Should Humans Lie to Machines? The Incentive Compatibility of Lasso and GLM Structured Sparsity Estimators

Mehmet Caner* Kfir Eliaz†

January 25, 2024

Abstract

We consider situations where a user feeds her attributes to a machine learning method that tries to predict her best option based on a random sample of other users. The predictor is incentive-compatible if the user has no incentive to misreport her covariates. Focusing on the popular Lasso estimation technique, we borrow tools from high-dimensional statistics to characterize sufficient conditions that ensure that Lasso is incentive compatible in the asymptotic case. We extend our results to a new nonlinear machine learning technique, Generalized Linear Model Structured Sparsity estimators. Our results show that incentive compatibility is achieved if the tuning parameter is kept above some threshold in the case of asymptotics.

Keywords: Moment oracle inequality, machine learning, overfit

1 Introduction

Rapid advances in machine learning methods for analyzing big data have given rise to automated systems that employ these methods to predict the best fitting outcomes for users based on their personal characteristics. For example, many online platforms try to predict which content - a song, a video, a post, or an article - is the best fit for each user. Medical providers have also begun using machine learning techniques to automate check-ups and test appointments for patients based on their medical history. Typically, these

*North Carolina State University, Nelson Hall, Department of Economics, NC 27695. Email:mcaner@ncsu.edu.

†School of Economics, Tel-Aviv University and David Eccles School of Business, the University of Utah. Email: kfire@tauex.tau.ac.il.

automated systems use data from past users to estimate a model that relates the best fit for a user (such as the most preferred content or the appropriate medical test) to her characteristics. These estimates are then applied to a new user's characteristics, which she discloses either actively or passively via her past online behavior (which may be reflected in her cookies or collected by her browser). Given the growing interaction of users with such automated systems, it is only natural to ask whether a user should truthfully disclose her characteristics?

If the information the user discloses is also used to exploit her (say, by providing it to third parties for advertising or price discrimination), then the user has an obvious reason not to reveal her private information. The question is whether special features of some popular machine learning methods introduce an incentive to misreport one's personal characteristics even when this information will be used *solely* for predicting her best outcome? This question is of crucial importance: If individuals submit false reports to systems that rely on these reports for estimation and predictions, then the conclusions drawn from such estimates and predictions will be wrong and may lead to quite undesirable outcomes (e.g., think of an automated medical platform that schedules tests for patients based on false reports on attributes such as smoking, drinking and physical exercise).

To address the above question, we begin by considering a stylized environment where each user i's ideal option is a function f of her privately observed attributes $Z_i = (Z_{i,1}, ..., Z_{i,p})'$. A "statistician", who represents some automated prediction platform, has a sample of the attributes of n users and *noisy* observations on their ideal options. For instance, suppose $f(Z_i)$ is the optimal dosage of some medication when taken immediately at the onset of symptoms, conditional on the patient's medical history Z_i, but the statistician observes the dosage that was given after some delay. Similarly, $f(Z_i)$ may be the mix of news and reality shows that a user with attributes Z_i actually watches, but the statistician observes only self reports by a user who may have forgotten exactly what she watched.

Since $f(Z_i)$ may have a complicated non-linear form, the statistician uses her sample to estimate a linear approximation of this function. To this end, she uses some transformation of Z_i (e.g., a polynomial transformation or splines), denoted X_i, to estimate the linear approximation to the function f by computing an estimate $X_i'\hat{\beta}$. The statistician wishes to apply these estimates to predict the ideal option of a new user, $n+1$, whose true attributes Z_{n+1} (and hence, true *transformed* attributes X_{n+1}) are not observed by the statistician. This new user must decide what vector of transformed attributes $R(X_{n+1})$ (which may *differ* from the truth) to report to the statistician. When making this decision, the new user does not know the vector of X_i's nor the values of the estimate $\hat{\beta}$, but he does know the statistician's estimation procedure.

The statistician then plugs the new user's reported attributes into the estimated function and gives the user the option $R(X_{n+1})'\hat{\beta}$, which is the statistician's estimate of the user's ideal option based on her report. The new user's expected loss from a report $R(X_{n+1})$ is given by the mean square error between her expectation of the linearly approximated ideal option $X_{n+1}'\beta_0$ (where β_0 is the true coefficient in the linear approximation), and her assigned option $R(X_{n+1})'\hat{\beta}$. That is, the function f is also too complicated for the new user to consider, and therefore she considers only its linear approximation. The statistician's estimator is *incentive − compatible*, if the new user has no incentive to deviate from truthful reporting whatever her attributes are, i.e., if for every possible value of β_0 and X_{n+1}, the expected value of $(X_{n+1}'\beta_0 - R(X_{n+1})'\hat{\beta})^2$ is minimized at the truth $R(X_{n+1}) = X_{n+1}$, where the expectation is taken with respect to the statistician's sample.

Intuition suggests that an individual cannot benefit from lying to a procedure that is meant to predict the best outcome for her. To counter this intuition, Eliaz and Spiegler (2019), and Eliaz and Spiegler (2020) use the above framework to illustrate that a user may have a strict incentive to lie about her attributes when the prediction is based on a linear regression that penalizes non-zero estimated coefficients. The rough intuition is that

the user believes that despite the statistician's good intentions, these estimation techniques lead to distortions, which she tries to undo by lying. For instance, she may be concerned that the estimator will admit too many irrelevant attributes, and hence, she reports a zero value for these attributes (see Eliaz and Spiegler (2019), and Eliaz and Spiegler (2020) for more details). However, these papers focus on particular examples in which attributes are *binary*, the statistician has the *same* (fixed) finite number of observations on each possible combination of attribute values, and the penalty parameter is *fixed* and does *not* adjust to the sample size. That is, these papers only raise the problem of incentive compatibility but do not provide an econometric solution. Hence, they leave open the following important question: For a general environment, are there conditions ensuring that a penalized regression model is incentive compatible in large samples?

Answering this question can potentially allow platforms, like those discussed above, to use machine-learning methods to predict users' most preferred options without worrying that their data is "contaminated" by non-truthful users. Put bluntly, estimates and predictions made by methods that are *not* incentive-compatible are possibly unreliable since they may be based on false data.

This paper addresses the above open question by first focusing on the most popular form of penalized regressions - the *Lasso* estimator. Borrowing tools from high-dimensional statistics, we establish sufficient conditions for incentive compatibility of the Lasso estimator with sufficiently large samples. In the special case of asymptotics these sufficient conditions simplify. We show that to achieve incentive compatibility, the tuning parameter must be *large* enough (i.e., it must remain above some threshold as sample size increases) so as to avoid overfitting, which is the main reason why a user may want to lie. This potential to lie implies that the standard way of choosing small enough tuning parameters to ensure consistency may violate incentive compatibility. In Appendix D we provide simulation results that illustrate how the tuning parameter can be chosen in practice to ensure incentive

compatibility. Incentive compatibility may therefore be viewed as an additional important property that should be imposed on estimators on top of consistency and unbiasedness.

We also extend our analysis to an environment in which the statistician uses a general-convex penalty function in high dimensions. This is different than our first model and estimation, since the Lasso estimator uses a least squares based loss and a l_1 norm based penalty. We assume that the statistician uses the Generalized Linear Models with Structured Sparsity Estimators that was recently proposed in Caner (2023), which is a new nonlinear machine learning technique. We propose a definition of incentive compatibility, and show that these estimators are also incentive compatible but require larger tuning parameter then in the Lasso case.

The motivation to focus first on the Lasso estimator stems from the fact that this estimator is the benchmark among all high dimensional statistical estimators that predict large scale models when the number of regressors exceeds the sample size. Following its original proposal by Tibshirani (1996), econometricians and statisticians have used Lasso-based estimators to push the boundaries of economics and finance. One of the most critical issues facing these Lasso type estimators is post-inference after estimation and model selection, which require uniformly valid confidence intervals. In a seminal series of papers, Belloni et al. (2012) and Belloni et al. (2014) solved these issues by introducing the idea of "partialling out" the regressors. A different, but complementary approach, via debiasing-desparsifying is proposed by van de Geer et al. (2014). Caner and Kock (2018) extended the debiasing of van de Geer et al. (2014) to heteroskedastic-non-sub-Gaussian data with strong oracle optimality properties, thereby proposing a high dimensional estimator that is robust to heteroskedasticity, and with uniformly valid confidence intervals. Lasso-based debiasing are used in panel data models (see, e.g., Chernozhukov et al. (2018), Kock (2016), Kock and Tang (2019)) and for addressing quantile treatment effects and text analysis (see, e.g., Chiang and Sasaki (2019) and Chiang (2020)).

The concern that statistical procedures such as estimation, forecasting and classification are vulnerable to manipulation, has been the subject of some recent papers in the computer science literature. In contrast to us, this literature assumes there is an explicit conflict of interest between the statistician and the data providers - either because the latter are concerned about their privacy, they have to incur a cost to provide a precise report, or they have a different objective than the statistician. These papers analyze the Nash equilibria of a game where users submit private values that are used for estimation/classification, and propose incentive schemes that induce truthful reporting. Some notable works in this literature include Cai et al. (2015), Cummings et al. (2015), Dekel et al. (2010), Gao et al. (2015), Hardt et al. (2016), Meir et al. (2012) and Perte and Perote-Pena (2004). *None* of these papers consider penalized regression methods, and *none* of them characterize conditions guaranteeing incentive compatibility of regression techniques when the statistician and users have *aligned interests* (as is the case in our model).

A different take on the effect of data usage on incentives is explored in Liang and Madsen (2023). They consider an agent who chooses a costly effort level in order to affect the market's perception of his private type. Their innovation is that the market's perception is based both on observation of the outcome (which depends stochastically on effort) and on a forecast of the agent's type given a set of covariates. In contrast to our work, the agent does not affect what covariates are used and the question is how the (exogenously given) covariates used by the forecast affect the agent's choice of effort.

The remainder of the paper is organized as follows. Section 2 considers the incentive compatibility of a Lasso estimator of a high dimensional linear approximation to a nonlinear model. Section 3 considers the incentive compatibility of a nonlinear model with GLM structured sparsity estimators. Section 4 concludes. Supplementary material is provides in the appendices. Appendix A contains the proofs of the results on the Lasso estimator when the number of regressors (p) exceed the number of observations (n). Appendix B

addresses the case of $p \leq n$ and shows how to extend our Lasso results when we relax our assumption on the signal to noise ratio. Appendix C contains the proofs for the GLM structured sparsity estimators, and Appendix D presents simulations.

2 A High Dimensional Linear Approximation to Nonlinear Model

We begin this section by describing our theoretical framework in the case of a nonlinear model approximated by a high dimensional linear model. We then specify our assumptions on the statistician's data and introduce our notion of incentive-compatibility.

Throughout the paper we will use the following notational conventions. Let $\mathbf{R}$ represent the real line. For any vector $\nu \in \mathbf{R}^d$, let $\|\nu\|_1, \|\nu\|_2, \|\nu\|_\infty$ denote its l_1, l_2, l_∞ norm respectively, and $\|\nu\|_0$ be the l_0 norm, which means the total number of nonzero entries. For a set $S \subseteq \{1, 2, \cdots, d\}$, let $|S| = s$ be the cardinality of the set. Let ν_S be the modified ν such that we put 0 when the index does not belong to S (i.e., say $S = \{1, 2, 6\}$ for a 10×1 vector ν, this means that ν is modified such that now all elements are zero except elements $1, 2, 6$). Let $\|A\|_{l_1}$ be the maximum absolute column-sum norm of a matrix of dimensions $m \times l$, i.e., $\|A\|_{l_1} = \max_{1 \leq k \leq l} \sum_{i=1}^{m} |A_{ik}|$ which is also called the induced l_1 norm of A. Let $\|A\|_{l_\infty} := \max_{1 \leq i \leq m} \sum_{k=1}^{l} |A_{ik}|$ which is the maximum absolute row sum norm.

Our environment consists of users who are characterized by a set of personal characteristics. For instance, in the context of medical decision making, a characteristic can represent a risk factor (obesity, smoking, etc.). For each user i, these characteristics are modeled as a vector of p explanatory variables, $Z_i = (Z_{i,1}, ..., Z_{i,p})$ drawn from some distribution over a subset of $\mathbf{R}^p$. These attributes determine the ideal option for a user according to the function $f(Z_i)$. This function applies to all users, who differ only in the (privately observed) realized values of their characteristics.

A *statistician* (representing the automated prediction systems described in the introduction) has *private* access to a sample of n observations. Each observation $i = 1, ..., n$ consists of the true attributes Z_i of user i and a noisy signal y_i of that user's ideal option:

$$y_i = f(Z_i) + u_i, \tag{1}$$

where u_i is random noise that is drawn $i.i.d$ from some distribution with zero mean. Since $f(Z_i)$ can be a complicated non-linear model, the statistician estimates a high-dimensional linear approximation of the true model. To do this, the statistician forms a $p \times 1$ vector X_i of the attributes Z_i, which can take the form of $X_i = l(Z_i)$ where $l(Z_i)$ is a polynomial or spline transformation of Z_i (see Belloni and Chernozhukov (2009)) to estimate the linear model,

$$y_i = X_i'\beta_0 + r_i + u_i, \tag{2}$$

where β_0 is the sparse $p \times 1$ vector of true parameters and $r_i := f(Z_i) - X_i'\beta_0$ is the approximation error. The X_i' s are i.i.d. across i and will be discussed in detail in Assumption 2 in the next subsection. The first element of the regressors is the intercept. The statistician does not know β_0 and needs to estimate it using his sample.

This type of linear approximation is used to model earnings regression, and cross country growth regressions in Belloni and Chernozhukov (2009). Note that the true vector of parameters β_0 represents the outcome of an oracle problem (ideal risk minimization) that balances the approximation error with a scaled variance. We define the average square error from approximating $f(Z_i)$ by $X_i'\beta_0$ as

$$c_s^2 := \frac{1}{n}\sum_{i=1}^{n} r_i^2.$$

Belloni and Chernozhukov (2009) show that to solve the oracle problem, we need to set

$$c_s^2 = O(s_0/n). \tag{3}$$

This is a deterministic rate, and the stochastic case is shown in Theorem 11.1 of Gyorfi et al. (2010). While the model and equations in Belloni and Chernozhukov (2009) use fixed regressors, the authors point out that random sampling is a subcase of their approach (see p.128).

2.1 The Lasso Estimator

Using her (privately observed) sample, the statistician estimates the function f, or equivalently, she estimates the coefficients $\beta_{0,1}, ..., \beta_{0,p}$. When $p > n$, the least squares estimator is infeasible due to singularity of the empirical Gram matrix. Hence, the statistician uses Lasso, the penalized regression procedure that assigns costs to including explanatory variables in the regression. Specifically, statistician solves the following minimization problem

$$\hat{\beta} = \underset{\beta \in \mathbf{R}^p}{\operatorname{argmin}} \left[\frac{\sum_{i=1}^{n} (y_i - X_i'\beta)^2}{n} + 2\lambda_n \|\beta\|_1 \right], \tag{4}$$

where $\lambda_n > 0$ is the tuning parameter that is local to zero (an explicit expression for the sequence λ_n is given in (A.16)). Note that even when $p \leq n$, Lasso is used mainly for model selection and estimation. We analyze both cases in this article. However, the main text and Theorems 1-2 are written for the case of $p > n$. Appendix B contains all theorems related to $p \leq n$.

The Lasso estimator has desirable properties when there is a high dimensional linear sparse model. It can mimic an oracle (i.e. a Maximum Likelihood Estimate (MLE) with knowledge of true nonzero parameters in β_0) in moments up to a logarithmic order in the number of parameters:

$$E\|\beta_{ora} - \beta_0\|_1 = O(s_0/n^{1/2}),$$

with β_{ora} as the oracle-MLE. In particular, Jankova and van de Geer (2018) shows that

$$E\|\hat{\beta} - \beta_0\|_1 = O(s_0\sqrt{lnp/n}).$$

where $\hat{\beta}$ is the Lasso estimator (see p.2342 of Jankova and van de Geer (2018)).

Lasso is also used for prediction in high dimensional problems in factor models. Section 6 of Bing et al. (2021) shows this using simulations against some of the other machine learning techniques such as ridge regression, generalized least squares (GLS) and principal components estimation. In particular, the second panel of Figure 1 in Bing et al. (2021) clearly shows that when the number of factors grow, Lasso can do better than the ridge, GLS, and principal components in terms of prediction error. In terms of an empirical example, p.437-438, Table 13.1 of Murphy (2012) shows that Lasso has the best MSE compared to ridge and least squares in a health related empirical real life example even with small number of predictors. In addition, Chen (2023) shows in equation (15) that Lasso can be used as an estimator in the synthetic control literature, which like our work is concerned with comparing prediction losses.

2.2 Assumptions

In this subsection we present our assumptions regarding the Lasso estimation. These assumptions will use the following notation. Let $S_0 = \{j : \beta_{0,j} \neq 0\}$ denote the set of relevant regressors with s_0 being the cardinality of the set S_0. (i.e., s_0 of the elements of β_0 are nonzero, and the rest are zero). The uniform l_0 ball is defined as $\mathcal{B}_{l_0}(s_0) := \{\|\beta_0\|_{l_0} \leq s_0\}$ which represents all nonzero coefficients including the local to zero coefficients.

We impose the following assumption on β_0:

Assumption 1 *(i). The average square of approximation error is given by (3), and the total number of nonzero parameters in β_0 is $s_0 \geq 1$, which is a nondecreasing function of n.*

(ii). $\|\beta_0\|_2 = O(1)$.

(iii). $\|\beta_0\|_2 \geq c_1 > 0$, *where c_1 is a positive constant.*

Assumption 1(i) is a standard assumption that allows the number of nonzero coefficients in the model to grow with the sample size. This assumption also requires the model to contain at least one nonzero coefficient. We also assume that the statistician uses an approximate sparse model to estimate β_0 which is sparse. We fix the approximation error rate but as discussed before, this rate comes from an ideal risk minimization, and is shown in Belloni and Chernozhukov (2009). In Appendix B we take a more flexible approach compared with Assumption 1(ii). There, we assume that $\|\beta_0\|_2 = O(\sqrt{s_0})$. From Assumption 1(iii), it is clear that we need a lower bound on the l_2 norm on β_0 and this should be positive and not converging to zero.

To specify our assumptions on the statistician's data, we introduce the following notations. Denote $\Sigma := EX_iX_i'$ for $i = 1, 2, \cdots, n$, let $\hat{\Sigma} := X'X/n$ be the sample counterpart, and let $\phi_{min}(\Sigma)$ denote the minimum eigenvalue of Σ.

Assumption 2 *(i)* $E(u_i|X_i) = 0$, *where* X_i, u_i *are i.i.d. across* $i = 1, \cdots, n$,
(ii) For some positive constant C,

$$\max_{1\leq j\leq p} E|X_{ij}|^4 \leq C < \infty$$

$$E|u_i|^{4k} \leq C < \infty$$

for all $k \geq 1$.
(iii) $\phi_{min}(\Sigma) \geq c > 0$.

This assumption essentially extends the sub-Gaussian data assumption used in the moment oracle inequality (Theorem 1) of Jankova and van de Geer (2018).

Since the intercept in the statistician's regression does not encode any personal attributes of the $n + 1$ user, we assume that she uses the correct value of $X_{n+1,1}$:

Assumption 3 $R(X_{n+1,1}) = X_{n+1,1} = 1$.

We define the following terms:

$$M_1 := \max_{1\leq i\leq n}\max_{1\leq j\leq p}|X_{ij}u_{ij}|,$$

$$M_2 := \max_{1\leq i\leq n}\max_{1\leq j\leq p}\max_{1\leq l\leq p}|X_{il}X_{ij} - EX_{il}X_{ij}|.$$

M_1 is the maximal covariance between the regressors and errors in a high dimensional context. M_2 is the maximal variance of the regressors in the sample. With large p and n, these covariance and variance terms can grow arbitrarily large , hence, we need a condition that restricts the growth rate of their moments.

Assumption 4

$$\max(\sqrt{EM_1^2}, \sqrt{EM_2^2})\frac{\sqrt{lnp}}{\sqrt{n}} = O(1).$$

Note that Assumption 4 allows for diverging moments of M_1 and M_2 but up to a point (this assumption is also made in Chernozhukov et al. (2017)). In the case of $p > n$ Assumptions 2-4 imply:

$$\lambda_n = O(\sqrt{lnp/n}), \tag{5}$$

where the rate is shown in Appendix A.3 (as mentioned earlier, the case of $p \leq n$ is discussed at the end of Appendix B.1).

The following assumption ensures that the Lasso estimator is consistent, and that the estimation error of the moments of Lasso converge to zero (by Lemmas A.1-A.4, and Theorem A.1):

Assumption 5

$$s_0\frac{\sqrt{lnp}}{\sqrt{n}} \to 0.$$

This assumption is standard in high dimensional statistics. It captures the tradeoff between the sparsity of the model and the sample size. It is clear from this assumption that the case of $p > n$ is allowed.

We next define the following deterministic terms:

$$M_3 := \max_{1\leq j\leq p} |X_{n+1,j}|,$$

$$M_4 := \max_{1\leq j\leq p} |R(X_{n+1,j}) - X_{n+1,j}|.$$

Since these terms can grow with n we assume the following:

Assumption 6

$$s_0^{3/2}\frac{\sqrt{lnp}}{\sqrt{n}}M_3M_4 \to 0.$$

To illustrate that Assumption 6 is plausible, consider the case of $p = 2n$ with $s_0 = O(ln\,(n))$, $M_3 = O(ln\,(n))$ and $M_4 = O(ln\,(n))$. Note that we allow for a diverging M_4, but then the model has to be sparse (although s_0 can be diverging).

We also define two events $\mathcal{A}_1, \mathcal{A}_2$ and their intersection $\mathcal{F} := \mathcal{A}_1 \cap \mathcal{A}_2$ in Section A.2.1 in the Appendix. The event $\mathcal{A}_1$ captures a noise inequality, and $\mathcal{A}_2$ captures an eigenvalue inequality.

2.3 Asymptotic Incentive Compatibility

Given her estimates $\hat{\beta}$, the statistician must take an action $a \in \mathbf{R}$ on behalf of a *new* user, $j = n + 1$. This action is the statistician's prediction of the ideal option of that user. The new user's payoff from action a is $-(a - X'_{n+1}\beta_0)^2$, i.e., the new user suffers a loss, which is proportional to the squared difference between the linear approximation of his *true* ideal option, and the option selected by the statistician. Note that we assume that just like the statistician the true function$f$ is also too complicated for the new user to compute, and hence, he also uses a linear approximation. We assume that each component in X_{n+1} can take any real value except for $\pm\infty$.

Since the statistician does not observe X_{n+1}, in order to make her prediction of the linear approximation of $f(Z_{n+1})$, she asks the $n+1$ user to report a $p \times 1$ vector, $R(X_{n+1})$,

which is interpreted as the transformation of that user's attributes Z_{n+1}. The reporting function $R(X_{n+1})$ assigns each (transformed) attribute one of its feasible values (hence, it cannot assign an infinite value to any attribute). The statistician then plugs $R(X_{n+1})$ into her estimated model and chooses the action $a = R(X_{n+1})'\hat{\beta}$.

When the $n+1$ user decides what values to report, she takes into account that she does not observe the statistician's sample, and hence, does not know the values of the estimated coefficients $\hat{\beta}$. She only knows the distribution from which the statistician's sample is drawn, and that given her sample, the statistician chooses $\hat{\beta}$ according to (4). Given this, the user chooses the report $R(X_{n+1})$ that minimizes his expected loss $E[(R(X_{n+1})'\hat{\beta} - X'_{n+1}\beta_0)^2]$.

2.3.1 Examples of the Setting

The above setting may be interpreted as one where the first n users are treated as "in-sample"(training sample) by an automated system, while the $(n+1)$user is treated as "out-sample" (test object). There are several real world examples mainly in healthcare and retail investments. Davenport and Kalakota (2019) discusses machine learning, including penalized regressions such as Lasso, in healthcare and its use in Electronic Health Records (EHR) and in "precision medicine", which aims to predict what a single patient may need based on their characteristics. For an additional list of applications see Table 1 of Habehh and Gohel (2021). Recent health apps such as Symptomate, Symptomchecker and Ada give patients surveys to try and learn their health attributes in order to recommend potentially useful medical tests.

The widespread use of robo-advisors by retail investors is another example. A client generally creates an account online by responding to a series of questions that may include risk preferences, assets, income, debt and investment goals. The robo-advisor uses machine-learning techniques to offer investment selections deemed appropriate in terms of asset allocation and diversification based on the information supplied by the client and data

gathered on past users (see Fisch et al. (2019)).

2.3.2 A Notion of Incentive Compatibility

To introduce our notion of incentive compatibility, consider a user who upon observing her vector of covariates decides which vector of values to report (which may differ from the true values). She may lie about her attributes if she thinks that the choice of λ_n biases the statistician's action. For example, if the new user suspects that λ_n is too small - and hence the $\hat{\beta}$ that the statistician estimates is overfit - she may try to correct for this by appropriately adjusting the values of her reported attributes. Indeed, a small value of λ_n may be chosen by a statistician who wants to ensure that his Lasso estimate is consistent. Consequently, the new user may not report her true attributes, i.e., $R(X_{n+1}) \neq X_{n+1}$. In particular, she may decide to "opt out" and submit a vector of zeros (except for the intercept).

An estimator is said to be (ex-post) incentive-compatible, if for *any* vector of covariates, and for *any* belief over the true model parameters, the user's expected payoff from truthful reporting is at least as high as her expected payoff from any misreport, where the expectation is taken with respect to the statistician's sample.

Definition 1 *An estimator is* ***asymptotically-uniformly incentive-compatible*** *if for every X_{n+1}, for every $R(X_{n+1})$ and for every β_0 that satisfy Assumptions 1-3, and for $p \to \infty$ when $n \to \infty$,*

$$\lim_{n\to\infty} \sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \{E[R(X_{n+1})'\hat{\beta} - X'_{n+1}\beta_0]^2 - E[X'_{n+1}\hat{\beta} - X'_{n+1}\beta_0]^2\} \geq 0 \tag{6}$$

where the expectation E is taken with respect to the possible realizations of the statistician's sample.

Asymptotic incentive compatibility means that at the limit, as $n, p \to \infty$, the new user is unable to perform better in the Mean Squared Error (MSE) sense by misreporting her

personal characteristics, *regardless* of her beliefs over the true model's parameters. We use MSE as the criterion for evaluating the payoff from a report since this criterion is used for prediction purposes as a measure involving bias and variance. MSE can also be thought as a regret type loss since we try to predict $X'_{n+1}\beta_0$ (true model linear component) with $X'_{n+1}\hat{\beta}$. See (3)-(6) of Chen (2023) for a similar type of regret loss in a synthetic control case.

How should we interpret our notion of incentive-compatibility, given that the user may not be sophisticated enough to think in these terms? One interpretation is that lack of incentive compatibility is merely a *normative* statement about the user's welfare - namely, given our model of how the statistician takes actions on the user's behalf, it would be advisable for her to misrepresent her personal characteristics. Furthermore, there are opportunities for new firms to enter and offer the user paid advice for how to manipulate the procedure - in analogy to the industry of search engine optimization. Incentive compatibility theoretically eliminates the need for such an industry. In the context of the online content provision story, some misreporting strategies take the form of "deleting cookies". This deviation is straightforward to implement, and the user can check if it makes her better off in the long run.

Note that incentive-compatibility is not a property that can be tested statistically. To see this, suppose each user is characterized by only a single covariate that is uniformly distributed on $\{0, 1\}$. If users are truthful, then one would expect a 50-50 distribution of 0's and 1's in the population. However, if each user lies about his covariate, then one would also observe a 50-50 distribution of 0's and 1's

2.3.3 Discussion

We now discuss the motivation for some key ingredients of our model, and also remark on the implications of making alternative modeling choices.

Truth telling versus lying. Recall that the statistician's sample contains the true attributes of n users. The idea is that the data on these users is obtained through a different process than the way the statistician obtains the data from the $n+1$ user. For instance, as mentioned earlier, this data may be obtained from a marketing survey where there is no incentive to lie. Alternatively, one may interpret our incentive compatibility requirement as a requirement that truth-telling is a Nash equilibrium among all participants - such that given that everyone else is telling the truth, no user has an incentive to lie.

If lying benefits the new $n+1$ user, whom does it hurt? As mentioned in the Introduction, the concern is that when truth-telling is not a Nash equilibrium among all users, the data that will be used by automated systems for predictions will be false, and hence, all conclusions drawn from it will be wrong. This could be detrimental in areas such as medicine.

The statistician's benevolence. Our paper addresses the issue raised in Eliaz and Spiegler (2019, 2020) that even if a statistician wants to make the best prediction for the user (so there is no a priori conflict of interest between them), the user may still have an incentive to lie because of the model selection component in Lasso (or any penalized regression for that matter), and because the user does not observe the statistician's sample. Since the source of lying in this no-conflict benchmark comes from the estimation procedure itself, the question is, how can we fix the procedure - without harming its estimation properties - so as to ensure truth-telling?

What if the user and the statistician did have a conflict of interests - say, the statistician uses the information that the user gives him in a way that may harm the user? Then obviously, the user will have an incentive to lie no matter which tuning parameter is chosen. In other words, in such an environment, Lasso (or any other estimator) will not be incentive-compatible unless the user is compensated, or the statistician uses an alternative estimation technique that is not optimal econometrically. Exploring this direction is clearly a separate

research agenda.

2.3.4 A Sufficient Condition

We now establish a sufficient condition for asymptotic incentive-compatibility. This condition involves the tuning parameter, the sparsity of the model and the *exception probability* for the event $\mathcal{F} := \{\mathcal{A}_1 \cap \mathcal{A}_2\}$, where $\mathcal{A}_1$ represents a noise inequality and $\mathcal{A}_2$ represents an eigenvalue inequality (these are formally defined in Section A.2.1 of the Appendix). The exception probability is the complement of the event $\mathcal{F}$, and is denoted by $\mathcal{F}^c$.

Theorem 1 *Under Assumptions 1-6, the Lasso is uniformly incentive compatible over* $\mathcal{B}_{l_0}(s_0)$ *if* $n \to \infty, p \to \infty$ *with*

$$\lambda_n s_0^{1/2} P(\mathcal{F}^c)^{-1/8} \to \infty. \tag{7}$$

Remarks.

1. When $n \to \infty$ a sufficient condition for the lower bound to be satisfied is given by (7). By Lemma A.4, when $p > n$,

$$P(\mathcal{F}^c) = O(\frac{1}{p^{C_1}} + \frac{EM_1^2 + EM_2^2}{nln\,(p)}). \tag{8}$$

where $C_1 > 0$ is a positive constant. It follows that when $n \to \infty$, and hence $p \to \infty$, and when λ_n is given by (5), condition (7) can be satisfied (in terms of rates only) when p is exponential in n. :

$$\frac{s_0^{1/2}}{[\frac{1}{p^{C_1}} + \frac{EM_1^2+EM_2^2}{nlnp}]^{1/8} \frac{\sqrt{n}}{\sqrt{ln(p)}}} \to \infty,$$

2. Note that Assumption 6 may require stricter sparsity than Assumption 5. If $M_3 = O(1)$ and $M_4 = O(1)$, then Assumption 6 amounts to $s_0^{3/2}\sqrt{\frac{ln(p)}{n}} \to 0$, which is a sparsity requirement stronger than Assumption 5.

3. Consistency requires a small λ_n, but incentive compatibility requires a large λ_n when $n \to \infty$, so are they compatible with each other? When we select a large λ_n to satisfy incentive compatibility, we should not sacrifice consistency - i.e. we need $s_0\lambda_n \to 0$. By Lemma A.4, the exception probability is upper bounded by the rate

$$P(\mathcal{F}^c)^{1/8} = O\left(max\left(\frac{1}{p^{C_1/8}}, \left(\frac{EM_1^2 + EM_2^2}{nlnp}\right)^{1/8}\right)\right),$$

If

$$s_0^{1/2}\frac{\sqrt{lnp}}{\sqrt{n}}p^{C_1/8} \to \infty,$$

and

$$\frac{s_0^{1/2}\sqrt{lnp}}{\sqrt{n}}\frac{(nlnp)^{1/8}}{(EM_1^2 + EM_2^2)^{1/8}} \to \infty,$$

then the consistency requirement in Lasso does not violate incentive compatibility (asymptotically), with large s_0 and p (both diverging).

4. When we relax Assumption 1 to $\|\beta_0\|_2 = O(\sqrt{s_0})$, the incentive compatibility is still satisfied but under the slightly stronger condition

$$s_0^2\sqrt{\frac{lnp}{n}}[M_3][M_4] \to 0.$$

The proofs are in Appendix B.2.

5. Consider the case of a fully dense model where $s_0 = p$ with $n \to \infty$ and $p \to \infty$. Clearly, by Assumptions 5-6, incentive compatibility is achievable only when $p = o(n^{1/3})$. Hence $p > n$ and asymptotic incentive compatibility are not compatible in the fully dense model - sparsity is essential. Incentive compatibility in the asymptotic case with a fully dense model extends only to the specific case of $p \leq n$, which is analyzed in Appendix B with $p = o(n^{1/3})$.

6. Since the lower bound for λ_n involves $P(\mathcal{F}^c)$ and s_0 terms, it is natural to ask whether this bound is feasible. In Appendix D we present an algorithm for selecting feasible choices of λ_n.

3 Incentive Compatibility in a Nonlinear Model

In this section we extend our analysis of incentive compatibility to a specific nonlinear machine learning model: the Generalized Linear Model (GLM) with structured sparsity estimators. Structured sparsity refers to knowledge of the sparsity patterns in the model. This class of estimators is introduced in van de Geer (2016) and Stucky and Van de Geer (2018) in the least squares linear-loss. A key feature of GLM is the following: When $p > n$ the plug-in estimate of second-order partial derivative matrix of GLM loss is singular, hence standard maximum likelihood estimators are not valid. When $p > n$, one solution is penalized estimators as in structured sparsity estimation.

Caner (2023) recently generalized the statistics literature with GLM loss under weaker assumptions. He provides the debiased version of these estimators, which allows to make inference on coefficients and also shows the superior power of these tests compared with the ones that use unstructured sparsity. GLM is a rather flexible non-linear model for categorical data outcomes. In particular, given the data gathered from the training sample of the first n users, this estimation procedure can rely on binary or multiple choice based answers provided by the $n + 1$ user. For the use of GLM in asset pricing, see Gu et al. (2020) (in particular, this paper demonstrates that in some cases, GLM may outperform deep neural networks).

Let (X_i, y_i), $i = 1, \cdots, n$ be i.i.d. across $i = 1, \cdots, n$, and $X_i \in \mathcal{X} \subseteq \mathbf{R}^p$ and $y_i \in \mathcal{Y} \subseteq \mathbf{R}$. Let $\rho(y_i, X_i'\beta)$ be the GLM loss (e.g., logistic loss), which is convex in $\beta \in \mathcal{B}$, where $\mathcal{B}$ is a convex subset of $\mathbf{R}^p$. Let $\lambda_n > 0$ be the positive tuning parameter as a sequence in n, and define $\Omega(\beta)$ as the norm (we give more information about the general norm $\Omega(.)$ below and in the Appendix). Let S_0 be the the indices of an active set (in the l_1 norm this is just number of nonzero elements in β_0, and in the case of weighted group norm, the number of nonzero groups in β_0 vector), and let $s_0 := |S_0|$, which is nondecreasing with n, and

satisfies $s_0 \geq 1$. For more information on this topic, see Section 2.6 in Caner (2023).

The GLM structured sparsity estimator is given by

$$\hat{\beta}_G := argmin_{\beta \in \mathcal{B}}[\frac{1}{n}\sum_{i=1}^{n} \rho(y_i, X_i'\beta) + \lambda_n \Omega(\beta)],$$

where the norm $\Omega(.)$ is weakly decomposable as will be defined in Appendix C.1. Weakly decomposable norms are introduced in Definition 6.3 and Section 6.9 of van de Geer (2016). They are generated from convex cones and include Lasso, weighted group Lasso and wedge norm. Our analysis will use a lower bound norm for $\Omega(.)$, which is defined as $\underline{\Omega}(\beta) \leq \Omega(\beta)$. For more on this lower bound norm, see Appendix C, and also Section 2.6 of Caner (2023) and section 6.4 of van de Geer (2016).

3.1 Assumptions

We now provide the assumptions that we use in our analysis of GLM structured sparsity estimators. Define $X_{i,j}$ as the j-th element in the X_i vector. Also define a scalar $a := X_i'\beta$.

Assumption G.1. $\max_{1 \leq j \leq p} E|X_{i,j}|^r \leq C < \infty$, *where* $r \geq 4$ *and* C *is a positive constant. In addition, the minimum eigenvalue of* EX_iX_i' *is bounded away from zero uniformly in* n.

Assumption G.2. *Define*

$$M_1 := \max_{1 \leq i \leq n} \max_{1 \leq j \leq p} \max_{1 \leq l \leq p} |X_{i,j}X_{i,l} - EX_{i,j}X_{i,l}|,$$

$$M_2 := \max_{1 \leq i \leq n} \max_{1 \leq j \leq p} |X_{i,j}|.$$

For $k \geq 1$,

(i).

$$\max\left(\frac{\sqrt{EM_1^2}\sqrt{lnp}}{\sqrt{n}}, \frac{\sqrt{EM_2^{2k}}\sqrt{lnp}}{\sqrt{n}}\right) = O(1).$$

(ii).

$$s_0\sqrt{lnp/n} \to 0.$$

(iii). The new user reports the truth on the intercept $R(X_{n+1,1}) = X_{n+1,1}$.

Assumption G.3. *There exists a positive constant,* C_ρ*, which depends on the shape of the second order partial derivative* $\ddot{\rho}(.)$*, such that* $\ddot{\rho}(y_i, X_i'\beta) \geq \frac{1}{C_\rho^2}$ *for all*

$$\sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} |X_i'(\beta - \beta_0)| \leq \kappa$$

where κ *is a positive constant, and*

$$\inf_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \ddot{\rho}(y_i, X_i'\beta_0) \geq \frac{1}{C_\rho^2}$$

Assumption G.4. *The derivatives* $\dot{\rho}(y_i, a) = \frac{\partial \rho(y_i,a)}{\partial a}$ *and* $\ddot{\rho}(y_i, a) = \frac{\partial \rho(y_i,a)}{\partial a^2}$ *exist for all* y_i, a*, and the following holds for some* $\delta > 0$*,*

(i). For $j = 1, \cdots, n, n+1$*,*

$$\sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \max_{a_0 \in \{X_i'\beta_0\}} \sup_{|a-a_0| \leq \delta} \sup_{y_j \in \mathcal{Y}} |\dot{\rho}(y_j, a)| = O(1).$$

(ii).

$$\sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \max_{a_0 \in \{X_i'\beta_0\}} \sup_{|a-a_0| \leq \delta} \sup_{y_i \in \mathcal{Y}} |\ddot{\rho}(y_i, a)| = O(1).$$

(iii).

$$\sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \max_{a_0 \in \{X_i'\beta_0\}} \sup_{|\hat{a}-a_0| \cup |a-a_0| \leq \delta} \sup_{y_i \in \mathcal{Y}} \frac{|\ddot{\rho}(y_i, a) - \ddot{\rho}(y_i, \hat{a})|}{|a - \hat{a}|} \leq 1.$$

Assumption G.5 $\sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \Omega(\beta_0) = g(s_0)$, *where* $g(.)$ *is a function, which depends on the norm* $\Omega(.)$*, and is non-decreasing in* n.

Assumption G.6.

$$M_3 M_4 s_0 g(s_0) \frac{\sqrt{lnp}}{\sqrt{n}} \to 0.$$

Assumption G.1 is about the data and the non-singularity of EX_iX_i. It is needed to get an oracle inequality (Caner (2023) uses a slightly weaker condition than the minimum eigenvalue of EX_iX_i' being positive and not bounded away from zero). Assumption G.2 is adopted from Chernozhukov et al. (2017), where it is used in the concentration inequality. We use a slightly stronger version for the condition on the second and higher moments for X_i, where the second moment is used in the concentration inequality (this does not change the concentration inequality as can be seen from (A.4)). Assumption G.3 is from Chapter 12 of van de Geer (2016), and is used to control the loss function in a neighborhood of β_0 (this assumption is also used in Caner (2023)). Assumption G.4 is used in Caner (2023) as well as in Assumption C.1 of van de Geer et al. (2014) for the first order partial derivative of the unpenalized GLM loss. We use a stronger version of that condition in part (i), since here, the first order partial derivative has to be bounded in a neighborhood of β_0 rather than only at β_0. Note that Assumption G.4(i) is also imposed on the new user. Assumption G.5 is a mild restriction that is used in the proof of Lemma 1 in Caner (2023). This assumption ties the penalty to known functions of sparsity. This can be thought of also as an upper bound on the norm $\Omega(.)$ in terms of a function of sparsity. In l_1 norm case if we take $\Omega(\beta_0) = \|\beta_0\|_1$, then by a norm inequality and Assumption 1, $\|\beta_0\|_1 \leq \sqrt{s_0}\|\beta_0\|_2 = O(\sqrt{s_0})$. So $g(s_0) = C\sqrt{s_0}$ where $C > 0$ a positive constant. Finally, Assumption G.6 will be used to obtain incentive compatible as a sufficient condition in the proof of Theorem 2 in Appendix B.

3.2 Asymptotic Incentive Compatibility

As in our benchmark model, we assume that the new user of the platform (the $n+1$ user), submits a report of her attributes (which may be false) $R(X_{n+1})$ to the statistician. Given this attribute vector, the statistician obtains the empirical risk function at the estimate $\rho(y_{n+1}, R(X_{n+1})'\hat{\beta}_G)$. We assume that the new user aims to minimize the expectation of $[\rho(y_{n+1}, R(X_{n+1})'\hat{\beta}_G) - \rho(y_{n+1}, X'_{n+1}\beta_0)]^2$, where the expectation is taken with respect to the possible realizations of the statistician's sample. The statistician's estimate is incentive compatible if the new user has no incentive to deviate from the truth for all β_0 and X_{n+1}, i.e., if the expected value of the mean squared error with respect to ρ is minimized at the truth $R(X_{n+1}) = X_{n+1}$. Our goal in this subsection is to characterize sufficient conditions for incentive compatibility in the asymptotic case of $n \to \infty, p \to \infty$. These conditions are simpler than those of the finite sample case, which are provided in Theorem C.2 in Appendix C.

Definition 2 *The GLM structured sparsity estimator is* ***asymptotically-uniformly incentive-compatible*** *if for every X_{n+1}, for every $R(X_{n+1})$ and for every β_0 that satisfy Assumptions G.1-G.6, and for $p \to \infty$ when $n \to \infty$,*

$$\lim_{n\to\infty} \sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \{ \int E[\rho(y_{n+1}, R(X_{n+1})'\hat{\beta}_G) - \rho(y_{n+1}, X'_{n+1}\beta_0)]^2 dP_{y_{n+1}} - \int E[\rho(y_{n+1}, X'_{n+1}\hat{\beta}_G) - \rho(y_{n+1}, X'_{n+1}\beta_0)]^2 dP_{y_{n+1}} \} \geq 0 \tag{9}$$

where the expectation E is taken with respect to the possible realizations of the statistician's sample. (the integral is taken over the distribution of y_{n+1}).

The following result characterizes sufficient conditions for asymptotic incentive compatibility of the GLM structured sparsity estimator. To the best of our knowledge this is the first treatment of incentive compatibility for a nonlinear machine learning model with a high dimensional regressor set. Our characterization uses the notion of an exception set $\mathcal{E}^c$

and its probability $P(\mathcal{E}^c)$, which are defined in a similar manner to the Lasso case (see the detailed discussion in Section C.1 of Appendix C). Comparison of $P(\mathcal{F}^c)$ and $P(\mathcal{E}^c)$ will be done below.

Theorem 2 *Under Assumptions G.1-G.6, the GLM structured sparsity estimator is asymptotically uniformly incentive compatible with the following conditions on* λ_n

(i).

$$\lambda_n (lnp)^{1/8} n^{-1/8} P(\mathcal{E}^c)^{-1/8} s_0^{1/2} \to \infty,$$

(ii).

$$\lambda_n P(\mathcal{E}^c)^{-1/4} s_0 [g(s_0)]^{-1} \to \infty.$$

Remarks.

1. We compare the lower bound in Theorem 2(i) with the one in the approximately linear case of Theorem 1. First, since $ln\,(p)/n \to 0$ by Assumption G.2(iii), we have $(ln\,(p)/n)^{1/8} \to 0$. Since $P(\mathcal{E}^c) \geq P(\mathcal{F}^c)$ by Lemma A.2 of Caner (2023) and Lemma A.4 here, we obtain that, $P(\mathcal{E}^c)^{-1/8} \leq P(\mathcal{F}^c)^{-1/8}$. Thus, to satisfy asymptotic incentive-compatibility, λ_n has to be larger in the nonlinear case than in the approximately linear one. This is due to the nonlinear loss function rather than the new penalty structure.

2. Because we have the weakly decomposable norm penalty $\Omega(.)$ instead of Lasso, there is an additional lower bound (ii) compared with the approximately linear case. This extra bound arises because of the moment bound in Theorem C.1.

3. Note that even with a fairly general penalized machine-learning method like GLM structured sparsity estimators, we still achieve asymptotic incentive-compatibility. However, both the nonlinear nature of the loss (as in GLM), and the general penalty function (as in structured sparsity inducing norms) make incentive-compatibility more difficult to establish compared to the linear least square loss with l_1 penalty.

4. If there is a dense model $s_0 = p$, then by Assumption G.2(ii), we need $p/n^{1/2} \to 0$, and by Assumption G.6, we need $g(p)/n^{1/2} \to 0$. Thus, if $p > n$ we cannot achieve asymptotic incentive-compatibility However, if $p << n$, then it is possible to attain asymptotic incentive-compatibility with $pg(p) = o(n^{1/2})$.

5. The statistician may use a more flexible deep learning estimator instead of GLM loss with structured sparsity estimator. However, the statistical theory behind deep neural networks is very recent, and the first econometric contribution came from Farrell et al. (2021). There, by Theorems 1-2, there is a possibility that the new user may report fewer attributes to the machine and hence try to avoid the curse of dimensionality, if she believes the deep neural network estimator may overfit. However, it is not immediately clear how a general misreport may affect incentive-compatibility. We plan to analyze such general questions in subsequent work, as it is beyond the scope of the current paper.

4 Conclusion

This paper takes a first small step towards understanding how estimation techniques based on machine-learning methods can be made immune to misreporting by users in the absence of an explicit conflict of interests between the users and the entity computing the estimator (more specifically, when the estimator tries to predict the best outcome for a user based on her reported attributes). We analyze two models and two machine learning techniques. First, we use an approximately linear model with a Lasso estimator. Then we consider GLM estimated by structured sparsity penalization based analysis. Our main contribution is showing that truthful reporting can be ensured by appropriately adjusting the tuning parameter to be larger than what is required for consistency.

Acknowledgement: We thank co-editor Ivan Canay and an associate editor and two referees for valuable comments that substantially changed the paper. We thank Anders Kock, José Luis Montiel Olea, Ran Spiegler and seminar participants at Simon Fraser University for their valuable comments. We are grateful for the hospitality of the Economics Department at Columbia University, where this research is initiated when both authors were visitors in 2018-2019. Eliaz gratefully acknowledges financial support from ISF grant 470/19.

Disclosure: The authors report that there are no competing interests to declare.

References

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*, 2369–2429.

Belloni, A. and V. Chernozhukov (2009). High dimensional sparse econometric models: An introduction. *Inverse Problems and High Dimensional Estimation Springer Verlag*, 121–156.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high dimensional controls. *Review of Economic Studies 81*, 608–650.

Bing, X., F. Bunea, S. Strimas-Mackey, and M. Wegkamp (2021). Prediction under latent factor regression: adaptive pcr, interpolating predictors and beyond. *Journal of Machine Learning Research 22*, 1–50.

Cai, Y., C. Daskalakis, and C. Papadimitrou (2015). Optimum statistical estimation with strategic data sources. *Proceedings of the 28 th Conference on Learning Theory 40*, 1–40.

Caner, M. (2023). Generalized linear models with structured sparsity estimators. *Journal of Econometrics 236-2*, 105478.

Caner, M. and A. B. Kock (2018). Asymptotically honest confidence regions for high dimensional parameters by the desparsified conservative lasso. *Journal of Econometrics 203*, 143–168.

Chen, J. (2023). Synthetic control as online linear regression. *Econometrica 91*, 465–493.

Chernozhukov, V., D. Chetverikov, and K. Kato (2017). Central limit theorems and bootstrap in high dimensions. *Annals of Probability 45*, 2309–2452.

Chernozhukov, V., M. Goldman, V. Semenova, and M. Taddy (2018). Orthogonal machine learning for demand estimation: High dimensional causal inference in dynamic panels. *arXiv:1712.09988*.

Chiang, H. (2020). Many average partial effects: with an application to text regression. *Working Paper*.

Chiang, H. and Y. Sasaki (2019). Causal inference by quantile regression kink designs. *Journal of Econometrics 210*, 405–433.

Cummings, R., S. Ioannidis, and K. Ligett (2015). Truthful linear regression. *Conference on Learning Theory 40*, 448–483.

Davenport, T. and R. Kalakota (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal 6*, 94–98.

Dekel, O., F. Fischer, and A. Procaccia (2010). Incentive compatible regression learning. *Journal of Computer System and Sciences 76*, 759–77.

Eliaz, K. and R. Spiegler (2019). The model selection curse. *American Economic Review-Insights 1*, 127–140.

Eliaz, K. and R. Spiegler (2020). On incentive compatible estimators. *Working Paper-Tel Aviv University*.

Farrell, M. H., T. Liang, and S. Misra (2021). Deep neural networks for estimation and inference. *Econometrica 89*(1), 181–213.

Fisch, J., M. Laboure, and J. Turner (2019). *The Disruptive Impact of FinTech on Retirement Systems*. The emergence of robo-advisor.

Gao, C., A. Van der Vaart, and H. Zhou (2015). A general framework for bayes structured linear models. *arXiv:1506.02174*.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical asset pricing via machine learning. *Review of Financial Studies 33*, 2223–2273.

Gyorfi, L., M. Kohler, A. Krzyzak, and H. Walk (2010). *A Distribution Free Theory of Nonparametric Regression*. Springer Verlag.

Habehh, H. and S. Gohel (2021). Machine learning in healthcare. *Current Genomics 22*, 291–300.

Hardt, M., N. Megiddo, C. Papadimitrou, and M. Wooters (2016). Strategic classification. *Proceedings. of the ACM Conference on Innovations in. Theoretical Computer Science*, 111–122.

Jankova, J. and S. van de Geer (2018). Semi-parametric efficiency bounds for high-dimensional models. *Annals of Statistics 46*, 2336–2359.

Kock, A. (2016). Oracle inequalities, variable selection and uniform inference in high-dimensional correlated random effects panel data models. *Journal of Econometrics 195*, 71–85.

Kock, A. and H. Tang (2019). Inference in high-dimensional dynamic panel data models. *Econometric Theory 35*, 295–359.

Liang, A. and E. Madsen (2023). Data and incentives. *Theoretical Economics Forthcoming*.

Meir, R., A. Procaccia, and J. Rosenschein (2012). Algorithms for strategyproof classification. *Artificial Intelligence 186*, 123–156.

Murphy, K. (2012). *Machine learning: a probabilistic prespective*. MIT press.

Perte, J. and J. Perote-Pena (2004). Strategy-proof estimators for simple regression. *Mathemitical Social Sciences 47*, 153–176.

Stucky, B. and S. Van de Geer (2018). Asymptotic confidence regions for high dimensional structured sparsity. *IEEE Trans. Signal Process. 66*, 2178–2189.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of The Royal Statistical Society Series B 58*, 267–288.

van de Geer, S. (2016). Estimation and testing under sparsity. *Springer Verlag*.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*.