# Algorithmic Signaling of Features in Auction Design

Shaddin Dughmi[1]$^{\star}$, Nicole Immorlica[2], Ryan O'Donnell[3]$^{\star\star}$, and Li-Yang Tan[4]$^{\star\star\star}$

[1] University of Southern California
[2] Microsoft Research
[3] Carnegie Mellon University
[4] Simons Institute, UC Berkeley

**Abstract.** In many markets, products are highly complex with an extremely large set of features. In advertising auctions, for example, an impression, i.e., a viewer on a web page, has numerous features describing the viewer's demographics, browsing history, temporal aspects, etc. In these markets, an auctioneer must select a few key features to signal to bidders. These features should be selected such that the bidder with the highest value for the product can construct a bid so as to win the auction. We present an efficient algorithmic solution for this problem in a setting where the product's features are drawn independently from a known distribution, the bidders' values for a product are additive over their known values for the features of the product, and the number of features is exponentially larger than the number of bidders and the number of signals. Our approach involves solving a novel optimization problem regarding the expectation of a sum of independent random vectors that may be of independent interest. We complement our positive result with a hardness result for the problem when features are arbitrarily correlated. This result is based on the conjectured hardness of learning $k$-juntas, a central open problem in learning theory.

## 1 Introduction

Much of the computer science literature on auction design assumes bidders have full knowledge of their own values. However, in many markets, this assumption is quite unrealistic in part because the item for sale is not fully observable by the bidders. In used car auctions, for example, the cars for sale are each unique items with a long list of features – make, model, year, mileage, color, etc. Time and communication constraints make it impractical for the auctioneer to provide bidders with a full description of each car. Similarly, in advertising auctions, the impressions for sale correspond to searchers, again with a long list of features – gender, age, income, zip code, search history, etc.

$^{\star}$ Supported in part by NSF CAREER Award CCF-1350900. Part of this work performed while the author was visiting Microsoft Research New England.

$^{\star\star}$ Supported by NSF grants CCF-0747250 and CCF-1116594. Part of this work performed while the author was visiting Microsoft Research New England. Part of this work performed at the Boğaziçi University Computer Engineering Department, supported by Marie Curie International Incoming Fellowship project number 626373.

$^{\star\star\star}$ Supported by NSF grants CCF-1115703 and CCF-1319788. Part of this research was done while visiting Carnegie Mellon University.

Again it is impractical for the auctioneer to communicate all these features for each search, let alone track them all. This raises a natural question: *which features should an auctioneer signal to bidders?*

We study this question in the context of a single item auction. The item is parameterized by a large feature vector drawn from some known distribution. A bidder's value for an item is a function of its features. The goal is to signal a small subset of features to bidders such that the welfare[5] generated by the resulting auction is maximized. Trivial brute-force search can solve this problem in time $O(nk \cdot m^k)$ where $n$ is the number of players, $k$ is the number of allowed signals, and $m$ is the number of features. Throughout this paper, we think of the number of bidders and allowable signals as small, whereas the number of features is exponentially larger, and thus seek running times at most linear in $m$.

We wish to focus attention on the algorithmic problem of selecting features, and so we make several simplifying assumptions. First we assume bidders' values are additively separable across features. This assumption is a reasonable approximation to valuations in many settings and is also a good first step in understanding general substitutable valuations. Second, as is common in much of the computer science literature on signaling [2,8,9,15], we assume bidders' values for features are known to the auctioneer. This information could be available to the auctioneer through historical data, and is also a first step in designing systems for the more common Bayesian setting.[6]

Even with these simplifying assumptions, we obtain strong negative results for the problem of finding a welfare-maximizing set of signals. We do this by relating the feature selection problem to the problem of *learning $k$-juntas* (i.e. $m$-variable boolean functions that depend only on $k \ll m$ of its coordinates) with respect to the uniform distribution[7]. Introduced by Blum in 1994 [3,7], the junta problem is a clean abstraction of learning in the presence of irrelevant information, and represents a necessary first step towards the notorious problems of learning polynomial-size decision trees and DNF formulas. Progress on the problem has been slow despite significant interest — the current best algorithm is due to G. Valiant and runs in time $O(m^{0.6k})$ [17], a polynomial improvement over brute-force search in time $O(m^k)$, and it is a generally accepted assumption that is no $m^{o(k)}$-time algorithm for the problem. (Indeed, it is known that the broad class *statistical query* learning algorithms require both time and sample complexity $m^{\Omega(k)}$ for the junta problem [6]). Assuming that the junta problem does in fact require time $m^{\Omega(k)}$, we show there is no $m^{o(k)}$-time algorithm that can find an $(1/n + \epsilon)$-approximately optimal set of signals.

On the positive side, we consider a setting where each feature is selected independently from a (not necessarily identical) distribution, and takes on only a constant number of values. In this case, we give an $(1 - \epsilon)$-approximate algorithm that runs in time $O(m) + 2^{O(k \log(k/\epsilon))}$ for all fixed values of $n$. This algorithm solves a general optimiza-

[5] The welfare of a single item auction is the value of the winning bidder.

[6] Clearly, if the auctioneer knowns the values of the bidders, he can maximize welfare by simply assigning the item to the highest-value bidder, circumventing the auction altogether. We assert that even if the auctioneer has this information, market constraints require the use of a second-price auction format as is the case in, e.g., ad auctions.

[7] See Section 3 for a definition.

tion problem of potentially independent interest: for any norm $\|\cdot\|$ on $\mathbb{R}^n$, given $\theta \in \mathbb{R}^n$ and $m$ independent mean-zero vector-valued random variables $\mathbf{X}_1, \ldots, \mathbf{X}_m$, find a subset $S \subseteq [m]$ of cardinality $k$ that approximately maximizes $\mathbf{E}[\|\theta + \sum_{i \in S} \mathbf{X}_i\|]$. Prior to our work there were no non-trivial algorithms even when $n = 1$ — given $m$ *real-valued* random variables $\mathbf{X}_1, \ldots, \mathbf{X}_m$ and $\theta \in \mathbb{R}$, find a $k$-subset $S \subseteq [m]$ that approximately maximizes $\mathbf{E}[|\theta + \sum_{i \in S} \mathbf{X}_i|]$ — and even under the assumption that all the $\mathbf{X}_i$'s are two-valued.

**Related Work.** Understanding the structure of optimal signaling schemes is a classical question in economics [14], and has recently generated great interest within the computer science community [2,8,9,15,13]. One line of prior work [2,9,15] studies unconstrained signaling schemes that maximize revenue. In such unconstrained settings, full information revelation is guaranteed to optimize welfare. Other prior work [8], more closely related to the current paper, studies constrained signaling schemes and seeks to maximize welfare. That work considered two settings: one in which goods were represented by high dimensional feature vectors and one where the goods were arbitrary and had to be partitioned into classes. The former setting is closely related to ours, but in the prior work the signaling schemes were arbitrary bounded-length bit strings. In this paper, we constrain our signaling schemes to announce subsets of features, an arguably more natural scheme for which the techniques of the prior work cannot be applied. For an overview of the junta problem and its role in learning theory see [16,4] and the references therein. As mentioned above its hardness is a generally accepted assumption in learning theory, and indeed it is commonly used as hardness primitive to establish the intractability of various other learning problems (e.g. [1,11,10,12]).

## 2 Preliminaries

We consider a setting in which there is a set of possible items $\Omega$ for sale, where each $\omega \in \Omega$ is summarized by an $m$-dimensional vector of *features* — formally, $\Omega = \prod_{j=1}^{m} \Omega_j$, where $\Omega_j$ is the set of possible values of the $j$'th feature. We assume that an item is drawn according to a distribution $\lambda \in \Delta_\Omega$. There is a set of $n$ players, each of whom is equipped with a *valuation* function $v_i : \Omega \to \mathbb{R}_+$ mapping items to the real numbers. We restrict attention to *linearly separable* valuation functions, of the form $v_i(\omega) = \sum_{j=1}^{m} v_{ij}(\omega_j)$, for functions $v_{ij} : \Omega_j \to \mathbb{R}_+$.

We assume that the features of the item being sold are a-priori unknown to the players, who learn them through a *signaling scheme* mapping an items to messages, known as *signals*. In this paper, we restict attention to signaling schemes which simply fix a set $S \subseteq [m]$ of feature indices of a given size $|S| = k$, and announces $\omega_S = \{(j, \omega_j) : j \in S\}$. After players learn this partial information, some protocol — typically an auction — is run to assign the item to one of the players. We focus on auctions, such as the second-price auction, which assign the item to the player with the highest posterior expected value for the item given the features revealed. In this case the expected *social welfare*, i.e. the expected value of the winning player, can be written as follows.

$$\text{welfare}(S) = \mathbf{E}[\max_{i=1}^{n} \mathbf{E}[v_i(\omega)|\omega_S]]$$

Where both expectations are over $\omega \sim \lambda$. Using $v_{ij}$ as shorthand for the random variable $v_{ij}(\omega_j)$, we can rewrite the above expression as follows.

$$\text{welfare}(S) = \mathbf{E}\left[\max_{i=1}^{n}\left(\sum_{j\in S} v_{ij} + \sum_{j\notin S}\mathbf{E}[v_{ij}|\omega_S]\right)\right]$$

In the special case in which the features are independently distributed, this reduces to

$$\begin{aligned}\text{welfare}(S) &= \mathbf{E}\left[\max_{i=1}^{n}\left(\sum_{j\in S} v_{ij} + \sum_{j\notin S}\mathbf{E}[v_{ij}]\right)\right]\\ &= \mathbf{E}\left[\max_{i=1}^{n}\left(\sum_{j\in S}(v_{ij} - \mathbf{E}[v_{ij}]) + \sum_{j=1}^{m}\mathbf{E}[v_{ij}]\right)\right]\\ &= \mathbf{E}\left[\left\|\sum_{j\in S}(\boldsymbol{v}_j - \mathbf{E}[\boldsymbol{v}_j]) + \sum_{j=1}^{m}\mathbf{E}[\boldsymbol{v}_j]\right\|_{\infty}\right]\end{aligned}$$

when $\boldsymbol{v}_j$ denotes the $n$-dimensional random vector $(v_{1j}, v_{2j}, \ldots, v_{nj})$. Note that the vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ are independent when the features are independently distributed.

We adopt the perspective of an auctioneer seeking to optimize his choice of signaling scheme, with the goal of maximizing the expected welfare. This is nontrivial when $0 < k < m$, and we focus on the algorithmic question of finding the best set of features $S \in \binom{[m]}{k}$. We consider this question when the distribution $\lambda$ is represented explicitly. The sets $\Omega_1, \ldots, \Omega_m$ are given explicitly, as are the functions $\{v_{ij}\}_{i=1}^{n}$. In the general (correlated) case, $\lambda$ is described explicitly by a list of items $\Omega' \subseteq \Omega$ with associated probabilities $\{p(\omega) : \omega \in \Omega'\}$ summing to $1$ — all other items in $\Omega$ assumed to have probability $0$. In the independent case, the marginal distribution of each feature $j$ is given explicitly by the associated probabilities $\{p_j(\mu) : \mu \in \Omega_j\}$. We also consider the oracle model whereby only oracle access is given to $\lambda$; however, uniform convergence arguments reduce the algorithmic task of signaling in the oracle model to that in the explicit model, up to an arbitrarily small additive error term. In fact, our hardness result is proved in the oracle model, and thus translates to the explicit model.

## 3 Hardness for General Distributions

We now prove that, in general, no nontrivial approximation is possible for the feature signaling problem when the features are arbitrarily correlated. Our starting point is the conjectured hardness of a special case of the $k$-junta learning problem. A $k$-junta on $m$ variables is a boolean function $f : \{-1,1\}^m \to \{-1,1\}$ which depends on only $k$ bits of its input. When the bits $S \subseteq [m]$ determining $f$ are unknown, and a learner is given access to sample access to evaluations $(x, f(x))$ of $f$ on bit strings $x \in \{-1,1\}^m$ drawn uniformly at random, it is widely believed that no algorithm can recover $S$ in polynomial time/samples. In fact, this is believed true even for $k$-junta functions which compute the majority function on $k/2$ of the input bits, and the parity function on another $k/2$ bits, and then xor the results (these are listed explicitly as candidate hard functions in Blum's surveys on the junta problem [4,5]) — those functions are "balanced" in the sense we describe below.

**Definition 1.** *A boolean function $f : \{-1, 1\}^m \to \{-1, 1\}$ is c-balanced if the following holds for every $T \subseteq [m]$ with $|T| \leq c$, and $y \in \{-1, 1\}^c$.*

$$\mathbf{Pr}[f(x) = 1 | x_T = y] = \frac{1}{2},$$

*where $x_T$ denotes the projection of $x$ onto the coordinates in $T$, and the probability is over $x$ drawn uniformly from $\{-1, 1\}^m$.*

**Definition 2.** *We say a randomized algorithm ($\epsilon$,$\delta$)-weakly learns a $k$-junta $f$ if it outputs $S \subseteq [m]$ with $|S| \leq k$ such that, with probability at least $1 - \delta$,*

$$advantage(S) := \mathop{\mathbf{E}}_{x_S}\left[\left|\mathop{\mathbf{Pr}}_{x}[f(x) = 1 | x_S] - \frac{1}{2}\right|\right] \geq \epsilon$$

*where $x$ is uniformly distributed on $\{-1, 1\}^m$.*

We use the following commonly believed conjecture.

*Conjecture 1 (see e.g. [4,5]).* There are functions $k = k(m) = o(m)$ and $c = c(m) = \Theta(k)$ such that $c$-balanced $k$-juntas on $m$ variables can not be ($\epsilon$,$\delta$)-weakly learned in time $m^{o(k)}$ under the uniform distribution, for any pair of constants $\epsilon, \delta > 0$.

The above conjecture implies the following corollary.

**Corollary 1.** *Assuming Conjecture 1, there are functions $k = k(m) = o(m)$ and $c = c(m) = \Theta(k)$ such that no $\mathrm{poly}(m^{o(k)}, \log \frac{1}{\delta})$-time learning algorithm, given sample access to a $c$-balanced $k$-Junta $f$ on $m$ variables, outputs with probability $1 - \delta$ a set of variables $S$ of size $O(k)$ intersecting more than $c$ of the relevant variables of $f$.*

*Proof.* We assume that such an algorithm $\mathcal{A}$, with runtime $m^{o(k)}$ and arbitrarily small failure probability $\delta = \exp(-\Omega(k))$, exists. To simplify the proof, we assume $\mathcal{A}$ recovers a set of size $2k$ which includes $k/2$ relevant variables of a $k/2$-balanced Junta $f$, though the choice of constants is unimportant. We now show how to weakly learn $f$ in time $m^{o(k)}$, and with constant success probability, violating Conjecture 1.

We learn the relevant variables $S^* \in \binom{m}{k}$ of $f$ as follows: first, run $\mathcal{A}$ to recover $S \subseteq [m]$ with $|S| = 2k$ and $|S \cap S^*| \geq k/2$. Then, for each possible setting $z$ of the bits $S$ (of which there are $2^{2k}$) recurse on the function $f_{S,z}$ — often referred to as a *restriction* of $f$ — which simply replaces the portion of its input at indices $S$ with $z$ and then evaluates $f$. Note that $f_{S,z}$ remains $k/2$-balanced, and hence also $k/4$-balanced, yet is now a $k/2$-Junta on $m$ variables. Assuming the recursive calls succeed, between them they return the set $S^* \setminus S$. To complete $S^*$, it then suffices to try all $2^{2k}$ subsets of $S$.

In the event all invokations of $\mathcal{A}$ in the recursion tree are successful, correctness follows by induction. It remains to bound the runtime. Note that each recursive call halves the number of variables of the Junta. Therefore, the number of recursive calls equals $2^{2k} + 2^{2k} \cdot 2^k + 2^{2k} \cdot 2^k \cdot 2^{k/2} + \ldots \leq \log 2k \cdot 2^{4k} \leq 2^{5k} = m^{o(k)}$. By essentially the same analysis, the runtime of the algorithm is also $2^{5k} \leq m^{o(k)}$. The success probability is at least $1-\delta$ raised to a power equal to the number of calls of $\mathcal{A}$, which is a constant when $\delta = \exp(-\Omega(k))$ is sufficiently small. □

### 3.1 Warmup: Two players

As a warmup, we prove our impossibility result for 2 players assuming Conjecture 1. Note that we do not need the balance assumption for the 2-player special case.

**Theorem 1.** *Assuming Conjecture 1, there is no $m^{o(k)}$-time $(\frac{1}{2}+\epsilon)$-approximation algorithm for the feature signaling problem with two players in the sample oracle model, for any constant $\epsilon > 0$. This holds for Monte Carlo approximation algorithms having a constant success probability.*

*Proof.* Given sample access to an $m$-bit $k$-Junta $f$, with $k = o(m)$, we construct an instance of the feature signaling problem in the sample oracle model as follows. We let $\Omega = \{-1,0,1\}^{2m}$, and consider two players Alice and Bob. Both players have no value for features 1 through $m$ — i.e. $v_{ij}(.) = 0$ for $i \in \{A,B\}$ and $1 \leq j \leq m$. For the remaining features $j \in [m+1, 2m]$, Alice has value 1 if $\omega_j = 1$ and 0 otherwise, and Bob has value 1 if $\omega_j = -1$ and 0 otherwise.

The distribution $\lambda$ is constructed as follows. The first $m$ features of $\omega \sim \lambda$, which we denote by $x$, are uniformly distributed in $\{-1,1\}^m$. The last $m$ features, which we denote by $y$, are all set to 0, except for a single feature $j^*$ chosen uniformly at random, which is set to $f(x)$.

Note that if $f$ is a $k$-Junta determined by the bits $S^* \subseteq [m]$ with $|S^*| = k$, then $welfare(S^*) = 1$ as those bits uniquely determine which of Alice or Bob values the item being sold. To complete the proof, we now show that if $T \subseteq [2m]$ is a set of $k$ features satisfying $welfare(T) \geq \frac{1}{2}+\epsilon$, then $S = T \cap [m]$ is a solution to the $k$-Junta problem with $advantage(S) = \Omega(\epsilon)$. Indeed:

$$\begin{aligned}
advantage(S) &= \mathop{\mathbf{E}}_{x_S}\left[\left|\mathop{\mathbf{Pr}}_{x}[f(x)=1|x_S] - \frac{1}{2}\right|\right] \\
&= \mathop{\mathbf{E}}_{x_S}\left[\max\left(\mathop{\mathbf{Pr}}_{x}[f(x)=1|x_S], \mathop{\mathbf{Pr}}_{x}[f(x)=-1|x_S]\right)\right] - \frac{1}{2} \\
&\geq \mathop{\mathbf{E}}_{\omega_T}\left[\max\left(\mathop{\mathbf{Pr}}_{x}[f(x)=1|\omega_T], \mathop{\mathbf{Pr}}_{x}[f(x)=-1|\omega_T]\right)\right] - \frac{|T\setminus[m]|}{m} - \frac{1}{2} \\
&\geq \text{welfare}(T) - \frac{k}{m} - \frac{1}{2}
\end{aligned}$$

where the next to last inequality is a consequence of the fact that, with probability at least $1 - \frac{|T\setminus[m]|}{m}$, the feature $j^*$ is not in $T$ and therefore $\omega_T$ provides no information on $f(x)$ beyond $x_S$. □

### 3.2 $n$ players

Next, we show that the feature signaling problem is hard to approximate to within any constant independent of the number of players, assuming Conjecture 1. Specifically, for $n$ players where $n$ is a constant independent of $m$, we show that it is hard to approximate the feature signaling problem to within any constant exceeding $1/n$, and this holds for both the oracle and explicit representation models.

**Theorem 2.** *Assuming Conjecture 1, there is no $m^{o(k)}$-time, $(\frac{1}{n} + \epsilon)$-approximation algorithm for the feature signaling problem with $n$ players in the sample oracle model, for any constant $\epsilon > 0$. This holds for Monte Carlo approximation algorithms having a constant success probability.*

*Proof.* Our reduction for $n$ players generalizes that for 2 players. Specifically, Given sample access to an $c$-balanced $k'$-Junta $f : \{-1, 1\}^m \to \{0, 1\}$, with $k' = k/\log n$ and $c = \theta(k')$, we construct an instance of the $k$-feature signaling problem in the sample oracle model as follows. We let $\Omega = \{-1, 1\}^{m \log n} \times \{0, 1, \ldots, n\}^m$, and consider players $[n] = \{1, \ldots, n\}$. All players have no value for features 1 through $m \log n$. For the remaining features $j \in [m \log n + 1, m \log n + m]$, player $i$ has value 1 if $\omega_j = i$ and 0 otherwise.

The distribution $\lambda$ is constructed as follows. The first $m \log n$ features of $\omega \sim \lambda$, which we denote by $x$, are uniformly distributed in $\{-1, 1\}^{m \log n}$. We partition $x$ into sub-vectors $x_1, \ldots, x_{\log n}$, of length $m$ each. The last $m$ features, which we denote by $y$, are all set to 0, except for a single feature $i^*$ chosen uniformly at random, which is set to the integer encoded by the bit-string $f(x_1)f(x_2)\ldots f(x_{\log n})$.

Note that since $f$ is determined by some bits $S^* \subseteq [m]$ with $|S^*| = k'$, signaling the $k = k' \log n$ bits corresponding to the $S^*th$ indices of each sub-vector $x_i$ yields a welfare of 1, since those bits uniquely determine the player who values the item. We now show that any signaling algorithm with nontrivial performance must violate Corollary 1.

Indeed, consider any set $T$ of $k$ features computed by some algorithm for the feature signaling problem which runs in $m^{o(k)} = m^{o(k')}$ time. By Corollary 1 and the fact that $|T| = k' \log n = O(k')$, on some inputs $T$ will not contain more than $c$ relevant features from any sub-vector among $x_1, \ldots, x_{\log n}$. By the balance property, such a set $T$ affords no information regarding the player who values the item for sale beyond that afforded by the features $T \setminus [m \log n]$. A similar analysis to that of Theorem 1 shows that the advantage of the signaling scheme which reveals $T$ over one which randomly assigns the item to one of the $n$ players is at most the probability that $i^* \in T$, which is at most $\frac{k \log n}{m} = o(1)$, as needed. □

Finally, we note that any Monte Carlo algorithm with constant success probability can be boosted to one with exponentially small (in $k$) failure probability, as needed to violate Corollary 1.

**Corollary 2.** *Assuming Conjecture 1, there is no $m^{o(k)}$-time $(\frac{1}{n} + \epsilon)$-approximation algorithm for the feature signaling problem with $n$ players in the explicit model, for any constant $\epsilon > 0$. This holds for Monte Carlo approximation algorithms having a constant success probability.*

## 4 An Approximation Algorithm for Independent Distributions

We cast the algorithmic task of feature selection as the following optimization problem. The inputs are $\theta \in \mathbb{R}^n$, $k \in [m]$, and independent $t$-valued $n$-dimensional random vectors $\mathbf{X}_1, \dots, \mathbf{X}_m$ with $\mathbf{E}[\mathbf{X}_i] = 0$ for all $i \in [m]$. We will assume that each $\mathbf{X}_i$ is specified as $\{(p_1, v_1), \dots, (p_t, v_t)\}$ where $\mathbf{Pr}[\mathbf{X}_i = v_j] = p_j$ and $\sum_{j=1}^{t} p_j = 1$, and that basic arithmetic can be done in constant time (e.g. we can compute $p_i + p_j$ in constant time, and $\|v_i\|_\infty$ in $O(n)$ time). Given $S \subseteq [m]$ we write

$$\mathsf{value}(S) = \mathbf{E}\left[\left\|\theta + \sum_{j \in S} \mathbf{X}_j\right\|_\infty\right], \tag{1}$$

and define

$$S^* = \underset{|S|=k}{\operatorname{argmax}} \{\mathsf{value}(S)\}, \qquad \mathsf{opt} = \mathsf{value}(S^*). \tag{2}$$

For $0 < \epsilon \leq \frac{1}{2}$, we say that a subset $S \subseteq [m]$ with $|S| \leq k$ is $\epsilon$-optimal if $\mathsf{value}(S) \geq (1-\epsilon)\mathsf{opt}$; the algorithmic task is to find an $\epsilon$-optimal $k$-subset $S \subseteq [m]$ efficiently.

To see that this does in fact capture the feature selection problem where each feature is selected independently, we recall the expression for welfare$(S)$ given in (1). Setting

$$\mathbf{X}_j := \boldsymbol{v}_j - \mathbf{E}[\boldsymbol{v}_j] \quad \forall j \in [m] \quad \text{and} \quad \theta := \sum_{j=1}^{m} \mathbf{E}[\boldsymbol{v}_j],$$

and noting that the $\mathbf{X}_j$'s do indeed satisfy $\mathbf{E}[\mathbf{X}_j] = 0$, we have that for all $S \subseteq [m]$,

$$\begin{aligned} \mathsf{value}(S) &= \mathbf{E}\left[\left\|\theta + \sum_{j \in S} \mathbf{X}_j\right\|_\infty\right] = \mathbf{E}\left[\left\|\sum_{j \in S}(\boldsymbol{v}_j - \mathbf{E}[\boldsymbol{v}_j]) + \sum_{j=1}^{m} \mathbf{E}[\boldsymbol{v}_j]\right\|_\infty\right] \\ &= \mathrm{welfare}(S). \end{aligned}$$

Note that for any $S \subseteq [m]$ of cardinality $k$, the quantity $\mathsf{value}(S)$ can be computed exactly in time $O(nk \cdot t^k)$. Hence the naive algorithm which computes $\mathsf{value}(S)$ for all $\binom{m}{k}$ possible $k$-subsets $S$ runs in time $O(nk \cdot (mt)^k)$ and finds $S^*$ achieving $\mathsf{value}(S^*) = \mathsf{opt}$. As mentioned in the introduction, we will be primarily interested in the setting where the number of players $n$ is constant, as is the number of values each feature takes, and so this runtime can be written as $m^{O(k)}$. To the best of our knowledge, prior to our work there were no known improvements to this trivial algorithm even when $n = 1$ — *given* $m$ real-valued *random variables* $\mathbf{X}_1, \dots, \mathbf{X}_m$ *and* $\theta \in \mathbb{R}$, *find a* $k$-*subset* $S \subseteq [m]$ *that approximately maximizes* $\mathbf{E}[|\theta + \sum_{i \in S} \mathbf{X}_i|]$ — and even under the assumption that all the $\mathbf{X}_i$'s are two-valued (i.e. $t = 2$).

We give an algorithm that finds an $\epsilon$-optimal set $S$ of cardinality $k$, running in time $O(m) + 2^{O(k \log(k/\epsilon))}$ for all fixed values of $n$ and $t$. (In particular, this is $\mathrm{poly}(m)$ for all $\epsilon \geq 1/\mathrm{polylog}(m)$ and $k \ll \frac{\log m}{\log\log m}$.)

**Theorem 3.** *There is an algorithm* $\mathcal{A}$ *which, given as input* $0 < \epsilon \leq \frac{1}{2}$, $k \in [m]$, $\theta \in \mathbb{R}^n$, *and independent* $t$-*valued* $d$-*dimensional random vectors* $\mathbf{X}_1, \dots, \mathbf{X}_m$ *with*

$\mathbf{E}[\mathbf{X}_i] = 0$ *for all* $i \in [m]$*, runs in time* $O(mnt) + \mathrm{poly}(kt/\epsilon)^{knt}$ *and outputs a* $k$*-subset* $S \subseteq [m]$ *satisfying* $\mathsf{value}(S) \geq (1-\epsilon)\mathsf{opt}$.

The techniques we develop to establish Theorem 3 are fairly general and robust. Indeed, we obtain Theorem 3 as a special case of our most general result which we now state. Given an arbitrary norm $\|\cdot\|$ on $\mathbb{R}^n$, we may define $\mathsf{value}(\cdot)$ and $\mathsf{opt}$ with respect to $\|\cdot\|$ instead of $\|\cdot\|_\infty$, and hence also an analogous optimization problem of finding an $\epsilon$-optimal $k$-subset. Our most general result is an efficient algorithm for this abstract optimization problem for any norm $\|\cdot\|$ on $\mathbb{R}^n$:

**Theorem 4.** *Fix a norm* $\|\cdot\|$ *on* $\mathbb{R}^n$*. Given* $\epsilon > 0$ *and* $k \in [m]$*, let* $\mathcal{N} = \mathcal{N}(\epsilon, k)$ *be an* $(\epsilon/k)$*-net within the ball* $\{v \in \mathbb{R}^n \colon \|v\| \leq k^2/\epsilon\}$ *with the property that for every vector* $v$ *in the ball, its closest point in* $\mathcal{N}$ *can be found in time* $r$*. There is an algorithm* $\mathcal{A}$ *which, given as input* $\epsilon > 0$*,* $k \in [m]$*,* $\theta \in \mathbb{R}^n$*, and independent* $t$*-valued* $n$*-dimensional random vectors* $\mathbf{X}_1, \ldots, \mathbf{X}_m$ *with* $\mathbf{E}[\mathbf{X}_i] = 0$ *for all* $i \in [m]$*, runs in time* $O(mt(r+n) + nk \cdot (4\ell t)^k)$ *where* $\ell = (|\mathcal{N}|k^3 t/\epsilon)^{O(t)}$*, and outputs a* $k$*-subset* $S \subseteq [m]$ *satisfying* $\mathsf{value}(S) \geq (1-\epsilon)\mathsf{opt}$*, where* $\mathsf{value}(\cdot)$ *and* $\mathsf{opt}$ *are defined with respect to* $\|\cdot\|$.

To see that Theorem 3 does in fact follow from Theorem 4, we note that for all $B, \delta > 0$, the grid points $\mathcal{N} = \big\{(\lambda_1\delta, \ldots, \lambda_n\delta) \colon \lambda_i \in \{0, 1, \ldots, \lfloor B/\delta \rfloor\}\big\}$ is a $\delta$-net of size $(\lfloor B/\delta \rfloor + 1)^n$ within the ball $\{v \in \mathbb{R}^n \colon \|v\|_\infty \leq B\}$. Furthermore, it is clear that given any vector $v$ in the ball, its closest vector within $\mathcal{N}$ can be computed in time $O(n)$. The remainder of this section will be devoted to proving Theorem 4. The following simple fact will be useful for us:

**Fact 1** *Let* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *be independent random vectors where* $\mathbf{E}[\mathbf{X}_1] = 0$*. Then* $\mathbf{E}[\|\mathbf{X}_1 + \mathbf{X}_2\|] \geq \mathbf{E}[\|\mathbf{X}_2\|]$*. Consequently, if* $S' \supseteq S$ *then* $\mathsf{value}(S') \geq \mathsf{value}(S)$ *(and in particular, it is equivalent to maximize over all* $|S| \leq k$ *in the definition of* $\mathsf{opt}$ *in (2)).*

*Proof.* The inequality holds pointwise for every possible outcome $\theta \in \mathbb{R}^d$ of $\mathbf{X}_2$ since $\|\theta\| = \|\mathbf{E}[\mathbf{X}_1 + \theta]\| \leq \mathbf{E}[\|\mathbf{X}_1 + \theta\|]$. □

*Overview of proof.* We assume for the sake of scaling that $\max(\|\theta\|, \max_i \mathbf{E}[\|\mathbf{X}_i\|]) = 1$, and hence $\mathsf{opt} \geq 1$ by Fact 1. Thus to find an $\epsilon$-optimal set $S$, it suffices to find one achieving value at least $\mathsf{opt} - \epsilon$; for notational simplicity, we will only achieve value at least $\mathsf{opt} - O(\epsilon)$. The main idea is to modify the random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_m$ in such a way that changes $\mathsf{value}(S)$ by at most an additive $\pm O(\epsilon)$ for all $k$-subsets $S \subseteq [m]$, and yet results in a total of only $\ell$ distinct random variables $\mathbf{Y}_1, \ldots, \mathbf{Y}_\ell$ where $\ell$ is independent of $m$ (i.e. many $\mathbf{X}_i$'s are modified to become the same $\mathbf{Y}_j$). If for each $j \in [\ell]$ we let $M_j$ denote the number of $\mathbf{X}_i$'s that are modified to become $\mathbf{Y}_j$, this reduces the problem of finding an $O(\epsilon)$-optimal $k$-subset $S \subseteq [m]$ to that of finding $\lambda \in \mathbb{Z}^\ell$ that maximizes

$$\mathbf{E}\left[\left\|\theta + \sum_{i=1}^{\ell} \lambda_i \mathbf{Y}_i\right\|\right] \tag{3}$$

$$\text{subject to } \lambda_i \in \{0, 1, \ldots, M_i\} \text{ and } \sum_{i=1}^{\ell} \lambda_i = k. \tag{4}$$

Since there are at most $4^k\binom{\ell}{k}$ many $\lambda \in \mathbb{Z}^\ell$ satisfying (4), and for each such $\lambda$ the quantity (3) can be computed in time $O(nk \cdot t^k)$, the optimal $\lambda$ can be found in time $O(nk \cdot (4\ell t)^k)$.

### 4.1 Transforming the $\mathbf{X}_i$'s

Given numbers $a, b \in \mathbb{R}$ and $\epsilon > 0$, we write $a \overset{\epsilon}{\approx} b$ as shorthand for $|a - b| \leq \epsilon$. By the triangle inequality, if $a \overset{\epsilon_1}{\approx} b$ and $b \overset{\epsilon_2}{\approx} c$ then $a \overset{\epsilon_1+\epsilon_2}{\approx} c$.

**Definition 3.** *Given a parameter $B > 1$ we say that an $n$-dimensional random vector $\mathbf{X}_i$ is $B$-*bounded *if $\|\mathbf{X}_i\| \leq B$ with probability $1$.*

We begin with the following proposition which states that the $\mathbf{X}_i$'s can be modified so that all of them are $B$-bounded; we defer its proof to the full version of this paper.

**Proposition 1.** *Fix a parameter $B > 1$ and assume $\mathbf{X}_i$ is not $B$-bounded. Then there is a $B$-bounded random vector $\mathbf{X}'_i$ such that*

$$\mathbf{E}[\|\mathbf{X}'_i + \mathbf{Y}\|] \overset{4(k+1)/B}{\approx} \mathbf{E}[\|\mathbf{X}_i + \mathbf{Y}\|]$$

*for all random vectors $\mathbf{Y}$ that are independent of $\mathbf{X}_i$, $\mathbf{X}'_i$ and satisfy $\mathbf{E}[\|\mathbf{Y}\|] \leq k$. Furthermore, $\mathbf{X}'_i$ can be defined from $\mathbf{X}_i$ in time $O(nt)$.*

As a corollary of Proposition 1, for all $k$-subsets $S \subseteq [m]$ containing $i$ we have

$$\mathsf{value}(S) \overset{4(k+1)/B}{\approx} \mathbf{E}\left[\left\|\theta + \mathbf{X}'_i + \sum_{j \in S \setminus \{i\}} \mathbf{X}_j\right\|\right].$$

In words, replacing $\mathbf{X}_i$ by $\mathbf{X}'_i$ in $\mathbf{X}_1, \ldots, \mathbf{X}_n$ changes $\mathsf{value}(S)$ by at most an additive $\pm O(k/B)$ for all $k$-subsets $S \subseteq [m]$. Consequently, by the union bound, we may make *all* of $\mathbf{X}_1, \ldots, \mathbf{X}_n$ $B$-bounded and change $\mathsf{value}(S)$ by at most an additive $\pm O(k^2/B)$.

We will need a simple numerical lemma for our next modification; we defer its proof to the full version of this paper.

**Lemma 2.** *Let $p_1, \ldots, p_t \in (0, 1)$ where $\sum_{j=1}^t p_i = 1$, and $0 < \eta \leq 1$ where $1/\eta \in \mathbb{Z}$. There exist nonnegative integer multiples $p'_1, \ldots, p'_t$ of $\eta$ also summing to $1$ and satisfying $|p'_j - p_j| < \eta$ for all $j$.*

**Proposition 2.** *Fix parameters $B > 1$, $\delta > 0$, and $0 < \eta \leq 1$, where $1/\eta \in \mathbb{Z}$. Let $\mathcal{N}$ denote a $\delta$-net within the ball $\{v \in \mathbb{R}^n : \|v\| \leq B\}$, and assume that for every vector $v$ in the ball, its closest vector in the $\delta$-net $\mathcal{N}$ can be computed in time $r$. Then for any $B$-bounded $n$-dimensional random vector $\mathbf{X}_i$, there is a random vector $\mathbf{X}'_i$, dependent on $\mathbf{X}_i$, such that:*

1. *all outcomes for $\mathbf{X}'_i$ are in $\mathcal{N}$;*
2. *all outcomes for $\mathbf{X}'_i$ occur with probability equal to an integer multiple of $\eta$;*
3. $\mathbf{E}[\|\mathbf{X}'_i - \mathbf{X}_i\|] \leq \delta + 2Bt\eta$.

*Furthermore, $\mathbf{X}'_i$ can be defined from $\mathbf{X}_i$ in time $O(rt)$.*

*Proof.* Let $\mathbf{X}_i \equiv \{(p_1, v_1), \ldots, (p_t, v_t)\}$ (i.e. $\mathbf{Pr}[\mathbf{X}_i = v_j] = p_j$ and $\sum_{j=1}^t p_j = 1$). We first consider $\mathbf{X}^*_i = \{(p_1, v^*_1), \ldots, (p_t, v^*_t)\}$, where $v^*_j$ is the vector in $\mathcal{N}$ closest to $v_j$, coupled to $\mathbf{X}_i$ in such a way that $\mathbf{Pr}[\mathbf{X}^*_i = v^*_j \mid \mathbf{X}_i = v_j] = 1$. Since

$$\mathbf{E}[\|\mathbf{X}^*_i - \mathbf{X}_i\|] = \sum_{j=1}^{t} p_j \cdot \|v^*_j - v_j\| \leq \delta$$

and all outcomes of $\mathbf{X}^*_i$ are in $\mathcal{N}$ (i.e. satisfying (1)), it remains to show how to achieve (2) while incurring error at most $2Bt\eta$ in (3). By Lemma 2 there exist nonnegative integer multiples $p'_1, \ldots, p'_t$ of $\eta$, summing to 1 and satisfying $|p'_j - p_j| < \eta$ for all $j$ (and it is straightforward to verify that $p'_1, \ldots, p'_t$ can be computed from $p_1, \ldots, p_t$ and $\eta$ in time $O(t)$). We can then define $\mathbf{X}'_i$ by $\mathbf{Pr}[\mathbf{X}'_i = v^*_j] = p'_j$, coupled to $\mathbf{X}^*$ in such a way that $\mathbf{Pr}[\mathbf{X}^*_i = \mathbf{X}'_i = v^*_j] = \min(p_j, p'_j)$ for all $j \in [t]$. It is clear then that

$$\mathbf{Pr}[\mathbf{X}'_i \neq \mathbf{X}^*_i] \leq \sum_{j=1}^{t} |p'_j - p_j| \leq t\eta,$$

and that whenever $\mathbf{X}'_i \neq \mathbf{X}^*_i$ we at least have $\|\mathbf{X}'_i - \mathbf{X}^*_i\| \leq 2B$ by the $B$-boundedness of $\mathbf{X}^*_i$ and $\mathbf{X}_i$. The lemma follows.

**Proof of Theorem 4** Applying Proposition 1 with $B := k^2/\epsilon$, we may assume that $\mathbf{X}_1, \ldots, \mathbf{X}_m$ are all $B$-bounded — this modification can be carried out in time $O(mnt)$, and changes $\mathsf{value}(S)$ by at most an additive $\pm O(\epsilon)$ for all $k$-subsets $S \subseteq [m]$. Next, applying Proposition 2 with $\delta := \epsilon/k$ and $\eta$ any number in $[\epsilon/(2kBt), \epsilon/(kBt)]$ such that $1/\eta \in \mathbb{Z}$, there is an algorithm which runs in time $O(mrt)$ (i.e. $O(rt)$ for each $\mathbf{X}_i$) and outputs $\mathbf{X}'_1, \ldots, \mathbf{X}'_m$ satisfying

$$\begin{aligned}\left|\mathsf{value}(S) - \mathbf{E}\left[\left\|\theta + \sum_{i \in S} \mathbf{X}'_i\right\|_\infty\right]\right| &\leq \mathbf{E}\left[\left\|\sum_{i \in S} \mathbf{X}'_i - \mathbf{X}_i\right\|\right] \\ &\leq \sum_{i \in S} \mathbf{E}[\|\mathbf{X}'_i - \mathbf{X}_i\|] \ \leq \ k(\delta + 2Bt\eta) = O(\epsilon)\end{aligned}$$

for all $k$-subsets $S \subseteq [m]$. Furthermore, by Proposition 2 each $\mathbf{X}'_i$ is of the form $\{(p_1, v_1), \ldots, (p_t, v_t)\}$ where every $p_i$ is an integer multiple of $\eta$, and every $v_i$ is in $\mathcal{N}$. It follows that there are in fact at most

$$\ell \leq \binom{|\mathcal{N}| \cdot \eta^{-1}}{t} = (|\mathcal{N}| k^3 t/\epsilon)^{O(t)}$$

many distinct random variables $\mathbf{Y}_1, \ldots, \mathbf{Y}_\ell$ in the multiset $\{\mathbf{X}'_1, \ldots, \mathbf{X}'_m\}$. Letting $M_i$ denote the multiplicity of $\mathbf{Y}_i$ in $\{\mathbf{X}'_1, \ldots, \mathbf{X}'_n\}$, we have reduced the problem of finding an $O(\epsilon)$-optimal $k$-subset $S \subseteq [m]$ to that of finding $\lambda \in \mathbb{Z}^\ell$ that maximizes

$$\mathbf{E}\left[\left\|\theta + \sum_{i=1}^{\ell} \lambda_i \mathbf{Y}_i\right\|\right] \tag{5}$$

$$\text{subject to } \lambda_i \in \{0, 1, \ldots, M_i\} \text{ and } \sum_{i=1}^{\ell} \lambda_i = k. \tag{6}$$

Since there are at most $4^k \binom{\ell}{k}$ many $\lambda \in \mathbb{Z}^\ell$ satisfying (6), and for each such $\lambda$ the quantity (5) can be computed in time $O(nk \cdot t^k)$, the optimal $\lambda$ can be found in time $O(nk \cdot (4\ell t)^k)$.

## References

1. Michael Alekhnovich, Mark Braverman, Vitaly Feldman, Adam R. Klivans, and Toniann Pitassi. Learnability and automatizability. In *FOCS*, pages 621–630, 2004. 1
2. N. Alon, M. Feldman, I. Gamzu, and M. Tennenholtz. The asymmetric matrix partition problem. In *Conference on Web and Internet Economics (WINE)*, 2013. 1
3. A. Blum. Relevant examples and relevant features: Thoughts from computational learning theory, 1994. In AAAI Fall Symposium on 'Relevance'. 1
4. A. Blum. Open problem: Learning a function of $r$ relevant variables. In *Proceedings of COLT*, pages 731–733, 2003. 1, 3, 1
5. A. Blum. Tutorial on Machine Learning Theory given at FOCS '03, 2003. Available at http://www.cs.cmu.edu/~avrim/Talks/FOCS03/. 3, 1
6. A. Blum, M. Furst, J. Jackson, M. Kearns, Y. Mansour, and S. Rudich. Weakly learning DNF and characterizing statistical query learning using Fourier analysis. In *Proceedings of STOC*, pages 253–262, 1994. 1
7. A. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245–271, 1997. 1
8. S. Dughmi, N. Immorlica, and A. Roth. Constrained signaling in auction design. In *ACM Symposium on Discrete Algorithms (SODA)*, 2014. 1
9. Y. Emek, M. Feldman, I. Gamzu, R. Paes-Leme, and M. Tennenholtz. Signaling schemes for revenue maximization. In *ACM Conference on Electronic Commerce (EC)*, 2012. 1
10. V. Feldman, H. Lee, and R. Servedio. Lower bounds and hardness amplification for learning shallow monotone formulas. In *Journal of Machine Learning Research - COLT Proceedings*, volume 19, pages 273–292, 2011. 1
11. Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. On agnostic learning of parities, monomials, and halfspaces. *SIAM J. Comput.*, 39(2):606–645, 2009. 1
12. Vitaly Feldman, Pravesh Kothari, and Jan Vondrák. Nearly tight bounds on $\ell_1$ approximation of self-bounding functions. *CoRR*, abs/1404.4702, 2014. 1
13. Mingyu Guo and Argyrios Deligkas. Revenue maximization via hiding item attributes. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 157–163. AAAI Press, 2013. 1
14. P. Milgrom and R.J. Weber. A theory of auctions and competitive bidding. *Econometrica*, 50:1089–1122, 1982. 1
15. Peter B. Miltersen and Or Sheffet. Send mixed signals: earn more, work less. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 234–247. ACM, 2012. 1
16. E. Mossel, R. O'Donnell, and R. Servedio. Learning functions of $k$ relevant variables. *Journal of Computer & System Sciences*, 69(3):421–434, 2004. Previously published as "Learning juntas". 1
17. Gregory Valiant. Finding correlations in subquadratic time, with applications to learning parities and juntas. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 11–20. IEEE, 2012. 1