1 Introduction

Designing machine learning algorithms that are robust to noise in training data has lately been a subject of intense research. A large body of work addresses stochastic noise [12, 7], while another one studies adversarial noise [11, 2] in which errors are introduced by an adversary with the explicit purpose of sabotaging the algorithm. This is often too pessimistic, and leads to negative results. The literature on game theory and mechanism design offers an interesting middle ground: strategic noise. In this paradigm, training data is provided by strategic sources that purposefully introduce errors for maximizing their own benefit. This is less pessimistic than adversarial noise where the errors are introduced for simply harming the algorithm.

There is a growing body of research on designing machine learning algorithms that are robust to strategic noise. This research can be categorized using three key axes: i) manipulable information, ii) goal of the agents, and iii) use of payments and incentive guarantee. On the first axis, most papers assume that independent variables (feature vectors) are public information, and dependent variables (labels) are private, manipulable information [6, 15, 18], though some papers also design algorithms robust to strategic feature vectors [8]. On the second axis, one body of research focuses on agents motivated by privacy concerns (with a tradeoff between accuracy and privacy) [5, 3], while another one focuses on agents who want the algorithm to make accurate assessment on their own sample, even if this reduced the overall accuracy. Such strategic manipulations have been studied for estimation [4], classification [13, 14, 15], and regression [19, 6] problems. On the third axis, the research differs on whether monetary payments to agents are allowed [3], and on how strongly to guarantee truthful reporting (the stronger “strategyproofness” requirement [18, 19, 15] versus the weaker Bayes-Nash equilibrium requirement [9, 5]).

In this paper, we focus on the problem of linear regression, i.e., fitting a hyperplane through given data, which is studied extensively in statistics and machine learning. We consider agents who can manipulate their dependent variables in order to increase the algorithm’s accuracy on their own samples, and design strategyproof mechanisms without payments.

Our contributions. In this work, we extend results about two existing families of strategyproof mechanisms, introduce a novel family of strategyproof mechanisms, and along the way, provide two useful (albeit non-constructive) characterizations of strategyproof mechanisms for linear regression. More specifically, Dekel et al. [6] show that the empirical risk minimization (ERM) with the $L_1$ loss (in short, $L_1$-ERM), coupled with a specific tie-breaking rule, is strategyproof. We extend this result and show that adding arbitrary agent-specific weights and convex regularization to the risk function preserves strategyproofness. Moreover, Perote and Perote-Peña [19] introduce the family of Clockwise Repeated Median (CRM) mechanisms, parametrized by two subsets of agents, $S$ and $S'$. They claim that CRM is strategyproof when $S \subseteq S'$ or $S \cap S' = \emptyset$. We identify a serious bug in their proof, and refute their claim by producing counterexamples violating strategyproofness. We then reclaim strategyproofness under more restrictive conditions on $S$ and $S'$. In an effort to provide a short and algebraic proof of strategyproofness (as opposed to the long and geometric proof by Perote and Perote-Peña [19]), we discover two novel characterizations of strategyproof mechanisms, which are also useful for “sensitivity analysis”. Finally, we introduce a novel family of strategyproof mechanisms by imposing a stronger condition known as impartiality, which requires that the out-
come for agent \( i \) be independent of her report. While impartiality trivially implies strategyproofness, it is apriori unclear if non-trivial impartial mechanisms even exist for linear regression. We provide a large family of impartial mechanisms, and prove that our family is complete for two dimensions (i.e., fitting a line through points in \( \mathbb{R}^2 \)).

2 Model

Let \( N = [n] \) be the set of agents. In a collection of data points \( \mathcal{D} = (x_i, y_i)_{i \in N} \), the independent variables \( x = (x_i)_{i \in N} \) is public information, where \( x_i \in \mathbb{R}^d \) for \( d \in \mathbb{N} \), and the dependent variable \( y_i \in \mathbb{R} \) is private to agent \( i \). For fixed public information \( x \), a linear regression mechanism \( M \) takes as input the reported private information \( \tilde{y} = (\tilde{y}_i)_{i \in N} \), and returns a hyperplane in the form of its normal vector \( M(\tilde{y}) \in \mathbb{R}^d \). The outcome for agent \( i \) is given by \( \tilde{y}_i(M(\tilde{y})) = M(\tilde{y})^T \alpha_i \). As the public information \( x \) is non-manipulable, the mechanisms, functions, and constants we define throughout the paper can depend on \( x \), which we omit from notation.

Each agent \( i \) has single peaked preferences over her outcome, denoted \( \succeq_i \), with peak at \( y_i \); \( \forall a, b \in \mathbb{R}, (b \geq a > y_i) \lor (b \leq a < y_i) \Rightarrow y_i \succeq_i a \succeq_i b \). The agent reports \( \tilde{y}_i \) in order to achieve the most preferred outcome. Mechanism \( M \) is called strategyproof if for each agent \( i \), reporting \( \tilde{y}_i = y_i \) results in the most preferred outcome, irrespective of the reports of the other agents.

3 Strategyproof Linear Regression

There are two known (claimed) approaches to designing strategyproof linear regression mechanisms.

**L1-ERM:** Dekel et al. [6] show that using empirical risk minimization (ERM) with the \( L_1 \) loss, and breaking ties by minimizing the \( L_2 \) norm of the regressor, yields a strategyproof mechanism. Their algorithm is a two step procedure. Let \( r(\beta) = \sum_{i \in N} |\tilde{y}_i - \beta^T \tilde{x}_i| \) be the empirical \( L_1 \) risk. Then: i) compute \( r^* \leftarrow \min \beta r(\beta) \), and ii) return \( \beta^* \leftarrow \arg \min \beta, r(\beta) = r^* \|\beta\|_2 \). In fact, they establish group-strategyproofness (i.e., even groups of agents cannot gain by misreporting collectively) for possibly nonlinear regression if the minimization is over a convex family of regression functions. We extend their result by showing that adding agent-specific weights independent of the private information and any convex regularization to the risk function preserves group-strategyproofness. For simplicity, we state the result only for strategyproofness and for linear regression. Our proof essentially follows the line of argument presented by Dekel et al. [6].

**Theorem 1.** Let \( w_i \in \mathbb{R} \) denote the only weight for agent \( i \in N \), and let \( h \) be a convex regularizer. Define the risk function \( r(\beta) = \sum_{i \in N} w_i \cdot |\tilde{y}_i - \beta^T \tilde{x}_i| + h(\beta) \). Then, the following mechanism is strategyproof: i) compute \( r^* \leftarrow \min \beta r(\beta) \), and ii) return \( \beta^* \leftarrow \arg \min \beta, r(\beta) = r^* \|\beta\|_2 \).

**Clockwise Repeated Median (CRM):** The second approach is suggested by Perote and Perote-Peña [19] for the special case of 2D linear regression, i.e., for fitting a line through points on a plane. They introduce the family of CRM mechanisms parametrized by two sets of agents \( S, S' \subseteq N \). Informally, the \((S, S')\)-CRM mechanism operates as follows. First, for each \( i \in S \), the mechanism finds the "clockwise angle"\(^2\) to each point \( j \in S' \), and keeps the median of those. Then, it picks \( i^* \in S \) by taking the median of the median clockwise angles, and draws the line passing through \( i^* \) and \( j^* \), where \( j^* \in S' \) produces the median of all clockwise angles from \( i^* \) to points in \( S' \). For a more rigorous definition, we refer the reader to the original paper [19]. Perote and Perote-Peña [19] claim that \((S, S')\)-CRM is strategyproof when \( S \subseteq S' \) or \( S \cap S' = \emptyset \). Their proof is long, geometric, and difficult to understand. We identify a serious bug in their proof, and refute their overall claim by producing an example with \( S \subseteq S' \) and another example with \( S \cap S' = \emptyset \) such that the corresponding \((S, S')\)-CRMs violate strategyproofness. Fortunately, we manage to reclaim strategyproofness for three interesting cases given by stricter constraints on \( S \) and \( S' \). Our proof is short, algebraic, and uses two novel characterizations of strategyproof mechanisms that we introduce (described later).

**Theorem 2.** \((S, S')\)-CRM is strategyproof when i) \( S = S' \), ii) \(|S| = 1 \) or \(|S'| = 1 \), or iii) \( S \) and \( S' \) are separable, i.e., \( \max_{i \in S} x_i < \min_{j \in S'} x_j \) or \( \min_{i \in S} x_i > \max_{j \in S'} x_j \).

---

1Note that a strictly convex regularizer \( h \) can be used to eliminate the need for tie-breaking as the first step now yields a unique minimizer.

2This belongs to \([0, 2\pi]\), where 0 means \( j \) is directly above \( i \), \( \pi/2 \) means \( j \) is exactly to the left of \( i \), etc.
The first case is a variant of the repeated median estimator of Siegel [20], and the third case precisely coincides with the well-known family of resistant line methods [10] from the statistics literature. Two popular examples of resistant line methods are the Brown-Mood [1] and Tukey estimators [21].

**Impartial mechanisms.** A mechanism is called impartial if the outcome for each agent \(i\) (in our case, \(M(\mathbf{y})^T x_i\)) is independent of the report of agent \(i\) (in our case, \(y_i\)). It is apriori unclear if non-trivial impartial mechanisms even exist for linear regression. We provide a large family of non-trivial impartial (thus strategyproof) mechanisms, and show that it characterizes impartial mechanisms in 2D (i.e., for fitting a line through points on a plane).

**Theorem 3.** For functions \(\{g_i : \mathbb{R} \to \mathbb{R}^d\}_{i \in N}\) and constant \(c \in \mathbb{R}\), mechanism \(M\) under which \(\hat{y}_i(M(\mathbf{y})) = \sum_{j \in N \setminus \{i\}} (g_j(\mathbf{y}_j), x_i - x_j) + c\) for each \(i \in N\) is a valid and impartial mechanism for linear regression, where \(\langle \cdot, \cdot \rangle\) is the inner product. This is the set of all impartial mechanisms for linear regression when \(d = 1\).

**Characterizations of strategyproof mechanisms.** We propose two characterizations of strategyproof mechanisms for linear regression. They build upon the characterization of strategyproof mechanisms for aggregating real-valued reports by Moulin [16]. We first need the following definition.

**Definition 1 (Locally Constant Function).** For \(A, B \subseteq \mathbb{R}\), function \(f : A \to B\) is called locally constant at \(x \in A\) if there exists \(\epsilon > 0\) such that \(f(x') = f(x)\) for all \(x' \in [x - \epsilon, x + \epsilon]\).

**Theorem 4.** Mechanism \(M\) for linear regression is strategyproof if and only if one of the following two conditions hold.

1. For every \(\mathbf{y}_{-i} \in \mathbb{R}^{n-1}\) and \(i \in N\), there exist \(\ell_i, h_i \in \mathbb{R} \cup \{-\infty, \infty\}\) such that for every \(\hat{y}_i \in \mathbb{R}\), we have \(\hat{y}_i(M(\mathbf{y})) = \text{med}(\hat{y}_i, \ell_i, h_i)\).

2. For every \(\mathbf{y}_{-i} \in \mathbb{R}^{n-1}\) and \(i \in N\), the function \(f_i(\cdot) = \hat{y}_i(M(\cdot, \mathbf{y}_{-i}))\) is continuous, and for every \(\hat{y}_i \in \mathbb{R}\), either \(f_i(\hat{y}_i) = \hat{y}_i\) or \(f_i\) is locally constant at \(\hat{y}_i\).

The second condition provides a useful way for checking strategyproofness of a proposed mechanism and we use it to establish strategyproofness of CRM mechanisms in Theorem 2. In contrast, the first condition provides a simple analytical form. Note that \(\ell_i\) and \(h_i\) define the influence region of agent \(i\); all other reports fixed, if the agent reports \(\hat{y}_i \in [\ell_i, h_i]\), she becomes a dictator \(\hat{y}_i = \hat{y}\), and if she reports a value less than \(\ell_i\) (resp. higher than \(h_i\)), her outcome remains fixed at \(\ell_i\) (resp. \(h_i\)). This, in particular, implies a weaker property: an agent on one side of the hyperplane cannot change her outcome as long as she remains on the same side. This weaker property was observed for \(L_1\)-ERM by Narula and Wellington [17], who argued that this makes \(L_1\)-ERM robust to fluctuations in the dependent variables. They computed the influence regions in \(L_1\)-ERM on certain examples, and termed it “sensitivity analysis”. As a corollary of our characterization, we provide an algorithm to compute exactly the influence bounds using \(O(d \cdot n \cdot d^2 + 1 \cdot \log n)\) calls to the mechanism in question. We omit the details due to lack of space.

### 4 Ongoing & Future Work

The results presented here are part of ongoing work, in which we are exploring a number of open questions. First, while the \(L_1\)-ERM is group-strategyproof, we do not know if (though we conjecture that) CRM mechanisms are group-strategyproof. Second, unlike the \(L_1\)-ERM and the impartial mechanisms, the CRM mechanisms are only defined for 2D. We are working on extending them to higher dimensions; however, this seems tricky because the “angle” in higher dimensions is a vector, so the “single-parameter” result of Moulin [16] cannot be applied directly. Finally, the most ambitious challenge is to find a constructive characterization of strategyproof mechanisms for linear regression. We hope to be able to derive a characterization using residuals, which seem to play a key role in all the existing families of strategyproof mechanisms. Such a characterization may allow us to identify the most efficient strategyproof mechanism, where efficiency is measured by the mean squared error.
References


