How to Sell High-Dimensional Data Optimally

Andrew A. Li[1], R. Ravi[1], Karan Singh[1], Zihong Yi[1], and Weizhong Zhang[1]

[1]Tepper School of Business, Carnegie Mellon University, Pittsburgh
{aali1,ravi,karansingh}@cmu.edu,{zihongy,weizhong}@andrew.cmu.edu

Abstract

Motivated by the problem of selling large, proprietary data, we consider an information pricing problem proposed by Bergemann et al. [6] that involves a decision-making buyer and a monopolistic seller. The seller has access to the underlying state of the world that determines the utility of the various actions the buyer may take. Since the buyer gains greater utility through better decisions resulting from more accurate assessments of the state, the seller can therefore promise the buyer supplemental information at a price. To contend with the fact that the seller may not be perfectly informed about the buyer's private preferences (or utility), we frame the problem of designing a data product as one where the seller designs a revenue-maximizing menu of statistical experiments.

Previous work by Cai et al. [13] showed that an optimal menu can be found in time polynomial in the size of the state space, whereas we observe that the latter quantity is naturally exponential in the dimension of the data. We propose an algorithm which, given only sampling access to the state space, provably generates a near-optimal menu with a number of samples independent of the size of the state space. We then analyze a special case of high-dimensional Gaussian data, showing that (a) it suffices to consider scalar Gaussian experiments, (b) the optimal menu of such experiments can be found efficiently via a semidefinite program, (c) full surplus extraction occurs if and only if a natural separation condition holds on the set of potential preferences of the buyer, and (d) deterministic experiments suffice for high-dimensional data.

1 Introduction

Consider the problem of selling proprietary data so as to maximize revenue. The prevalence and importance of this problem in practice are by this point firmly-established. In fact, the Electronic Frontier Foundation identified 1,400 separate data *brokers* operating in the United States as of 2025,[1] to say nothing of the number of individual *sellers* on these platforms. And while the training of machine learning models is just one use of data, the race for ever-increasing size and sophistication all but guarantees that the value of high-quality data will rise in step.

Any principled approach to this problem must deal with a set of challenges that include, at the very least, (a) designing a set of potential data "products," (b) identifying the value that potential buyers place on these products, and (c) selecting the optimal subset of products to offer, together with their corresponding optimal prices. Much of the prior work on selling data addresses strict subsets of these challenges, often in the context of stylized or parametric models.

[1]https://www.eff.org/document/appendix-b-databrokerfullregistry2025

A line of work that began with Bergemann et al. [6] allows all of these challenges to be modeled cleanly and simultaneously. That model treats data (and information, more generally) as valuable to a buyer insofar as it improves the eventual selection of a utility-maximizing action. There is assumed to be an abstract underlying **state** $\omega \in \Omega$, which is unknown to buyers and governs the resulting utility of each abstract **action** $a \in \mathcal{A}$. Buyers of different **types** have their own prior belief distributions on the true state, and the space of potential products to sell to them consists of statistical **experiments**, which is entirely generic in representing the manner in which buyers can update their beliefs.

It is worth introducing a few applications at this point to make the discussion more concrete:

- *Data-driven decision-making:* Buyers may face a real decision and be interested in reducing uncertainty with data. For example, live traffic data can be sold to buyers who intend to solve (implicitly or explicitly) different shortest path problems. The actions are the possible paths, the true state is the set of travel times on every road, and greater utility corresponds to a shorter travel time.

- *Training ML models:* Buyers may be using data to train machine learning models. In this case, the true state is the data itself (encoded numerically), the set of actions is a set of machine learning models (encoded parametrically), and greater utility corresponds to higher accuracy.

- *Features for ML predictions:* Another use of data lies in machine learning inference. Buyers may use the data as feature vectors while using, for example, a linear regression model to estimate an uncertain quantity of interest. Here, the action is scalar, and the true state is the data itself, and greater utility results from ever more accurate estimates of the uncertain target. We will soon analyze a special version of this application in detail.

In addition to illustrating the wide-ranging applications of the model we will study, these examples underscore a critical limitation of work to date: *scalability with respect to the state space* Ω, which is often high-dimensional as in the applications above. This model is only useful if the resulting *algorithmic* problem of selecting and pricing data products is solvable and, unfortunately, previous work in this regard is insufficient. The original work of Bergemann, Bonatti and Smolin [6] solved smaller special cases, including $|\Omega| = 2$, and general $|\Omega|$ but limited to two types of buyers. Cai and Velegkas [13] then proposed a linear programming algorithm, although its size scales polynomially with $|\Omega|$. We contend that $|\Omega|$ is the limiting factor in practice, as it grows exponentially in the dimension of the data.[2]

1.1 Main Contributions

An Efficient Algorithm for Data Pricing: Thus motivated, the primary contribution of this paper is an algorithm for data pricing that obtains near-optimal revenue with a runtime that is *independent* of the potentially infinite state space:

> **Theorem 1 (Informal).** *Given samples of the state space drawn according to the buyers' belief distributions, there is an algorithm that computes a near-optimal menu of*

[2] Incidentally, [13] discuss the same shortest-path application as here, but in their case they describe Ω as a two-element set that encodes "low" and "high" traffic.

Paper	Buyer Types	State Space Size	Runtime
[6]	n	2	$O(n)$ for $\lvert\mathcal{A}\rvert = 2$
[6]	2	Finite	$O(\lvert\Omega\rvert\lvert\mathcal{A}\rvert)$
[13]	n	Finite	$\text{poly}(n, \lvert\Omega\rvert, \lvert\mathcal{A}\rvert)$
[13]	n	Finite	$\text{poly}(n, \lvert\Omega\rvert^{\lvert\Omega\rvert^2}, 1/\epsilon^{\lvert\Omega\rvert^2})$*
This Paper	n	Infinite	$\text{poly}(n, \lvert\mathcal{A}\rvert, 1/\epsilon)$**

Table 1: Comparison to prior work. Reported runtimes that include ϵ are for algorithms which approximate the optimal value up to an additive factor. *Requires an oracle that gives the expected-utility-maximizing action for any given distribution over states. **Requires a sampling oracle over the state space.

> *experiments and prices with high probability, and both its runtime and sample complexity are independent of the size of the state space and polynomial in the number of actions.*

Table 1 highlights the key contrasts between this result and previous algorithmic work. For example, in addition to scalability with respect to the state space, our algorithm's polynomial dependence on the action space compares well to the choice in existing work between (a) polynomial dependence and (b) independence at the cost of super-exponential dependence on the state space.

High-Dimensional Gaussian Data. Our second contribution is to sharpen our results for an important special case where the state space is $\mathbb{R}^d$ and the state ω follows a Gaussian distribution. Here, each buyer type has a private preference vector $\theta_i \in \mathbb{R}^d$, which differentially weighs the various features (coordinates) of ω along which they wish to estimate the unknown state. The utility of type i is $u^i(\omega, a) = -(\theta_i^\top \omega - a)^2$. For this setting, we have the following results.

1. We show that it is sufficient, without any loss of revenue, to consider scalar Gaussian experiments of the form $\mathcal{N}(v^\top \omega, \sigma^2)$, which project the state ω along a direction before adding noise, although this direction may not align with any of the preference vectors.
2. We provide an intuitive condition characterizing when the seller can extract the full surplus from the buyer. This happens if and only if, for each type i, its preference vector θ_i is longer than the shadow that other θ_j's cast on it. We also show for high-dimensional states that there exists a revenue-optimal menu composed entirely of deterministic experiments ($\sigma_E = 0$).
3. We give a semidefinite program, solvable in time scaling polynomially in the number of buyers and the *dimension* of the underlying state space, that computes such a revenue-maximizing menu composed entirely of scalar Gaussian experiments.

This special case reveals important qualitative properties of the broader model we study, including the relative value of offering multiple data products to extract revenue from heterogeneous buyers, and the sufficiency of deterministic data products (i.e. those which do not introduce "artificial" randomness).

We conclude this section with a brief review of related work. We introduce the model and problem in Section 2. In Section 3, we present the algorithm and prove our main result. In Section 4, we analyze the specialized Gaussian setting and conclude in Section 5.

1.2 Related Literature

The scope of our work lies within the intersection of monopoly pricing, information design, and information pricing.

Monopoly Pricing: A critical but distinct issue is the monopoly pricing problem associated with the sale of goods, where the buyer's valuation is derived from a known distribution. Complete characterizations are known in the single-dimensional setting [29, 30], and for specific cases in multidimensional settings [19, 16, 21, 20]. Furthermore, [18] explored the simultaneous design of mechanisms for selling an *indivisible* good, where the auctioneer can, without prior observation, release additional signals about the item. Relative to goods, pricing for data/information differs significantly in two primary ways: First, while goods are typically indivisible, the pricing of information has more complexity due to its intangible nature. Second, buyers with varying preferences do not only assign different values to various products; their preferences also influence the ranking of these products. Consequently, the value of information inherently encompasses both a vertical element (the quality of the information) and a horizontal element (the position of the information).

Information Design: The primary goal in information design is for the designer to disclose signals about the state to influence the actions of the agents involved [7, 25, 24, 4, 17]. A fundamental distinction arises in our model of information selling: unlike traditional information design, where the designer's revenue is typically derived from the actions taken by the buyers, in our model the seller's revenue is accrued directly from the payments made by the buyers.

Information Pricing: Two primary research streams have emerged: *ex ante* pricing and *ex post* pricing, in parallel to real-life information pricing models, specifically subscription-based (ex ante) and one-time payment systems (ex post). In ex ante pricing, the seller commits to revealing information about the state ω without prior observation, while in ex post pricing, the pricing decision is contingent upon the state ω observed at the time of the transaction.

Ex-ante Pricing: [2, 3] analyzed the sale of information to a continuum of ex ante homogeneous buyers, demonstrating that it is optimal to provide noisy and idiosyncratic information, thereby ensuring that the seller maintains a local monopoly. The context in their work is the rational expectations equilibrium, which requires interaction between buyers. This contrasts with our model where there is no interaction among buyers, and buyers are inherently heterogeneous due to their utilities. As mentioned earlier, we use the model introduced in [6], where the state space is discrete and the seller's information is conveyed through experiments [8, 9]. They provide an explicit construction of the optimal menu in the case of binary states and actions. Based on their framework, [13] proposed a linear program to calculate the revenue-maximizing menu. They also highlighted that when a best-action oracle is available, that is, an oracle that receives a distribution of states as input and returns the optimal action, there exists an algorithm capable of efficiently (in the number of actions) computing the revenue maximizing menu.

Ex-post Pricing: A distinct body of research focuses on ex-post Pricing where the value function post-realization depends on two state variables - one held by the data seller and the other by the buyer [5, 27, 14]. Ex-post pricing highlights substantial differences from ex-ante pricing, as it permits the seller to tailor experiments based on the realized state.

2 Preliminaries

2.1 Model

We begin by formalizing the model: The buyer has a state-dependent utility that they want to maximize by taking an action from an action space. The buyer has a private *type* that captures their (ex-post) utility as a function of the state-action pairs. Although the seller is not aware of the buyer's private type, they know that the type follows a known distribution. The seller determines a menu of experiments with corresponding prices to offer for sale to the buyer. Each experiment maps the state to a distribution over a space of signals. Here is a list of notation that we will use.

- The *state* ω is drawn from a distribution μ, a publicly known common prior supported on the (possibly uncountable) state space Ω.
- $\mathcal{A}$ denotes the finite *action* space. Let $m = |\mathcal{A}|$.
- The buyer's *type* i, belonging to $[n]$, determines their utility function $u^i(\omega, a) : \Omega \times \mathcal{A} \mapsto [0, 1]$. The distribution over types is $f \in \Delta[n]$; this is public knowledge.
- $E = (S, \pi)$ denotes a statistical *experiment*, where S is the finite set of *signals*, and $\pi : \Omega \mapsto \Delta S$ is the signaling function which maps each state ω to a distribution $\pi(\cdot|\omega)$ over the signals. The fact that S is finite holds without loss of generality, as we will discuss momentarily.
- $\mathcal{M} = \{\mathcal{E}, t\}$ denotes a *menu* of experiments, where $\mathcal{E}$ is a set of experiments, and $t : \mathcal{E} \mapsto \mathbb{R}_{\geqslant 0}$ is the price function that denotes the price $t(E)$ of each experiment $E \in \mathcal{E}$.

The game modeling the design, sale and usage of the menu of experiments proceeds as follows: (1) The seller posts a menu $\mathcal{M} = \{\mathcal{E}, t\}$; (2) The type of the buyer, i, is determined; (3) The buyer chooses an experiment $E = (S, \pi) \in \mathcal{E}$ and agrees to pay a (randomized) price $t(E)$; (4) Next, the state ω is realized, and the seller sends a signal s that is drawn from $\pi(\cdot \mid \omega)$; (5) Then the price $t(E)$ is fully determined and paid; (6) The buyer updates their belief based on the signal s received and chooses an optimal action a based on it.

In this setup, both the buyer and the seller are expected utility-maximizing agents. Thus, in the absence of any privileged information, a buyer of type i has a baseline utility of $u^i := \max_{a \in \mathcal{A}} \mathbb{E}_{\omega \sim \mu}[u^i(\omega, a)]$. resulting from taking the action $a^i := \arg\max_{a \in \mathcal{A}} \mathbb{E}_{\omega \sim \mu}[u^i(\omega, a)]$.

Now, if the same buyer receives the signal s from experiment $E = (S, \pi)$, they revise their belief before selecting an action. Hence, their conditional expected utility will be

$$u^i(s, E) \coloneqq \max_{a \in \mathcal{A}} \mathbb{E}_{\omega \sim \mu|(s,E)}[u^i(\omega, a)] = \max_{a \in \mathcal{A}} \mathbb{E}_{\omega \sim \mu} \left[\frac{\pi(s \mid \omega)}{\mathbb{E}_{\omega' \sim \mu}[\pi(s \mid \omega')]} u^i(\omega, a) \right],$$

and they will select an action, denoted by $a^i(s, E)$, which achieves this maximum. Together with μ, each experiment $E = (S, \pi)$ produces a marginal distribution $\pi \circ \mu$ in the signal space.

Finally, we can calculate the *value* of an experiment $E = (S, \pi)$ for a buyer of type i:

$$V(E, i) \coloneqq \mathbb{E}_{s \sim \pi \circ \mu}[u^i(s, E)] = \sum_{s \in S} \mathbb{E}_{\omega \sim \mu}[\pi(s \mid \omega)] u^i(s, E) = \sum_{s \in S} \max_{a \in \mathcal{A}} \mathbb{E}_{\omega \sim \mu}[\pi(s \mid \omega) u^i(\omega, a)].$$

Note that since $V(E, i)$ is a point-wise maximum of linear functions, it is in fact convex in π.

2.2 The Revenue Maximization Problem

As argued in [6], by the revelation principle, the seller can restrict their attention to direct mechanisms, where the menu has n entries such that each type i favors E^i over other experiments. Generically, a buyer of type i is incentivized to purchase an experiment E^i if and only if $V(E^i, i) - t(E^i) \geqslant u^i$. Viewed as a constraint on the menu, this is referred to as *individual rationality (IR)*. Moreover, every buyer selects the experiment that maximizes their own expected net utility, so for any $i, j \in [n]$, we have $V(E^i, i) - t(E^i) \geqslant V(E^j, i) - t(E^j)$. These are the *incentive compatibility (IC)* constraints.

Following [6], the seller can further take $S = \mathcal{A}$, which means that the signal and action spaces are the same, and restrict their search to *responsive* menus without loss of generality. A responsive menu is one in which every signal in each experiment corresponds to the optimal action it results in for the intended buyer. Concretely, a responsive menu requires that $a = a^i(a, E^i)$ for all types i in $[n]$ and signals a in $\mathcal{A}$, which can be enforced via *obedience* constraints [7] as

$$\sum_{a \in \mathcal{A}} \mathbb{E}_{\omega \sim \mu}[\pi(a \mid \omega) u^i(\omega, a)] \geqslant \sum_{a \in \mathcal{A}} \max_{a' \in \mathcal{A}} \mathbb{E}_{\omega \sim \mu}[\pi(a \mid \omega) u^i(\omega, a')], \qquad \forall i \in [n].$$

Conveniently, since every signal from $E^i = (\mathcal{A}, \pi^i)$ coincides with the utility-maximizing action for type i, it follows that $V(E^i, i)$ can further be written as a linear function of π_i as

$$\overline{V}(E^i, i) = \sum_{a \in \mathcal{A}} \mathbb{E}_{\omega \sim \mu}[\pi^i(a \mid \omega) u^i(\omega, a)].$$

Compiling the IC, IR, and obedience constraints, we can now formulate the design of the revenue-maximizing menu as the following optimization problem.

$$\begin{array}{lll} \max\limits_{\{\pi^i, t^i\}} & \sum\limits_{i \in [n]} f_i t^i & \\ \text{subject to} & \overline{V}(E^i, i) - t^i \geqslant V(E^j, i) - t^j, & \forall i, j \in [n] \times [n] \\ & \overline{V}(E^i, i) - t^i \geqslant u^i, & \forall i \in [n]. \end{array} \tag{1}$$

Since $V(E^i, i)$ and $V(E^j, i)$ are linear and convex, respectively, in π^i and π^j, the formulation is a convex (in fact, linear) program.

Finally, it is important to emphasize that we have taken the buyers here to be heterogeneous in their utility functions u^i, but assumed that they share a common prior distribution μ on the state. This is different from [6, 13], where the reverse is assumed: buyers have different priors, but share a utility function ([13] also note briefly that generalizing their approach to allow buyers to vary on utility functions is straightforward). We conclude here by noting that, under mild conditions, our optimization problem is equivalent to that under the most general model with buyer-specific prior distributions. Specifically, if $\mu^1, \ldots, \mu^n$ are prior distributions (over Ω) corresponding to each buyer type, the resulting optimization problem takes the same form of (1), although with expectations taken over the type-specific priors μ^i as opposed to μ:

$$u^i = \max_{a \in \mathcal{A}} \mathbb{E}_{\omega \sim \mu^i}[u^i(\omega, a)], \qquad V(E^j, i) = \sum_{s \in S} \max_{a \in \mathcal{A}} \mathbb{E}_{\omega \sim \mu^i}[\pi^j(s \mid \omega) u^i(\omega, a)].$$

These expectations over μ^i can be replaced with ones over μ, so long as the changes of measure $d\mu^i/d\mu$ exist (for example, if μ^i is absolutely continuous with respect to μ) which can then be encoded into the utility functions. Thus, the problem with prior distributions μ^i and utility functions $u^i(\omega, a)$ is equivalent to the problem with a shared prior μ and utility functions $u^i(\omega, a)\frac{d\mu^i}{d\mu}(\omega)$.

2.3 The Value of Differentiated Data Products

Before we introduce our main results, we make an observation that underscores the benefits (and limits) of offering multiple data products on a menu. Let R_{menu} be the revenue generated by an optimal menu. Let R_{one} be the maximum revenue generated by a menu with a single experiment; hence the only lever available to the seller is the price. Finally, let $R_{\text{full-info}}$ be the revenue obtainable in an oracular setting where the buyer's type is known to the seller; this is therefore the revenue in a setting without any information asymmetry. It is easy to see that $R_{\text{one}} \leqslant R_{\text{menu}} \leqslant R_{\text{full-info}}$, and that these equalities are realized for certain settings, for example, if all types share the same utility function. We are interested in lower bounds on: $R_{\text{one}}/R_{\text{menu}}$ that captures the benefit of offering differentiated products; $R_{\text{menu}}/R_{\text{full-info}}$ that captures the cost of information asymmetry.

Proposition 1. For any setting of type-dependent preferences, $R_{\text{one}}/R_{\text{menu}}$ and and $R_{\text{menu}}/R_{\text{full-info}}$ lie in $[1/n, 1]$. Furthermore, for any $\varepsilon > 0$, there exists a setting in which $R_{\text{one}}/R_{\text{menu}} \leqslant 1/n + \varepsilon$ and $R_{\text{menu}} = R_{\text{full-info}}$ hold simultaneously.

Our key takeaway from this observation (proved in Appendix A) is that the ability to provide multiple products can be a substantial benefit and, in certain settings, completely overcome the information asymmetry to which the seller is subject.

3 Near-Optimal Menus for Large State Spaces

We are now prepared to state our main result. Recall that n and m denote, respectively, the number of buyer types and actions.

Algorithm 1 Lazy Experiments via Linear Programming

Input: Buyer's type $i \in [n]$, the realized state $\overline{\omega}$.
Parameters: Sample budget K.

1: Draw $K-1$ samples $\omega_{2:K} \sim \mu$, and set $\omega_1 \coloneqq \overline{\omega}$.
2: Solve the LP below to obtain an optimal solution $\{\pi^i, t^i\}$.

$$
\begin{aligned}
\max \quad & \textstyle\sum_{i\in[n]} f_i t^i \\
\text{s.t.} \quad & \textstyle\sum_{a\in\mathcal{A}} \frac{1}{K} \sum_{k=1}^{K} \pi^i(a \mid \omega_k) u^i(a, \omega_k) - t^i \geqslant \sum_{a\in\mathcal{A}} v_{a,i,j} - t^j && \forall i, j \in [n] \\
& v_{a,i,j} \geqslant \frac{1}{K} \textstyle\sum_{k=1}^{K} \pi^j(a \mid \omega_k) u^i(a', \omega_k) && \forall a, a' \in \mathcal{A},\ i, j \in [n] \qquad (2) \\
& \textstyle\sum_{a\in\mathcal{A}} \frac{1}{K} \sum_{k=1}^{K} \pi^i(a \mid \omega_k) u^i(a, \omega_k) - t^i \geqslant u^i && \forall i \in [n] \\
& \textstyle\sum_{a\in\mathcal{A}} \pi^i(a \mid \omega_k) = 1, && \forall i \in [n], k \in [K]
\end{aligned}
$$

3: Output a signal sampled from $\pi^i(\cdot \mid \overline{\omega})$ and the corresponding price t^i.

Remark. The LP in Equation (2) is naturally symmetric with respect to its inputs $\omega_{1:K}$. We assume that the LP solver respects this symmetry, that is, if the solver is run with two different permutations of $\omega_{1:K}$, for any fixed $a \in \mathcal{A}$ and $\omega \in \omega_{1:K}$, the output $\pi^i(a \mid \omega)$, or its distribution, is identical. This can always be ensured by first randomly permuting $\omega_{1:K}$ before the LP is constituted.

Theorem 1. *Let E^i be the experiment that Algorithm 1 executes for input type i with sample budget K, and let t^i be the corresponding (randomized) price. For any $\varepsilon, \delta \in (0,1)$, there exists K satisfying*

$$K = O\left(\min\left\{\frac{n^2}{\varepsilon^2}, \frac{1}{\varepsilon^4}\right\}\left(m \log \frac{mn}{\delta}\right)\right)$$

such that $\{E^i, t^i\}$ is a incentive-compatible direct menu whose revenue is within ε of the optimal with probability at least $1-\delta$, that is, assuming R^* is the revenue of an optimal menu, we have

$$\sum_{i=1}^{n} f_i t^i \;\geqslant\; R^* - \varepsilon.$$

Algorithm 1 gives an implicit representation of the experiment, along with a stochastic price, that a buyer would receive if they declare their type as being i. Internally, the algorithm simply solves the LP from [13] while restricting the state space to a small randomly chosen subset $S \subseteq \Omega$ pretending that the observed empirical distribution is the common prior. This computation can be performed in time polynomial in m, n and $1/\varepsilon$, which crucially does not scale with the size of the state space.[3] Since the realized state is part of S by design, this allows us to generate the signal for it. We show that such an implicit menu enforces truth-telling. Notice that the price extracted depends on the state, but the way this is generated is agreed to when the buyer chooses an experiment-price pair. Although random sampling here is reminiscent of sample average approximation methods [26], one can not hope that the resolution of this sampling is such that the entire menu can be reconstructed, instead we show that a few samples are enough to *locally decode* the LP and sample the signals for the realized state.

More sophisticated variants of this argument occur in the literature on mechanism design [22, 11, 31] when dealing with rich-type spaces and mechanism-to-algorithm reductions. More recently, such techniques were applied to the Bayesian persuasion setting in [17], where, unlike in our setting, the receiver (or in our language, the buyer) does not have private preferences. Beyond differences in the problem setting, a key qualitative distinction of our work is that the proposed mechanism is exactly incentive compatible (and exactly responsive). Previous applications of similar arguments resulted in approximate incentive compatibility (for example, in [22, 17]), which means that the resultant implementations require further assumptions on the agent's (or the buyer's) behavior, specifically that it must be willing to follow suboptimal non-utility-maximizing recommendations. In contrast, in our setup, buyers are exact utility maximizers; for this, our proof of Lemma 2 relies on adjusting prices to ensure exact incentive compatibility and *coarsening* experiments correctly in the Blackwell sense [8] to ensure exact obedience.

Finally, we remark that, as presented, the (randomized) price extracted by the mechanism depends on the realized state, which might not be desirable (see the discussion in [5]). However, this is easy to fix. Consider a variant of the mechanism in which the buyer agrees to pay a (randomized) price determined by running Algorithm 1 with the input state independently sampled from μ. The signal is released by running Algorithm 1 separately on the realized state as usual. By linearity of expectation and the affine dependence of the constraints and the objective in the prices, the marginal distribution of the revenue of this mechanism is the same as that of the original and

[3] Technically, *naming* or addressing any object taking N possible values requires $\log N$ space and time, but we treat this as a unit cost operation.

incentive compatibility (Lemma 1) continues to hold.[4]

Proof of Theorem 1: We begin by showing that the procedure constructs an incentive-compatible menu. This hinges on the fact that the random variables sampled algorithmically, namely $\omega_{2:K}$ and the realized state, ω, are a priori exchangeable. As a consequence of this result, we know that the type i agrees to pay t^i.

Lemma 1. *Let (E^i, t^i) be the experiment-price pair obtained by executing Algorithm 1 with input type i. Then the menu $\{E^i, t^i\}$ implements an incentive-compatible direct mechanism.*

By elementary concentration and then fixing constraints violations, we show that the LP built on sampled states has a feasible solution with an optimal value close to R^*. Since the prices set by the algorithm are selected to maximize the objective value, $\sum_{i=1}^n f_i t^i$ can only be greater.

Lemma 2. *Let R^* be the revenue of an optimal menu. Then there exists K satisfying $K = O(\min\{n^2/\varepsilon^2, 1/\varepsilon^4\}\,(m \log mn/\delta))$ such that, with probability at least $1-\delta$, the LP formulated in Equation (2) has a feasible solution $\{\pi^i, t^i\}$ with value $\sum_i f_i t^i \geqslant R^* - \epsilon$.*

To conclude, we note that although the prices $\{t^i\}$ generated via the LP depend on the realized state, they are realized independently of the buyer's type. □

3.1 Proof of Lemma 1

We wish to establish that IC and IR constraints hold for this menu, specifically that

$$\mathop{\mathbb{E}}_{\substack{\omega_{2:K}\sim\mu\\ \overline{\omega}\sim\mu}}\left[\sum_{a\in\mathcal{A}}\pi^i(a\mid\overline{\omega})u^i(a,\overline{\omega})-t^i\right]\geqslant\max_j\left\{u^i,\mathop{\mathbb{E}}_{\omega_{2:K}\sim\mu}\left[\sum_{a\in\mathcal{A}}\max_{a'\in\mathcal{A}}\mathop{\mathbb{E}}_{\overline{\omega}\sim\mu}\left[\pi^j(a\mid\overline{\omega})u^i(a',\overline{\omega})-t^j\right]\right]\right\}$$

for all i and j in $[n]$. From the definition of the LP in Equation (2), we have for all i and j in $[n]$:

$$\sum_{a\in\mathcal{A}}\frac{1}{K}\sum_{k=1}^K\pi^i(a\mid\omega_k)u^i(a,\omega_k)-t^i\geqslant\max_j\left\{u^i,\sum_{a\in\mathcal{A}}\max_{a'\in\mathcal{A}}\sum_{a\in\mathcal{A}}\frac{1}{K}\sum_{k=1}^K\pi^j(a\mid\omega_k)u^i(a',\omega_k)-t^j\right\}.$$

Thus, a natural strategy is to take the expectations of this inequality with respect to the randomness in $\omega_{1:K}$, while utilizing that $\omega_{1:K}$ are exchangeable random variables. Indeed, this attempt succeeds, but one has to take care that, being the result of an optimization process, π^i itself is not fixed, but a function of $\omega_{1:K}$. Fix any i and j in $[n]$ and a and a' in $\mathcal{A}$. Observe that

$$\begin{aligned}
&\mathbb{E}_{\omega_{1:K}\sim\mu}\left[\frac{1}{K}\sum_{k=1}^K\pi^j(a\mid\omega_k)u^i(a',\omega_k)\right]=\frac{1}{K}\sum_{k=1}^K\mathbb{E}_{\omega_{1:K}\sim\mu}\left[\pi^j(a\mid\omega_k)u^i(a',\omega_k)\right]\\
&=_{(i)}\frac{1}{K}\sum_{k=1}^K\mathbb{E}_{\omega_{1:k-1},\overline{\omega},\omega_{k+1:K}\sim\mu}\left[\pi^j(a\mid\overline{\omega})u^i(a',\overline{\omega})\right]=_{(ii)}\frac{1}{K}\sum_{k=1}^K\mathbb{E}_{\overline{\omega},\omega_{2:K}\sim\mu}\left[\pi^j(a\mid\overline{\omega})u^i(a',\overline{\omega})\right]\\
&=\mathbb{E}_{\overline{\omega}\sim\mu}\mathbb{E}_{\omega_{2:K}\sim\mu}\left[\pi^j(a\mid\overline{\omega})u^i(a',\overline{\omega})\right],
\end{aligned}$$

[4] In fact, a variant of this argument can be used to the make the prices entirely deterministic (when the buyer sees them) at the cost of increased sample complexity by running Algorithm 1 a number of times with input states sampled from μ and posting the average prices thus obtained.

where (i) follows by renaming ω_k to $\overline{\omega}$ for each term within the sum, and (ii) utilizes the fact that all ω's are identically distributed and, hence, can be permuted. Note that in step (ii) this reordering of ω's does not alter $\pi^j(a \mid \overline{\omega})$, since although π_j depends on the set of ω's, it is not affected by their ordering, as we have previously remarked. For any random variable X, index set $\mathcal{I}$ and functions $\{f_i : i \in \mathcal{I}\}$, we have $\mathbb{E}_X \max_{i\in\mathcal{I}} f_i(X) \geqslant \max_{i\in\mathcal{I}} \mathbb{E}_X f_i(X)$. Using the last observation, and this fact to deal with max operators in the IC constraints, the desired IC and IR constraints follow.

3.2 Proof of Lemma 2

We prove that there exists a candidate solution to the LP described in Equation (2) that achieves the optimal revenue, which, however, is only approximately feasible. Next, we repair the constraint violations to construct a feasible solution which achieves an ever so slightly smaller net revenue.

Lemma 3. *Let R^* be the revenue of an optimal menu. Fix any $\varepsilon, \delta > 0$. For large enough K satisfying $K = O((m \log mn\delta^{-1})/\varepsilon^2)$, with probability at least $1-\delta$, there exist $\{\widetilde{\pi}^i, \widetilde{t}^i\}$ such that $\sum_i f_i \widetilde{t}^i = R^*$ and $\{\widetilde{\pi}^i, \widetilde{t}^i\}$ violates each IC and IR constraint in the LP (Equation (2)) by at most ε.*

Proof. Consider a responsive menu $\{\pi^{*i}, t^{*i}\}$ that maximizes revenue. Set $\widetilde{t}^i = t^{*i}$ and $\widetilde{\pi}^i(\cdot \mid \omega) = \pi^{*i}(\cdot \mid \omega)$ for all i in $[n]$ and $\omega \in \omega_{1:K}$, where $\omega_{1:K}$ define the LP. Fix any i and j in $[n]$ and a map $\sigma : \mathcal{A} \to \mathcal{A}$. Then $\mathbb{E}_{\omega_{1:K}\sim\mu}[\frac{1}{K}\sum_{k=1}^K \sum_{a\in\mathcal{A}} u^i(\omega_k, \sigma(a))\widetilde{\pi}^j(\mid \omega_k)] = \mathbb{E}_{\omega\sim\mu}[\sum_{a\in\mathcal{A}} u^i(\omega, \sigma(a))\pi^{*j}(a \mid \omega)]$. Since $\omega_{1:K}$ are independently sampled, and utilities always lie within $[0,1]$, by Hoeffding's inequality, with probability $1-\delta$, we have

$$\left| \frac{1}{K}\sum_{k=1}^K \sum_{a\in\mathcal{A}} u^i(\omega_k, \sigma(a))\widetilde{\pi}^j(a \mid \omega_k) - \mathbb{E}_{\omega\sim\mu}\left[\sum_{a\in\mathcal{A}} u^i(\omega, \sigma(a))\pi^{*j}(a \mid \omega)\right] \right| \leqslant \sqrt{\frac{\log 2\delta^{-1}}{2K}}.$$

We also know for all i in n that

$$\begin{aligned}\mathbb{E}_{\omega\sim\mu}\left[\sum_{a\in\mathcal{A}} u^i(\omega, a)\pi^{*i}(a \mid \omega)\right] - t^{*i} &\geqslant \max_j \left\{ u^i, \sum_{a\in\mathcal{A}} \max_{a'\in\mathcal{A}} \mathbb{E}_{\omega\sim\mu}\left[u^i(\omega, a)\pi^{*j}(a' \mid \omega)\right] - t^{*j} \right\} \\ &= \max_j \left\{ u^i, \max_{\sigma:\mathcal{A}\to\mathcal{A}} \mathbb{E}_{\omega\sim\mu}\left[\sum_{a\in\mathcal{A}} u^i(\omega, a)\pi^{*j}(\sigma(a) \mid \omega)\right] - t^{*j} \right\}.\end{aligned}$$

where the last line uses the fact that for any bivariate function $f : \mathcal{A}\times\mathcal{A} \to \mathbb{R}$, we have that $\sum_{a\in A} \max_{a'\in\mathcal{A}} f(a, a') = \max_{\sigma:\mathcal{A}\to\mathcal{A}} \sum_{a\in A} f(a, \sigma(a))$. Translating all terms that involve expectations into their empirical variants, along with union bounds over i, j in n and all mappings σ, we see that the equivalent IC and IR constraints in the LP are satisfied by $\{\widetilde{\pi}^i, \widetilde{t}^i\}$ up to an error of $2\sqrt{(m \log 2mn\delta^{-1})/K}$. □

From Lemma 3, we know there exists a menu $\{\widetilde{\pi}^i, \widetilde{t}^i\}$ such that $\sum_{i=1}^n f_i \widetilde{t}^i = R^*$ and

$$\widehat{V}(\widetilde{\pi}^i, i) - \widetilde{t}^i \geqslant \widehat{V}(\widetilde{\pi}^j, i) - \widetilde{t}^j - \varepsilon, \quad \text{and} \quad \widehat{V}(\widetilde{\pi}^i, i) - \widetilde{t}^i \geqslant u^i - \varepsilon, \quad \forall i, j \in [n],$$

where $\widehat{V}(\pi, i)$ is the value the type i assigns to an experiment π when the common prior has the probability mass function $\widehat{\mu}(\omega) = \frac{1}{K}\sum_{k=1}^K \mathbf{1}_{\omega=\omega_k}$.

Without loss of generality, we may assume that $\widetilde{t}^i$ form a nondecreasing sequence. To construct a truly feasible solution, we begin by modifying the prices to $t^i = \widetilde{t}^i - i\varepsilon$ for all i. With these prices,

the IR constraints are satisfied. Although the IC constraints are not yet satisfied, we claim that with this menu, each type i cannot prefer an experiment with an index lower than its own, since for all $j < i$, $\widehat{V}(\widetilde{\pi}^i, i) - \widehat{V}(\widetilde{\pi}^j, i) \geqslant \widetilde{t}^i - \widetilde{t}^j - \varepsilon = t^i - t^j + \varepsilon(i - j - 1) \geqslant t^i - t^j$. Therefore, with this menu, which is possibly not incentive compatible, the seller extracts a price of at least $\widetilde{t}^i - n\varepsilon$ from the type i, and therefore a net revenue of at least $R^* - n\varepsilon$. Using Lemma 9 in [13], but also see [12, 15], a similar transformation of prices through $t^i = (1 - \sqrt{\varepsilon})\widetilde{t}^i - \varepsilon$ can guarantee a revenue of $R^* - 3\sqrt{\varepsilon}$. To compile the lemma, we use the better of the two.

Finally, by relabeling the experiment that the type i actually chooses in this world, where $\widehat{\mu}$ is the common prior, as π^i and along with labeling the associated price as t^i we can transform this into an incentive-compatible menu with the same revenue. Similarly, by relabeling the signals in all experiments to match the actions they induce for their corresponding type, the menu can be made responsive. Note that such relabeling might *coarsen* the experiments in the Blackwell sense [9] and decrease their values. However, by construction, the type i values the new π^i the same, since its actions are unchanged. Thus, the IC and IR constraints in the LP are still satisfied.

4 Pricing High-dimensional Gaussian Data

We now turn to a special case that models the pricing and design of data products for high-dimensional data. The restriction we impose in the section allows us to construct explicit signaling schemes and derive a number of structural results that capture the nature of data markets today.

We consider a d-dimensional state space $\Omega = \mathbb{R}^d$, where the state ω is drawn from a Gaussian distribution $\mu = \mathcal{N}(0, I)$.[5] While the assumption of a Gaussian prior is a strong restriction compared to the algorithmic results in Section 3, its utility here is driven by the simplification of the signaling scheme and the specificity of structural results it allows. The buyer here is interested in a certain aspect of this high-dimensional state, represented by $\theta_i \in \mathbb{R}^d$ for the type i. This vector θ_i determines the utility function of type i over a real-valued action space as $u^i(\omega, a) = -(\theta_i^\top \omega - a)^2$. Thus, in this setup, the learner is interested in estimating the projection of a high-dimensional state in a certain privately known direction. The squared loss is indeed commonly found and widely used in statistics, economics, and machine learning (see, for example, [23, 1, 28]).

To concretely motivate the setting, consider the case when a data seller has a dataset comprising of a large number of observations, perhaps capturing health habits, travel logs, medical history, income history and financial dealings, about some number of individuals. It is not unreasonable to imagine that given access to such characteristics, a bank could tailor the interest rate for a loan, a tour company could offer more targeted discounts, or that an insurance company could personalize the premium for health insurance. However, it is clear that these entities care about distinct yet overlapping attributes that can best inform their decisions. The bank may rely on income history and past financial status; the tour company may base its decisions on income and the propensity for travel as expressed in travel logs; the insurance company might care about health and income related attributes. This is the diversity of interests that θ's capture. If such decisions are automated through machine learning, θ's can also be interpreted as the coefficients of the linear regression model that such entities use to make business decisions.

[5] These specific choices of the mean vector and the covariance matrix do not impede generality.

4.1 Scalar Gaussian Experiments Suffice

First, we observe that the value of any experiment is captured by its (expected) posterior covariance. To establish this, for an arbitrary experiment $E = (S, \pi)$, let $\omega_{s,E} = \mathbb{E}_{\omega\sim\mu|(s,E)}[\omega]$ be the posterior mean and $\mathrm{Cov}[\omega|s, E] = \mathbb{E}_{\omega\sim\mu|(s,E)}[(\omega - \omega_{s,E})(\omega - \omega_{s,E})^\top]$ be the posterior covariance upon observing the signal s from experiment E, and thus averaging over the signal s itself, let $\mathrm{Cov}[\omega|E] = \mathbb{E}_{s\sim\pi\circ\mu}[\mathrm{Cov}[\omega|s, E]]$ be the expected posterior covariance associated with E.

Proposition 2. For a Gaussian prior $\mu = \mathcal{N}(0, I)$ and utility functions $u^i(\omega, a) = -(\theta_i^\top\omega - a)^2$, for any experiment E, we have $V(E, i) = -\theta_i^\top \mathrm{Cov}[\omega|E]\theta_i$.

Proof. Since $\mathbb{E}_{\omega\sim\mu|(s,E)}[\theta_i^\top\omega] = \theta_i^\top\omega_{s,E}$, we have $u^i(s, E) = -\min_{a\in\mathcal{A}} \mathbb{E}_{\omega\sim\mu|(s,E)}[(\theta_i^\top\omega - a)^2] = -\mathbb{E}_{\omega\sim\mu|(s,E)}[(\theta_i^\top(\omega - \omega_{s,E}))^2] = -\theta_i^\top \mathrm{Cov}[\omega|s, E]\theta_i$. Taking the expectation over s being sampled from $\pi \circ \mu$ concludes the proof. □

The above proposition implies that the expected posterior covariance is a sufficient statistic to characterize the value of any experiment. We use this fact to prove that scalar Gaussian experiments are expressive enough to maximize revenue.[6]

Theorem 2. *For a Gaussian prior $\mu = N(0, I)$ and utility functions $u^i(\omega, a) = -(\theta_i^\top\omega - a)^2$, there exists a revenue maximizing menu whose each constituent experiment E generates signals as $\mathcal{N}(v_E^\top\omega, \sigma_E^2)$ for some $v_E \in \mathbb{R}^d$ and some $\sigma_E^2 \in \mathbb{R}_{\geqslant 0}$. In addition, the same parameter choices also satisfy $V(E, i) = (v_E^\top\theta_i)^2/(\|v_E\|^2 + \sigma_E^2) - \|\theta_i\|^2$.*

Proof. Let $\{E^i, t^i\}$ be a revenue-maximizing menu, which we generically assume to implement an incentive-compatible direct mechanism. For each experiment E^i we will construct v_i and σ_i^2 so that the scalar Gaussian experiment – let us call it G^i – representing $\mathcal{N}(v_i^\top\omega, \sigma_i^2)$ satisfies $V(E^i, i) = V(G^i, i)$ and $V(E^i, j) \geqslant V(G^i, j)$ for all i and j in $[n]$. Thus, $\{G^i, t^i\}$ also satisfies IC and IR constraints and is therefore a direct menu that generates the same revenue.

Lemma 4 (see e.g. [10])**.** *Consider a Gaussian random vector $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ with mean $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ and covariance $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$, then $(X_1|X_2 = a) \sim \mathcal{N}(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$.*

Using this folklore result, we see that $\mathrm{Cov}[\omega|G^i] = I - v_i v_i^\top/(\|v_i\|^2 + \sigma_i^2)$. For any vector-valued random variables Y and random variable X, we know $\mathrm{Cov}(Y) = \mathbb{E}(\mathrm{Cov}(Y|X)) + \mathrm{Cov}(\mathbb{E}(Y|X))$. For notational brevity, define $M_i \coloneqq I - \mathrm{Cov}[\omega|E^i]$, for which we know $I \geq M_i \geq 0$ using the previous fact. Set $v_i = M_i\theta_i$ and $\sigma_i^2 = \theta_i^\top(M_i - M_i^2)\theta_i$, which is valid since $M_i \geq M_i^2$. Using Proposition 2:

$$\begin{aligned} V(G^i, i) &= -\theta_i^\top \mathrm{Cov}[\omega|G^i]\theta_i = (\theta_i^\top M_i\theta_i)^2/(\theta_i^\top M_i^2\theta_i + \theta_i^\top(M_i - M_i^2)\theta_i) - \|\theta_i\|^2 \\ &= \theta_i^\top(M_i - I)\theta_i = -\theta_i^\top \mathrm{Cov}[\omega|E^i]\theta_i = V(E^i, i). \end{aligned}$$

Finally, we invoke Proposition 2 to observe that

$$\begin{aligned} V(G^i, j) &= -\theta_j^\top \mathrm{Cov}[\omega|G^i]\theta_j = (\theta_j^\top M_i\theta_i)^2/(\theta_i^\top M_i^2\theta_i + \theta_i^\top(M_i - M_i^2)\theta_i) - \|\theta_j\|^2 \\ &= (\theta_j^\top M_i\theta_i)^2/(\theta_i^\top M_i\theta_i) - \|\theta_j\|^2 \leqslant \theta_j^\top M_i\theta_j - \|\theta_j\|^2 = \theta_j^\top(M_i - I)\theta_j = V(E^i, j), \end{aligned}$$

[6] The revenue-optimal menus we consider in this section might not be *responsive*.

where the sole inequality follows from the generalized Cauchy-Schwarz inequality that for any $M \succeq 0$, it holds that $(\theta_i^\top M \theta_i)(\theta_j^\top M \theta_j) \geqslant (\theta_i^\top M \theta_j)^2$. The addendum can be independently verified by combining the expression for $\mathrm{Cov}[w|G^i]$ and Proposition 2. □

4.2 Optimal Menu Design via Semidefinite Programming

We first formulate the optimal menu design problem as a non-convex quadratically constrained quadratic program (NC-QCQP). Further, we prove that this NC-QCQP has an exact SDP relaxation and thus can be solved in polynomial time.

Proposition 3. Let $\{v_i, t^i\}$ be an optimal solution to the following (non-convex) QCQP.

$$\begin{array}{lll} \max\limits_{\{v_i, t^i\}} & \sum\limits_{i \in [n]} f_i t^i & \\ \text{subject to} & (\theta_i^\top v_i)^2 - t^i \geqslant (\theta_i^\top v_j)^2 - t^j, & \forall i, j \in [n] \times [n] \\ & (\theta_i^\top v_i)^2 - t^i \geqslant 0 \quad \text{and} \quad \|v_i\| \leqslant 1, & \forall i \in [n] \end{array} \tag{3}$$

Then (E^i, t^i) is a revenue-maximizing menu, where E^i represents the experiment $\mathcal{N}(v_i^\top \omega, 1 - \|v_i\|^2)$.

Proof. First, we note that any scaling of the outcome (signal) of an experiment does not alter its value because scaling operations do not change the posterior. Let $\mathcal{N}(v_E^\top \omega, \sigma_E^2)$ be witnesses to the statement of Theorem 2. Consider $v'_E = v_E / \sqrt{\|v_E\|^2 + \sigma_E^2}$ and ${\sigma'_E}^2 = \sigma_E^2 / (\|v_E\|^2 + \sigma_E^2)$. Thus, without loss of generality, we can operate with Gaussian experiments (say, E) $\mathcal{N}({v'_E}^\top \omega, {\sigma'_E}^2)$ where we are guaranteed that $\|v'_E\|^2 + {\sigma'_E}^2 = 1$, as long as we limit $\|v'_E\| \leqslant 1$. Furthermore, for such Gaussian experiments, from the latter part of Theorem 2, we know $V(E, i) = ({v'_E}^\top \theta_i)^2 + u^i$. Substituting this into Equation (1), we arrive at the claimed QCQP. □

Theorem 3. *The following SDP has an optimal rank-one solution. Furthermore, given a solution $\{V_i, t^i\}$ to the SDP, $\{v_i \coloneqq V_i \theta_i / \sqrt{\theta_i^\top V_i \theta_i}, t^i\}$ forms a solution to the NC-QCQP in Equation (3).*

$$\begin{array}{lll} \max\limits_{\{V_i, t^i\}} & \sum\limits_{i \in [n]} f_i t^i & \\ \textit{subject to} & \langle V_i, \theta_i \theta_i^\top \rangle - t^i \geqslant \langle V_j, \theta_i \theta_i^\top \rangle - t^j, & \forall i, j \in [n] \times [n] \\ & \langle V_i, \theta_i \theta_i^\top \rangle - t^i \geqslant 0, \quad \textit{and} \quad 0 \preceq V_i \preceq I, & \forall i \in [n] \end{array} \tag{4}$$

Proof. Let us first establish the second part of the claim. Note that $(v_i^\top \theta_i)^2 = (\theta_i^\top V_i \theta_i)^2 / (\theta_i^\top V_i \theta_i) = \langle V_i, \theta_i \theta_i \rangle$ and that $(\theta_j^\top v_i)^2 = (\theta_j^\top V_i \theta_i)^2 / (\theta_i^\top V_i \theta_i) \leqslant \langle V_i, \theta_j \theta_j^\top \rangle$ using the generalized Cauchy-Schwarz inequality. Moreover, $\|v_i\|^2 = \theta_i^\top V_i^2 \theta_i / \theta_i^\top V_i \theta_i \leqslant 1$ since $0 \preceq V_i \preceq I$ implies $0 \preceq V_i^2 \preceq V_i$. Thus, we have shown that all the NC-QCQP constraints are satisfied by the proposed solution. For the first part, we note that for any $\{v_i, t^i\}$ feasible for the NC-QCQP, $\{v_i v_i^\top, t^i\}$ is feasible for the SDP. □

4.3 A Characterization of Full Surplus Extraction

As a consequence of having private preferences, the types in our setup can extract information rents. We give a necessary and sufficient characterization of θ_i's under which the seller can perfectly screen the buyer and extract the entire revenue as if the buyer's type were publicly revealed. In the absence of this information asymmetry, the seller's revenue is $\sum_{i=1}^n f_i \|\theta_i\|^2$. Without loss of generality, we assume henceforth that $f_i > 0$ for all types i.

Theorem 4. *The seller extracts the full surplus from the buyer, with revenue totaling* $\sum_{i=1}^{n} f_i \|\theta_i\|^2$, *if and only if* $\{\theta_i\}$ *is well separated, concretely that for all* i *and* j *in* $[n]$, *we have* $|\theta_i^\top \theta_j| \leqslant \|\theta_i\|^2$.

Proof. Let us say that $\{\theta_i\}$ is well separated. Then, we claim $v_i = \theta_i/\|\theta_i\|$ and $t_i = \|\theta_i\|^2$ is feasible for the NC-QCQP in Equation (3). Note that $(\theta_i^\top v_i)^2 - t^i = 0$ for all i, while $(\theta_i^\top v_j)^2 - t^j = (\theta_i^\top \theta_j)^2/\|\theta_j\|^2 - \|\theta_j\|^2 \leqslant 0$. This establishes sufficiency.

Conversely, suppose that there exists a menu that extracts the full surplus; such a menu must be optimal. Given Theorem 2 and Proposition 3, we can generically assume that the experiments in such an optimal menu are of the kind used whose values are linked in Equation (3). The IR constraints, together with $\|v_i\|^2 \leqslant 1$, imply that $v_i = \theta_i/\|\theta_i\|$. Now, for any i and j, from the IC constraints, we arrive at $0 \geqslant (\theta_i^\top \theta_j)^2/\|\theta_j\|^2 - \|\theta_j\|^2$, as needed. □

4.4 Deterministic Experiments Suffice for High-dimensional Data

Based on previous work on data pricing (for example, [6, 13], and indeed the previous section of this work, also), one may form the expectation that randomizing signals is crucial to extract optimal revenue. Although true from a theoretical viewpoint, a puzzling aspect of modern data markets is that most products being sold involve little to no explicit randomization. In our simplified setting, although this is not uniformly true in all dimensions, we prove that there always exist deterministic signaling schemes for high-dimensional state spaces.

Theorem 5. *For a Gaussian prior* $\mu = \mathcal{N}(0, I)$ *and utility functions* $u^i(\omega, a) = -(\theta_i^\top \omega - a)^2$, *if the dimension of the state space is large enough, concretely, if* $d \geqslant n$, *there exists a revenue-maximizing menu composed entirely of deterministic experiments.*

Proof. Consider a setting where $d \geqslant n$ and an optimal solution composed of scalar Gaussian experiments of the form $\mathcal{N}(v_i^\top \omega, 1 - \|v_i\|^2)$ where $\{v_i\}$ are derived from an optimal solution to the NC-QCQP in Equation (3). Either $\|v_i\| = 1$ for all i, in which case we are done, since the variance of the scalar Gaussian experiments $1 - \|v_i\|^2$ is zero. If not, there exists some i with $\|v_i\| < 1$. Since $n - 1 < d$, there must exist a nonzero vector $\Delta_i \in \mathbb{R}^d$ that is perpendicular to $\{\theta_j : j \neq i\}$ and $\theta_i^\top v_i \theta_i^\top \Delta_i \geqslant 0$; the latter can be ensured by flipping the sign of Δ_i if necessary. Consider setting $v_i' = v_i + \alpha \Delta_i$ where α is nonnegative and set so that $\|v_i'\| = 1$, while retaining all other variables unchanged in the NC-QCQP. We claim the solution remains feasible, and in doing so, we haven't altered the transfers. To see this, note that $(\theta_j^\top v_i') = (\theta_j^\top v_i)^2$ for all $j \neq i$, and hence the other types retain their opinion of v_i (or correctly, E^i). However, for type i, $(\theta_i^\top v_i')^2 = (\theta_i^\top v_i)^2 + (\theta_i^\top \Delta_i)^2 + 2\theta_i^\top v_i' \theta_i^\top \Delta_i \geqslant (\theta_i^\top v_i)^2$, and thus the new experiment E^i is at least as attractive to type i as the one being replaced. This procedure can be repeated until all $\{v_i\}$ have unit norms. □

5 Conclusion

We revisited the information-selling framework of [6] and its algorithmic study in [13], noting that the size of the state space grows exponentially when used to model the sale of high-dimensional data, and proposed an algorithm that produces a near-optimal menu whose sample complexity and runtime are independent of the size of the state space. We analyzed the special case of Gaussian data, for which we found a compact description of the space of data products, an efficient algorithm to compute an optimal menu, and an intuitive necessary and sufficient condition for full surplus extraction.

References

[1] D. Acemoglu, A. Makhdoumi, A. Malekian, and A. Ozdaglar. Too much data: Prices and inefficiencies in data markets. *American Economic Journal: Microeconomics*, 14(4):218–256, 2022.

[2] A. R. Admati and P. Pfleiderer. A monopolistic market for information. *Journal of Economic Theory*, 39(2):400–438, 1986.

[3] A. R. Admati and P. Pfleiderer. Direct and indirect sale of information. *Econometrica: Journal of the Econometric Society*, pages 901–928, 1990.

[4] R. Alonso and O. Câmara. Bayesian persuasion with heterogeneous priors. *Journal of Economic Theory*, 165:672–706, 2016.

[5] M. Babaioff, R. Kleinberg, and R. Paes Leme. Optimal mechanisms for selling information. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 92–109, 2012.

[6] D. Bergemann, A. Bonatti, and A. Smolin. The design and price of information. *American Economic Review*, 108(1):1–48, 2018.

[7] D. Bergemann and S. Morris. Information design: A unified perspective. *Journal of Economic Literature*, 57(1):44–95, 2019.

[8] D. Blackwell. Comparison of experiments. In *Proceedings of the Second Berkeley symposium on Mathematical Statistics and Probability*, volume 2, pages 93–103. University of California Press, 1951.

[9] D. Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, pages 265–272, 1953.

[10] G. E. Box and G. C. Tiao. *Bayesian inference in statistical analysis*. John Wiley & Sons, 2011.

[11] Y. Cai, C. Daskalakis, and S. M. Weinberg. Understanding incentives: Mechanism design becomes algorithm design. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 618–627. IEEE, 2013.

[12] Y. Cai, A. Oikonomou, G. Velegkas, and M. Zhao. An efficient ϵ-bic to bic transformation and its application to black-box reduction in revenue maximization. In *Proceedings of the 2021 acm-siam symposium on discrete algorithms (soda)*, pages 1337–1356. SIAM, 2021.

[13] Y. Cai and G. Velegkas. How to sell information optimally: An algorithmic study. In *Proceedings of the 12th Innovations in Theoretical Computer Science Conference*, volume 185, 2021.

[14] J. Chen, M. Li, and H. Xu. Selling data to a machine learner: Pricing via costly signaling. In *International Conference on Machine Learning*, pages 3336–3359. PMLR, 2022.

[15] C. Daskalakis and S. M. Weinberg. Symmetries and optimal multi-dimensional mechanism design. In *Proceedings of the 13th ACM conference on Electronic commerce*, pages 370–387, 2012.

[16] N. R. Devanur, K. Goldner, R. R. Saxena, A. Schvartzman, and S. M. Weinberg. Optimal mechanism design for single-minded agents. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 193–256, 2020.

[17] S. Dughmi and H. Xu. Algorithmic bayesian persuasion. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 412–425, 2016.

[18] P. Eső and B. Szentes. Optimal information disclosure in auctions and the handicap auction. *The Review of Economic Studies*, 74(3):705–731, 2007.

[19] A. Fiat, K. Goldner, A. R. Karlin, and E. Koutsoupias. The fedex problem. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 21–22, 2016.

[20] Y. Giannakopoulos and E. Koutsoupias. Duality and optimality of auctions for uniform distributions. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 259–276, 2014.

[21] N. Haghpanah and J. Hartline. Reverse mechanism design. In *Proceedings of the sixteenth ACM conference on Economics and Computation*, pages 757–758, 2015.

[22] J. D. Hartline, R. Kleinberg, and A. Malekian. Bayesian incentive compatibility via matchings. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pages 734–747. SIAM, 2011.

[23] C. I. Jones and C. Tonetti. Nonrivalry and the economics of data. *American Economic Review*, 110(9):2819–2858, 2020.

[24] E. Kamenica. Bayesian persuasion and information design. *Annual Review of Economics*, 11:249–272, 2019.

[25] E. Kamenica and M. Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

[26] A. J. Kleywegt, A. Shapiro, and T. Homem-de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502, 2002.

[27] S. Liu, W. Shen, and H. Xu. Optimal pricing of information. *arXiv preprint arXiv:2102.13289*, 2021.

[28] W.-Y. Loh. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(1):14–23, 2011.

[29] R. B. Myerson. Optimal auction design. *Mathematics of Operations Research*, 6(1):58–73, 1981.

[30] J. Riley and R. Zeckhauser. Optimal selling strategies: When to haggle, when to hold firm. *The Quarterly Journal of Economics*, 98(2):267–289, 1983.

[31] S. M. Weinberg. *Algorithms for strategic agents.* PhD thesis, Massachusetts Institute of Technology, 2014.

A Proof of Proposition 1

First, we establish that $R_{\text{one}}/R_{\text{full-info}} \geqslant 1/n$: Full surplus extraction earns a revenue $R_{\text{full-info}} = \sum_{i=1}^{n} f_i(\max_{a\in\mathcal{A}} u^i(\omega, a) - u^i)$. In the one-item menu, we propose to reveal the state (without noise) at a price of $t^* = \max_{a\in\mathcal{A}} u^{i^*}(\omega, a) - u^{i^*}$ where i^* is the type for which $f_i(\max_{a\in\mathcal{A}} u^i(\omega, a) - u^i)$ is maximized. This menu of composed of a single item generates a revenue r^* that is at least $\max_{i\in[n]} f_i(\max_{a\in\mathcal{A}} u^i(\omega, a) - u^i)$, since the type i^* agrees with this transaction, demonstrating this claim. The first part of the proposition now follows directly from this claim, since $R_{\text{one}}/R_{\text{full-info}} = R_{\text{one}}/R_{\text{menu}} \times R_{\text{menu}}/R_{\text{full-info}}$ while both terms on the right are at most one.

For the existence part, we in fact provide a setting that conforms to the specialized Gaussian setting of Section 4: The type i with the preference vector $\theta_i = \alpha_i e_i$ occurs with probability f_i, where e_i's are the canonical basis vectors. We will soon specify the values of α_i and f_i. Clearly, $R_{\text{full-info}} = \sum_{i=1}^{n} f_i \alpha_i^2$. Similarly, $R_{\text{menu}} = \sum_{i=1}^{n} f_i \alpha_i^2$ can be verified by plugging $v_i = e_i$ and $t_i = \alpha_i^2$ into the NC-QCQP in Equation (3). Now, consider an optimal one-item menu that results in revenue R_{one}. We can generically assume that the product on offer here reveals the state entirely since this choice cannot decrease any type's willingness to pay. Let t^* be the price at which this product is sold. Now, $R_{\text{one}} = \sum_{i=1}^{n} f_i t^* \mathbf{1}_{\alpha_i^2 \geqslant t^*}$. Set $\alpha_i = \alpha^{-i/2}$ and $f_i = \alpha^i / \sum_{i=1}^{n} \alpha^i$, for some $\alpha \in (0, 1]$. Now, $\sum_{i=1}^{n} f_i \alpha_i^2 = n / \sum_{i=1}^{n} \alpha_i$, while $R_{\text{one}} = \max_{j\in[n]} \sum_{i=j}^{n} \alpha^{i-j} / \sum_{i=1}^{n} \alpha^i < 1/((1-\alpha)\sum_{i=1}^{n} \alpha^i)$. Thus $R_{\text{one}}/R_{\text{menu}} < 1/(n(1-\alpha))$. Taking a small enough α, in particular $\alpha = \varepsilon n/(1+\varepsilon n)$, suffices.