = (0, 0, 1, 0, \ldots, 0)$ and let $\pi_{\boldsymbol{w}}$ be a finite uniform prior over rational points in $\mathbb{R}^d$. It is NP-hard to decide whether there exists an equilibrium partition into at most k reports with expected user loss at most B.*

Proof of Theorem D.3. It suffices to prove hardness for the admissible special case $\nu = \delta_{\mathbf{0}}$, for which $\kappa = 0$. We reduce from planar k-means, which is NP-hard [22]. An instance consists of points $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in \mathbb{Q}^2$, an integer k, and a threshold $L \in \mathbb{Q}$. The question is whether there exists a partition

$$\mathcal{R} = \{R_1, \ldots, R_m\}, \qquad m \leq k,$$

such that

$$\mathrm{cost}(\mathcal{R}) := \sum_{\alpha=1}^{m} \sum_{i \in R_\alpha} \|\boldsymbol{y}_i - \bar{\boldsymbol{y}}_\alpha\|_2^2 \leq L, \qquad \bar{\boldsymbol{y}}_\alpha = \frac{1}{|R_\alpha|} \sum_{i \in R_\alpha} \boldsymbol{y}_i.$$

Construction. Set $d = 3$ and $\boldsymbol{c} = (0, 0, 1)$. For each point $\boldsymbol{y}_i \in \mathbb{Q}^2$, define

$$\boldsymbol{w}_i = (\boldsymbol{y}_i, 0) \in \mathbb{Q}^3.$$

Let $\pi_{\boldsymbol{w}}$ be the uniform prior over $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n\}$, and set $B = \frac{L+n}{n}$. The learner has quadratic utility $\widetilde{U}^\ell(\boldsymbol{w}_i, \boldsymbol{w}) = -\|\boldsymbol{w} - \boldsymbol{w}_i\|_2^2$, so after region R_α, the learner best-responds with the posterior mean

$$\Phi(R_\alpha) = \frac{1}{|R_\alpha|} \sum_{\boldsymbol{w}_i \in R_\alpha} \boldsymbol{w}_i = (\bar{\boldsymbol{y}}_\alpha, 0).$$

The user has quadratic utility $\widetilde{U}^r(\boldsymbol{w}_i, \boldsymbol{w}) = -\|\boldsymbol{w} - (\boldsymbol{w}_i + \boldsymbol{c})\|_2^2$.

Loss identity. For any partition $\mathcal{R} = \{R_1, \dots, R_m\}$ of $\{\boldsymbol{w}_1, \dots, \boldsymbol{w}_n\}$, if $\boldsymbol{w}_i \in R_\alpha$, then the user's loss from the learner response $\Phi(R_\alpha)$ is

$$\|\Phi(R_\alpha) - (\boldsymbol{w}_i + \boldsymbol{c})\|_2^2 = \|(\bar{\boldsymbol{y}}_\alpha, 0) - (\boldsymbol{y}_i, 1)\|_2^2 = \|\bar{\boldsymbol{y}}_\alpha - \boldsymbol{y}_i\|_2^2 + 1.$$

Therefore the expected user loss is

$$\mathrm{Loss}(\mathcal{R}) = \frac{1}{n} \sum_{\alpha=1}^{m} \sum_{\boldsymbol{w}_i \in R_\alpha} \|\Phi(R_\alpha) - (\boldsymbol{w}_i + \boldsymbol{c})\|_2^2 = \frac{\mathrm{cost}(\mathcal{R}) + n}{n}.$$

Equilibrium condition. A partition $\mathcal{R}$ is an equilibrium if and only if no user model prefers the report associated with another region. Thus, for every $\boldsymbol{w}_i \in R_\alpha$ and every region R_β,

$$\|\Phi(R_\alpha) - (\boldsymbol{w}_i + \boldsymbol{c})\|_2^2 \le \|\Phi(R_\beta) - (\boldsymbol{w}_i + \boldsymbol{c})\|_2^2.$$

Substituting the construction gives

$$\|\bar{\boldsymbol{y}}_\alpha - \boldsymbol{y}_i\|_2^2 \le \|\bar{\boldsymbol{y}}_\beta - \boldsymbol{y}_i\|_2^2,$$

which is exactly the nearest-centroid condition for k-means.

If the planar k-means instance is a yes-instance, then there exists a partition with cost at most L. Starting from such a partition, repeatedly reassign any point to a nearest current centroid whenever this strictly decreases the k-means objective, recomputing centroids after each reassignment. This process terminates at a nearest-centroid partition with cost at most L. Lifting this partition to $\mathbb{R}^3$ gives an equilibrium partition $\mathcal{R}$ with at most k regions and

$$\mathrm{Loss}(\mathcal{R}) = \frac{\mathrm{cost}(\mathcal{R}) + n}{n} \le \frac{L + n}{n} = B.$$

Conversely, if there exists an equilibrium partition $\mathcal{R}$ with at most k regions and $\mathrm{Loss}(\mathcal{R}) \le B$, then projecting each cell onto the first two coordinates gives a planar clustering $\mathcal{R}$ with at most k clusters and

$$\mathrm{cost}(\mathcal{R}) = n \cdot \mathrm{Loss}(\mathcal{R}) - n \le nB - n = L.$$

Thus the planar k-means instance is a yes-instance.

The reduction is polynomial, so the information-release problem is NP-hard for $d = 3$. For $d > 3$, append zero coordinates to all $\boldsymbol{w}_i$ and set

$$\boldsymbol{c} = (0, 0, 1, 0, \dots, 0).$$

The same argument applies unchanged. □

D.2 Missing Proofs of Section 4.2.1

We provide an example of a factorized prior.

Example D.2. *Consider the HFT problem with two learned models, the HFT system's sensitivity to a tranche's visible size and to its submission frequency. Let the private dataset D result in $\boldsymbol{w}_\ell^*(D) = (0.6, 0.4)$, and let the user's conflict vector be $\boldsymbol{c} = (0.3, 0.1)$. Then the user's ideal hypothesis is $\boldsymbol{w}_\ell^*(D) + \boldsymbol{c} = (0.9, 0.5)$, representing less predatory execution terms, and the conflict direction is $\hat{\boldsymbol{c}} = \boldsymbol{c}/\|\boldsymbol{c}\|_2 = (3/\sqrt{10}, 1/\sqrt{10})$. This represents the particular combination of model changes over which the user and learner disagree: the desired change in favorability attributed to the size-sensitivity coefficient is three times the desired increase in the frequency-sensitivity coefficient. The scalar conflict coordinate is $s = \hat{\boldsymbol{c}}^\top \boldsymbol{w}_\ell^*(D) = 2.2/\sqrt{10}$, and the orthogonal agreement component is $\boldsymbol{z} = \boldsymbol{w}_\ell^*(D) - s\hat{\boldsymbol{c}} = (-0.06, 0.18)$. Thus s measures the component of the learned hypothesis in the user's preferred direction, while $\boldsymbol{z}$ records the remaining information.*

Proof of Theorem 4.4. Let $\beta := \|\boldsymbol{c}\|_2$ and $\hat{\boldsymbol{c}} := \boldsymbol{c}/\|\boldsymbol{c}\|_2$, so that $\boldsymbol{c} = \beta\hat{\boldsymbol{c}}$. Every model has the orthogonal decomposition

$$\boldsymbol{w} = s\hat{\boldsymbol{c}} + \boldsymbol{z}, \qquad s = \hat{\boldsymbol{c}}^\top \boldsymbol{w}, \qquad \boldsymbol{z} = \boldsymbol{w} - s\hat{\boldsymbol{c}},$$

where $\boldsymbol{z} \perp \hat{\boldsymbol{c}}$. By assumption, $\pi_{\boldsymbol{w}} = \pi_s \otimes \pi_{\boldsymbol{z}}$.

Let $R_i = [t_{i-1}, t_i]$, $i = 1, \ldots, n$, be a one-dimensional equilibrium partition for s with conflict β, with boundary points assigned according to the fixed tie-breaking rule. Let

$$\Phi_i := \mathbb{E}_{\pi_s}[s \mid s \in R_i]$$

be the corresponding scalar learner response. For each i and agreement component $\boldsymbol{z}$, choose a report $M_{i,\boldsymbol{z}}$ and let the user select $M_{i,\boldsymbol{z}}$ whenever $\boldsymbol{w}_\ell^*(D) = s\hat{\boldsymbol{c}} + \boldsymbol{z}$ with $s \in R_i$.

Because the report reveals $\boldsymbol{z}$ and restricts s to R_i, independence implies that a regular conditional posterior satisfies, for $\pi_{\boldsymbol{z}}$-almost every $\boldsymbol{z}$,

$$\mathbb{E}_{\pi_{\boldsymbol{w}}}[\boldsymbol{w} \mid M_{i,\boldsymbol{z}}] = \Phi_i\hat{\boldsymbol{c}} + \boldsymbol{z}.$$

By Lemma C.1, the learner therefore responds to $M_{i,\boldsymbol{z}}$ with $\Phi_i\hat{\boldsymbol{c}} + \boldsymbol{z}$. Any remaining report may be assigned a learner response already specified above, so it adds no new response.

We next verify the user's strategy. Fix a private dataset D with $\boldsymbol{w}_\ell^*(D) = s\hat{\boldsymbol{c}} + \boldsymbol{z}$ and $s \in R_i$. The user's preferred model is

$$\boldsymbol{w}_\ell^*(D) + \boldsymbol{c} = (s + \beta)\hat{\boldsymbol{c}} + \boldsymbol{z}.$$

By Lemma C.1, the common learning-error term cancels from comparisons between reports. Under the prescribed report $M_{i,\boldsymbol{z}}$, the relevant squared loss is

$$\left\|\Phi_i\hat{\boldsymbol{c}} + \boldsymbol{z} - \big((s + \beta)\hat{\boldsymbol{c}} + \boldsymbol{z}\big)\right\|_2^2 = (\Phi_i - s - \beta)^2.$$

Consider a deviation to $M_{j,\boldsymbol{z}'}$. The learner then responds with $\Phi_j\hat{\boldsymbol{c}} + \boldsymbol{z}'$. Since $\boldsymbol{z}' - \boldsymbol{z} \perp \hat{\boldsymbol{c}}$,

$$\begin{aligned}&\left\|\Phi_j\hat{\boldsymbol{c}} + \boldsymbol{z}' - \big((s + \beta)\hat{\boldsymbol{c}} + \boldsymbol{z}\big)\right\|_2^2 \\ &\quad = (\Phi_j - s - \beta)^2 + \|\boldsymbol{z}' - \boldsymbol{z}\|_2^2.\end{aligned}$$

For a fixed j, this expression is minimized at $\boldsymbol{z}' = \boldsymbol{z}$. Moreover, because $\{R_i\}_{i=1}^n$ is a scalar equilibrium partition,

$$(\Phi_i - s - \beta)^2 \le (\Phi_j - s - \beta)^2 \qquad \text{for every } j.$$

Consequently,

$$(\Phi_i - s - \beta)^2 \le (\Phi_j - s - \beta)^2 + \|\boldsymbol{z}' - \boldsymbol{z}\|_2^2$$

for every report $M_{j,\boldsymbol{z}'}$. Thus, the user cannot benefit by changing either the conflict-coordinate report or the revealed agreement component. Together with the learner's posterior-mean responses, this proves that the lifted strategies form an equilibrium.

It remains to establish optimality within the lifted class. Consider any equilibrium that reveals $\boldsymbol{z}$ exactly and uses a scalar equilibrium response rule $\phi(s)$ along $\hat{\boldsymbol{c}}$. Its learner response has the form $\phi(s)\hat{\boldsymbol{c}} + \boldsymbol{z}$, and its user loss is

$$\begin{aligned}&\left\|\phi(s)\hat{\boldsymbol{c}} + \boldsymbol{z} - \big((s + \beta)\hat{\boldsymbol{c}} + \boldsymbol{z}\big)\right\|_2^2 + \kappa \\ &\quad = (\phi(s) - s - \beta)^2 + \kappa.\end{aligned}$$

Hence, its expected multidimensional user utility is the same as that of its underlying one-dimensional equilibrium, up to the common learning-error term. Therefore, if $\{R_i\}_{i=1}^n$ is user-optimal among one-dimensional equilibria, its lift is user-optimal among equilibria that reveal $\boldsymbol{z}$ exactly and modify information only along $\hat{\boldsymbol{c}}$. □

Approximate Equilibrium. The factorization assumption is used to construct the lifted strategy. When it does not hold, we can evaluate the returned strategy under any finite true prior using a computable incentive-compatibility (IC) error that measures the largest profitable deviation available to any model within the returned report space.

Proposition D.4. *For finite support $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n\}$, let $\mathcal{M}_{\mathrm{ret}} = \{M_j : j \in [n]\}$ be the returned reports (duplicates identified), M_j the report assigned to $\boldsymbol{w}_j$, $\boldsymbol{u}_j := \boldsymbol{w}_j + \boldsymbol{c}$ its ideal, and $\Phi(R_M) = \mathbb{E}_{\pi_{\boldsymbol{w}}}[\boldsymbol{w} \mid \boldsymbol{w} \in R_M]$ the response to M. Then the returned strategy is an $\varepsilon_{\mathrm{IC}}$-equilibrium of the returned reporting game on the support, with the $O(n^2 d)$-time computable certificate*

$$\varepsilon_{\mathrm{IC}} = \max_j \max_{M \in \mathcal{M}_{\mathrm{ret}}} \big[\|\Phi(R_{M_j}) - \boldsymbol{u}_j\|_2^2 - \|\Phi(R_M) - \boldsymbol{u}_j\|_2^2\big]_+.$$

Because the certificate ranges only over deviations within the returned report space $\mathcal{M}_{\mathrm{ret}}$, $\varepsilon_{\mathrm{IC}} = 0$ certifies that the strategy is an exact equilibrium of the returned game under the true prior, even when factorization fails. More generally, $\varepsilon_{\mathrm{IC}} \le \varepsilon$ certifies an ε-equilibrium of that game.

Proof of Proposition D.4. By Lemma C.1, the learner responds to each returned report M with the posterior mean $\Phi(R_M)$, and learning error does not affect comparisons between reports. It therefore remains to verify the user's incentives.

Fix a support point $\boldsymbol{w}_j$ whose prescribed report is M_j. With the learner's strategy fixed, the gain from selecting any returned report $M \in \mathcal{M}_{\mathrm{ret}}$ instead of M_j is

$$\begin{aligned}&\left\|\Phi(R_{M_j}) - (\boldsymbol{w}_j + \boldsymbol{c})\right\|_2^2 \\ &\quad - \left\|\Phi(R_M) - (\boldsymbol{w}_j + \boldsymbol{c})\right\|_2^2 .\end{aligned}$$

The positive part of the largest such gain is the inner maximum in the definition of $\varepsilon_{\mathrm{IC}}$, and the outer maximum ranges over all support points. Hence, no user can gain more than $\varepsilon_{\mathrm{IC}}$ by switching to another report in $\mathcal{M}_{\mathrm{ret}}$. The returned strategy is therefore an $\varepsilon_{\mathrm{IC}}$-equilibrium of the returned reporting game on the support.

The regions $\{R_M : M \in \mathcal{M}_{\mathrm{ret}}\}$ partition the support. Their posterior means can be computed in one pass over the n support points in $O(nd)$ time. The certificate then evaluates at most $n|\mathcal{M}_{\mathrm{ret}}| \le n^2$ squared-distance differences, each in $O(d)$ time. Thus, its total computation time is $O(n^2 d)$.

To conclude, suppose the factorization assumption of Theorem 4.4 holds and the returned strategy is its lift. Write $\boldsymbol{w}_j = s_j \hat{\boldsymbol{c}} + \boldsymbol{z}_j$, and suppose $s_j \in R_i$. Then

$$\Phi(R_{M_j}) = \Phi_i \hat{\boldsymbol{c}} + \boldsymbol{z}_j, \qquad \Phi_i := \mathbb{E}[s \mid s \in R_i].$$

Any returned report has the form $M_{\ell,\boldsymbol{z}'}$, with learner response $\Phi_\ell \hat{\boldsymbol{c}} + \boldsymbol{z}'$. Therefore,

$$\left\|\Phi(R_{M_j}) - (\boldsymbol{w}_j + \boldsymbol{c})\right\|_2^2 = (\Phi_i - s_j - \beta)^2,$$

whereas

$$\begin{aligned}&\left\|\Phi_\ell \hat{\boldsymbol{c}} + \boldsymbol{z}' - (\boldsymbol{w}_j + \boldsymbol{c})\right\|_2^2 \\ &\quad = (\Phi_\ell - s_j - \beta)^2 + \|\boldsymbol{z}' - \boldsymbol{z}_j\|_2^2.\end{aligned}$$

Because $\{R_i\}$ is a one-dimensional equilibrium partition,

$$(\Phi_i - s_j - \beta)^2 \le (\Phi_\ell - s_j - \beta)^2$$

for every ℓ, and the orthogonal term is nonnegative. Hence, every deviation gain is nonpositive, so $\varepsilon_{\mathrm{IC}} = 0$, consistent with Theorem 4.4. □

E Data-Dependent Error in Learning

So far, we have investigated data-independent error in learning (Section 2). We now investigate settings where the learner's error depends on the dataset. In HFT, the user may choose a dataset report in order to convince the learner to learn a more favorable model of its intended trades. However, the learner may not train on exactly the dataset report that the user intended. Thus, error in learning can make the learner behave as if it had trained on a perturbed dataset. We assume data-dependent error in learning can be viewed equivalently as perfect learning but from a *noisy* dataset. In this section, we investigate these settings and propose algorithms for computing the user's optimal dataset report.

Framework under Data-Dependent Error. To investigate strategic information release under data-dependent error, we extend the framework of Section 2. The user sends a dataset report $M \in \mathcal{M}$, *but the learner may use a* ***noisy interpretation*** $\tilde{M} \in \widetilde{\mathcal{M}}$. Since a report matters only through the model it induces, we work with the reduced report $\boldsymbol{w}_\ell^*(M)$, and take $\widetilde{\mathcal{M}}$ to be the space of perturbed models; a natural example is an additive zero-mean Gaussian perturbation. More generally, conditional on M, the learner receives $\tilde{M}$ according to an interpretation rule $G(\cdot \mid M)$ with density $g(\tilde{M} \mid M)$.

The key difference from the data-independent error framework is that the learner no longer conditions on the exact report M. It conditions only on $\tilde{M}$, and therefore cannot, in general, distinguish

M from other reports M' for which $g(\tilde{M} \mid M') > 0$. Two consequences follow. First, Bayes' rule no longer reduces to conditioning on an event, since every dataset is consistent with every interpretation; the learner's posterior reweights the prior by the likelihood of what it observed, and under strictly positive noise every $\tilde{M}$ is on path, so beliefs are pinned down everywhere. Second, the user faces uncertainty even though its strategy is deterministic, because $G(\cdot \mid M)$ randomizes what the learner sees. For a fixed private dataset D, the user's expected utility from sending report M is $\mathbb{E}_{\tilde{M}\sim G(\cdot|M)}\big[U^r(D, \sigma_\ell(\tilde{M}))\big]$, and for a fixed noisy interpretation $\tilde{M}$, the learner's expected utility from choosing hypothesis h is $\mathbb{E}_{D\sim\pi(\cdot|\tilde{M})}\big[U^\ell(D, h)\big]$. The definition of equilibrium in this setting follows after updating the expected utilities in Definition 2.1, with the learner's strategy now a mapping $\sigma_\ell : \widetilde{\mathcal{M}} \to \mathcal{H}$.

E.1 Single Dimension

We retain the utility functions U^ℓ, U^r and an arbitrary prior π_w over $w^*_\ell(D)$ from Section 4.1. The user submits a dataset report $M \in \mathcal{M}$, but the learner does not observe M exactly. Instead, the learner observes a noisy interpretation $\widetilde{M} \in \widetilde{\mathcal{M}}$.

Noisy statistics oracle. We reformulate to model-reduced strategies. With a slight abuse of notation, we continue to write σ_r for the user strategy on the scalar model space,

$$\sigma_r : [0, 1] \to [0, 1].$$

Formally, $\sigma_r(w) = M$ means that $\sigma_r(D) = M$ for datasets satisfying $w^*_\ell(D) = w$. Thus, $\sigma_r(w)$ denotes the report sent by user whose learned model is w. The learner's posterior-mean response after observing $\widetilde{M}$ is

$$\Phi(\widetilde{M}; \sigma_r) = \frac{\int_0^1 w\, G(\widetilde{M} \mid \sigma_r(w))\, d\pi_w(w)}{\int_0^1 G(\widetilde{M} \mid \sigma_r(w))\, d\pi_w(w)}. \tag{7}$$

When π_w has finite support

$$\mathrm{supp}(\pi_w) = \{w^{(1)} < \cdots < w^{(n)}\}, \qquad \pi_i := \pi_w(w^{(i)}),$$

we write the **report tuple** as

$$\mathbf{M} = (M_1, \ldots, M_n), \qquad M_i = \sigma_r(w^{(i)}).$$

Then

$$\Phi(\widetilde{M}; \mathbf{M}) = \frac{\sum_{i=1}^n \pi_i w^{(i)} G(\widetilde{M} \mid M_i)}{\sum_{i=1}^n \pi_i G(\widetilde{M} \mid M_i)}. \tag{8}$$

Error Distribution. We take $\mathcal{M} = [0, 1]$ and use the additive Gaussian rule

$$G(\widetilde{M} \mid M) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\widetilde{M} - M)^2}{2\sigma^2}\right), \tag{9}$$

equivalently,

$$\widetilde{M} = M + \zeta, \qquad \zeta \sim \mathcal{N}(0, \sigma^2).$$

Any compact report interval can be reduced to $[0, 1]$ by an affine rescaling of the reports and σ. Since G is strictly positive, every noisy interpretation has positive density under every reporting strategy. We use a Gaussian distribution because it is smooth, strictly positive on $\mathbb{R}$, and satisfies a strict monotone likelihood-ratio property. Similar arguments may extend to additive error distribution with a log-concave density [28] satisfying the strict monotone likelihood-ratio property [23].

Equilibrium under data-dependent error. For a fixed reporting strategy σ_r, the user's expected utility from sending a candidate report M when the true model is w is

$$V_w(M; \sigma_r) = -\int_{\mathbb{R}} \Big(w + c - \Phi(\widetilde{M}; \sigma_r)\Big)^2 G(\widetilde{M} \mid M)\, d\widetilde{M}. \tag{10}$$

For finite support, define

$$V_i(M;\mathbf{M}) = -\int_{\mathbb{R}} \left(w^{(i)} + c - \Phi(\widetilde{M};\mathbf{M})\right)^2 G(\widetilde{M} \mid M)\, d\widetilde{M}. \tag{11}$$

Expanding the square in (10) gives

$$V_w(M;\sigma_r) = -(w+c)^2 + 2(w+c)g_{\sigma_r}(M) - h_{\sigma_r}(M), \tag{12}$$

where

$$g_{\sigma_r}(M) = \mathbb{E}\Big[\Phi(\widetilde{M};\sigma_r) \mid M\Big], h_{\sigma_r}(M) = \mathbb{E}\Big[\Phi(\widetilde{M};\sigma_r)^2 \mid M\Big].$$

Here the conditional expectation is taken over $\widetilde{M} \sim G(\cdot \mid M)$. Thus, the private model w affects the deviation payoff only through the linear term $2(w+c)g_{\sigma_r}(M)$. Based on this, we get that a user's reporting strategy σ_r is an equilibrium if

$$V_w(\sigma_r(w);\sigma_r) \geq V_w(M;\sigma_r), \qquad \pi_w\text{-a.e. } w, \quad \forall M \in [0,1]. \tag{13}$$

For finite support, condition (13) becomes

$$V_i(M_i;\mathbf{M}) \geq V_i(M;\mathbf{M}), \qquad i \in [n], \quad \forall M \in [0,1]. \tag{14}$$

Monotone equilibria. Without further structure, (13) is a continuum of inequalities over all model-report pairs. Moreover, $\Phi(\cdot;\sigma_r)$, g_{σ_r}, and h_{σ_r} depend on the entire reporting strategy. We therefore restrict attention to monotone equilibria, as in the data-independent error framework.

Call σ_r monotone if it is nondecreasing and influential if it is not π_w-almost surely constant. Under Gaussian location error, a monotone influential strategy induces a posterior mean that is strictly increasing in the noisy interpretation.

Lemma E.1. *If σ_r is nondecreasing and not π_w-almost surely constant, then $\Phi(\widetilde{M};\sigma_r)$ is strictly increasing in $\widetilde{M}$.*

Proof. For $\widetilde{M}' > \widetilde{M}$, the Gaussian likelihood ratio

$$\frac{G(\widetilde{M}' \mid M)}{G(\widetilde{M} \mid M)}$$

is strictly increasing in the intended report M. Because $\sigma_r(w)$ is nondecreasing in w, a larger noisy interpretation shifts the posterior distribution of w upward in the monotone-likelihood-ratio order and hence in first-order stochastic dominance. Since σ_r is not almost surely constant, the ordering is strict on a positive-probability set. Therefore, the posterior mean $\Phi(\widetilde{M};\sigma_r)$ is strictly increasing in $\widetilde{M}$. □

By (12), the deviation payoff has increasing differences in (w, M): higher models have relatively stronger incentives to send higher reports. This produces the ordered equilibrium structure characterized next.

Structure of equilibria.

Proposition E.2. *Let σ_r be a monotone influential noisy equilibrium strategy. There exist $0 \leq w_0 \leq w_1 \leq 1$ such that*

$$\sigma_r(w) = 0 \quad \textit{for } w < w_0, \qquad \sigma_r(w) = 1 \quad \textit{for } w > w_1,$$

and σ_r is strictly increasing on (w_0, w_1). Equivalently, if two distinct models send a common report M, then $M \in \{0,1\}$. With finite support, the corresponding tuple satisfies

$$\begin{aligned} M_1 = \cdots = M_t = 0 < M_{t+1} < \cdots < M_{n-t'} \\ < M_{n-t'+1} = \cdots = M_n = 1 \end{aligned} \tag{15}$$

for integers $t, t' \geq 0$ with $t + t' \leq n$, where the middle chain is omitted when $t + t' = n$.

Proof of Proposition E.2. By Lemma E.1, $\Phi(\widetilde{M};\sigma_r)$ is strictly increasing in $\widetilde{M}$.

Suppose two models $w < w'$ send the same interior report $M_0 \in (0,1)$. Since M_0 is interior, local optimality requires the derivative of each model's deviation utility to vanish there.

Using the Gaussian location identity

$$\frac{\partial}{\partial M'} G(\widetilde{M} \mid M') = -\frac{\partial}{\partial \widetilde{M}} G(\widetilde{M} \mid M')$$

and integration by parts gives

$$\begin{aligned} \frac{\partial}{\partial M'} V_w(M';\sigma_r)\Big|_{M'=M_0} &= 2\int_{\mathbb{R}} \Big(w + c - \Phi(\widetilde{M};\sigma_r)\Big) \\ &\qquad \cdot \frac{\partial}{\partial \widetilde{M}} \Phi(\widetilde{M};\sigma_r) G(\widetilde{M} \mid M_0)\, d\widetilde{M}. \end{aligned} \tag{16}$$

where the boundary term vanishes because the integrand is bounded.

Subtracting the derivative for w from the derivative for w' yields

$$\begin{aligned} &\frac{\partial}{\partial M'} V_{w'}(M';\sigma_r)\Big|_{M'=M_0} - \frac{\partial}{\partial M'} V_w(M';\sigma_r)\Big|_{M'=M_0} \\ &\quad = 2(w'-w)\int_{\mathbb{R}} \frac{\partial}{\partial \widetilde{M}} \Phi(\widetilde{M};\sigma_r) G(\widetilde{M} \mid M_0)\, d\widetilde{M} > 0. \end{aligned}$$

The strict inequality follows from Lemma E.1 and the strict positivity of G. Both derivatives therefore cannot vanish at the same interior report, a contradiction.

Hence, distinct models can pool only at the boundary reports 0 and 1. Monotonicity places the lower models in the lower pool, the upper models in the upper pool, and makes the strategy strictly increasing between them. □

Proposition E.2 says that an interior common report would have to be locally optimal for two different models. However, a higher model always has a strictly stronger marginal incentive to increase its report. Two distinct models therefore cannot both be locally satisfied at the same interior report. Pooling remains possible only at the endpoints, where deviations are one-sided.

Example E.1. *Let $\{w^{(1)}, w^{(2)}\} = \{0.8, 0.9\}$ with $\pi_1 = \pi_2 = \frac{1}{2}$. In the data-independent error framework, if both models send a pooling report M_{pool} with $R_{M_{\text{pool}}} = \{0.8, 0.9\}$, the learner responds with $\mathbb{E}[w_\ell^*(D) \mid w_\ell^*(D) \in \{0.8, 0.9\}] = 0.85$.*

Now let $c = 0.03$ and $\sigma = 0.2$. For the noisy endpoint pattern $\mathbf{M} = (m, 1)$, solving

$$\frac{\partial}{\partial M'} V_1(M';\mathbf{M})\Big|_{M'=m} = 0$$

gives $m \approx 0.2823$. Hence $\mathbf{M}^* = (0.2823, 1)$, so the user reports 0.2823 when $w_\ell^*(D) = 0.8$ and 1 when $w_\ell^*(D) = 0.9$. In the former case, the learner observes $\widetilde{M} = 0.2823 + \zeta$, where $\zeta \sim \mathcal{N}(0, 0.2^2)$, rather than the intended report.

Exact local characterization. Condition (13) is a continuum of global inequalities. Proposition E.2 reduces the possible form of a monotone influential equilibrium. We now show that, within this class, global incentive compatibility can be characterized by first-order conditions.

Define

$$z(w;\sigma_r) := \frac{\partial}{\partial M'} V_w(M';\sigma_r)\Big|_{M'=\sigma_r(w)}. \tag{17}$$

For finite support, define

$$z_i(\mathbf{M}) := \frac{\partial}{\partial M'} V_i(M';\mathbf{M})\Big|_{M'=M_i}. \tag{18}$$

Proposition E.3. *Let σ_r be monotone and influential with thresholds $w_0 \le w_1$. Then σ_r satisfies* (13) *if and only if*

$$z(w;\sigma_r) = 0, \qquad w \in (w_0, w_1), \tag{19}$$
$$z(w_0;\sigma_r) \le 0, \qquad \text{if } w_0 > 0, \tag{20}$$
$$z(w_1;\sigma_r) \ge 0, \qquad \text{if } w_1 < 1. \tag{21}$$

With finite support and endpoint pattern (15)*, these conditions become*

$$z_i(\mathbf{M}) = 0, \qquad i = t+1, \dots, n-t',$$

together with

$$z_t(\mathbf{M}) \le 0 \quad \text{when } t \ge 1, \; z_{n-t'+1}(\mathbf{M}) \ge 0 \quad \text{when } t' \ge 1.$$

Proof of Proposition E.3. Necessity follows from first-order optimality at an interior report and one-sided optimality at a boundary report.

For sufficiency, we invoke Lemma 4 of [5]. For a fixed model, if the learner's response is strictly increasing in the noisy interpretation and the user's pointwise utility is bounded and strictly single-peaked in that response, then expected utility is strictly quasi-concave in the intended report. Consequently, an interior stationary point is the unique global maximizer, while an endpoint is globally optimal when the corresponding one-sided derivative has the appropriate sign.

By Lemma E.1, $\Phi(\cdot;\sigma_r)$ is strictly increasing. Moreover, $-(w+c-\Phi)^2$ is bounded on the learner's response range and strictly single-peaked in Φ, with ideal point $w+c$. Therefore, $V_w(\cdot;\sigma_r)$ is strictly quasi-concave on $[0,1]$.

It remains to verify the models pooled at the endpoints. From (12),

$$\frac{\partial}{\partial M'} V_w(M';\sigma_r)\Big|_{M'=0} = 2(w+c)g'_{\sigma_r}(0) - h'_{\sigma_r}(0),$$

which is strictly increasing in w. Hence, (20) implies the required lower-endpoint inequality for every $w < w_0$. Symmetrically, (21) implies the required upper-endpoint inequality for every $w > w_1$. Thus every model globally prefers its prescribed report. □

By Proposition E.3, a monotone influential strategy satisfying (19)–(21) requires no additional verification against (13).

However, if σ_r is constant, then $\Phi(\widetilde{M};\sigma_r) \equiv \mathbb{E}_{\pi_w}[w]$, so $V_w(M;\sigma_r)$ is constant in M. Every candidate report is therefore a best response, and every constant strategy is an equilibrium. Constant strategies are not covered by Proposition E.3, whose hypothesis requires an influential strategy.

Optimal dataset report. Propositions E.2 and E.3 characterize the monotone influential equilibria. It remains to rank them. The user's ex ante expected utility admits a variance representation.

Lemma E.4. *For any strategy σ_r,*

$$\mathbb{E}[V_w(\sigma_r(w);\sigma_r)] = \mathrm{Var}\big(\Phi(\widetilde{M};\sigma_r)\big) - \mathrm{Var}_{\pi_w}(w) - c^2,$$

where $w \sim \pi_w$, $\widetilde{M} \sim G(\cdot \mid \sigma_r(w))$ and $\mathrm{Var}(.)$ is variance.

Proof. Since $\Phi(\widetilde{M};\sigma_r) = \mathbb{E}[w \mid \widetilde{M}]$, expanding the quadratic loss and using $\mathbb{E}[w - \Phi(\widetilde{M};\sigma_r)] = 0$ gives

$$\mathbb{E}[V_w(\sigma_r(w);\sigma_r)] = -\mathbb{E}[\mathrm{Var}(w \mid \widetilde{M})] - c^2.$$

The result follows from the total variance decomposition. □

Since $\mathrm{Var}_{\pi_w}(w)$ and c^2 do not depend on the strategy, selecting the user-optimal monotone equilibrium is equivalent to maximizing the variance of the learner's posterior mean. The conflict parameter affects which strategies satisfy incentive compatibility, but it does not affect how feasible strategies are ranked.

Algorithm 3 Finite Noisy User-Optimal Dataset Report

Input: private dataset D; conflict c; error variance σ; prior π_w on $\{w^{(1)} < \cdots < w^{(n)}\}$
$\mathbf{M}^* \leftarrow (0, \ldots, 0)$
$U^* \leftarrow \sum_{i=1}^{n} \pi_i V_i(0; \mathbf{M}^*)$ ▷ constant equilibrium
for all $t, t' \geq 0$ such that $t + t' \leq n$ **do**
    Set $M_i = 0$ for $i \leq t$ and $M_i = 1$ for $i \geq n - t' + 1$
    Solve $z_i(\mathbf{M}) = 0$ for $i = t+1, \ldots, n-t'$, subject to

$$0 < M_{t+1} < \cdots < M_{n-t'} < 1$$

    for all returned solutions $\mathbf{M}$ **do**
        if $t = 0$ or $z_t(\mathbf{M}) \leq 0$ **then**
            if $t' = 0$ or $z_{n-t'+1}(\mathbf{M}) \geq 0$ **then**
                $U(\mathbf{M}) \leftarrow \sum_{i=1}^{n} \pi_i V_i(M_i; \mathbf{M})$
                if $U(\mathbf{M}) > U^*$ **then**
                $\mathbf{M}^* \leftarrow \mathbf{M}$
                $U^* \leftarrow U(\mathbf{M})$
                end if
            end if
        end if
    end for
end for
Compute $w_\ell^*(D)$ and find i such that $w_\ell^*(D) = w^{(i)}$
return M_i^*

Theorem E.5. *Let $\mathcal{E}_{\text{error}} := \{\sigma_r : \sigma_r$ is nondecreasing and satisfies* (13)*$\}$. Then $\mathcal{E}_{\text{error}} \neq \varnothing$, and a user-optimal monotone noisy equilibrium solves*

$$\sup_{\sigma_r \in \mathcal{E}_{\text{error}}} \operatorname{Var}\big(\Phi(\widetilde{M}; \sigma_r)\big). \tag{22}$$

For finite support, let $\mathcal{E}_{\text{error}}^{(n)} \subseteq [0,1]^n$ be the corresponding set of nondecreasing equilibrium tuples. Then the supremum is attained and

$$\mathbf{M}^* \in \arg\max_{\mathbf{M} \in \mathcal{E}_{\text{error}}^{(n)}} \sum_{i=1}^{n} \pi_i V_i(M_i; \mathbf{M}). \tag{23}$$

If $w_\ell^*(D) = w^{(i)}$, the user-optimal dataset report is $M^*(D) = M_i^*$.

Proof of Theorem E.5. A constant strategy is nondecreasing and always satisfies (13), so $\mathcal{E}_{\text{error}} \neq \varnothing$. Equivalence of (22) with maximization of ex ante user utility follows from Lemma E.4.

Under finite support, $\mathcal{E}_{\text{error}}^{(n)}$ is a closed subset of the compact box $[0,1]^n$, and the objective in (23) is continuous. Therefore, a maximizer exists. Once $\mathbf{M}^*$ is selected, the private dataset determines which component M_i^* is returned. □

Theorem E.5 shows that the noisy problem is not a finest-partition problem. In the data-independent error framework, each report induces a disjoint posterior region. Under data-dependent error, the interpretations of different reports overlap, so the relevant equilibrium object is the entire user strategy.

Algorithm. For a finite prior, the search exploits the two structural results. By Proposition E.2, a monotone tuple has the endpoint pattern (15), so it is determined by a pair (t, t') and the $q = n - t - t'$ interior reports, leaving $O(n^2)$ patterns. By Proposition E.3, the interior reports of a given pattern solve the square stationarity system $z_i(\mathbf{M}) = 0$, where $z_i(\mathbf{M})$ is the derivative of $V_i(\cdot; \mathbf{M})$ at M_i, subject to the boundary inequalities; together these form a mixed complementarity system [12]. Algorithm 3 enumerates the patterns, retains the ordered solutions satisfying the boundary conditions, and returns the highest-utility solution found. It is initialized at the constant equilibrium, which is always feasible, so it always returns a certified equilibrium.

Complexity. There are $O(n^2)$ endpoint patterns. For a fixed pattern (t, t'), the nonlinear solver handles $q = n - t - t' \le n$ interior reports. Writing $T_{\mathrm{solve}}(q)$ for the cost of solving one fixed-pattern system, the total running time is

$$O\big(n^2 T_{\mathrm{solve}}(n)\big)\,.$$

The factor T_{solve} is not bounded in closed form because the stationarity system is nonlinear in the reports.

Proposition E.6. *For a fixed endpoint pattern (t, t'), the stationarity system need not have a unique solution.*

Proof of Proposition E.6. It suffices to provide an example. Let $n = 2, w^{(1)} = 0, w^{(2)} = 1, \pi_1 = 0.7, \pi_2 = 0.3$, and let

$$c = 0.303, \qquad \sigma = 0.5.$$

Fix the pattern $(t, t') = (0, 1)$, so $M_2 = 1$ and $M_1 = m \in (0, 1)$. Write

$$\mathbf{M}(m) = (m, 1).$$

The stationarity system reduces to $F(m) = 0$, where

$$F(m) = \frac{\partial}{\partial M'} \int_{\mathbb{R}} -\Big(c - \Phi(\widetilde{M}; \mathbf{M}(m))\Big)^2 G(\widetilde{M} \mid M')\, d\widetilde{M}\Big|_{M'=m},$$

and

$$\Phi(\widetilde{M}; \mathbf{M}(m)) = \frac{0.3 G(\widetilde{M} \mid 1)}{0.7 G(\widetilde{M} \mid m) + 0.3 G(\widetilde{M} \mid 1)}.$$

The map F is continuous, and direct evaluation gives

$$F(0.35) > 0, F(0.45) < 0, F(0.75) < 0, F(0.85) > 0.$$

By the intermediate value theorem, F has one root in $(0.35, 0.45)$ and another in $(0.75, 0.85)$. □

Guarantees. The pattern enumeration is exhaustive and, by Proposition E.3, every returned tuple is an exact equilibrium. What is not guaranteed is finding every root of a fixed pattern's stationarity system, since the system is nonlinear in the reports through the ratio-of-Gaussians posterior and a numerical solver returns one root per initialization; an MCP solver such as PATH [12] has the same limitation. Algorithm 3 is therefore sound but not complete: every tuple it returns is a certified equilibrium, while an equilibrium of higher user utility may go undetected, so the output is optimal among the equilibria found. In implementation, we formulate each pattern's stationarity system as a box-constrained nonlinear least-squares problem, use random restarts under a fixed time budget, and retain only ordered solutions satisfying the stationarity and boundary conditions up to numerical tolerance.

Remark E.7. *Lemma E.4, Proposition E.2, Proposition E.3, and Theorem E.5 hold for an arbitrary prior π_w, so the equilibrium characterization and the user-optimal program are valid for continuous priors as well. What is specific to finite support is the algorithm. With a continuum, the endpoint thresholds (w_0, w_1) range over an interval rather than over $O(n^2)$ integer patterns, and* (19) *becomes a condition at every $w \in (w_0, w_1)$, so the finite stationarity system is unavailable and* (22) *must be approximated by discretization.*

E.2 User's Optimal Release for Data-Dependent Learning Error in Multiple Dimensions

We extend the noisy one-dimensional result from Section E.1. We retain the factorized-prior setup of Section 4.2.1, with $\boldsymbol{c} \neq 0$, $\hat{\boldsymbol{c}} = \boldsymbol{c}/\|\boldsymbol{c}\|_2$, $\beta = \|\boldsymbol{c}\|_2$, and $w_\ell^*(D) = s\hat{\boldsymbol{c}} + z$. We additionally assume $\mathrm{supp}(\pi_s) = \{s^{(1)} < \cdots < s^{(n)}\}$ with $\pi_i := \pi_s(s^{(i)}) > 0$, and that data-dependent error acts only on the conflict coordinate. The lifted report space is

$$\mathcal{M}_{\mathrm{lift}} = \{M\hat{\boldsymbol{c}} + \boldsymbol{z} : M \in [0, 1],\ \boldsymbol{z} \in \mathrm{supp}(\pi_{\boldsymbol{z}})\}.$$

If the user sends $M\hat{\boldsymbol{c}} + \boldsymbol{z}$, the learner observes $\widetilde{M}\hat{\boldsymbol{c}} + \boldsymbol{z}$, where $\widetilde{M} = M + \zeta$ and $\zeta \sim \mathcal{N}(0, \sigma^2)$. Thus, the error affects only the conflict coordinate, while $\boldsymbol{z}$ is observed exactly. When $\pi_{\boldsymbol{z}}$ is continuous, conditioning on $\boldsymbol{z}$ is understood through a regular conditional distribution. Let $\mathbf{M} = (M_1, \ldots, M_n)$ be a one-dimensional noisy equilibrium tuple for scalar state s, conflict β, prior π_s, and error level σ. We lift it by assigning $s^{(i)}\hat{\boldsymbol{c}} + \boldsymbol{z}$ the report $M_i\hat{\boldsymbol{c}} + \boldsymbol{z}$.

Proposition E.8. *Under the factorized-prior and conflict-coordinate error assumptions, any one-dimensional noisy equilibrium tuple* $\mathbf{M} = (M_1, \ldots, M_n)$ *lifts to a multidimensional noisy equilibrium on* $\mathcal{M}_{\text{lift}}$. *In particular, if* $\boldsymbol{w}_\ell^*(D) = s^{(i)}\hat{\boldsymbol{c}} + \boldsymbol{z}$, *the user reports* $M_i\hat{\boldsymbol{c}} + \boldsymbol{z}$. *Moreover, if* $\mathbf{M}$ *is user-optimal among the one-dimensional noisy equilibria characterized in Theorem E.5, then its lift is user-optimal among equilibria that reveal* $\boldsymbol{z}$ *exactly and use noisy reports only along* $\hat{\boldsymbol{c}}$.

Proof. Fix $\boldsymbol{w}_\ell^*(D) = s^{(i)}\hat{\boldsymbol{c}} + \boldsymbol{z}$, so the user's ideal hypothesis is $(s^{(i)} + \beta)\hat{\boldsymbol{c}} + \boldsymbol{z}$. Consider a deviation to $M\hat{\boldsymbol{c}} + \boldsymbol{z}'$. The learner observes $\widetilde{M} = M + \zeta$ and $\boldsymbol{z}'$.

Because s and $\boldsymbol{z}$ are independent, observing $\boldsymbol{z}'$ does not change the posterior of s. The learner therefore responds with $\Phi(\widetilde{M}; \mathbf{M})\hat{\boldsymbol{c}} + \boldsymbol{z}'$, where

$$\Phi(\widetilde{M}; \mathbf{M}) = \frac{\sum_{j=1}^n \pi_j s^{(j)} G(\widetilde{M} \mid M_j)}{\sum_{j=1}^n \pi_j G(\widetilde{M} \mid M_j)}.$$

Since $\boldsymbol{z}' - \boldsymbol{z} \perp \hat{\boldsymbol{c}}$, the expected squared loss from the deviation is

$$\mathbb{E}_{\widetilde{M}|M}\Big[\big(\Phi(\widetilde{M}; \mathbf{M}) - s^{(i)} - \beta\big)^2\Big] + \|\boldsymbol{z}' - \boldsymbol{z}\|_2^2. \tag{24}$$

For every M, the second term is minimized by $\boldsymbol{z}' = \boldsymbol{z}$. The remaining deviation problem is exactly the one-dimensional noisy problem for $s^{(i)}$. Since $\mathbf{M}$ is a scalar equilibrium, M_i is optimal against every $M \in [0, 1]$. Hence no multidimensional deviation is profitable. The learner's posterior-mean response is optimal under quadratic utility, proving equilibrium.

For optimality, consider any competing lifted equilibrium. Conditional on a revealed $\boldsymbol{z}$, let $\mathbf{M}(\boldsymbol{z})$ denote its induced scalar tuple. Factorization implies that the conditional prior of s remains π_s, so $\mathbf{M}(\boldsymbol{z})$ is a feasible equilibrium of the same scalar game for $\pi_{\boldsymbol{z}}$-almost every $\boldsymbol{z}$. Scalar optimality of $\mathbf{M}$, followed by integration over $\pi_{\boldsymbol{z}}$, proves optimality within the lifted class. □

Algorithm. Given D, compute $\boldsymbol{w}_\ell^*(D)$, then set $s = \hat{\boldsymbol{c}}^\top \boldsymbol{w}_\ell^*(D)$ and $\boldsymbol{z} = \boldsymbol{w}_\ell^*(D) - s\hat{\boldsymbol{c}}$. Run Algorithm 3 on π_s with conflict $\beta = \|\boldsymbol{c}\|_2$ and error level σ, obtaining $\mathbf{M}^* = (M_1^*, \ldots, M_n^*)$. If $s = s^{(i)}$, return $M^*(D) = M_i^*\hat{\boldsymbol{c}} + \boldsymbol{z}$. The projection and decomposition cost $O(d)$, so the equilibrium search takes $O(n^2 T_{\text{solve}}(n) + d)$.

F Missing Details of Section 5

This appendix supports the three experiments of Section 5. Appendix F.2 reports the scalability and equilibrium-certification results behind Table 1; Appendix F.3 reports the results for the setting of data-dependent error in learning from Appendix E.

F.1 Shared Experiment Setup

Datasets and preprocessing. We use four real-world datasets: Credit Approval [21, 26], School Admission [24, 34], Census [3, 21], and Prosper Loan[3] [25]. Table 2 summarizes sizes, feature dimensions, and prediction tasks. Numerical features are imputed with the training median and standardized; categorical features are imputed with the training mode and one-hot encoded. Preprocessing is fit separately on each train split, so the number of one-hot categories and hence the processed feature dimension varies slightly across splits, which is why several entries below are ranges.

Learners. We evaluate two learner classes. The first is a linear SVM (`LinearSVC`); here $\boldsymbol{w}_\ell^*(D)$ is the learned linear model including the intercept, fit one-vs-rest for the multi-class dataset. The second is a two-layer neural network with architecture $\text{Linear}(p, 32)$–ReLU–$\text{Linear}(32, K_y)$; here $\boldsymbol{w}_\ell^*(D)$ is the flattened vector of all learned parameters. The SVM uses default hyperparameters with a fixed random seed. The network is trained with Adam, learning rate 10^{-3}, batch size 64, and cross-entropy loss, with a 10% validation split and early stopping after 10 epochs without validation-loss improvement. Table 3 reports the resulting parameter dimensions d, which range from 47 to 6443.

[3] `https://www.prosper.com/`

Dataset	# Samples	# Raw Feat.	# Proc. Feat.	# Labels	Prediction Task
Credit Approval	690	15	46	2	Binary: approval
School Admission	22,407	30	54	2	Binary: pass exam
Census	32,561	14	105	2	Binary: income $> \$50{,}000$
Prosper Loan	84,853	64	189	11	Multi-class: loan risk score

Table 2: Datasets used in the experiments. Proc. Feat. is the feature dimension after preprocessing.

Dataset	Linear SVM	MLP
Credit Approval	47	1570
School Admission	52–57	1762–1826
Census	104–106	3426–3458
Prosper Loan	2090	6443

Table 3: Model parameter dimension $d = \dim(\boldsymbol{w}_\ell^*(D))$. For processed feature dimension p and K_y labels, the linear SVM has $p+1$ parameters when $K_y = 2$ and $K_y(p+1)$ under one-vs-rest otherwise, and the network has $32p + 32 + 32K_y + K_y$. Ranges arise from split-dependent one-hot encoding.

Empirical priors. The prior over private datasets induces a prior $\pi_{\boldsymbol{w}}$ over $\boldsymbol{w}_\ell^*(D)$. We approximate $\pi_{\boldsymbol{w}}$ by stratified bootstrap resampling: for each bootstrap dataset we retrain the learner and store the resulting model. Support sizes differ by experiment and are stated in each subsection. We additionally use synthetic Gaussian priors with support size up to 3000 to isolate algorithm cost from training cost.

Conflict magnitude. The conflict $\boldsymbol{c}$ has both a direction and a magnitude, and our results should not depend on the particular direction chosen. We therefore sample $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, I_d)$ and set $\boldsymbol{u} = \boldsymbol{g}/\|\boldsymbol{g}\|_2$, which is uniform on the unit sphere, and average over several such draws per setting. Only the magnitude is then swept. Let $\rho(\boldsymbol{u}) = \max_i \boldsymbol{u}^\top \boldsymbol{w}_i - \min_i \boldsymbol{u}^\top \boldsymbol{w}_i$ be the spread of the prior support along $\boldsymbol{u}$. We set $\boldsymbol{c} = \|\boldsymbol{c}\|_2\, \boldsymbol{u}$ and sweep $\|\boldsymbol{c}\|_2$ directly.

Strategic and non-strategic benchmarks. For each private dataset D, the *strategic* benchmark is the user-side strategy computed by our algorithm, which selects the report maximizing the user's expected utility. The *non-strategic* benchmark is the non-influential baseline of Section 3: the learner ignores the report and chooses

$$\boldsymbol{w}_{\text{uninf}} \in \arg\max_{\boldsymbol{w} \in \mathcal{W}} \mathbb{E}_{D' \sim \pi}\left[U^\ell(D', h_{\boldsymbol{w}})\right],$$

so the user's report does not affect the learned model. This is not an arbitrary baseline: by Section 3 a non-influential equilibrium always exists, so it is always available to the learner. The *truthful* benchmark reports D unmodified; under a nonzero conflict it is not an equilibrium and serves only as a no-conflict upper bound.

Hardware and software. All experiments ran on a MacBook Pro with an Apple M1 Pro processor, 16 GB of memory, and CPU-only execution, using Python 3.11.4, NumPy 1.24.3, Pandas 1.5.3, SciPy 1.10.1, scikit-learn 1.3.0, and PyTorch 2.0.1.

F.2 User-Optimal Report

This subsection reports how the dynamic program scales beyond the support sizes used in Table 1, and how often the factorized-prior construction is exact equilibrium under the true empirical prior.

Instance. Empirical priors with support size 50 and conflict $\|\boldsymbol{c}\|_2 \in \{0.005, 0.02, 0.05, 0.10\}$, with three independent train/test splits per level, giving 12 trials. We stop at $\|\boldsymbol{c}\|_2 = 0.10$ because larger conflicts increasingly yield non-influential equilibria: in our sweep the selected partition already collapses to a single region in several $\|\boldsymbol{c}\|_2 = 0.10$ trials, including all three linear-SVM runs on School Admission and one linear-SVM run on Prosper Loan. The range therefore spans weak to moderate conflict while retaining informative reports in most settings.

Support	Time (s)	Non-strategic	Strategic
100	0.06 ± 0.00	-0.93 ± 0.01	$\mathbf{-0.02 \pm 0.04}$
200	0.45 ± 0.01	-0.93 ± 0.01	$\mathbf{-0.03 \pm 0.04}$
500	7.01 ± 0.07	-1.03 ± 0.01	$\mathbf{-0.05 \pm 0.06}$
1000	55.97 ± 1.41	-0.96 ± 0.01	$\mathbf{-0.06 \pm 0.08}$
2000	448.65 ± 13.69	-1.00 ± 0.01	$\mathbf{-0.07 \pm 0.08}$
3000	1516.79 ± 41.36	-0.99 ± 0.01	$\mathbf{-0.07 \pm 0.08}$

Table 4: Scaling of the dynamic program on synthetic Gaussian priors, averaged over the four conflict levels $\|\boldsymbol{c}\|_2 \in \{0.005, 0.02, 0.05, 0.10\}$. Truthful reporting is an upper bound rather than an equilibrium and yields 0.00 ± 0.01 for all rows.

Model	Dataset	Exact Rate	$\varepsilon_{\mathrm{IC}}$ Mean	$\varepsilon_{\mathrm{IC}}$ Max
Linear SVM	Credit	0.00	0.89	1.35
	School	0.25	0.03	0.12
	Census	0.08	1.61	4.79
	Prosper	0.58	0.15	1.01
MLP	Credit	0.92	0.01	0.17
	School	0.67	0.92	5.24
	Census	0.50	5.59	28.74
	Prosper	1.00	0.00	0.00

Table 5: Equilibrium verification under the true empirical prior. The exact rate is the fraction of the 12 trials with $\varepsilon_{\mathrm{IC}} = 0$. Values of $\varepsilon_{\mathrm{IC}}$ are in the utility units of Table 1, scaled by 10^3.

Scalability on larger prior support. Table 4 evaluates the exact finite-prior dynamic program on synthetic Gaussian priors with up to 3000 support points. Runtime tracks the cubic dependence proved in Section 4.2.1 closely: each doubling of the support multiplies runtime by 7.5–8.0 against a predicted 8, and the $2000 \to 3000$ step costs $3.38\times$ against a predicted $3.38\times$. The strategic equilibrium yields higher user utility than the non-strategic equilibrium at every support size, and the margin is stable as the support grows, so the gains in Table 1 are not an artifact of the support size used there.

Robustness to factorization violations. Table 5 evaluates the certificate of Proposition D.4 under the true empirical prior, that is, without the factorization assumption the strategy was constructed under. A value $\varepsilon_{\mathrm{IC}} = 0$ certifies an exact equilibrium of the returned reporting game; a positive value bounds the largest profitable deviation. The returned strategy is exact in a majority of trials on Prosper Loan for both learners and on Credit Approval for the MLP. Violations remain small in several other settings but are substantial for Census–MLP and some School–MLP trials, indicating larger departures from the factorization assumption in those cases.

F.3 User-Optimal under Data-Dependent Error

In this section, we evaluate the multidimensional algorithm of Appendix E.2.

Instance. The learner observes the report with data-dependent Gaussian error $\zeta \sim \mathcal{N}(0, \sigma^2)$, distinct from the data-independent deployment error $\boldsymbol{\xi}$ of Section 2. We evaluate the full grid $\|\boldsymbol{c}\|_2 \in \{0.005, 0.02, 0.05, 0.10\}$ by $\sigma \in \{0.005, 0.02, 0.05, 0.10\}$ with three independent train/test splits per pair, giving 48 trials, on empirical priors of support size 50.

Solver. We use `scipy.optimize.least_squares`. For each candidate report pattern the nonlinear first-order conditions are solved with box constraints restricting reports to $[0, 1]$, and a candidate is accepted only if the maximum unilateral best-response gain is at most 10^{-4}. Posterior expectations under Gaussian error are approximated by Gauss–Hermite quadrature with 41 nodes. We use three random restarts per pattern, the most that complete within a 30-minute per-instance limit.

Restart ablation. Table 6 evaluates the effect of the number of random restarts. At $\sigma = 0.05$, a single restart already attains the best objective value found with $R = 50$, and additional restarts do not

R	$\sigma = 0.05$			$\sigma = 0.20$			$\sigma = 0.50$		
	Obj. gap	**Cand.**	**Time (s)**	**Obj. gap**	**Cand.**	**Time (s)**	**Obj. gap**	**Cand.**	**Time (s)**
1	0.000	2.0	13.4	0.238	3.0	5.1	0.024	11.0	1.6
2	0.000	2.0	23.2	0.148	4.0	9.1	0.017	19.2	3.1
5	0.000	2.0	56.0	0.055	6.6	20.5	0.013	42.0	7.6
10	0.000	2.0	109.5	0.054	10.0	39.5	0.013	81.4	15.3
20	0.000	2.0	217.5	0.002	18.6	76.8	0.013	157.4	30.6
30	0.000	2.0	327.7	0.000	28.4	116.6	0.009	228.6	45.8
50	0.000	2.0	546.8	0.000	44.8	193.2	0.000	384.0	76.7

Table 6: Solver restart ablation on synthetic finite priors with support size $n = 12$, conflict $\|\boldsymbol{c}\|_2 = 0.303$, and error levels $\sigma \in \{0.05, 0.20, 0.50\}$. We vary the number of random restarts R and average over five solver seeds. For each seed, let J_R denote the largest objective value among the verified equilibria found with R restarts. The reported objective gap is $10^3(J_{50} - J_R)$, averaged over the five seeds. Smaller values are better, and zero means that R restarts attain the best objective value found with 50 restarts for that seed. *Cand.* is the number of verified candidate equilibria.

Model	**Dataset**	**# Params.**	**Acc. (%)**	**Time (s)**	**Non-strat.**	**Strategic**
Linear SVM	Credit Approval	47	85.4 ± 2.3	1436 ± 435	-54.3 ± 15.4	$\mathbf{-22.3 \pm 17.7}$
	Census	104–106	85.0 ± 0.3	1455 ± 412	-33.8 ± 6.5	$\mathbf{-17.0 \pm 12.0}$
	School Admission	52–57	95.0 ± 0.1	1378 ± 488	-4.1 ± 4.1	$\mathbf{-4.0 \pm 4.2}$
	Prosper Loan	2090	36.7 ± 0.3	1628 ± 377	-10.0 ± 4.2	$\mathbf{-7.9 \pm 5.8}$
MLP	Credit Approval	1570	83.8 ± 1.5	1501 ± 389	-23.8 ± 4.7	$\mathbf{-13.9 \pm 9.5}$
	Census	3426–3458	85.7 ± 0.1	1425 ± 434	-84.2 ± 5.6	$\mathbf{-29.5 \pm 24.7}$
	School Admission	1762–1826	94.8 ± 0.1	1477 ± 405	-37.5 ± 4.4	$\mathbf{-17.5 \pm 14.3}$
	Prosper Loan	6443	49.4 ± 0.3	1391 ± 448	-453.9 ± 74.1	$\mathbf{-122.0 \pm 122.2}$

Table 7: Reporting under data-dependent error, averaged over the full grid $\|\boldsymbol{c}\|_2 \in \{0.005, 0.02, 0.05, 0.10\}$ and $\sigma \in \{0.005, 0.02, 0.05, 0.10\}$ on empirical priors of support size 50. Average user utility is scaled by 10^3. Truthful reporting yields -3.2 ± 4.2 in every setting and is an upper bound rather than an equilibrium. Bold indicates the larger utility between the non-strategic and strategic strategies.

increase the number of verified candidate equilibria. At larger error levels, additional restarts discover more candidates and reduce the best-found objective gap, while runtime grows approximately linearly in R. The ablation includes error levels above the main sweep to make this effect visible. These results suggest that few restarts suffice at low error, while larger error may require more extensive search.

User Optimal Report. Table 7 reports the multidimensional algorithm under data-dependent error, in the same format as Table 1 of Section 5. The strategic report improves on the non-strategic baseline in every dataset–learner setting, with the largest gains on Census and Prosper Loan for the MLP, where it recovers roughly two-thirds and three-quarters of the gap to the truthful upper bound respectively. The improvement is smallest on School Admission for the linear SVM, where the non-strategic baseline is already near that bound and there is little room to gain. Comparing with Table 1, the strategic advantage persists under data-dependent error.