Should Humans Lie to Machines? The Incentive Compatibility of Lasso and General Weighted Lasso*

Mehmet Caner† Kfir Eliaz‡

April 13, 2023

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 with sufficiently large sample size. We also provide simplification of some conditions for incentive compatibility in the asymptotic case. We extend our results to the Conservative Lasso estimator and provide new moment bounds for this generalized weighted version of Lasso. Our results show that incentive compatibility is achieved if the tuning parameter is kept above some threshold in the case of asymptotics. We present simulations that illustrate how this can be done in practice.

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 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

*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.

†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.

methods introduce an incentive to misreport one's personal characteristics even when this information will be used *solely* for predicting her best outcome?[1] 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 consider a stylized environment where each user i's ideal option is a linear function f of her privately observed attributes $X_i = (X_{i,1}, ..., X_{i,p})'$ such that $f(X_i) = X_i'\beta_{\mathbf{0}}$. 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 $f(X_i)$ is the optimal dosage of some medication when taken immediately at the onset of symptoms, conditional on the patient's medical history X_i, but the statistician observes the dosage that was given after some delay. Similarly, $f(X_i)$ may be the mix of news and reality shows that a user with attributes X_i actually watches, but the statistician observes only self reports by a user who may have forgotten exactly what she watched.

The statistician uses her sample to estimate the function f by computing an estimate $\hat{\beta}$ of the true coefficients $\beta_{\mathbf{0}}$. The statistician wishes to apply these estimates to predict the ideal option of a new user, $n+1$, whose true attributes X_{n+1} are not observed by the statistician. This new user must decide what vector of attributes $R(X_{n+1})$ (which may *differ* from the truth) to report to the statistician. In making this decision, the new user takes into account her beliefs about the statistician's sample (the new user only knows the distribution from which the sample is drawn, but she does not observe its realization), and her beliefs about the true parameters β_0.

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 ideal option $X_{n+1}'\beta_0$ and her assigned option $R(X_{n+1})'\hat{\beta}$. The statistician's estimator is *incentive-compatible*, if the new user has no incentive to deviate from truthful reporting whatever her attributes are, 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

[1] In a recent interview of Brian Christian, the author of *The Alignment Problem*, he notes that "computers may one day be able not only to learn our behavior but also intuit our values - figure out from our actions what it is we're trying to optimize. ... What if an algorithm intuits the 'wrong' values, based on its best read of who we currently are but not of who we aspire to be? Do we really want our computers inferring our values form browser histories? See Shaywitz (2020) for this interview.

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.[2] 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 (see Remark 1, Corollary 1). This potential to lie implies that the standard way of choosing small enough tuning parameters to ensure consistency may violate incentive compatibility. 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.

Next, we extend our results to a general weighted Lasso, also known as the "Conservative Lasso". Caner and Kock (2018) develop this estimator as a data-dependent weighted penalized estimator. Conservative Lasso better differentiates between relevant and irrelevant variables, which results in better l_2 norm errors. The superior model selection properties of the Conservative Lasso (compared to the standard Lasso) is shown in Caner and Kock (2018) analytically as well as in simulations. We characterize the conditions for ensuring the incentive-compatibility of the Conservative Lasso in large samples, and show this may require a higher (relative to the standard Lasso) lower bound for the tuning parameter under certain scenarios.

We also offer a new technical contribution by extending the oracle-moment-inequalities of Jankova and van de Geer (2018) from sub-Gaussian to i.i.d. data. Using a different proof technique, we derive less conservative bounds on the moments of the Lasso estimator and relax the bounded signal to noise ratio assumption in Jankova and van de Geer (2018). We also extend Jankova and van de Geer (2018) from Lasso moment estimation to generalized weighted Lasso (Conservative Lasso). It is shown that moment bound estimation results cover also this general class of penalty. These are all new results for general weighted Lasso.

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. (2012,2014) 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

[2]Our results can be extended to apply to the debiased Lasso estimator, but this involves a different proof technique, and hence, is beyond the scope of the current paper.

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).

The remainder of the paper is organized as follows. Section 2 introduces our model and assumptions. Section 3 provides new oracle inequalities. Section 4 characterizes the sufficient conditions for ensuring that Lasso is incentive compatible with sufficiently large samples, and special case of sufficient condition for asymptotics is shown as well. Section 5 extends these results to general weighted Lasso. Section 6 provides simulation results and Section 7 concludes. 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. Finally, Appendix C contains the proofs for the general weighted Lasso. Appendix D analyzes the special case of asymptotics in general weighted Lasso.

2 The model

We begin this section by describing our theoretical framework. We then specify our assumptions on the statistician's data and introduce our notion of incentive-compatibility. We conclude by discussing the key ingredients of our model.

Throughout the paper we will use the following notational conventions. For any vector $\nu \in R^d$, 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 p 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 p explanatory variables, $X_{i,1}, ..., X_{i,p}$, drawn from some distribution over a subset of $\mathbb{R}^p$. These attributes determine

the ideal option for a user according to the function

$$f(X_{i,1}, ..., X_{i,p}) = \sum_{k=1}^{p} X_{i,k}\beta_{0,k}$$

where β_0 is a $p \times 1$ vector, representing the true parameters in f. This function applies to all users, who differ only in the values of their characteristics. The realized values of $(X_{i,1}, ..., X_{i,p})$ are privately observed by user i. β_0 is unknown to the users.

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 $X_i = (X_{i,1}, ..., X_{i,p})$ of user i and a noisy signal y_i of that user's ideal option,

$$y_i = X_i'\beta_0 + u_i, \tag{1}$$

where u_i is random noise that is drawn $i.i.d$ from some distribution with zero mean.[3] The X_i' s are i.i.d. across i and exogenous 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.

We 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).* $s_0 \geq 1$ *and is a nondecreasing function of* n.

(ii).

$$\inf_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \|\beta_0\|_2^2 \geq c_1^2 > 0,$$

where c_1 *is a positive constant.*

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

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.

Assumption 1(ii) rules out the case of all local to zero coefficients in a model as well. While allow local to zero coefficients, we need at least one non-zero coefficient and that coefficient cannot be local to zero. Letting $\beta_{0,j}$ represent the j-th coefficient of β_0, Assumption 1(ii) allows $\beta_{0,j} = c/l_n$, for some $j \in S_0$ and positive constant c, but we cannot have an all local to zero β_0 vector, where $n \to \infty, l_n \to \infty$ for all $j = 1, \cdots, p$. From Assumptions 1(i)-(ii), it is clear that we need $\beta_{0,j} = c$ for at least one j.

We will show in the next subsection that Assumption 1(iii) ensures that the signal to noise ratio is bounded (see p.2343 of Jankova and van de Geer (2018)). The empirical implication of this is that only a fixed number of nonzero coefficients can be constants, and the other nonzero coefficients

[3] Access to such observations is a necessary condition for any platform that tries to learn about users (say, Netflix, Spotify). In the introduction, we gave examples for such data, which may be obtained from a third party, or from marketing surveys.

have to be local to zero. To see this, note that

$$\|\beta_0\|_2 = \sqrt{\sum_{j=1}^{p} \beta_{0,j}^2} = \sqrt{\sum_{j \in S_0} \beta_{0,j}^2} = O(1).$$

since in the case of s_0 growing with n

$$\sqrt{\sum_{j \in S_0} \beta_{0,j}^2} = \sqrt{\sum_{j \in D_1} \beta_{0,j}^2 + \sum_{j \in S_0 - D_1} \beta_{0,j}^2} = \sqrt{d_1 C^2 + (s_0 - d_1)\frac{C^2}{s_0 - d_1}} = O(1),$$

where $D_1 := \{j : |\beta_{0,j}| = C\}$ with $|D_1| = d_1$ being a fixed number, C is a generic positive constant and $D_2 := \{j : |\beta_{0,j}| = \frac{C}{\sqrt{s_0 - d_1}}\}$ with $|D_2| = s_0 - d_1$. For ease of exposition, we set all coefficients in D_1 and D_2 to be the same constants, C and $C/\sqrt{s_0 - d_1}$, respectively. D_2 contains indices of all local to zero coefficients. This can easily be generalized without affecting our results.

In Appendix B we take a more flexible approach compared with Assumption 1(iii). There, we assume that $\|\beta_0\|_2 = O(\sqrt{s_0})$. In this case, all nonzero coefficients can be large (i.e., none of them are local to zero, as in set D_2 above). In other words, there is no index set D_2 as above, but all nonzero coefficients (their indices) are in the set D_1 above.

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, she solves the following minimization problem

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

where $\lambda_n > 0$ is the tuning parameter that is local to zero (an explicit expression for the sequence λ_n is given in Remark 3 of Theorem 1). 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 1-6 and Corollaries 1-2 are written for the case of $p > n$, Appendix B contains all theorems related to $p \leq n$.

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).

Assumption 2 together with Assumption 1(iii) ensure that the signal to noise ratio is bounded. To see this, set $\sigma_u^2 := var(u_i)$, the variance of the errors, such that $\sigma_u^2 \geq c > 0$, where c is a generic positive constant that is weakly below the minimum eigenvalue of Σ (which is positive by Assumption 2 in the next subsection). Hence, when $E(u_i|X_i) = 0$, which is imposed in Assumption 2,

$$\frac{var(y_i)}{var(u_i)} = \frac{\beta_0'\Sigma\beta_0}{\sigma_u^2} + 1,$$

However,

$$\frac{\beta_0'\Sigma\beta_0}{\sigma_u^2} + 1 \geq \frac{\|\beta_0\|_2^2\phi_{min}(\Sigma)}{\sigma_u^2} + 1.$$

where $\phi_{min}(\Sigma) \geq c > 0$. Hence, if Assumption 1(iii) holds, then the signal to noise ratio satisfies $var(y_i)/var(u_i) \geq C_0 + 1 > 0$, with C_0 being a positive constant, and defined as $C_0 := \frac{\|\beta_0\|_2^2\phi_{min}(\Sigma)}{\sigma_u^2}$.

Given her estimates $\hat{\beta}$, the statistician must take an action $a \in \mathbb{R}$ on behalf of a *new* user, $j = n + 1$. This action is just the statistician's prediction of the ideal option of that user. The new user's payoff from action a is $-(a - f(X_{n+1}))^2$, where $f(X_{n+1})$ is the true ideal option associated with her personal attributes X_{n+1}. We assume that each attribute 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 $f(X_{n+1})$, she asks the $n+1$ user to report a $p \times 1$ vector, $R(X_{n+1})$, which is interpreted as that user's attributes. The reporting function $R(X_{n+1})$ assigns each 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 attribute values to report, this new user 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 (2). Given this, the user chooses the report $R(X_{n+1})$ that minimizes her expected loss $E[(R(X_{n+1})'\hat{\beta} - X_{n+1}'\beta_0)^2]$, where the expectation is taken with respect to the user's prior beliefs about the true parameters β_0, and her beliefs about the estimate $\hat{\beta}$. Thus, the user may decide not to lie about her covariates.

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

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

2.2 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).[4]

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.[5] Define a positive integer n_0.

Definition 1. *An estimator is* ***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 $p \geq 2$ and $n \geq n_0 > 0$,*

$$\sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \{E[(R(X_{n+1})'\hat{\beta} - X'_{n+1}\beta_0)^2] - E[(X'_{n+1}\hat{\beta} - X'_{n+1}\beta_0)^2]\} \geq 0 \tag{3}$$

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

Incentive compatibility means that the user is unable to perform better in the mean squared sense by misreporting her personal characteristics, *regardless* of her beliefs over the true model's parameters. This definition is for sufficiently large n ($n \geq n_0$), and uniform over $\mathcal{B}_{l_0}(s_0)$ ball. In our Theorem 3 below, we show that n_0 can be as low as 8. We allow p to change with n, but to save notation we do not subscript p.

How should we interpret this requirement, given that we do not necessarily want to think of the user as being 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.

To see that our definition of incentive-compatibility is not vacuous, simply add and subtract the term $X'_{n+1}\hat{\beta}$ inside the squared brackets on the left side term of (3), such that, with $n \geq n_0$, and uniformly over the ball $\mathcal{B}_{l_0}(s_0)$

$$\begin{aligned} E[R(X_{n+1})'\hat{\beta} - X'_{n+1}\beta_0]^2 &- E[X'_{n+1}\hat{\beta} - X'_{n+1}\beta_0]^2 \\ &= E\left[\|(R(X_{n+1}) - X_{n+1})'\hat{\beta}\|_2^2\right] + E[X'_{n+1}\hat{\beta} - X'_{n+1}\beta_0]^2 \\ &+ 2E[\hat{\beta}'(R(X_{n+1}) - X_{n+1})X'_{n+1}(\hat{\beta} - \beta_0)] \\ &- E[X'_{n+1}\hat{\beta} - X'_{n+1}\beta_0]^2 \end{aligned}$$

[4] In the case in which the individual's attributes are collected "passively" from her browsing history, then reporting a vector of zero attributes can be interpreted as the act of deleting cookies.

[5] Recall that the $n + 1$ user does not observe the statistician's realized sample, and hence does not know the precise values of the estimated $\hat{\beta}$.

Canceling common terms reduces incentive-compatibility to the sign of the following term

$$2E[\hat{\beta}'(R(X_{n+1}) - X_{n+1})X'_{n+1}(\hat{\beta} - \beta_0)]. \tag{4}$$

Note that this sign can go either way. For example, if all elements of the vectors, $\hat{\beta}$, X_{n+1} and $\hat{\beta} - \beta_0$ are positive, and for every realized $R(X_{n+1})$, the difference $R(X_{n+1}) - X_{n+1}$ is also positive, then incentive-compatibility holds. If, however, $\hat{\beta} - \beta_0 < 0$, while all the other terms are positive, then incentive-compatibility may be violated.

A weaker ex-ante notion of incentive-compatibility considers a user, who *prior* to observing her covariates, commits to a strategy that maps every possible realization of the covariates to a (possibly non-truthful) report of these realized values. This notion fits situations in which the user either automates her reports to the statistician, or delegates the reporting to a third party. According to this notion, the estimator is ex-ante incentive-compatible if on average (over the different realizations of the user's covariates), the $n+1$ user has no incentive to misreport, with $n \geq n_0$

$$\sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \left[\int E[(R(X_{n+1})'\hat{\beta} - X'_{n+1}\beta_0)^2] dP_{X_{n+1}} - \int E[(X'_{n+1}\hat{\beta} - X'_{n+1}\beta_0)^2] dP_{X_{n+1}} \right] \geq 0,$$

where the integral is computed with respect to the distribution of the new user's attributes. Clearly, if an estimator is (ex-post) incentive-compatible, then it is also ex-ante incentive compatible. Thus, the sufficient conditions for (ex-post) incentive-compatibility of the Lasso estimator, which we establish in Section 4, also guarantee ex-ante incentive-compatibility. While ex-ante incentive-compatibility can be achieved with weaker conditions, the proof of these conditions follows from our proof of (ex-post) incentive-compatibility. In light of this, we shall focus on the ex-post notion henceforth.

2.3 Discussion

In this subsection we discuss the motivation for some key ingredients of our model, and we 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 truthtelling 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 choice of the Lasso estimator. We chose to focus on Lasso because it is the most basic machine learning technique that engages in model selection. Since this is the first paper to ask, under what conditions are such techniques incentive-compatible, it makes sense to start with the most basic textbook technique. Once we understand whether and how to ensure incentive compatibility in the simplest penalized regression model, we move on to explore the weighted general penalized estimator (the Conservative Lasso) in Section 5.

Nevertheless, it is worth mentioning that Lasso has several desirable properties. First, its prediction error is of the same order of magnitude as if there were an oracle, who could make predictions based on the true model. This is shown in Theorem 6.4 and Corollary 6.3 of Buhlmann and van de Geer (2011), who provide general oracle inequalities for convex loss with Lasso penalty. Second, James et al. (2013) shows (see p.26) that despite being less flexible than non-linear models such as random forests and deep learning, the Lasso estimator can prevent overfitting, which is clearly a major issue in out-of-sample contexts. In addition, Lasso is a continuous subset selection, which has good prediction properties as shown in p.61-69 of Hastie et al. (2011).

Since we also consider the Conservative Lasso in Section 5, we briefly mention its properties here. Conservative Lasso is a two-step algorithm, where in the first step, standard Lasso is run and all the variables are kept, and then in the second step, a general weighted algorithm is run to select and estimate the relevant variables. Conservative Lasso is therefore a general weighted version of Lasso: when all weights are equal to one in the penalty, it reduces to standard Lasso. Compared with the standard Lasso, the data-dependent penalties of the Conservative Lasso allow for better differentiation of relevant and irrelevant variables as seen in Lemma 1 of Caner and Kock (2018).[6] Further details on the Conservative Lasso will be provided in Section 5.

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 (say, he deliberately adds noise to it). Exploring this direction is clearly a separate research agenda.

The user's loss function. As explained above, incentive-compatibility means that the user cannot profit by misreporting. Suppose the user had a generic loss function $g(\cdot)$, such that $Eg(R(X_{n+1}), X_{n+1})$ denoted the expected payoff of a user whose true characteristics are given by X_{n+1}, but she reports the values $R(X_{n+1})$. Then incentive-compatibility requires that $Eg(R(X_{n+1}), X_{n+1}) \geq Eg(X_{n+1}, X_{n+1})$ for any realization of X_{n+1} and for any report $R(X_{n+1})$. Note that in general, the user's expected payoff is completely independent of the statistician's loss function. However, without imposing any structure on $g(\cdot)$, it is impossible to characterize a condition that ensures the incentive-compatibility of Lasso.

Given our focus on the no-conflict-of-interests benchmark (which we discussed in the previous point), it is only natural to let the user and the statistician have the same loss function that measures how far (in expectation) the estimate is from the truth. For any loss function one chooses for the statistician, the user has no incentive to lie if the expected loss from lying (i.e., the distance between the predicted best outcome based on lying and the actual ideal outcome for the agent) is higher than under truth-telling. Hence, the definition of incentive-compatibility clearly extends to

[6] The Adaptive Lasso is an alternative estimator that also uses a data-dependent weighted penalty (see Zou (2006)). However, in high dimensional econometrics, the first step of the Adaptive Lasso can cut off relevant variables, which can be undesirable as discussed in p.144-145 of Caner and Kock (2018).

any loss function shared by the statistician and the user. Of course, for each candidate loss function one would need to find the exact sufficient condition. We chose to focus on the mean squared error since it is the most commonly used loss function.

If the user and the statistician evaluated the estimates using different loss functions, then the incentive compatibility condition will apply only to the user's loss function, and again, the precise sufficient condition for incentive-compatibility will depend on the specification of this function.

3 New Oracle Inequalities for Lasso

Oracle inequalities in high dimensional statistics are upper bounds on prediction and estimation errors. In this section we establish new oracle inequalities, which are different from those that are given in the literature for $\|\hat{\beta} - \beta_0\|_1$. These inequalities will serve an important role in proving our main result in the next section (Theorem 3). They are also of independent interest as they extend previous results on sub-Gaussian data to *heteroskedastic* (conditionally) data sets that are commonly used in econometrics. Our proof technique will use a less conservative bound compared with Jankova and van de Geer (2018). Hence, our new inequalities contribute to the literature on high-dimensional econometrics where they can be used for proving generalized semiparametric efficiency of Lasso-type-estimators (as, e.g., in Jankova and van de Geer (2018)).

Our first result in this section is a k-th moment bound for the l_1 norm of the Lasso bias, with $k \geq 1$. A key concept used in this result is the *exception probability* for the event $\mathcal{F} := \{\mathcal{A}_1 \cap \mathcal{A}_2\}$, where $\mathcal{A}_1$ and $\mathcal{A}_2$ are defined in (A.5) and (A.8), which represent the empirical process-noise, and the eigenvalue condition, respectively. The exception probability is the complement of the event $\mathcal{F}$, and is denoted by $P(\mathcal{F}^c)$. An explicit upper bound for the exception probability is calculated in Lemma A.4.

Theorem 1. *Suppose Assumptions 1-2 hold and that $n \geq 8$. Let L_{11}, L_{12}, L_2 be some positive constants. If λ_n is chosen such that*

$$\lambda_n \geq max\left(L_{11}\frac{P(\mathcal{F}^c)^{1/4k}}{s_0^{1/2}}, L_{12}\frac{P(\mathcal{F}^c)^{1/2k}}{s_0^{1/2}}\right), \tag{5}$$

then

$$[E\|\hat{\beta} - \beta_0\|_1^k]^{1/k} \leq L_2 s_0 \lambda_n. \tag{6}$$

This result is valid uniformly over $\mathcal{B}_{l_0}(s_0) = \{\|\beta_0\|_{l_0} \leq s_0\}$.

Remarks.

1. Setting $k = 1$ allows us to learn whether the Lasso estimator is unbiased.

2. A natural question that arises is whether there are plausible situations where the derived λ_n exceeds the lower bound in (5). To see that the answer is yes, first note that the tuning parameter is given by (this is shown in Lemma A.2-(A.13):

$$\lambda_n = K_2[\sqrt{\frac{lnp}{n}} + \frac{\sqrt{EM_1^2}lnp}{n} + \sqrt{\frac{lnp}{n}}], \tag{7}$$

where K_2 is a positive constant, and

$$M_1 := \max_{1\le i\le n}\max_{1\le j\le p}|X_{ij}u_i|, \tag{8}$$

It is easy to see that the bounds in (5) can be satisfied for a sufficiently large p or a sufficiently large s_0. We also provide a numerical example with $p = 2$ and $n = 142$ that achieves the lower bound (5) (this is shown in the appendix following the proof of Theorem 1).

3. Theorem 1 can be reformulated such that the lower bound on λ_n is expressed in terms of the moments of the data. However, this necessitates a higher n. To show this, we first derive an upper bound, expressed in terms of the moments of the data, on the exception probability $P(\mathcal{F}^c)$ (Lemma A.4 derives this bound):[7]

$$P(\mathcal{F}^c) \le [\frac{2}{p^{C_1}} + K_1'\frac{EM_1^2 + EM_2^2}{nlnp}], \tag{9}$$

where K_1' is a positive constant and

$$M_2 := \max_{1\le i\le n}\max_{1\le j\le p}\max_{1\le l\le p}|X_{il}X_{ij} - EX_{il}X_{ij}|. \tag{10}$$

However, to satisfy the upper bound in (9) we need a larger n. In particular, we need $n \ge max(8, N)$, where N is given in (A.14) (this is shown in Lemmas A.3-A.4 in the appendix).

Using (9), we can reformulate Theorem 1 to be expressed in terms of the moments of the data: for $n \ge max(8, N)$, if λ_n can be chosen to be higher than the following two bounds,

$$\frac{L_{11}}{s_0^{1/2}}[\frac{2}{p^{C_1}} + K_1'\frac{EM_1^2 + EM_2^2}{nlnp}]^{1/4k}, \tag{11}$$

and

$$\frac{L_{12}}{s_0^{1/2}}[\frac{2}{p^{C_1}} + K_1'\frac{EM_1^2 + EM_2^2}{nlnp}]^{1/2k}. \tag{12}$$

then (6) is satisfied. It is easy to verify that for $n \ge max(8, N)$, the derived λ_n in (7) exceeds the the bounds in (11) and (12) for either large enough p or large enough s_0. In addition, the numerical example given after the proof of Theorem 1 shows that the lower bounds, (11)-(12) for λ_n are satisfied with $p = 2$ and $n = 142$. Also note that since (11)-(12) are larger than the ones in (5) due to (9), our result is also valid for the lower bounds in (5).

4. Consider the case of a fully dense model with $s_0 = p$, where all regressors are relevant (i.e. all coefficients of β_0 vector are non zero). In this case, the lower bound (5) is easier to achieve compared with $s_0 < p$, but the moment upper bound in Theorem 1 will be larger.

Our second result in this section is a moment bound on the Lasso estimator.

Theorem 2. *Suppose Assumptions 1-2 hold, that $n \ge 8$ and that $P(\mathcal{F}^c) \le min(2^{3-2k}, 1)$ with $k \ge 1$. Let L_3, L_4, L_5 be some positive constants. If*

$$\frac{L_3}{s_0^{1/2}} \ge \lambda_n \ge \frac{L_4 P(\mathcal{F}^c)^{1/2k}}{s_0^{1/2}} \tag{13}$$

then

[7] We will use this upper bound to select the tuning parameter selection in Section 6.

$$[E\|\hat{\beta}\|_1^k]^{1/k} \leq L_5 s_0^{1/2}.$$

This result is valid uniformly over $\mathcal{B}_{l_0}(s_0) = \{\|\beta_0\|_{l_0} \leq s_0\}$.

Remarks.

1. This is a new result and a simple extension of Theorem 1 above (it does not immediately follow from Theorem 1). Note that all the bounds in Theorem 2 are functions of sparsity and the tuning parameter.

2. For large enough p, the choice of λ_n given in (7) satisfies the upper and lower bounds in Theorem 2.

3. As in Remark 3 of Theorem 1, we can use an upper bound on the exception probability that involves moments of the data. Using this bound, the sufficient condition for the lower and upper bounds in Theorem 2 become:

$$\frac{L_3}{s_0^{1/2}} \geq \lambda_n \geq \frac{L_4}{s_0^{1/2}} \left[\frac{2}{p^{C_1}} + K_1' \frac{EM_1^2 + EM_2^2}{nlnp} \right]^{1/2k},$$

To reformulate Theorem 2 with these alternative bounds we need to require that $n \geq max(8, N)$. It can be verified that for such n and for large enough p, the choice of λ_n given in (7) satisfies the alternative bounds given above.

4. In the appendix (after the proof of Theorem 2) we present an example with $p = 2, n = 142$, and $s_0 = 1$ that shows that lower and upper bounds for λ_n are satisfied. The key observation is that with a large signal, which is represented by L_3, it is plausible to satisfy the bounds.

5. Consider the case of a fully dense model where $s_0 = p$. In this case, the upper bound on λ_n in Theorem 2 is more difficult to achieve (compared to $s_0 < p$) but the lower bound for λ_n is smaller compared with the sparse case. Given small K_2 and p slightly larger than n, this upper bound for λ_n can hold in the dense case. Note that in the moment upper bound of Theorem 2, the bound gets larger in the fully dense case compared with the sparse case.

4 Incentive Compatibility of Lasso

Our first main result, which is new in the literature on penalized regressions, characterizes sufficient conditions for the Lasso estimator to be incentive-compatible for a sufficiently large sample size $n \geq n_0$.

We begin by defining the misreport vector $(p \times 1)$:

$$D_{n+1} := R(X_{n+1}) - X_{n+1}.$$

One of the key terms in our proof is the scalar term $\hat{\beta}' D_{n+1}$. In the finite sample characterization of incentive compatibility, we need a mild technical condition for the case in which $\hat{\beta}' D_{n+1} \neq 0$ (this condition is not needed in the asymptotic characterization of incentive compatibility). The condition says that for any misreport by the $n+1$ user, we can find a sequence $c_n > 0$, which can be local to zero sequence and is independent of β_0 and $\hat{\beta}$. More formally,

Condition 1. *For any misreport satisfying $\hat{\beta}' D_{n+1} \neq 0$, there exists $c_n > 0$ such that*

$$E[\hat{\beta}' D_{n+1} D_{n+1}' \hat{\beta}] \geq c_n E[\|\hat{\beta}\|_2^2] > 0. \tag{14}$$

Note that $E[\hat{\beta}' D_{n+1} D_{n+1}' \hat{\beta}] = E[\sum_{j=1}^{p} (\hat{\beta}_j D_{n+1,j})^2]$. Note also that we allow for any misreport, one where the user lies about all attributes ($D_{n+1} \neq 0_p$) as well as partial misreports where D_{n+1} contain some zero values and some non-zero values. Note that 0_p represents a $p \times 1$ vector of zeros. Our characterization of incentive compatibility will also allow for the case in which the new user reports truthfully ($D_{n+1} = 0_p$). The case of $\hat{\beta}' D_{n+1} = 0$ is shown in the proof of Theorem 3.

Next, we define the following terms which will be used in our sufficient condition for incentive compatibility:

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

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

Since no attribute value can get an infinite value, $M_3 \neq \infty, M_4 \neq \infty$. However, since our incentive compatibility notion is ex-post, M_3 and M_4 are deterministic but can grow with n.

Theorem 3. *Suppose all of the following conditions hold: Assumptions 1-3, Condition 1, $n \geq 8$ and $P(\mathcal{F}^c) \leq 1/2$. Then the Lasso estimator is incentive compatible if the tuning parameter satisfies the following:*

$$min\left(\frac{L_3}{s_0^{1/2}}, \frac{c_1^2 c_n}{2c_2 L_2 s_0 c_n + 2L_2 L_5 M_3 M_4 s_0^{3/2}}\right) \geq \lambda_n \geq max\left(\frac{L_{11} P(\mathcal{F}^c)^{1/8}}{s_0^{1/2}}, \frac{\max(L_{12}, L_4) P(\mathcal{F}^c)^{1/4}}{s_0^{1/2}}\right) \tag{15}$$

where c_2 is the largest nonzero coefficient $\beta_{0,j}$, and c_1 is a positive constant defined in Assumption 1. Furthermore, incentive compatibility is valid uniformly over $\mathcal{B}_{l_0}(s_0) = \{\|\beta_0\|_{l_0} \leq s_0\}$.

Remarks.

1. As in Remark 3 of Theorems 1, and Remark 2 of Theorem 2, if we want to use moments of the data rather than the exception probability $P(\mathcal{F}^c)$ in the lower bound for the tuning parameter λ_n, we need $n \geq max(8, N)$.

2. When we relax Assumption 1 to $\|\beta_0\|_2 = O(\sqrt{s_0})$, the incentive compatibility is still satisfied but conditions change slightly. The details are in Appendix B.2.

3. A natural question that arises is whether the fully dense case of $s_0 = p$ is compatible with the concept of uniform incentive compatibility. In this remark, we exclude the asymptotic case of $n \to \infty, p \to \infty$, which will be discussed below following Corollary 1. Clearly, the condition for the lower bound on λ_n is easier to achieve when $s_0 = p$ compared with sparse case. However, the opposite is true for the upper bound since this bound gets smaller with $s_0 = p$. Nevertheless, it is possible to achieve this bound with large c_1^2 and L_3. This is an important observation as it implies that a fully dense model accommodates uniform incentive compatibility.

4.1 Incentive Compatibility and the choice of λ_n

In this subsection, we address the question of how to choose λ_n for a finite sample (sufficiently large n, but not $n \to \infty$) such that incentive compatibility is satisfied. In the literature, λ_n is

typically chosen using cross-validation or information criterion to choose λ_n (e.g., see Chetverikov et al. (2021) and Caner and Kock (2018)).

Our case is more challenging since λ_n has to satisfy the bounds in Theorem 3 to ensure incentive compatibility. The first difficulty is that both bounds involve s_0, and this is infeasible. We offer two solutions for this problem. The first is demonstrated in the above Remarks, where we set $s_0 = 1$ for the lower and upper bounds and then try the dense model with $s_0 = p$ for both bounds, with n, p being finite. If both results agree the estimator is incentive compatible and robust to model selection. The second solution uses Theorem 5 of Caner and Kock (2018), which shows that running conservative Lasso with Generalized Information Criterion provides $P(s_0 = \hat{s}) \to 1$, where $\hat{s}$ is the estimated number of nonzero coefficients. Hence, we can use $\hat{s}$ in the lower and upper bounds to check incentive compatibility.

The second issue is the choice of constants in the bounds. We need L_3 and c_1 to be large, and the remaining constants to be small. All these constants are explained in Appendix A.4, where we provide explicit formulas. We also present a numerical example in the remark following the proofs of Theorems 1-3, which show how λ_n can satisfy incentive compatibility with $p = 2$ and $n = 142$.

A better feasible incentive compatible λ_n is provided for the asymptotic case in Remark 6 of Corollary 1 below. The choice of tuning parameter for the asymptotic case of incentive compatibility will be our preferred choice since it involves only a lower bound for λ_n, and achieves good results in the in simulations of the finite sample case (see Section 5).

4.2 Incentive Compatibility: Asymptotics

Theorem 3 above applies to all $n \geq 8$, and in the case of using an upper bound on the exception probability $P(\mathcal{F}^c)$, incentive compatibility applies when $n \geq max(8, N)$, where N is defined in Lemma A.3. We next show that in the asymptotics case of $n \to \infty, p \to \infty$ we can simplify the sufficient conditions for incentive compatibility by using standard sparsity assumptions for Lasso. In addition, When $n \to \infty, p \to \infty$ we can specify a rate for λ_n. To achieve all this we require following assumptions and definitions.

Recall that M_1 is the maximal covariance between the regressors and errors in a high dimensional context. Roughly speaking, when this covariance is small, it captures exogeneity of the regressors in the sample. Recall also that 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. Because we are allowing for heteroskedastic data and unbounded regressors, we need to consider the growth rate of *higher-order* 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 M_1, 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:

$$\lambda_n = O_p(\sqrt{lnp/n}), \tag{16}$$

where the rate is shown in Appendix A.3. Case of $p \leq n$ is discussed at the end of Appendix B.1.

The following assumption ensures the following: (i) the consistency of the Lasso estimator, (ii) the estimation error of the moments of Lasso converge to zero (by Lemmas A.1-A.4, and Theorem 1), and (iii) the upper bound in Theorem 2 is met when $n \to \infty$ (we expand on this point when we discuss the Theorem 2 bounds below).

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 show that the bounds in Theorems 1-3 can be simplified in the asymptotics case. We start with the analysis of the lower bounds in Theorem 1. Note that by Lemma A.4., when $n \to \infty$ we have that $P(\mathcal{F}^c) \to 0$. Hence, with $k \geq 1$

$$\frac{P(\mathcal{F}^c)^{1/4k}}{P(\mathcal{F}^c)^{1/2k}} \geq 1. \tag{17}$$

When $n \to \infty$ a sufficient condition for the lower bound to be satisfied is that

$$\lambda_n s_0^{1/2} P(\mathcal{F}^c)^{-1/8} \to \infty. \tag{18}$$

By Lemma A.4, when $p > n$,

$$P(\mathcal{F}^c) = O(\frac{1}{p^{C_1}} + \frac{EM_1^2 + EM_2^2}{nlnp}). \tag{19}$$

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 (16), condition (18) can be satisfied (in terms of rates only):

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

when p is exponential in n.

Using the same logic as in (17), it can be verified that for Theorem 2, the lower bound $L_4 P(\mathcal{F}^c)^{1/2k}/s_0^{1/2}$ is smaller asymptotically than $L_{11} P(\mathcal{F}^c)^{1/4k}/s_0^{1/2}$. The upper bound in Theorem, 2 $L_3 \geq \lambda_n s_0^{1/2}$, is satisfied asymptotically since $\lambda_n s_0^{1/2} \to 0$ by (16) and Assumption 5 with L_3 being a positive constant.

To satisfy incentive compatibility when $n \to \infty$, we also need the following assumption, which is related to the upper bound in Theorem 3:

Assumption 6.

$$s_0^{3/2} \frac{\sqrt{lnp}}{\sqrt{n}} M_3 M_4 \to 0.$$

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

Definition 2. *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 $n \to \infty$,*

$$\lim_{n\to\infty} \sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \{E[(R(X_{n+1})'\hat{\beta} - X'_{n+1}\beta_0)^2] - E[(X'_{n+1}\hat{\beta} - X'_{n+1}\beta_0)^2]\} \geq 0 \tag{20}$$

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

We can now restate Theorem 3 for the asymptotic case:

Corollary 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*

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

Remarks.

1. The typical concern with Lasso is the consistency of the estimator ($\|\hat{\beta} - \beta_0\|_1 = o_p(1)$), which can be achieved by making sure that λ_n goes to zero at a relatively fast rate. However, if λ_n gets too small, the Lasso estimator may admit many nonzero variables incorrectly (i.e., it creates an *overfit*). Consequently, when the number of regressors p is very large, the expectation of the sum of l_1 errors ($E\|\hat{\beta} - \beta_0\|_1$) can grow arbitrarily large, and incentive compatibility may be violated. Put differently, *consistency does not imply incentive compatibility in large samples.* Thus, simply using the l_1 estimator bound on its own does *not* imply a bound for the *expectation* of l_1 error.

We illustrate this point with a simple example. Suppose we take a value for λ_n below the lower bound in Theorem 1 (i.e., $k = 2$, and $\lambda_n = o(P(\mathcal{F}^c)^{1/8}/s_0^{1/2})$). In particular, take $\lambda_n = P(\mathcal{F}^c)^{1/4}/s_0^{1/2}$. Then from the proof of Theorem 1, reversing the inequality in (A.33) and hence the rate, we obtain that

$$E\|\hat{\beta} - \beta_0\|_1^2 = O(\lambda_n^{-2} P(\mathcal{F}^c)^{1/2}),$$

But given our choice of λ_n,

$$\lambda_n^{-2} P(\mathcal{F}^c)^{1/2} = [P(\mathcal{F}^c)^{1/4}/s_0^{1/2}]^{-2} P(\mathcal{F}^c)^{1/2} = s_0 \to \infty.$$

Hence, even though there is consistency ($s_0\lambda_n \to 0$) under this choice of λ_n, the moment bound estimation error is *diverging.*

$$E\|\hat{\beta} - \beta_0\|_1^2 \to \infty.$$

Why is overfitting a significant issue for incentive compatibility? The intuition is as follows. Suppose the tuning parameter is sufficiently small so that given the user's prior on the true coefficients, she expects that many irrelevant variables will be included in the estimator. To correct this bias, she can report that these variables are equal to zero.

2. Note that Assumption 6 can 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 $s_0\lambda_n \to 0$.

The exception probability is upper bounded by the rate

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

by Lemma A.4. 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 full dense model - sparsity is essential. Incentive compatibility in the asymptotic case with a fully dense model extends only to the specific case of $p \le n$, 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. We can always set $s_0 = 1$, which makes the bound feasible, and $P(\mathcal{F}^c)$ can be calculated up to a constant as shown in the Simulation Section. This results in a choice of a tuning parameter. In high dimensional problems, calculation of tuning problem is a key question. The problem of selecting a tuning parameter in Lasso with respect to prediction is in Chetverikov et al. (2021), and model selection in conservative Lasso is shown in Caner and Kock (2018).

5 Incentive Compatibility Under a General Weighted Penalty: The Conservative Lasso

In this section we extend our analysis of incentive compatibility to a general weighted penalty function. Caner and Kock (2018) propose the Conservative Lasso, which has superior model selection properties relative to the standard Lasso. This is achieved by using a data-weighted penalty function. Specifically, the Conservative Lasso is a two-step estimator

$$\hat{\beta}_w = argmin_{\beta\in R^p}\{\|Y-X\beta\|_n^2 + 2\lambda_n\sum_{j=1}^{p}\hat{w}_j|\beta_j|\},$$

where $Y = (y_1,\cdots,y_i,\cdots,y_n)'$ is an $n\times 1$ vector, X is an $n\times p$ matrix, and with the prediction norm for a generic vector v defined as $\|v\|_n^2 := n^{-1}\sum_{i=1}^n v_i^2$. The weights $\hat{w}_j$ are defined as follows: for each $j = 1,\cdots,p$

$$\hat{w}_j = \frac{\lambda_{prec}}{|\hat{\beta}_j|\cup\lambda_{prec}},$$

where $\hat{\beta}_j$ is the Lasso estimator of Section 2 for variable j, and λ_{prec} is a positive sequence defined in Lemma D.1.

Roughly speaking, the Conservative Lasso may be viewed as giving excluded variables in Lasso a "second chance". For instance, when $\hat{\beta}_j = 0$, the weight will be one (in contrast to a weight of infinity in Adaptive Lasso). When $\hat{\beta}_j > \lambda_{prec}$, the weight is less than one, so there is a differentiation of weights based on Lasso estimation in the first step. Define the estimated minimum weight across all $j = 1, \cdots, p$ as $\hat{w}_{min}$. Note that this has to be larger than zero and smaller or equal to one by the definition of the weight above.

A formal argument for weight properties, and differentiation of relevant and irrelevant coefficients, is given in Lemma 1 of Caner and Kock (2018). For problems with the Adaptive Lasso in high dimensional settings, see p.144-145 of Caner and Kock (2018).

In order to analyze the incentive-compatibility of the Conservative Lasso, we will need the following mild assumption:

Assumption 7.

$$\hat{w}_{min}^{-1} \leq (C_{w1} d_n)^{-1},$$

where C_{w1} is a positive constant and d_n is a positive sequence that is local to zero.

This assumption is relaxed when we analyze the asymptotic case, and a lower bound for d_n is specified in Theorem 5.

Theorem 4. *Suppose Assumptions 1-2 and 7 hold and that $n \geq 8$. If λ_n is chosen such that*

$$\lambda_n \geq max(\frac{L_{w11} P(\mathcal{F}^c)^{1/4k}}{s_0^{1/2} d_n^{1/2}}, \frac{L_{w12} P(\mathcal{F}^c)^{1/2k}}{d_n s_0^{1/2}})$$

where L_{w11}, L_{w12}, L_{w2} are positive constants, then

$$\left[E \| \hat{\beta}_w - \beta_0 \|_1^k \right]^{1/k} \leq L_{w2} s_0 \lambda_n.$$

This result is valid uniformly over $\mathcal{B}_{l_0}(s_0)$.

Remarks.

1. To the best of our knowledge, this is a new result for general weight functions such as the Conservative Lasso. This extends a result of Jankova and van de Geer (2018) from Lasso with subgaussian data to Conservative Lasso with non subgaussian data. The crux of the proof involves finding the rate for the minimal estimated weight.

2. Since $d_n > 0$, and is a local to zero sequence, the lower bounds on λ_n in Theorem 4 may be larger than the lower bounds for Lasso in Theorem 1 when we ignore the constants. This stems from the weights used in the conservative Lasso estimator. A small weight may generate the possibility of overfit in conservative Lasso, and to prevent that we need a larger λ_n in the compared with the standard unweighted Lasso. Note that the standard Lasso estimator is a special case of the conservative Lasso where the weights are all set to one, such that $d_n = 1$.

3. Remark 3 of Theorem 1 on the upper bound for $P(\mathcal{F}^c)$ applies here as well. Hence, an alternative formulation of Theorem 4 in terms of moments of the data would require $n \geq max(8, N)$.

4. We can also have a result for a fully dense model with $s_0 = p$. It is clear that the lower bound for λ_n will get smaller and will be easier to achieve. The cost will be the larger upper bound on the moments.

Our next result, which is also new in the literature, provides a moment estimator for the Conservative Lasso.

Theorem 5. *Suppose Assumptions 1-2 and 7 hold, and that* $n \geq 8$. *If*

$$\frac{L_{w3}}{s_0^{1/2}} \geq \lambda_n \geq \frac{L_{w41} P(\mathcal{F}^c)^{1/2k}}{s_0^{1/2} d_n}, \tag{21}$$

and

$$d_n \geq L_{w42} P(\mathcal{F}^c)^{1/4k} \tag{22}$$

where $L_{w3}, L_{w41}, L_{w42}, L_{w5}$ *are positive constants, then*

$$\left[E \|\hat{\beta}_w\|_1^k \right]^{1/k} \leq L_{w5} s_0^{1/2}$$

This result holds uniformly over $\mathcal{B}_{l_0}(s_0)$.

Remarks.

1. The lower bounds on λ_n and d_n can be combined to have the following bounds for λ_n:

$$\frac{L_{w3}}{s_0^{1/2}} \geq \lambda_n \geq \frac{L_{w41} P(\mathcal{F}^c)^{1/4k}}{L_{w42} s_0^{1/2}}.$$

2. If we ignore the constants, then compared to the Lasso result in Theorem 2, the interval for λ_n may be narrower in conservative Lasso due to the lower bound being larger than the one in Theorem 2 for λ_n (note that $P(\mathcal{F}^c)^{1/4k} \geq P(\mathcal{F}^c)^{1/2k}$).

3. We can put an upper bound on $P(\mathcal{F}^c)$ as in Remark 2 of Theorem 2, in which case we would need $n \geq max(8, N)$.

4. In the fully dense model $s_0 = p$, the moment bound will get larger but the interval (upper and lower bounds) for λ_n will shift compared to the sparse case in Theorem 2: both the lower and upper bounds will get smaller.

We are now ready to characterize the sufficient conditions for incentive-compatibility in large samples of the Conservative Lasso. For this we require the following counterpart of Condition 1:

Condition 2. *For any misreport satisfying* $\hat{\beta}_w' D_{n+1} \neq 0$, *there exists a sequence* $c_{wn} > 0$ *such that*

$$E[\hat{\beta}_w' D_{n+1} D_{n+1}' \hat{\beta}_w] \geq c_{wn} E[\|\hat{\beta}_w\|_2^2] > 0. \tag{23}$$

Note that the incentive compatibility definition of Lasso applies here with $\hat{\beta}_w$ replacing $\hat{\beta}$ in Definition 1. We do not provide the proof of Theorem 6, since given Theorems 4-5, applying exactly the same proof for Theorem 3, we obtain Theorem 6.

Theorem 6. *Suppose Assumptions 1-3 and 7 hold, that Condition 2 is satisfied and that* $n \geq 8$. *Then the Conservative Lasso estimator is incentive compatible if the following bounds hold. The upper bound:*

$$min \left(\frac{L_{w3}}{s_0^{1/2}}, \frac{c_1^2 c_{wn}}{2c_2 L_{w2} s_0 c_{wn} + 2 L_{w2} L_{w5} M_3 M_4 s_0^{3/2}} \right) \geq \lambda_n$$

and the lower bound

$$\lambda_n \geq max \left(\frac{L_{w1} P\left(\mathcal{F}^c\right)^{1/8}}{s_0^{1/2} d_n^{1/2}}, \frac{\max(L_{w2}, L_4) P\left(\mathcal{F}^c\right)^{1/4}}{s_0^{1/2} d_n,}, \frac{L_{w41} P(\mathcal{F}^c)^{1/8}}{L_{w2} s_0^{1/2}} \right) \tag{24}$$

This result is uniform over l_0 ball $B_{l_0}(s_0)$.

Remarks.

1. Note that we use Remark 1 of Theorem 5 to obtain the third term in the lower bound for λ_n.

2. The upper bounds for λ_n are the same (up to the constants L_{w3}, L_{w2}, L_{w5}) in Theorem 3 and 6. However, again ignoring the constants and analyzing the sequences $s_0, P(\mathcal{F}^c)$ and d_n, the lower bound for Theorem 6 has one more extra term compared with Theorem 3, and all the lower bound terms in Theorem 6 are larger or equal to the ones in Theorem 3. This follows from the fact that d_n is local to zero in n (Assumption 7), and s_0 is non-decreasing in n by assumption. Clearly, it may be more difficult to satisfy incentive compatibility in conservative Lasso compared with Lasso.

3. It is possible to have a fully dense model $s_0 = p$, $p > n$ and incentive compatibility for $n \geq 8$ (excluding the case of $n \to \infty$, which we analyze below after Corollary 2). With $s_0 = p$, both the lower and upper bounds for λ_n are getting smaller, so it is not clear whether it is easier or more difficult to achieve incentive compatibility in a fully dense case versus a sparse case. The case of $p \leq n$ with fully dense model is possible.

5.1 Incentive Compatibility of Conservative Lasso: Asymptotics

This subsection shows that the conditions for incentive compatibility are simpler in the asymptotic case of $n \to \infty, p \to \infty$. To establish this we need the following assumption which replaces Assumption 7 of the finite sample case.

Assumption 8. *Define the precision matrix, $\Theta := \Sigma^{-1}$. Then*
(i).

$$\|\Theta\|_{l_\infty} = O(s_1),$$

with s_1 a non-decreasing positive sequence in n.
(ii). $s_1 \lambda_n = o(1)$.

Assumption 8(i) is a major relaxation of the assumptions in Lemma A.7 of Caner and Kock (2018) (where it is used to derive l_∞ bounds for Conservative Lasso estimators) and in Lemma 4.1 of van de Geer (2016). These papers assume that the l_∞ matrix norm of the precision matrix is constant, which is quite restrictive since in many realistic environments, the dimension of the matrix is $p \times p$ and its maximum row-sum can grow with n.

Assumption 8(ii) is needed for the minimum weights in the Conservative Lasso to be bounded above by 1 (as prescribed by Caner and Kock (2018)), which constraints the growth rate of s_1 in Assumption 8(i).

We are now ready to characterize the sufficient conditions for incentive compatibility in the asymptotic case (note that asymptotic incentive compatibility in Definition 2 applies here with $\hat{\beta}_w$ replacing $\hat{\beta}$.). Theorem 6 above is also valid for $n \to \infty, p \to \infty$, but the upper bounds and the proof are simpler for the case of $n \to \infty, p \to \infty$. We set $d_n := s_1 \lambda_n$ in Theorem 6 with Assumption 8 in mind.

Corollary 2. *Suppose Assumptions 1-6 and 8 hold. When $n \to \infty, p \to \infty$ the Lasso estimator is uniformly incentive compatible over $\mathcal{B}_{l_0}(s_0)$ if*

$$\frac{\lambda_n}{max\left(\frac{P(\mathcal{F}^c)^{1/8}}{s_0^{1/4} s_1^{1/2}}, \frac{P(\mathcal{F}^c)^{1/12}}{s_0^{1/3} s_1^{1/3}}\right)} \to \infty.$$

Remarks.

1. A simple way to satisfy the lower bound for the tuning parameter is to choose

$$\lambda_n := \text{u}pperbound\,(P(\mathcal{F}^c)^{1/12}),$$

by setting $s_0 = 1, s_1 = 1$.

2. A natural question that arises is whether a lower bound on λ_n is compatible with consistency. Applying the lower bound in Corollary 2, let us start with the first lower bound

$$s_0\lambda_n = s_0^{2/3}s_1^{-1/3}P(\mathcal{F}^c)^{1/12}.$$

Using Lemma A.4 with $l = 1, 2$

$$\frac{s_0^8}{p^{C_1}s_1^4} \to 0, \quad \frac{s_0^8 max_l EM_l^2}{nln\,(p)\,s_1^4} \to 0.$$

Note that the same exercise with the second bound in Corollary 2 results in weaker conditions. These conditions are

$$\frac{s_0^6}{p^{C_1}s_1^4} \to 0, \; \frac{s_0^6 \max_l EM_l^2}{n(lnp)s_1^4} \to 0.$$

3. We can also relax Assumption 1(iii) to $\|\beta_0\|_2 = O(\sqrt{s_0})$. The analysis will be very similar to that of Lasso (see Appendix B). The case of $p \leq n$ is the same as in Lasso, hence, the proofs are skipped.

4. Note that a fully dense model with $s_0 = p$ and $n \to \infty$ violates incentive compatibility since the assumptions will be violated as shown in Remark 5 to Corollary 1 in Lasso. However, the case of $p = o(n^{1/3})$ and $p < n$ in the fully dense model $s_0 = p$ does satisfy incentive compatibility since the discussion in Remark 5 of Corollary 1 also applies to Conservative Lasso.

6 Simulations

This section has three objectives. First, it illustrates how in practice the tuning parameter can be chosen to ensure incentive compatibility of the Lasso estimator. Second, it demonstrates that by appropriately choosing the tuning parameter (in line with the conditions in Corollaries 1-2), incentive compatibility is analyzed through the lens of "small" and "large" lies. Finally, we show that incentive compatibility is not vacuous, it is possible for a new user to gain from lying.

In choosing the tuning parameter we use the asymptotic case since this involves only a lower bound, and will be shown to perform well even with $p = 100$ and $n = 100$.

We provide a simple simulation setup. Let

$$y_i = X_i'\beta_0 + u_i,$$

where $\beta_0 = (1, 0'_{p-s_0}, 1'_{s_0-1})'$, 0_{p-s_0} is a $p - s_0$ column vector of all zero elements, and 1_{s_0-1} is a $s_0 - 1$ dimensional column vector of all ones. The term s_0 represent the sparsity of the above model and we set $s_0 = 5$. The error term u_i has t distribution with 5 degrees of freedom.

In our design we introduce a multivariate normal distribution for the attributes of users $i = 1, \cdots, n$, such that the covariance between the j and m-th random variables are governed by

$$\Sigma_{j,m} = 0.5^{|j-m|},$$

for $j = 1, \ldots, p$ and $m = 1, \cdots, p$. Thus, the correlation between the adjacent random variables is 0.5, and this declines when the random variables are further apart. This Toeplitz type structure is commonly used in the high dimensional literature (see Caner and Kock (2018)). The new user has a draw from a t distribution with three degrees of freedom. It is drawn from t_3 and that is kept fixed through the iterations so that we can compare between Lasso and Conservative Lasso.

The results are presented in Tables 1-4. Tables 1-2 consider Lasso with a "large" lie (the difference between the truth and the new user's report is 2 across all attributes) and with a "small" lie (the difference between the truth and thel report is 0.2 across all attributes) respectively. Tables 3-4 consider the Conservative Lasso for the same setup.

For Lasso, we aim to demonstrate that with a "large" tuning parameter as in Corollary 1, incentive compatibility can be achieved when the sample size n is large enough. As mentioned in the previous section, one possible choice of a tuning parameter that satisfies Corollary 1 is the upper bound on the exception probability, in case s_0 is increasing in n

$$\lambda_n \geq P(\mathcal{F}^c)^{1/8}.$$

The issue is to make the exception probability, $P(\mathcal{F}^c)$ operational and usable. We need $P(\mathcal{F}^c) < 1$ and close to zero given Lemma A.4. Note that an upper bound on this probability is (with positive constants $C_1 > 0, C_\lambda > 0, K_1 > 0$)

$$P(\mathcal{F}^c) \leq \frac{2}{p^{C_1}} + \frac{K_1'[EM_1^2 + EM_2^2]}{nlnp} \leq \frac{2}{p^{C_1}} + \frac{C_\lambda}{(lnp)^2}, \tag{25}$$

by observing that for $l = 1, 2$

$$\begin{aligned}
\frac{K_1' \max_l EM_l^2}{nlnp} &= \left[\frac{\sqrt{K_1'}\sqrt{\max_l EM_l^2}}{\sqrt{n}\sqrt{lnp}}\right]^2 \\
&= \left[\frac{\sqrt{K_1'}\sqrt{\max_l EM_l^2}\sqrt{lnp}}{\sqrt{n}}\right]^2 (\frac{1}{lnp})^2 \\
&\leq \frac{C_\lambda}{(lnp)^2},
\end{aligned}$$

where we use Assumption 4. Hence, we can write the upper bound of the exception probability by using $p \geq 2$

$$P(\mathcal{F}^c) \leq \frac{2}{p^{C_1}} + \frac{C_\lambda}{(lnp)^2}.$$

The tuning parameter is as follows

$$\lambda_n := \left[\frac{2}{p^{C_1}} + \frac{C_\lambda}{(lnp)^2}\right]^{1/8}, \tag{26}$$

and to have the probability $P(\mathcal{F}^c)$ close to zero, even for $p = 2$, we need a large C_1, and a small

C_λ.

In our experiments we set $C_1 = 6$ (we also tried $C_1 = 8, 10$, which delivered very similar simulation results). On the one hand, to control C_λ, we need a small value. On the other hand, a very small value for C_λ can create an overfit. We therefore use a criterion to choose C_λ. We put more weight on the choice of C_λ than on C_1 since the exception probability term, $P(\mathcal{F}^c)$ depends more on the second term, $\frac{C_\lambda}{(lnp)^2}$ through slow convergence to zero in (26). We select the values for C_λ and λ_n according to the Generalized Information Criterion (GIC) as in Caner and Kock (2018), which ensures consistent model selection that prevents overfit as well as underfit with weighted Lasso choices in the least squares framework (the choice of tuning parameter with GIC in least squares with Conservative Lasso is shown to be selection consistent in Theorem 5 of Caner and Kock (2018)).

Note that the criterion for choosing the tuning parameter should take incentive compatibility into account. Hence, we choose only C_λ with GIC, but the structure of our tuning parameter is determined by our characterization of incentive compatibility. Therefore, our choice of λ_n is *above* a lower bound, which prevents overfitting (this is the novel insight of Corollary 1). Note that in the literature for inference in high dimensional parameters, the tuning parameter is either selected by cross-validation or information criterion. Consistency of the Lasso estimators is the key in these types of selection, hence it can force the researcher to select a low value for tuning parameter. But a low tuning parameter can create overfit which may violate incentive compatibility. Our choice in (26) prevents this problem.

Define

$$\lambda_n^* := argmin_{\lambda_n \in \Lambda} \left[ln(\hat{\sigma}^2(\lambda_n)) + \frac{\hat{s}(\lambda_n)}{n} ln(n) ln(ln(p)) \right],$$

where $\hat{s}(\lambda_n)$ is the number of nonzero elements in the Lasso estimator, given a choice of λ_n in a grid Λ, and $\hat{\sigma}^2(\lambda_n)$ is the mean squared residuals from the Lasso regression, given a choice of λ_n in a grid Λ. We form Λ as follows: we take C_λ in a grid of values $C_\lambda := [0.01, 0.05, 0.1, 0.2]$, so Λ is the grid of values of λ_n depending on C_λ. The number of iterations is 1,000.

For the Conservative Lasso, the same type of tuning parameter analysis is used, but with Corollary 2, instead of Corollary 1. Hence, the tuning parameter choice for conservative Lasso is

$$\lambda_n := \left[\frac{2}{p^{C_1}} + \frac{C_\lambda}{(lnp)^2} \right]^{1/12}, \tag{27}$$

The Choice of C_λ is done in the same way as in Lasso above.

The "Report" columns in Tables 1-4 display $E[R(X_{n+1})'\hat{\beta} - X'_{n+1}\beta_0]^2$ as the mean squared error from a false report by the user. "Truth" refers to $E[X'_{n+1}(\hat{\beta} - \beta_0)]^2$. The difference between $R(X_{n+1}) - X_{n+1}$ is kept at two levels: 2 and 0.2 (for all p variables), which represent large, and small deviations from the truth. We have $p = 100, 200, 300$, and for each p level we analyze $n = 100, 200, 300$.

The numbers in each cell of the tables correspond to the disutility of the user (i.e., the mean square difference between the statistician's estimate and the optimal action). Hence, smaller numbers correspond to higher payoffs. Let us compare the tables when $p = 300$ and $n = 200$. In Table 1, which corresponds to a large magnitude of a lie, the user's disutility from reporting the truth is 0.40, while the disutility from lying is 51.68. Hence, the $n+1$ user prefers to be truthful. In Table 2, for a small lie, truth-telling induces a disutility of 0.40, while lying induces a lower disutility of 0.15. Hence a lie is preferred. Thus, even with our lower bound, it is possible to profit from a "small" lie. Note that some of the small lies are prevented by our lower bound as can be seen

Table 1: Lasso-Incentive Compatibility:

Difference 2	$n = 100$		$n = 200$		$n = 300$	
Dimension	Truth	Report	Truth	Report	Truth	Report
$p = 100$	0.51	45.88	0.38	49.86	0.33	50.77
$p = 200$	0.30	56.52	0.25	63.10	0.11	64.00
$p = 300$	0.73	42.30	0.40	51.68	0.37	52.52

Note: "Truth" refers to $E[X'_{n+1}(\hat{\beta} - \beta_0)]^2$ and "Report" refers to $E[R(X_{n+1})'\hat{\beta} - X'_{n+1}\beta_0]^2$ in Incentive Compatibility Definition. Smaller values of these average squared errors are desirable.

Table 2:Lasso-Incentive Compatibility:

Difference 0.2	$n = 100$		$n = 200$		$n = 300$	
Dimension	Truth	Report	Truth	Report	Truth	Report
$p = 100$	0.51	0.19	0.38	0.11	0.33	0.09
$p = 200$	0.30	1.28	0.15	1.12	0.11	1.10
$p = 300$	0.73	0.26	0.40	0.15	0.37	0.11

Note: "Truth" refers to $E[X'_{n+1}(\hat{\beta} - \beta_0)]^2$ and "Report" refers to $E[R(X_{n+1})'\hat{\beta} - X'_{n+1}\beta_0]^2$ in Incentive Compatibility Definition. Smaller values of these average squared errors are desirable.

in $p = 200$ with different n in Table 1. So for small lies, guaranteeing incentive-compatibility is more difficult. However, as predicted, all large lies are prevented by our lower bound for the tuning parameter.

Tables 3-4 show the same pattern for Conservative Lasso. The lower bound on the tuning parameter prevents large lies, but dissuading small lies depend on (p, n) combination. Also, when we move from small to large lie, the mean squared error from lying gets very large. This is evident by comparing Table 1 with Table 2, and comparing Table 3 with Table 4. To give an example, for Conservative Lasso with $p = 100$ and $n = 100$, in Table 3 the new user prefers to lie with a mean squared error of 0.19 from lying compared to 0.74 from truth-telling. However, with a larger lie, the mean squared error from lying increases to 35.12 making it not profitable to lie.

Table 3: Conservative Lasso-Incentive Compatibility:

Difference 2	$n = 100$		$n = 200$		$n = 300$	
Dimension	Truth	Report	Truth	Report	Truth	Report
$p = 100$	0.74	35.12	0.66	35.99	0.59	36.55
$p = 200$	0.35	49.45	0.21	51.67	0.18	52.16
$p = 300$	0.92	34.79	0.68	37.74	0.66	37.90

Note: "Truth" refers to $E[X'_{n+1}(\hat{\beta} - \beta_0)]^2$ and "Report" refers to $E[R(X_{n+1})'\hat{\beta} - X'_{n+1}\beta_0]^2$ in Incentive Compatibility Definition. Smaller values of these average squared errors are desirable.

Table 4: Conservative Lasso-Incentive Compatibility:

Difference 0.2	$n = 100$		$n = 200$		$n = 300$	
Dimension	Truth	Report	Truth	Report	Truth	Report
$p = 100$	0.74	0.19	0.66	0.09	0.59	0.06
$p = 200$	0.35	1.30	0.21	1.16	0.18	1.13
$p = 300$	0.92	0.29	0.68	0.12	0.66	0.08

Note: "Truth" refers to $E[X'_{n+1}(\hat{\beta} - \beta_0)]^2$ and "Report" refers to $E[R(X_{n+1})'\hat{\beta} - X'_{n+1}\beta_0]^2$ in Incentive Compatibility Definition. Smaller values of these average squared errors are desirable.

7 Conclusion

The growing reliance on machine learning in automating decisions raises the question of how would users interact with these automated systems. In particular, would users have an incentive to act strategically in order to manipulate such automated systems? This strategic interaction will become particularly important when these automated systems start playing a more prominent role in medical decision-making or even in driving.

This paper takes only a small preliminary step towards addressing this question by studying whether a user would want to lie to an automated system that uses Lasso or Conservative Lasso to predict that user's ideal outcome based on her reported attributes. 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. Our result is also significant from a pure econometrics point of view: just concentrating on oracle inequalities and post-selection inference can lead to a small tuning parameter, which in turn, can lead to model overfitting, which then introduces an incentive to misreport. If users have an incentive to provide false input to algorithms used for estimation and prediction, then it is no longer clear that one can rely on the output of these algorithms.

APPENDIX

In the next part, Appendix A considers the proofs when $p > n$, and Appendix B considers the case $p \leq n$, and relaxing Assumption 1(iii). Appendix C covers Conservative Lasso proofs of Theorems 4-6. Appendix D covers proof of asymptotics of Conservative Lasso incentive compatibility in Corollary 2.

A Appendix A

A.1 Notation

In this section, we show some results that will help us in proofs. Define random vector of variables $F_i := (F_{i1}, \cdots, F_{ij}, \cdots, F_{ip})'$. Also define $\sigma_F^2 := n(max_{1\leq j\leq p} var F_{ij})$, and $M_F := \max_{1\leq i\leq n} \max_{1\leq j\leq p} |F_{ij} - EF_{ij}|$. Note that $\hat{\mu}_j := n^{-1}\sum_{i=1}^n F_{ij}$, and $\mu_j := EF_{ij}$.

A.2 Maximal Inequalities

We use an assumption that will provide us maximal inequalities.

Assumption A.1. Assume F_i are iid random vectors across $i = 1, 2, \cdots, n$ with $\max_{1\leq j\leq p} var F_{ij} \leq C < \infty$ for a positive constant $C > 0$.

We use the following maximal inequality. With Assumption A.1, Lemma E.2(ii) of Chernozhukov et al. (2017) is: (see (A.2) of Caner and Kock (2019))

$$P\left[\max_{1\leq j\leq p} |\hat{\mu}_j - \mu_j| \geq 2E \max_{1\leq j\leq p} |\hat{\mu}_j - \mu_j| + \frac{t}{n}\right] \leq exp(-t^2/3\sigma_F^2) + K_1 \frac{EM_F^2}{t^2}, \tag{A.1}$$

for a positive constant $K_1 > 0$. With Assumption A.1 here, Caner and Kock (2019) or Lemma E.1 of Chernozhukov et al. (2017) provides

$$E \max_{1\leq j\leq p} |\hat{\mu}_j - \mu_j| \leq \frac{K_2}{2}[\frac{\sqrt{lnp}}{\sqrt{n}} + \frac{\sqrt{EM_F^2 lnp}}{n}], \tag{A.2}$$

for a positive constant $K_2 > 0$.
Define the sequence $\kappa_n = lnp$. Set $t = t_n = K_2(n\kappa_n)^{1/2}$ to have (A.1) as

$$\begin{aligned} P\left[\max_{1\leq j\leq p} |\hat{\mu}_j - \mu_j| \geq 2E \max_{1\leq j\leq p} |\hat{\mu}_j - \mu_j| + K_2\frac{\sqrt{\kappa_n}}{\sqrt{n}}\right] &\leq exp(-C_1\kappa_n) + K_1 \frac{EM_F^2}{K_2^2 n\kappa_n} \\ &= \frac{1}{p^{C_1}} + \frac{K_1 EM_F^2}{K_2^2 nlnp} \end{aligned} \tag{A.3}$$

where $C_1 > 0$, is a positive constant.

Now combine (A.2) with (A.3) to have

$$
\begin{aligned}
P(\max_{1\leq j\leq p}|\hat{\mu}_j-\mu_j| &\geq K_2[\frac{\sqrt{lnp}}{\sqrt{n}}+\frac{(EM_F^2)^{1/2}lnp}{n}+\frac{\sqrt{lnp}}{\sqrt{n}}]) \\
&\leq \frac{1}{p^{C_1}}+\frac{K_1EM_F^2}{K_2^2n(lnp)} \\
&\leq \frac{1}{p^{C_1}}+\frac{K_1'EM_F^2}{n(lnp)}, \qquad \text{(A.4)}
\end{aligned}
$$

with $K_1' \geq K_1/K_2^2$, and by Assumption A.1.

A.2.1 Events

Before the assumptions, we need to define events that will be helpful. The first event is:

$$
\mathcal{A}_1 = \left\{ 2\left\| \frac{u'X}{n} \right\|_\infty \leq \lambda_n \right\}, \qquad \text{(A.5)}
$$

which controls the noise. This is the maximal correlation between regressors and errors.

We start with defining first population counterparts of restricted eigenvalue conditions and then show the empirical version also. These are standard in high dimensional econometrics and statistics and can be seen from Assumption 1 of Caner and Kock (2018).

We define the population adaptive restricted eigenvalue of Σ

$$
\phi_\Sigma^2(s) = \min\left\{ \frac{\delta'\Sigma\delta}{\|\delta_S\|_2^2} : \delta \in R^p - \{0\}, \|\delta_{S_c}\|_1 \leq 3\sqrt{s}\|\delta_S\|_2, |S| \leq s \right\}. \qquad \text{(A.6)}
$$

Note that if $\Sigma = EX_iX_i'$ has full rank, the population adaptive restricted eigenvalue being positive is satisfied by Assumption 2. Also instead of minimizing all over R^p, we minimize vectors that satisfy $\|\delta_{S^c}\|_1 \leq 3\|\delta_S\|_2$. Even in the cases that Σ does not have full rank, it is possible that minimal adaptive restricted eigenvalue condition is satisfied due to optimization over a restricted set. The parameter δ will be related to structural parameter β in the proofs.

First define the empirical adaptive restricted eigenvalue condition, which is empirical counterpart of the population version in Assumption 2:

$$
\hat{\phi}_{\hat{\Sigma}}^2(s) = \min\left\{ \frac{\delta'\hat{\Sigma}\delta}{\|\delta_S\|_2^2} : \delta \in R^p - \{0\}, \|\delta_{S_c}\|_1 \leq 3\sqrt{s}\|\delta_S\|_2, |S| \leq s \right\}. \qquad \text{(A.7)}
$$

We are interested in behavior of the minimal empirical adaptive restricted eigenvalue condition evaluated for set S_0 at cardinality s_0. The second event is:

$$
\mathcal{A}_2 = \left\{ \hat{\phi}_{\hat{\Sigma}}^2(s_0) \geq \phi_\Sigma^2(s_0)/2 \right\}. \qquad \text{(A.8)}
$$

Empirical adaptive restricted eigenvalue condition is needed since in case of $p > n$, $X'X$ is singular and the minimal eigenvalue of $X'X$ is zero. Set $\mathcal{F} = \mathcal{A}_1 \cap \mathcal{A}_2$, and the complement event as $\mathcal{F}^c$.

A.2.2 Proofs of Lemmata

The following four Lemmata are the intermediate results that are used for Theorems.

Lemma A.1. *Under the joint event* $\mathcal{F} := \{\mathcal{A}_1 \cap \mathcal{A}_2\}$ *we have*

$$\|\hat{\beta} - \beta_0\|_1 \leq \frac{24\lambda_n s_0}{\phi^2_\Sigma(s_0)}.$$

This is also valid uniformly over $\mathcal{B}_{l_0}(s_0) = \{\|\beta_0\|_{l_0} \leq s_0\}$.

Proof of Lemma A.1. Using $\hat{\beta}$ definition

$$\|Y - X\hat{\beta}\|_n^2 + 2\lambda_n \sum_{j=1}^p |\hat{\beta}_j| \leq \|Y - X\beta_0\|_n^2 + 2\lambda_n \sum_{j=1}^p |\beta_{0,j}|.$$

Use the model $Y = X\beta_0 + u$ on the first left side term as well as the first right side term to simplify the inequality above combining with Holder's Inequality

$$\begin{aligned}\|X(\hat{\beta} - \beta_0)\|_n^2 + 2\lambda_n \sum_{j=1}^p |\hat{\beta}_j| &\leq 2\left|\frac{u'X}{n}(\hat{\beta} - \beta_0)\right| + 2\lambda_n \sum_{j=1}^p |\beta_{0,j}| \\ &\leq 2\|\frac{u'X}{n}\|_\infty \|\hat{\beta} - \beta_0\|_1 + 2\lambda_n \sum_{j=1}^p |\beta_{0,j}|\end{aligned}$$

On the right side assuming we are on the event $\mathcal{A}_1$

$$2\|\frac{u'X}{n}\|_\infty \|\hat{\beta} - \beta_0\|_1 \leq \lambda_n \|\hat{\beta} - \beta_0\|_1.$$

So we have

$$\|X(\hat{\beta} - \beta_0)\|_n^2 + 2\lambda_n \sum_{j=1}^p |\hat{\beta}_j| \leq \lambda_n \|\hat{\beta} - \beta_0\|_1 + 2\lambda_n \sum_{j=1}^p |\beta_{0,j}|.$$

Use $\|\hat{\beta}\|_1 = \|\hat{\beta}_{S_0}\|_1 + \|\hat{\beta}_{S_0{}^c}\|_1$ on the second term for the left side of the inequality immediately above

$$\|X(\hat{\beta} - \beta_0)\|_n^2 + 2\lambda_n \sum_{j \in S_0^c} |\hat{\beta}_j| \leq \lambda_n \|\hat{\beta} - \beta_0\|_1 + 2\lambda_n \sum_{j=1}^p |\beta_{0,j}| - 2\lambda_n \sum_{j \in S_0} |\hat{\beta}_j|.$$

By assumption of sparsity $\sum_{j \in S_0^c} |\beta_{0,j}| = 0$, and using the reverse triangle inequality we have

$$\|X(\hat{\beta} - \beta_0)\|_n^2 + 2\lambda_n \sum_{j \in S_0^c} |\hat{\beta}_j| \leq \lambda_n \|\hat{\beta} - \beta_0\|_1 + 2\lambda_n \sum_{j \in S_0} |\hat{\beta}_j - \beta_{0,j}|.$$

Next by $\|\hat{\beta} - \beta_0\|_1 = \|\hat{\beta}_{S_0} - \beta_{0,S_0}\|_1 + \|\hat{\beta}_{S_0^c}\|_1$ for the first term on the right side of the inequality immediately above

$$\|X(\hat{\beta} - \beta_0)\|_n^2 + \lambda_n \sum_{j \in S_0^c} |\hat{\beta}_j| \leq 3\lambda_n \sum_{j \in S_0} |\hat{\beta}_j - \beta_{0,j}|.$$

Use $\|\hat{\beta}_{S_0} - \beta_{0,S_0}\|_1 \leq \sqrt{s_0}\|\hat{\beta} - \beta_{0,S_0}\|_2$ above on the right side to have

$$\|X(\hat{\beta} - \beta_0)\|_n^2 + \lambda_n \sum_{j \in S_0^c} |\hat{\beta}_j| \leq 3\lambda_n \sqrt{s_0}\|\hat{\beta}_{S_0} - \beta_{0,S_0}\|_2. \tag{A.9}$$

Ignoring the first term on the left of (A.9), (A.9) shows that we satisfy the restricted set condition

in empirical adaptive restricted eigenvalue condition, so we have

$$\|\hat{\beta}_{S_0{}^c}\|_1 \leq 3\sqrt{s_0}\|\hat{\beta}_{S_0} - \beta_{0,S_0}\|_2.$$

Using $\delta = \hat{\beta} - \beta_0$ in the empirical adaptive restricted eigenvalue condition (A.7) in (A.9)

$$\|X(\hat{\beta} - \beta_0)\|_n^2 + \lambda_n \sum_{j \in S_0^c} |\hat{\beta}_j| \leq 3\lambda_n \sqrt{s_0} \frac{\|X'(\hat{\beta} - \beta_0)\|_n}{\hat{\phi}_{\hat{\Sigma}}(s_0)}.$$

Then use $3uv \leq u^2/2 + 9v^2/2$ with $u = \lambda_n\sqrt{s_0}/\hat{\phi}_{\hat{\Sigma}}(s_0)$, $v = \|X(\hat{\beta} - \beta_0)\|_n$ to get

$$\|X(\hat{\beta} - \beta_0)\|_n^2 + \lambda_n \sum_{j \in S_0^c} |\hat{\beta}_j| \leq \frac{\|X(\hat{\beta} - \beta_0)\|_n^2}{2} + \frac{9}{2} \frac{\lambda_n^2 s_0}{\hat{\phi}_{\hat{\Sigma}}^2(s_0)}.$$

Simplify above

$$\|X(\hat{\beta} - \beta_0)\|_n^2 + 2\lambda_n \sum_{j \in S_0^c} |\hat{\beta}_j| \leq \frac{9\lambda_n^2 s_0}{\hat{\phi}_{\hat{\Sigma}}^2(s_0)}.$$

Use the event $\mathcal{A}_2$ we get the following

$$\|X(\hat{\beta} - \beta_0)\|_n^2 + 2\lambda_n \sum_{j \in S_0^c} |\hat{\beta}_j| \leq \frac{18\lambda_n^2 s_0}{\phi_{\Sigma}^2(s_0)}.$$

This implies the oracle inequality

$$\|X(\hat{\beta} - \beta_0)\|_n^2 \leq \frac{18\lambda_n^2 s_0}{\phi_{\Sigma}^2(s_0)}. \tag{A.10}$$

To get to the l_1 bound ignore the first term in (A.9) and add both sides $\lambda_n\|\hat{\beta}_{S_0} - \beta_{0,S_0}\|_1$ to have

$$\lambda_n \sum_{j \in S_0^c} |\hat{\beta}_j| + \lambda_n \sum_{j \in S_0} |\hat{\beta}_j - \beta_{0,j}| = \lambda_n \|\hat{\beta} - \beta_0\|_1 \leq \lambda_n \|\hat{\beta}_{S_0} - \beta_{0,S_0}\|_1 + 3\lambda_n\sqrt{s_0}\|\hat{\beta}_{S_0} - \beta_{0,S_0}\|_2,$$

by seeing also $\sum_{j \in S_0^c} |\beta_{0,j}| = 0$. Now use the norm inequality $\|\hat{\beta}_{S_0} - \beta_{0,S_0}\|_1 \leq \sqrt{s_0}\|\hat{\beta}_{S_0} - \beta_{0,S_0}\|_2$ to have

$$\lambda_n \|\hat{\beta} - \beta_0\|_1 \leq 4\lambda_n\sqrt{s_0}\|\hat{\beta}_{S_0} - \beta_{0,S_0}\|_2.$$

Use the empirical adaptive restricted eigenvalue condition with $\delta = \hat{\beta} - \beta_0$

$$\|\hat{\beta} - \beta_0\|_1 \leq 4\sqrt{s_0} \frac{\|X(\hat{\beta} - \beta_0)\|_n}{\hat{\phi}_{\hat{\Sigma}}(s_0)}.$$

Use (A.10) and the event $\mathcal{A}_2$ to have

$$\|\hat{\beta} - \beta_0\|_1 \leq 4\sqrt{s_0}\left[\frac{3\sqrt{2}\lambda_n\sqrt{s_0}}{\phi_\Sigma(s_0)}\right]\left[\frac{1}{\hat{\phi}_{\hat{\Sigma}}(s_0)}\right]$$
$$\leq \frac{24\lambda_n s_0}{\phi^2_\Sigma(s_0)}. \tag{A.11}$$

Note that uniformity over $\mathcal{B}_{l_0}(s_0)$ follows since the upper bound in (A.11) depends on β_0 only through s_0. **Q.E.D**

Lemma A.2. . *Under Assumptions 1-2, and since* $\kappa_n = lnp$

$$P(\mathcal{A}_1) \geq 1 - exp(-C_1\kappa_n) - \frac{K_1' EM_1^2}{(n\kappa_n)} = 1 - \frac{1}{p^{C_1}} - \frac{K_1' EM_1^2}{nlnp}$$

Proof of Lemma A.2. Establish the probability bound on $\mathcal{A}_1$ via Assumption 2, using (A.3)(A.4) with $F_i = X_i u_i$ there and $\kappa_n = lnp$, we have, with a universal positive constant $K_1' > 0$

$$P(\mathcal{A}_1) \geq 1 - exp(-C_1\kappa_n) - K_1'\frac{EM_1^2}{(n\kappa_n)} = 1 - \frac{1}{p^{C_1}} - \frac{K_1' EM_1^2}{nlnp}, \tag{A.12}$$

with a universal positive constant $K_2 > 0$

$$\lambda_n = K_2[\sqrt{\frac{lnp}{n}} + \frac{\sqrt{EM_1^2}lnp}{n} + \sqrt{\frac{lnp}{n}}]. \tag{A.13}$$

Q.E.D.

Define $N > 0$, a positive integer, which we will specify in Remark after Lemma A.3. Also remember that we have $p \geq 2$.

Lemma A.3. *Under Assumptions 1-2, with* $\kappa_n = lnp$*, and for* $n \geq N$ *(sufficiently large n)*

$$P(\mathcal{A}_2) \geq 1 - exp(-C_1\kappa_n) - \frac{K_1' EM_2^2}{(n\kappa_n)} = 1 - \frac{1}{p^{C_1}} - \frac{K_1' EM_2^2}{nlnp}.$$

Remark. We set, with $p \geq 2$, with a universal positive constant $K_3 > 0$, where $K_2 > K_3 > 0$, $n^{1/2}K_3 > K_2$

$$N \geq (32s_0)^2\left[\frac{K_3\sqrt{lnp^2} + K_3\sqrt{EM_2^2}lnp^2 + K_3\sqrt{lnp}}{\phi^2_\Sigma(s_0)}\right]^2. \tag{A.14}$$

We also show how we derive N expression in the proof of Lemma A.3. If the lower bound in (A.14) is not an integer, N is the smallest integer that is larger than the lower bound.

Proof of Lemma A.3. Start with

$$\left|\delta'\frac{X'X}{n}\delta\right| = \left|\delta'(\frac{X'X}{n} - \Sigma + \Sigma)\delta\right|$$
$$\geq |\delta'\Sigma\delta| - |\delta'(\hat{\Sigma} - \Sigma)\delta|. \tag{A.15}$$

The second term on the right side of (A.15) can be bounded by repeated application of Holders inequality

$$|\delta'(\hat{\Sigma} - \Sigma)\delta| \leq \|\delta\|_1^2\|\hat{\Sigma} - \Sigma\|_\infty.$$

So (A.15) becomes

$$|\delta'\hat{\Sigma}\delta| \geq |\delta'\Sigma\delta| - \|\delta\|_1^2\|\hat{\Sigma} - \Sigma\|_\infty. \tag{A.16}$$

Now we digress a bit to simplify (A.16). Note that we have the restriction set definition

$$\|\delta_{S_0^c}\|_1 \leq 3\sqrt{s_0}\|\delta_{S_0}\|_2,$$

where we add $\|\delta_{S_0}\|_1$ to both sides

$$\begin{aligned}\|\delta\|_1 &\leq 3\sqrt{s_0}\|\delta_{S_0}\|_2 + \|\delta_{S_0}\|_1 \\ &\leq 3\sqrt{s_0}\|\delta_{S_0}\|_2 + \sqrt{s_0}\|\delta_{S_0}\|_2 \\ &= 4\sqrt{s_0}\|\delta_{S_0}\|_2,\end{aligned}$$

where we used the norm inequality $\|\delta_{S_0}\|_1 \leq \sqrt{s_0}\|\delta_{S_0}\|_2$ in the second inequality above. So we get

$$\frac{\|\delta\|_1^2}{\|\delta_{S_0}\|_2^2} \leq 16s_0.$$

Now divide (A.16) by $\|\delta_{S_0}\|_2^2 > 0$ to have

$$\frac{|\delta'\hat{\Sigma}\delta|}{\|\delta_{S_0}\|_2^2} \geq \frac{|\delta'\Sigma\delta|}{\|\delta_{S_0}\|_2^2} - 16s_0\|\hat{\Sigma} - \Sigma\|_\infty.$$

Minimize over δ on the both sides

$$\hat{\phi}_{\hat{\Sigma}}^2(s_0) \geq \phi_\Sigma^2(s_0) - 16s_0\|\hat{\Sigma} - \Sigma\|_\infty. \tag{A.17}$$

Define $\epsilon_{n1} = 16s_0t_1$, where

$$t_1 = K_2[\sqrt{\frac{lnp^2}{n}} + \frac{\sqrt{EM_2^2}lnp^2}{n} + \sqrt{\frac{lnp}{n}}]. \tag{A.18}$$

By (A.3)(A.4), via Assumption 2

$$\begin{aligned}P[16s_0\|\hat{\Sigma} - \Sigma\|_\infty > \epsilon_{n1}] &= P[\|\hat{\Sigma} - \Sigma\|_\infty > t_1] \\ &\leq exp(-C_1lnp) + \frac{K_1'EM_2^2}{(nlnp)}.\end{aligned} \tag{A.19}$$

See that by multiplying the second term on right side of (A.18) by $n^{1/2}$ and hence define another positive sequence

$$\epsilon_{n2} := 16s_0\left[K_3[\sqrt{\frac{lnp^2}{n}} + \frac{\sqrt{EM_2^2}lnp^2}{\sqrt{n}} + \sqrt{\frac{lnp}{n}}]\right].$$

If $n \geq N$ using (A.14) we have

$$\epsilon_{n1} \leq \epsilon_{n2} \leq \phi_\Sigma^2(s_0)/2, \tag{A.20}$$

since both $\epsilon_{n1}, \epsilon_{n2}$ are positive and $t_1 > 0$. Then with $n \geq N$ and by (A.17)(A.19)

$$P[\hat{\phi}^2_{\hat{\Sigma}}(s_0) \geq \phi^2_{\Sigma}(s_0)/2] \geq 1 - exp(-C_1\kappa_n) - \frac{K_1' E M_2^2}{(n\kappa_n)}$$
$$= 1 - \frac{1}{p^{C_1}} - \frac{K_1' E M_2^2}{nlnp} \tag{A.21}$$

Q.E.D.

We need the following Lemma for the exception set $\mathcal{F}^c := \{A_1 \cap A_2\}^c$ upper bound probability.

Lemma A.4. *Under Assumptions 1-2, with $\kappa_n = lnp$, and $n \geq N$*

$$P(\mathcal{F}^c) \leq 2exp(-C_1\kappa_n) + \frac{K_1'[EM_1^2 + EM_2^2]}{(n\kappa_n)}$$
$$= \frac{2}{p^{C_1}} + \frac{K_1'(EM_1^2 + EM_2^2)}{nlnp}.$$

Proof of Lemma A.4.

Now we provide an upper bound for the probability $P(\mathcal{F}^c)$ in our case under Assumption 2, by using Lemmata A.2-A.3

$$P(\mathcal{F}^c) = P(\mathcal{A}_1 \cap \mathcal{A}_2)^c = P(\mathcal{A}_1^c \cup \mathcal{A}_2^c) \leq P(\mathcal{A}_1^c) + P(\mathcal{A}_2^c)$$
$$\leq 2exp(-C_1\kappa_n) + \frac{K_1'[EM_1^2 + EM_2^2]}{(n\kappa_n)}$$
$$= \frac{2}{p^{C_1}} + \frac{K_1'[EM_1^2 + EM_2^2]}{nlnp}. \tag{A.22}$$

Q.E.D.

A.2.3 New Oracle Inequality Proofs

We start with proof of Theorems 1-2, where they are used as inputs to proof of Theorem 3. Theorems 1-2 consider the new oracle inequalities.

Proof of Theorem 1. We proceed in four steps.

Denote the joint event $\mathcal{F} = \{\mathcal{A}_1 \cap \mathcal{A}_2\}$. $\mathcal{F}^c$ is $\mathcal{F}$'s complement. See that

$$E\|\hat{\beta} - \beta_0\|_1^k = E\|\hat{\beta} - \beta_0\|_1^k 1_{\{\mathcal{F}\}} + E\|\hat{\beta} - \beta_0\|_1^k 1_{\{\mathcal{F}^c\}}. \tag{A.23}$$

We want to form rates for the right side terms in (A.23). Define a positive constant $C_2 := [24/\phi^2_{\Sigma}(s_0)]^k$.

Step 1. Note that by Lemma A.1, the first term on the right side of (A.23) is:

$$E\|\hat{\beta} - \beta_0\|_1^k 1_{\{\mathcal{F}\}} \leq C_2 s_0^k \lambda_n^k. \tag{A.24}$$

Now we want to evaluate the second term on the right side of (A.23). But before that we need the following intermediate step.

Step 2. We use Nemirowski's moment inequality, Lemma 14.24 with Lemma 14.14 in Buhlmann and van de Geer (2011), with for all $k \geq 1$, for the first inequality below. Then for the second inequality we use Loeve's c_r inequality, and for the equality we use u_i being iid, also the definition

of $\sigma^2 := Eu_i^2$ Then define a positive constant $C_3 := 2^{2k-1}[8ln2]^k C'$, with $Eu_i^{4k} \leq C' < \infty$, $C' > 0$ a positive constant, $k \geq 1$

$$E\left|\frac{\sum_{i=1}^n u_i^2 - \sigma^2}{n}\right|^{2k} \leq [8ln(2)]^k E\left[\frac{\sum_{i=1}^n (u_i^4)}{n^2}\right]^k$$
$$\leq \frac{[8ln2]^k n^{k-1}}{n^{2k}} \sum_{i=1}^n Eu_i^{4k}$$
$$= [8ln2]^k [Eu_i^{4k}] n^{-k} \leq [8ln2]^k C' n^{-k} = \frac{C_3}{2^{2k-1}} n^{-k},$$

by Assumption 2. Before the next result we provide the inequality,

$$|x+y|^{2k} \leq 2^{2k-1}(|x|^{2k} + |y|^{2k}), \tag{A.25}$$

for $k \geq 1$, and x, y being generic scalars, and σ^2 being bounded above by Assumption 2 and using (A.25) with a positive constant defined as $C_4 := 2^{2k-1}\sigma^{4k}$

$$E\left|\frac{1}{n}\sum_{i=1}^n u_i^2\right|^{2k} = E\left|\frac{1}{n}\sum_{i=1}^n (u_i^2 - \sigma^2) + \sigma^2\right|^{2k}$$
$$\leq 2^{2k-1}\left[E\left|\frac{1}{n}\sum_{i=1}^n (u_i^2 - \sigma^2)\right|^{2k} + (\sigma^2)^{2k}\right]$$
$$\leq \frac{C_3}{n^k} + C_4 \leq 2C_4, \tag{A.26}$$

and last inequality is by $C_3/n^k \leq C_4$, which is implied by $n \geq 8$, and C_3, C_4 definitions. To see that last point $n \geq [C_3/C_4]^{1/k} = [8ln2]\frac{(C')^{1/k}}{\sigma^4} \geq 8ln2$, with choice of $C' \geq \sigma^{4k}$.

Step 3. Now we have to form another l_1 expectation bound for Lasso that will be key to the second right side term analysis in (A.23). This step 3 modifies the proof of Theorem 1, supplement, p.4 of Jankova and van de Geer (2018). We extend their proof to non-sub-Gaussian case and show that their bound is very conservative, and we provide a new less conservative bound. Start with the definition of Lasso.

$$\|Y - X\hat{\beta}\|_n^2 + 2\lambda_n \|\hat{\beta}\|_1 \leq \|Y - X\beta_0\|_n^2 + 2\lambda_n \|\beta_0\|_1.$$

Ignore the first term and use the model $u = Y - X\beta_0$ to have

$$\|\hat{\beta}\|_1 \leq \frac{\|u\|_n^2}{2\lambda_n} + \|\beta_0\|_1.$$

Then use triangle inequality and then the inequality above

$$\|\hat{\beta} - \beta_0\|_1 \leq \|\hat{\beta}\|_1 + \|\beta_0\|_1 \leq \frac{\|u\|_n^2}{2\lambda_n} + 2\|\beta_0\|_1. \tag{A.27}$$

Next taking the $2k$ th moment of the sampling error in l_1 norm, and using (A.25) by taking

expectations there for the second inequality below

$$E\|\hat{\beta}-\beta_0\|_1^{2k} \leq E\left[\frac{\|u\|_n^2}{2\lambda_n}+2\|\beta_0\|_1\right]^{2k} \leq 2^{2k-1}\{E\left[\frac{\|u\|_n^2}{2\lambda_n}\right]^{2k}+2\|\beta_0\|_1^{2k}\} \tag{A.28}$$

We use Assumption 1 $\|\beta_0\|_2 = O(1)$ to have, specifically with C_5 defined here as an upper bound constant, $\|\beta_0\|_2^{2k} \leq C_5$, C_5 in which C_5 is a positive constant

$$\|\beta_0\|_1^{2k} \leq (\sqrt{s_0}\|\beta_0\|_2)^{2k} \leq C_5 s_0^k. \tag{A.29}$$

Then use the last equation with (A.26) in (A.28) to have

$$E\left[\frac{\|u\|_n^2}{2\lambda_n}\right]^{2k}+2\|\beta_0\|_1^{2k} \leq 2C_4\lambda_n^{-2k}+2C_5 s_0^k. \tag{A.30}$$

Note that proof of Jankova and van de Geer (2018) use $s_0^k\lambda_n^{-2k}$ but this is very conservative upper bound since both two terms in multiplication can diverge with n. But a better bound is given below.

We get the rough bound for expectation using (A.30) in (A.28)

$$E\|\hat{\beta}-\beta_0\|_1^{2k} \leq 2^{2k}C_4\lambda_n^{-2k}+2^{2k}C_5 s_0^k. \tag{A.31}$$

Note that rates in (A.24)(A.31) are different and the last rate in this step is a rough bound which will be helpful in the next step.

Step 4. Rewrite the expectation using event $\mathcal{F}, \mathcal{F}^c$.

$$\begin{aligned} E\|\hat{\beta}-\beta_0\|_1^k &= E\|\hat{\beta}-\beta_0\|_1^k 1_{\{\mathcal{F}\}} + E\|\hat{\beta}-\beta_0\|_1^k 1_{\{\mathcal{F}^c\}} \\ &\leq C_2 s_0^k\lambda_n^k + \sqrt{E\|\hat{\beta}-\beta_0\|_1^{2k}}\sqrt{E1_{\{\mathcal{F}^c\}}} \\ &\leq C_2 s_0^k\lambda_n^k + 2^k max(\sqrt{C_5}s_0^{k/2}, \sqrt{C_4}\lambda_n^{-k}))\sqrt{P(\mathcal{F}^c)} \end{aligned} \tag{A.32}$$

where we use (A.24) and Cauchy-Schwartz inequality for the first inequality, and the second equality is by (A.31).

We can get the rate with the following condition

$$C_2 s_0^k\lambda_n^k \geq 2^k\sqrt{C_4}\lambda_n^{-k}P(\mathcal{F}^c)^{1/2}. \tag{A.33}$$

By (A.32)(A.33)

$$E\|\hat{\beta}-\beta_0\|_1^k \leq 2C_2 s_0^k\lambda_n^k.$$

We can simplify further (A.33),

$$\lambda_n \geq \sqrt{2}[\sqrt{C_4}/C_2]^{1/2k}[P(\mathcal{F}^c)^{1/4k}/s_0^{1/2}]. \tag{A.34}$$

So if $\lambda_n \geq \sqrt{2}[\sqrt{C_4}/C_2]^{1/2k}\frac{P(\mathcal{F}^c)^{1/4k}}{s_0^{1/2}}$ then

$$E\|\hat{\beta}-\beta_0\|_1^k \leq 2C_2 s_0^k\lambda_n^k. \tag{A.35}$$

If $C_2 s_0^k \lambda_n^k \geq 2^k \sqrt{C_5} s_0^{k/2} P(\mathcal{F}^c)^{1/2}$ which is possible by $\lambda_n \geq 2[\sqrt{C_5}/C_2]^{1/k} \frac{P(\mathcal{F}^c)^{1/2k}}{s_0^{1/2}}$ we have

$$E\|\hat{\beta} - \beta_0\|_1^k \leq 2C_2 s_0^k \lambda_n^k.$$

Set $L_{11} := \sqrt{2}[\sqrt{C_4}/C_2]^{1/2k}$ as a positive constant, with $L_{12} := 2(C_5^{1/2}/C_2)^{1/k}$ as another positive constant, and define $L_2 := (2C_2)^{1/k}$ as a positive constant.

The uniformity over $\mathcal{B}_{l_0}(s_0)$ follows since the rates in (A.24)(A.31)-(A.33) depends on β_0 only by s_0. **Q.E.D.**

Remarks.

1. Now we provide a numerical example to show that indeed even with $p = 2, n = 142$ it is possible that λ_n is greater than equal to the lower bound in (5). In that respect, set $K_1' = 1, K_2 = 0.55, K_3 = 0.075, k = 2, \phi_\Sigma(s_0) = 1, EM_1^2 = EM_2^2 = 1, C_1 = 10, C_2 = 576, C_4 = 1, s_0 = 1, C_5 = 625$ which provides $L_{11} = 0.288, L_{12} = 0.417, P(\mathcal{F}^c) \leq 0.003$, hence

$$\lambda_n = 0.143 > \max(L_{11}\frac{P(\mathcal{F}^c)^{1/8}}{s_0^{1/2}}, L_{12}\frac{P(\mathcal{F}^c)^{1/4}}{s_0^{1/2}}) = (0.138, 0.096).$$

Note that we take the case of $s_0 = 1$ for the lower bound, with the other possible case $s_0 = p = 2$, we get a smaller lower bound. Also we have $n = 142 \geq max(8, N = 142)$

2. Proof of Theorem 1 in Jankova and van de Geer (2018), in their appendix, p.5, shows that they use assumption with $P(\mathcal{F}^c)$ bound chosen as in (A.37) below

$$\lambda_n \geq \frac{P(\mathcal{F}^c)^{1/4k}}{s_0^{1/4}}, \tag{A.36}$$

which is equivalent to the following condition as shown in p.3 of proof of Theorem 1 in Jankova and van de Geer (2018)

$$\tau^2 > 2kln[(\sqrt{s_0}\lambda_n^2)^{-1}]/lnp,$$

given that $\lambda_n \geq C\tau\sqrt{lnp/n}$ and $C > 0, \tau > 1$ with

$$P(\mathcal{F}^c) \leq \frac{2}{(2p)^{\tau^2/2}} \tag{A.37}$$

by Lemma 7 in appendix of Jankova and van de Geer (2018). Our result and theirs are not comparable in terms of λ_n since they assume sub-Gaussian data, and ours is more general, and their upper bound in (A.37) is different than our Lemma A.4.

Proof of Theorem 2.

We start with

$$E\|\hat{\beta}\|_1^k = E\|\hat{\beta}\|_1^k 1_{\{\mathcal{F}\}} + E\|\hat{\beta}\|_1^k 1_{\{\mathcal{F}^c\}} \leq E\|\hat{\beta}\|_1^k 1_{\{\mathcal{F}\}} + \sqrt{E\|\hat{\beta}\|_1^{2k}}\sqrt{P(\mathcal{F}^c)}, \tag{A.38}$$

by using Cauchy-Schwartz inequality. Then use triangle inequality on set $\mathcal{F}$ and by Lemma A.1, and norm inequality to have

$$\|\hat{\beta}\|_1 \leq \|\hat{\beta} - \beta_0\|_1 + \|\beta_0\|_1 \tag{A.39}$$

$$\leq \frac{24\lambda_n s_0}{\phi_\Sigma^2(s_0)} + \sqrt{s_0}\|\beta_0\|_2 \tag{A.40}$$

by Assumption 2. This last rate in (A.40) shows that with (A.29)

$$E\|\hat{\beta}\|_1^k 1_{\{\mathcal{F}\}} \leq C_2 s_0^k \lambda_n^k + C_5^{1/2} s_0^{k/2} \leq 2C_5^{1/2} s_0^{k/2}, \tag{A.41}$$

with $\lambda_n \leq L_3 \frac{1}{s_0^{1/2}}$, with

$$L_3 := [C_5^{1/2}/C_2]^{1/k}. \tag{A.42}$$

To handle the second right side term in (A.38) we start with the second inequality in (A.27) and ignore $\|\beta_0\|_1$ in the middle to have

$$\|\hat{\beta}\|_1 \leq \frac{\|u\|_n^2}{2\lambda_n} + \|\beta_0\|_1.$$

then follow (A.32) to obtain

$$\sqrt{E\|\hat{\beta}\|_1^{2k}} P(\mathcal{F}^c)^{1/2} \leq 2^k \max((\frac{C_5}{2})^{1/2} s_0^{k/2}, C_4^{1/2} \lambda_n^{-k}) P(\mathcal{F}^c)^{1/2} \tag{A.43}$$

Now use (A.41) with (A.43) in (A.38)

$$E\|\hat{\beta}\|_1^k \leq 2C_5^{1/2} s_0^{k/2} + 2^k \max((C_5^{1/2}/2^{1/2}) s_0^{k/2}, C_4^{1/2} \lambda_n^{-k}) P(\mathcal{F}^c)^{1/2}. \tag{A.44}$$

There are two possibilities. First, with

$$2C_5^{1/2} s_0^{k/2} \geq 2^k C_4^{1/2} \lambda_n^{-k} P(\mathcal{F}^c)^{1/2}.$$

we obtain

$$E\|\hat{\beta}\|_1^k \leq 4C_5^{1/2} s_0^{k/2}.$$

This is possible with

$$\lambda_n \geq L_4 \frac{P(\mathcal{F}^c)^{1/2k}}{s_0^{1/2}},$$

with the positive constant L_4 defined as

$$L_4 := 2^{1-1/k} [C_4/C_5]^{1/2k}. \tag{A.45}$$

The second possibility is

$$2C_5^{1/2} s_0^{k/2} \geq 2^{k-1/2} C_5^{1/2} s_0^{k/2} P(\mathcal{F}^c)^{1/2},$$

which implies

$$min(2^{3-2k}, 1) \geq P(\mathcal{F}^c). \tag{A.46}$$

So with $\frac{L_3}{s_0^{1/2}} \geq \lambda_n \geq L_4 \frac{P(\mathcal{F}^c)^{1/2k}}{s_0^{1/2}}$ the rate is

$$E\|\hat{\beta}\|_1^k \leq 4C_5^{1/2} s_0^{k/2}. \tag{A.47}$$

So we have the desired result by setting $L_5 := (4C_5^{1/2})^{1/k}$. **Q.E.D.**

Remark. We follow the structure of the example after the proof of Theorem 1. So we set $s_0 = 1$ for the lower bound and upper bounds first. Hence, we obtain $L_3 = 0.208, \lambda_n = 0.143$, so the upper bound of $L_3 = 0.208 \geq \lambda_n = 0.143$ is satisfied. Then for the lower bound, we obtain $L_4 = 0.282$

and

$$\lambda_n = 0.143 \geq L_4 P(\mathcal{F}^c)^{1/4}/s_0^{1/2} = 0.065.$$

Alternatively we can put $s_0 = p = 2$, in both lower and upper bounds, to have the upper bound $L_3/p^{1/2} = 0.1473 \geq \lambda_n = 0.143$ is satisfied, and the lower bound of $\lambda_n = 0.143 \geq 0.046$ is satisfied. Clearly $P(\mathcal{F}^c) \leq 0.003 \leq 1/2$.

Q.E.D.

A.2.4 Main Theorem Proof: Incentive Compatibility

Proof of Theorem 3.

By Theorems 1 and 2 we can choose the larger of λ_n lower bounds in those theorems, with $s_0 \geq 1$, and since it is non-decreasing with n, start with the condition

$$\frac{L_3}{s_0^{1/2}} \geq \lambda_n \geq max(\frac{L_{11}P(\mathcal{F}^c)^{1/4k}}{s_0^{1/2}}, \frac{\max(L_{12}, L_4)P(\mathcal{F}^c)^{1/2k}}{s_0^{1/2}}). \tag{A.48}$$

Add and subtract $X'_{n+1}\hat{\beta}$ inside the right hand side of the incentive compatibility definition:

$$\begin{aligned} E[R(X_{n+1})'\hat{\beta} - X'_{n+1}\beta_0]^2\} &= \{E[R(X_{n+1})'\hat{\beta} - X'_{n+1}\hat{\beta} + X'_{n+1}\hat{\beta} - X'_{n+1}\beta_0]^2\} \\ &= \{E[R(X_{n+1})'\hat{\beta} - X'_{n+1}\hat{\beta}]^2 + E[X'_{n+1}\hat{\beta} - X'_{n+1}\beta_0]^2 \\ &+ E[\hat{\beta}'(R(X_{n+1}) - X_{n+1})X'_{n+1}(\hat{\beta} - \beta_0)] \\ &+ E[(\hat{\beta} - \beta_0)'X_{n+1}(R(X_{n+1})' - X'_{n+1})\hat{\beta}]. \end{aligned} \tag{A.49}$$

Using the definition of incentive compatibility, with defining $D_{n+1} := R(X_{n+1}) - X_{n+1}$, we have

$$\sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \{E[R(X_{n+1})'\hat{\beta} - X'_{n+1}\beta_0]^2 - E[X'_{n+1}\hat{\beta} - X'_{n+1}\beta_0]^2\} = \sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \{E[\hat{\beta}' D_{n+1} D'_{n+1}\hat{\beta}] \tag{A.50}$$

$$+ E[\hat{\beta}' D_{n+1} X'_{n+1}(\hat{\beta} - \beta_0)] \tag{A.51}$$

$$+ E[(\hat{\beta} - \beta_0)' X_{n+1} D'_{n+1}\hat{\beta}]\}. \tag{A.52}$$

The proof depends on $\hat{\beta}' D_{n+1}$ scalar term. Note that

$$\hat{\beta}' D_{n+1} = \sum_{j=1}^{p} \hat{\beta}_j D_{n+1,j}.$$

First, in case of $\hat{\beta}' D_{n+1} = 0$, since right side terms, (A.50)-(A.52) are all zero, the Lasso estimator is incentive compatible. $\hat{\beta}' D_{n+1} = 0_p$ can be zero in three scenarios, either $\hat{\beta} = 0$, or $\hat{\beta} \neq 0_p$, and $D_{n+1} = 0_p$ (case of full truth), or $\hat{\beta} \neq 0_p$, $D_{n+1} \neq 0_p$ but $\hat{\beta}' D_{n+1} = 0$.
In the case of $\hat{\beta}' D_{n+1} \neq 0$, we have the following analysis. We start with (A.50)

$$\begin{aligned} \sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} E[\hat{\beta}' D_{n+1} D'_{n+1}\hat{\beta}] &\geq \inf_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} E[\hat{\beta}' D_{n+1} D'_{n+1}\hat{\beta}] \\ &\geq [\inf_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} E\|\hat{\beta}\|_2^2] c_n, \end{aligned} \tag{A.53}$$

where $c_n > 0$, and can be a local to zero sequence (our proofs go through with $c_n = c$ being a positive constant as well) and we use condition (14). Then in (A.53) we analyze the first term on the right side. In that respect we start with

$$\begin{aligned} \|\hat{\beta}\|_2^2 &= \|\hat{\beta} - \beta_0 + \beta_0\|_2^2 \\ &= (\hat{\beta} - \beta_0)'(\hat{\beta} - \beta_0) + \beta_0'\beta_0 + 2(\hat{\beta} - \beta_0)'\beta_0 \\ &\geq \|\beta_0\|_2^2 + 2(\hat{\beta} - \beta_0)'\beta_0 \\ &\geq \|\beta_0\|_2^2 - 2\|(\hat{\beta} - \beta_0)\|_1\|\beta_0\|_\infty, \end{aligned} \tag{A.54}$$

where the first inequality is obtained by observing the first term in the second equality is non-negative, and dropping that first term in the second equality, and the second inequality is observed by Holder's inequality. Use (A.54) in (A.53) to have, by $\sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \|\beta_0\|_\infty \leq c_2 < \infty$, and $c_2 > 0$ is a positive constant

$$\begin{aligned} \sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} E[\hat{\beta}' D_{n+1} D_{n+1}' \hat{\beta}] &\geq \inf_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \|\beta_0\|_2^2 c_n \\ &- 2 \sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} E\|\hat{\beta} - \beta_0\|_1 \sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \|\beta_0\|_\infty c_n \\ &\geq c_n[c_1^2 - 2c_2 L_2 s_0 \lambda_n], \end{aligned} \tag{A.55}$$

by Assumption 1, Theorem 1.
Now analyze (A.51), the analysis of (A.52) is the same and thus omitted. See that

$$\begin{aligned} \hat{\beta}' D_{n+1} X_{n+1}'(\hat{\beta} - \beta_0) &\leq |\hat{\beta}' D_{n+1} X_{n+1}'(\hat{\beta} - \beta_0)| \\ &\leq |\hat{\beta}' D_{n+1}||X_{n+1}'(\hat{\beta} - \beta_0)| \\ &\leq \|\hat{\beta}\|_1 \|D_{n+1}\|_\infty \|X_{n+1}\|_\infty \|\hat{\beta} - \beta_0\|_1, \end{aligned} \tag{A.56}$$

where we use Holder's inequality. Then

$$\sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} E[\hat{\beta}' D_{n+1} X_{n+1}'(\hat{\beta} - \beta_0)] \leq \|D_{n+1}\|_\infty \|X_{n+1}\|_\infty \sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \|E\left[\|\hat{\beta}\|_1 \|\hat{\beta} - \beta_0\|_1\right] \tag{A.57}$$

$$\leq [M_3][M_4] \sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \left[E\|\hat{\beta}\|_1^2\right]^{1/2} \sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \left[E\|\hat{\beta} - \beta_0\|_1^2\right]^{1/2} \tag{A.58}$$

where we apply (A.56) for the first inequality and Holder's Inequality in the second inequality above, and M_3, M_4 definitions. Then we apply Theorems 1-2 with $k = 2$, and $P(\mathcal{F}^c) \leq 1/2$. By (A.48) (A.55),(A.58) in (A.50)-(A.52)

$$\begin{aligned} \sup_{\beta \in \mathcal{B}_{l_0}(s_0)} \{E[\hat{\beta}' D_{n+1} D_{n+1}' \hat{\beta}] + 2E[\hat{\beta}' D_{n+1} X_{n+1}'(\hat{\beta} - \beta_0)\} &\geq \inf_{\beta \in \mathcal{B}_{l_0}(s_0)} \{E[\hat{\beta}' D_{n+1} D_{n+1}' \hat{\beta}]\} \\ &- 2 \sup_{\beta \in \mathcal{B}_{l_0}(s_0)} |E[\hat{\beta}' D_{n+1} X_{n+1}'(\hat{\beta} - \beta_0)]| \\ &\geq c_1^2 c_n - 2c_2 L_2 s_0 \lambda_n c_n \\ &- 2L_2 L_5 M_3 M_4 s_0^{3/2} \lambda_n. \end{aligned} \tag{A.59}$$

Next, choose

$$\lambda_n \leq \frac{c_1^2 c_n}{2c_2 L_2 s_0 c_n + 2L_2 L_5 M_3 M_4 s_0^{3/2}}, \tag{A.60}$$

to have (A.59) to be non-negative, hence left side of (A.50)

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

We can see that λ_n has to be in the following bound for incentive compatibility by (A.48)(A.60)

$$min\left(\frac{L_3}{s_0^{1/2}}, \frac{c_1^2 c_n}{2c_2 L_2 s_0 c_n + 2L_2 L_5 M_3 M_4 s_0^{3/2}}\right) \geq \lambda_n \geq max\left(\frac{L_{11} P\left(\mathcal{F}^c\right)^{1/8}}{s_0^{1/2}}, \frac{\max(L_{12}, L_4) P\left(\mathcal{F}^c\right)^{1/4}}{s_0^{1/2}}\right).$$

Q.E.D.

Remark. We continue with the same example as in Remark 1 after the proofs of Theorems 1-2. Here in this theorem, we need an additional upper bound. Take the case of $s_0 = p = 2$. So with $c_1 = 10, c_2 = 4, c_n = 1$, $M_3 = 0.15$ (maximum absolute attribute), $M_4 = 1.1$ (maximum absolute lie across all attributes), $L_5 = 10$, $s_0 = p$ for the upper bound, $L_2 = 33.94$,

$$\frac{c_1^2 c_n}{2c_2 L_2 p c_n + 2L_2 L_5 M_3 M_4 p^{3/2}} = 0.167 \geq \lambda_n = 0.143.$$

Here given all the results of the proofs of Theorems 1-3, Incentive compatibility is satisfied with $p = 2, n = 142$. Case of the upper bound with $s_0 = 1$, is implied by the case of $s_0 = p = 2$.

A.3 Asymptotics

In this part we show how allowing $n \to \infty$ may affect the maximal inequality result in Section A.2. Assume the following assumption in Chernozhukov et al. (2017). The moment for iid data EM_F is defined in Assumption A.1.

Assumption A.2.

$$\frac{\sqrt{EM_F^2}\sqrt{lnp}}{\sqrt{n}} \to 0.$$

Then in (A.4) by Assumptions A.1-A.2, with $p > n$

$$\max_{1 \leq j \leq p} |\hat{\mu}_j - \mu_j| = O_p(\sqrt{lnp/n}). \tag{A.61}$$

In case of $p \leq n$, $\kappa_n = lnn$ in Section A.2 and

$$\max_{1 \leq j \leq p} |\hat{\mu}_j - \mu_j| = O_p(\sqrt{lnn/n}). \tag{A.62}$$

See that $P(\mathcal{F}^c) \to 0$ due to upper bound in the proof of Lemma A.4, (A.22) converging to zero under Assumptions A.1-A.2. Also see that

$$\lambda_n = O_p(\sqrt{lnp/n}),$$

by Assumption 4 with definition of M_1.

Proof of Corollary 1. Here we only provide the case of $\hat{\beta}' D_{n+1} \neq 0$ from the proof of Theorem 3. In that respect since (A.49) is non-negative, there will be no need for condition (14). We consider the left side of (A.59) in combination with (A.58), when $n \to \infty$

$$\begin{aligned} \sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \{E[\hat{\beta}' D_{n+1} D'_{n+1} \hat{\beta}] &+ 2E[\hat{\beta}' D_{n+1} X'_{n+1}(\hat{\beta} - \beta_0)\} \\ &\geq \inf_{\beta \in \mathcal{B}_{l_0}(s_0)} \{E[\hat{\beta}' D_{n+1} D'_{n+1} \hat{\beta}]\} - 2L_2 L_5 M_3 M_4 s_0^{3/2} \lambda_n \\ &\geq 0, \end{aligned}$$

by Assumption 6. Hence incentive compatibility is established, and there is no need for an upper bound for the tuning parameter.**Q.E.D.**

A.4 Constants

There are several positive constants, and all of them are defined explicitly. K_1, K_2, K_3 are positive constants, and are defined in Lemma A.2, Lemma A.3. C, C' are positive constants, and used in C_1, C_3 definitions below.

Constants	Definitions
C_1	$C_1 := K_2^2/3C$, C_1 is defined in (A.3),
C_2	$C_2 := [24/\phi_\Sigma^2(s_0)]^k$, C_2 is defined immediately after (A.23),
C_3	$C_3 := 2^{2k-1}[8ln2]^k C'$, C_3 is in step 2 of Theorem 1 proof,
C_4	$C_4 := 2^{2k-1}(\sigma^2)^{4k}$, C_4 is defined before (A.26),
C_5	$\infty > C_5 \geq \|\beta_0\|_2^{2k}$, C_5 is defined before (A.29).
L_{11}	$L_{11} := \sqrt{2}[C_4^{1/2}/C_2]^{1/2k}$,$L_{11}$ is defined after (A.35)
L_{12}	$L_{12} := 2[C_5^{1/2}/C_2]^{1/k}$,$L_{12}$ is defined after (A.35)
L_2	$L_2 := (2C_2)^{1/k}$,L_2 is defined after (A.35)
L_3	$L_3 := [C_5^{1/2}/C_2]^{1/k}$, L_3 is defined in (A.42),
L_4	$L_4 := (1/\sqrt{2})[C_4/C_5]^{1/2k}$, L_4 is defined in (A.45),
L_5	$L_5 := (4C_5^{1/2})^{1/k}$, L_5 is defined at the end of the proof of Theorem 2.

B Appendix B

Here we consider results when $p \leq n$, and relaxing Assumption 1(iii).

B.1 When $p \leq n$

There are minor modifications in the proofs compared to $p > n$. We consider them here. One major change is since $p \leq n$, we set $\kappa_n = lnn$.

We provide the maximal inequality here. Now take the case of $p \leq n$, and combine (A.2) with (A.3) to have with $\kappa_n = lnn$ in that case

$$\begin{aligned} P(\max_{1 \leq j \leq p} |\hat{\mu}_j - \mu_j| \geq K_2[\frac{\sqrt{lnp}}{\sqrt{n}} + \frac{(EM_F^2)^{1/2} lnp}{n} + \frac{\sqrt{lnn}}{\sqrt{n}}]) \\ \leq \frac{1}{n^{C_1}} + \frac{EM_F^2}{n(lnn)}, \end{aligned} \tag{B.1}$$

by Assumptions A1-A.2. There are three main differences in the proofs compared to the case of $p > n$. We describe them here. First (A.1)-(A.4) changes and hence we have

$$\lambda_n := K_2[\frac{\sqrt{lnp}}{\sqrt{n}} + \frac{\sqrt{EM_1^2}}{n} lnp + \frac{\sqrt{lnn}}{\sqrt{n}}],$$

so compared with (A.13) the third right side term changes. Then in Lemma A.3, N changes to

$$N \geq (32s_0)^2[\frac{K_3\sqrt{lnp^2} + K_3\sqrt{EM_1^2} lnp^2 + K_3\sqrt{ln(C_+p)}}{\phi_\Sigma^2(s_0)}],$$

where $C_+p \geq n$, and C_+ is a large positive constant, when $p = an$ with $0 < a < 1$ case. Other cases can be handled similarly. Next in Lemma A.4, the exception probability changes to

$$P(\mathcal{F}^c) \leq \frac{2}{n^{C_1}} + \frac{K_1'[EM_1^2 + EM_2^2]}{nlnn}.$$

After these three changes Theorems 1-3 carry as before. Asymptotically we have $\lambda_n = O(\sqrt{ln/n})$ when $p \leq n$ in Corollary 1.

B.2 Relaxing Assumption 1(iii)

In this subsection we relax Assumption 1(iii) from $\|\beta_0\|_2 = O(1)$ to $\|\beta_0\|_2 = O(\sqrt{s_0})$ and we explain the logic and meaning of this new assumption.

Assumption 1(iv).

$$\|\beta_0\|_2 = O(\sqrt{s_0}).$$

Assumption 1(iii) which is suggested by Jankova and van de Geer (2018) and simplifies their paper in semiparametric efficient estimators. Our Assumption 1(iv) here generalizes that assumption and in the case of s_0 being constant becomes Assumption 1(iii). The implication of Assumption 1(iv) is that all nonzero coefficients can be constant and none of them has to be local to zero.

$$\|\beta_0\|_2 = \sqrt{\sum_{j=1}^{p} \beta_{0,j}^2} = \sqrt{\sum_{j \in S_0} \beta_{0,j}^2} = O(\sqrt{s_0}).$$

In terms of discussion after Assumption 1, this implies $S_0 = D_1$, and D_2 is an empty set. So Assumption 1(iv) can simultaneously allow s_0 increasing with n, and all large nonzero coefficients in S_0. Previously in Assumption 1(iii), there can be only a fixed number of large coefficients, and increasing $(s_0 - f_1)$ number of local to zero (small) coefficients.

We have the counterpart to Theorem 1 with this new assumption.

Corollary B.1. *Under Assumptions 1(i)(ii)(iv), Assumption 2, with $n \geq 8$, and L_{11}, L_{12}, L_2 are positive constants if λ_n is chosen to reflect the following lower bound*

$$\lambda_n \geq max(L_{11}\frac{P(\mathcal{F}^c)^{1/4k}}{s_0^{1/2}}, L_{12}P(\mathcal{F}^c)^{1/2k}),$$

then

$$[E\|\hat{\beta} - \beta_0\|_1^k]^{1/k} \leq L_2 s_0 \lambda_n.$$

Remark. The lower bound for the tuning parameter λ_n in Corollary B.1 is larger than the one

in Theorem 1 due to $s_0 \geq 1$.

Proof of Corollary B.1. We proceed in a way that we only change the proofs in Appendix A, when necessary. All lemmata in Appendix A goes through, there is no usage of Assumption 1(iii) there. The first change comes in step 3 of Theorem 1 proof. First (A.29) changes to $\|\beta_0\|_1^{2k} \leq C_5 s_0^{2k}$ under Assumption 1(iv) instead of Assumption 1(iii). Then (A.30) becomes

$$E\left[\frac{\|u\|_n^2}{2\lambda_n}\right]^k + 2\|\beta_0\|_1^{2k} \leq 2C_4\lambda_n^{-2k} + 2C_5 s_0^{2k} \tag{B.2}$$

Then (A.32) changes to following

$$E\|\hat{\beta} - \beta_0\|_1^k \leq C_2 s_0^k \lambda_n^k + 2^k max(C_5^{1/2} s_0^k, C_4^{1/2}\lambda_n^{-k})\sqrt{P(\mathcal{F}^c)}. \tag{B.3}$$

(A.33)-(A.35) follows since they do not require Assumption 1(iv). To end the proof, if

$$C_2 s_0^k \lambda_n^k \geq 2^k \sqrt{C_5} s_0^k P(\mathcal{F}^c)^{1/2k},$$

which is possible by

$$\lambda_n \geq 2\left[\frac{C_5^{1/2}}{C_2}\right]^{1/k} P(\mathcal{F}^c)^{1/2k}.$$

Then we have

$$E\|\hat{\beta} - \beta_0\|_1^k \leq 2C_2 s_0^k \lambda_n^k.$$

L_{11}, L_{12}, L_2 are the same as before, and the uniformity over $\mathcal{B}_{l_0}(s_0)$ still holds. **Q.E.D.**

Here we provide the counterpart of Theorem 2, with Assumption 1(iv) instead of Assumption 1(iii).

Corollary B.2. *Under Assumptions 1(i)(ii)(iv),Assumption 2, with* $n \geq 8$, $P(\mathcal{F}^c) \leq min(2^{3-2k}, 1)$, $k \geq 1$, *and* L_3, L_4, L_5 *are positive constants if* λ_n *is chosen to reflect the following lower and upper bound*

$$L_3 \geq \lambda_n \geq L_4 \frac{P(\mathcal{F}^c)^{1/2k}}{s_0},$$

then

$$[E\|\hat{\beta}\|_1^k]^{1/k} \leq L_5 s_0.$$

Remarks. 1. The upper bound for moments of estimator is larger here compared with Theorem 2.

2. However, the upper bound here-L_3- for λ_n is larger compared with Theorem 2 upper bound for tuning parameter-$L_3/s_0^{1/2}$, where $s_0 \geq 1$. The lower bound here for λ_n is smaller than the one in Theorem 2 for the tuning parameter. Hence here with larger signal, choice of λ_n can take in a larger region compared with Theorem 2. But this comes with the price of larger moment bound for estimator as discussed in Remark 1.

Proof of Corollary B.2. With new Assumption 1(iv), (A.41) changes to

$$E\|\hat{\beta}\|_1^k 1_{\{\mathcal{F}\}} \leq 2C_5^{1/2} s_0^k. \tag{B.4}$$

with $\lambda_n \leq L_3$. Then follow proof of Corollary B.1, (B.2), so (A.43) changes to

$$\sqrt{E\|\hat{\beta}\|_1^{2k}}P(\mathcal{F}^c)^{1/2} \leq 2^k max((C_5^{1/2}/\sqrt{2})s_0^k, C_4^{1/2}\lambda_n^{-k})P(\mathcal{F}^c)^{1/2}. \tag{B.5}$$

Then by (B.4)(B.5)

$$E\|\hat{\beta}\|_1^k \leq 2C_5^{1/2}s_0^k + 2^k max((C_5^{1/2}/\sqrt{2})s_0^k, C_4^{1/2}\lambda_n^{-k})P(\mathcal{F}^c)^{1/2}. \tag{B.6}$$

There are two possibilities. First, with

$$2C_5^{1/2}s_0^k \geq 2^k C_4^{1/2}\lambda_n^{-k}P(\mathcal{F}^c)^{1/2},$$

we obtain

$$E\|\hat{\beta}\|_1^k \leq 4C_5^{1/2}s_0^k,$$

and this is possible with

$$\lambda_n \geq L_4\frac{P(\mathcal{F}^c)^{1/2k}}{s_0}.$$

The second possibility proceeds as in the proof of Theorem 2. So we have with the bounds

$$L_3 \geq \lambda_n \geq L_4\frac{P(\mathcal{F}^c)^{1/2k}}{s_0},$$

the following rate

$$E\|\hat{\beta}\|_1^k \leq 4C_5^{1/2}s_0^k,$$

hence the rate (upper bound) in Theorem 2 changes, and becomes larger here compared with Theorem 2. Uniformity over $\mathcal{B}_{l_0}(s_0)$ follows through as in the proof of Theorem 2. **Q.E.D.**

Corollary B.3.*Incentive compatibility is achieved by Assumptions 1(i)(ii)(iv) and Assumptions 2-3, and Condition 1, with $n \geq 8$, $P(\mathcal{F}^c) \leq 1/2$, and with the following bounds on λ_n*

$$min(L_3, \frac{c_1^2 c_n}{2c_2L_2s_0c_n + 2L_2L_5M_3M_4s_0^2}) \geq \lambda_n \geq max(\frac{L_{11}P(\mathcal{F}^c)^{1/8}}{s_0^{1/2}}, L_{12}P(\mathcal{F}^c)^{1/4}, \frac{L_4P(\mathcal{F}^c)^{1/4}}{s_0}).$$

Remarks. 1. To compare the case of Assumption 1(iii) with Assumption 1(iv) for $n \geq 8$ it is difficult to see whether λ_n bounds are different in these two cases. To see this, for the upper bound, the first term is L_3 in Corollary B.3 (case of larger signal), and this term is $L_3/s_0^{1/2}$ in the upper bound of Theorem 3. But, the second term in the upper bound is smaller in Corollary B.3 compared with the one in Theorem 3. Hence it is not clear which upper bound is larger.

2. For the lower bounds comparison in Corollary B.3 and Theorem 3, again the issue in Remark 1 is valid. The bounds are different and it is not clear which one is smaller. To see this, the second lower bound term in Corollary B.3 is always larger than the same condition in Theorem 3, but the third lower bound term in Corollary B.3 is smaller than the one in Theorem 3.

3. We consider the case of asymptotics. Assumption 5 does not change. But Assumption 6 has to change and strengthened to

$$s_0^2M_3M_4\sqrt{lnp/n} \to 0, \tag{B.7}$$

to have asymptotic incentive compatibility with the larger signal Assumption 1(iv) instead of Assumption 1(iii) in Theorem 3. Since in the asymptotic case, there is only lower bound, we consider the lower bound only. Assumptions 3-5 show that upper bounds are satisfied with $n \to \infty$ case.

First we analyze the second and third terms on the lower bounds for Corollary B.3, and it is clear that with $s_0 \to \infty$ with $n \to \infty$, the second term dominates the third term in Corollary B.3. So in the asymptotic case the lower bound condition for incentive compatibility is:

$$\lambda_n \geq max(\frac{L_{11}P(\mathcal{F}^c)^{1/8}}{s_0^{1/2}}, L_{12}P(\mathcal{F}^c)^{1/4}).$$

This last bound clearly shows that it is larger than equal to the bound in Theorem 3 lower bound in Corollary 1. Hence given we have to strengthen Assumption 6 and the lower bound for λ_n is larger here with the larger signal Assumption 1(iv). So it is more difficult to achieve incentive compatibility asymptotically with the larger signal than the case in Corollary 1.

Proof of Corollary B.3 The bounds for λ_n in the statement of Theorem 3 change to

$$L_3 \geq \lambda_n \geq max\left(\frac{L_{11}P(\mathcal{F}^c)^{1/4k}}{s_0^{1/2}}, L_{12}P(\mathcal{F}^c)^{1/2k}, \frac{L_4P(\mathcal{F}^c)^{1/2k}}{s_0}\right). \tag{B.8}$$

So the lower bound may be larger or smaller than the lower bound in Theorem 3, so the comparison of the lower bounds for λ_n here in (B.8) and in Theorem 3 does not provide a clear ranking. However, the upper bound for λ_n here in (B.8) is larger than the one in Theorem 3, since $L_3 \geq L_3/s_0^{1/2}$ and $s_0 \geq 1$. Our Assumption 1(iv) does not affect the proof of Theorem 3 until (A.60). Then by Corollaries B.1-B.2 equation (A.60) changes to the following

$$\lambda_n \leq \frac{c_1^2 c_n}{2c_2L_2s_0c_n + 2L_2L_5M_3M_4s_0^2}. \tag{B.9}$$

Hence (B.9) upper bound is smaller than the one in (A.60) in the proof of Theorem 3, since the second term in the denominator here is larger by a factor of $s_0^{1/2}$, and $s_0 \geq 1$. The rest of the proof follows as in the proof of Theorem 3.So the incentive compatibility is achieved with the following bound with $P(\mathcal{F}^c) \leq 1/2$,

$$min(L_3, \frac{c_1^2 c_n}{2c_2L_2s_0c_n + 2L_2L_5M_3M_4s_0^2}) \geq \lambda_n \geq max(\frac{L_{11}P(\mathcal{F}^c)^{1/8}}{s_0^{1/2}}, L_{12}P(\mathcal{F}^c)^{1/4}, \frac{L_4P(\mathcal{F}^c)^{1/4}}{s_0}),$$

uniformity over $\mathcal{B}_{l_0}(s_0)$ is achieved since all the bounds depend on β_0 only through s_0.**Q.E.D.**

C Appendix C

This section provides the proofs for the incentive compatibility of conservative Lasso, which is explained in Section 5. Let wpa1 denote with probability approaching one.

We have the following l_1 norm result, which is Lemma A.1 in Caner and Kock (2018).

Lemma C.1. *Let $0 < a_n \leq 1$, where a_n is a deterministic-positive sequence in n, then on the event $\mathcal{A}_1 \cap \mathcal{A}_2$*

$$\|\hat{\beta}_w - \beta_0\|_1 \leq 4(a_n + 1)(2a_n + 1)\frac{\lambda_n s_0}{\phi_\Sigma^2(s_0)} \leq C_{w2}\lambda_n s_0,$$

where $C_{w2} := \frac{24}{\phi_\Sigma^2(s_0)}$, and is a positive constant.

We start with the proof of moments for conservative Lasso's moments. This is extending

Theorem 1 to a more general weighted penalty.

Proof of Theorem 4. The proof will mirror proof of Theorem 1 above. We show the places that will differ.

Step 1. Using Lemma C.1 above

$$E\|\hat{\beta}_w - \beta_0\|_1^k 1_{\{\mathcal{F}\}} \leq C_{w2} s_0^k \lambda_n^k. \tag{C.1}$$

Step 2. This is exactly the same in Step 2, Theorem 1. It only involved error terms not the penalty. (A.25)-(A.26) are valid here as well, but instead of $2k$ moments we need $4k$ moments. This implies

$$E[\|u\|_n^2]^{4k} := E|\frac{1}{n}\sum_{i=1}^n u_i^2|^{4k} \leq 2C_4',$$

with $C_3' := 2^{4k-1}[8ln2]^{2k}C', C' > \sigma^{8k}, C_4' := 2^{4k-1}\sigma^{8k}$, and $n \geq 8$.

Step 3. This step is a major extension of step 3 for Theorem 1, and extends the Lasso penalty and its moments to a more general-data dependent weighted-conservative Lasso. Using the definition for conservative Lasso

$$\|Y - X\hat{\beta}_w\|_n^2 + 2\lambda_n \sum_{j=1}^p \hat{w}_j|\hat{\beta}_{w,j}| \leq \|Y - X\beta_0\|_n^2 + 2\lambda_n \sum_{j=1}^p \hat{w}_j|\beta_{0,j}|.$$

Ignoring the first term since its nonnegative and $u := Y - X\beta_0$

$$\sum_{j=1}^p \hat{w}_j|\hat{\beta}_{w,j}| \leq \frac{\|u\|_n^2}{2\lambda_n} + \sum_{j=1}^p \hat{w}_j|\beta_{0,j}|. \tag{C.2}$$

Thus since $\hat{w}_{max} := \max_{1\leq j\leq p} \hat{w}_j \leq 1$, and define $\hat{w}_{min} := \min_{1\leq j\leq p} \hat{w}_j$. Note that

$$\sum_{j=1}^p \hat{w}_j|\hat{\beta}_{w,j}| \geq \sum_{j=1}^p \hat{w}_{min}|\hat{\beta}_{w,j}|,$$

$$\sum_{j=1}^p \hat{w}_j|\beta_{0,j}| \leq \sum_{j=1}^p \hat{w}_{max}|\beta_{0,j}| \leq \sum_{j=1}^p |\beta_{0,j}|.$$

we can rewrite (C.2)

$$\|\hat{\beta}_w\|_1 \leq \frac{\|u\|_n^2}{2\lambda_n \hat{w}_{min}} + \frac{\|\beta_0\|_1}{\hat{w}_{min}}. \tag{C.3}$$

Use triangle inequality

$$\|\hat{\beta}_w - \beta_0\|_1 \leq \|\hat{\beta}_w\|_1 + \|\beta_0\|_1.$$

Then take expectations above and use (C.3) and (A.25)

$$
\begin{aligned}
E\|\hat{\beta}_w - \beta_0\|_1^{2k} &\leq 2^{2k-1}\left\{ E\left[\frac{\|u\|_n^2}{2\lambda_n \hat{w}_{min}}\right]^{2k} + E\left[\frac{2\|\beta_0\|_1}{\hat{w}_{min}}\right]^{2k}\right\} \\
&= 2^{2k-1}\left\{ \frac{1}{(2\lambda_n)^{2k}} E\left[\frac{\|u\|_n^2}{\hat{w}_{min}}\right]^{2k} + [2\|\beta_0\|_1]^{2k} E\left[\frac{1}{\hat{w}_{min}}\right]^{2k}\right\} \\
&\leq 2^{2k-1}\left\{ \frac{1}{(2\lambda_n)^{2k}} [E[\|u\|_n^2]^{4k}]^{1/2} [E(\hat{w}_{min})^{-4k}]^{1/2} + [2\|\beta_0\|_1]^{2k} E\left[\frac{1}{\hat{w}_{min}}\right]^{2k}\right\}, \quad \text{(C.4)}
\end{aligned}
$$

where we use Cauchy-Schwartz inequality for the first term on the right side to get the last inequality. We consider the term $\hat{w}_{min}^{-1}$ in (C.4)

$$
\hat{w}_{min}^{-1} = \frac{\max_{1\leq j\leq p} |\hat{\beta}_j| \cup \lambda_{prec}}{\lambda_{prec}}.
$$

If $\max_{1\leq j\leq p} |\hat{\beta}_j| \leq \lambda_{prec}$ then $\hat{w}_{min}^{-1} = 1$. With that estimated minimum weight, the proofs of Theorem 1 can go forward, but unfortunately since estimated minimum weight can take another value and make the problem and the proofs more complicated. Now we show this issue. If $\max_{1\leq j\leq p} |\hat{\beta}_j| > \lambda_{prec}$ then

$$
\hat{w}_{min}^{-1} = \frac{\max_{1\leq j\leq p} |\hat{\beta}_j|}{\lambda_{prec}}.
$$

By Assumption 7,

$$
\left\{ E\left[\left(\frac{1}{\hat{w}_{min}}\right)^{4k}\right]\right\}^{1/2} \leq \left(\frac{1}{C_{w1} d_n}\right)^{2k}.
$$

On the right-side of (C.4)

$$
\begin{aligned}
(\frac{1}{2\lambda_n})^{2k} [E(\|u\|_n^2)^{4k}]^{1/2} [E(\hat{w}_{min})^{-4k}]^{1/2} \quad &+ \quad [2\|\beta_0\|_1]^{2k} E[\frac{1}{\hat{w}_{min}}]^{2k} \\
&\leq \quad (\frac{1}{2\lambda_n})^{2k} (2C_4')^{1/2} \frac{1}{(C_{w1} d_n)^{2k}} + 2^{2k} \frac{C_5 s_0^k}{(C_{w1} d_n)^{2k}}.
\end{aligned}
$$

This last inequality implies by (C.4)

$$
E\|\hat{\beta}_w - \beta_0\|_1^{2k} \leq \frac{\sqrt{C_4'}}{\sqrt{2}\lambda_n^{2k}} \frac{1}{(C_{w1} d_n)^{2k}} + 2^{4k-1} C_5 s_0^k \frac{1}{(C_{w1} d_n)^{2k}}. \quad \text{(C.5)}
$$

Step 4. Now merge the rates in (C.1)(C.5)

$$
\begin{aligned}
E\|\hat{\beta}_w - \beta_0\|_1^k &= E\|\hat{\beta}_w - \beta_0\|_1^k 1_{\mathcal{F}} + E\|\hat{\beta}_w - \beta_0\|_1^k 1_{\mathcal{F}^c} \\
&\leq E\|\hat{\beta}_w - \beta_0\|_1^k 1_{\mathcal{F}} + \sqrt{E\|\hat{\beta}_w - \beta_0\|_1^{2k}} \sqrt{P\mathcal{F}^c} \\
&\leq C_{w2} s_0^k \lambda_n^k + \max\left(\left[\frac{C_4'^{1/4}}{2^{1/4} \lambda_n^k (C_{w1} d_n)^k}, \frac{2^{2k-1/2} C_5^{1/2} s_0^{k/2}}{(C_{w1} d_n)^k} \right] \right) P(\mathcal{F}^c)^{1/2}. \quad \text{(C.6)}
\end{aligned}
$$

There are two possibilities by following the proof of Theorem 1. The first one is:

$$C_{w2}s_0^k\lambda_n^k \geq \frac{C_4'^{1/4}}{2^{1/4}\lambda_n^k(C_{w1}d_n)^k}P(\mathcal{F}^c)^{1/2},$$

if the second term on the right side of (C.6) is larger or equal to the third term on the right side of the same inequality (C.6). This last inequality is implied by the choice of the tuning parameter

$$\lambda_n \geq \frac{L_{w11}P(\mathcal{F}^c)^{1/4k}}{s_0^{1/2}d_n^{1/2}}, \tag{C.7}$$

with $L_{w11} := (\frac{C_4'}{2})^{1/4k}(\frac{1}{C_{w2}})^{1/2k}\frac{1}{C_{w1}^{1/2}}$ is a positive constant. The second possibility is:

$$C_{w2}s_0^k\lambda_n^k \geq \frac{2^{2k-1/2}C_5^{1/2}s_0^{k/2}}{(C_{w2}d_n)^k}P(\mathcal{F}^c)^{1/2},$$

when the third term on the right side of (C.6) is larger than or equal to the second term on the right side of (C.6). That last inequality is implied by the following lower bound on the tuning parameter

$$\lambda_n \geq \frac{L_{w12}P(\mathcal{F}^c)^{1/2k}}{d_n s_0^{1/2}}, \tag{C.8}$$

where $L_{w12} := 2^{2-1/2k}C_5^{1/2k}/(C_{w2})^{1/k+1}$. Then under the two lower bounds (C.7)(C.8) for λ_n we obtain

$$E\|\hat{\beta}_w - \beta_0\|_1^k \leq 2C_{w2}s_0^k\lambda_n^k,$$

and define $L_{w2} := (2C_{w2})^{1/k}$. The proof is uniform over $\mathcal{B}_{l_0}(s_0)$ ball, since the upper bound depends on β_0 only through s_0 definition. **Q.E.D**

Proof of Theorem 5. See that

$$E\|\hat{\beta}_w\|_1^k \leq E\|\hat{\beta}_w\|_1^k 1_{\{\mathcal{F}\}} + \sqrt{E\|\hat{\beta}_w\|_1^{2k}}\sqrt{P(\mathcal{F}^c)}. \tag{C.9}$$

By Lemma C.1 on the event $\mathcal{F}$

$$\|\hat{\beta}_w\|_1 \leq \|\hat{\beta}_w - \beta_0\|_1 + \|\beta_0\|_1 \leq C_{w2}s_0\lambda_n + \sqrt{s_0}\|\beta_0\|_2.$$

Then use (A.25)(A.29)

$$E\|\hat{\beta}_w\|_1^k 1_{\{\{\mathcal{F}\}\}} \leq 2\max\left[C_{w2}^k s_0^k\lambda_n^k, C_5^{1/2}s_0^{k/2}\right] \tag{C.10}$$

Then given the upper bound $\lambda_n \leq L_{w3}/s_0^{1/2}$ with $L_{w3} := (C_5^{1/2k}/C_{w2})$ a positive constant,

$$E\|\hat{\beta}_w\|_1^k 1_{\{\mathcal{F}\}} \leq 2C_5^{1/2}s_0^{k/2}. \tag{C.11}$$

To handle the second right side term on (C.9), use (C.3) and (A.25) to obtain

$$E\|\hat{\beta}_w\|_1^{2k} \leq 2^{2k-1}\left\{E\left[\frac{\|u\|_n^2}{2\lambda_n \hat{w}_{min}}\right]^{2k} + E\left[\frac{\|\beta_0\|_1}{\hat{w}_{min}}\right]^{2k}\right\}$$
$$\leq 2^{2k-1}\left\{\frac{1}{(2\lambda_n)^{2k}}E\left[(\|u\|_n^2)^{4k}\right]^{1/2}E[\hat{w}_{min}^{-4k}]^{1/2} + \|\beta_0\|_1^{2k}E[\hat{w}_{min}^{-2k}]\right\}, \tag{C.12}$$

where we use Cauchy-Schwartz inequality for the first right side term in second inequality. Then use (C.12) via (A.29) and the step 2 of proof of Theorem 4

$$E\|\hat{\beta}_w\|_1^{2k} \leq \frac{\sqrt{C_4'}}{\sqrt{2}}\frac{1}{\lambda_n^{2k}}\frac{1}{(C_{w1}d_n)^{2k}} + 2^{2k-1}\frac{C_5 s_0^k}{(C_{w1}d_n)^{2k}}. \tag{C.13}$$

Next using (C.13)

$$\sqrt{E\|\hat{\beta}_w\|_1^{2k}}\sqrt{P(\mathcal{F}^c)} \leq max\left(\frac{C_4'^{1/4}}{2^{1/4}}\frac{1}{\lambda_n^k}\frac{1}{(C_{w1}d_n)^k}, 2^{k-1/2}\frac{C_5^{1/2}s_0^{k/2}}{(C_{w1}d_n)^k}\right)\sqrt{P(\mathcal{F}^c)}. \tag{C.14}$$

Merge the rates in (C.11) and (C.14)

$$E\|\hat{\beta}_w\|_1^{k} \leq 2C_5^{1/2}s_0^{k/2} + max\left(\frac{C_4'^{1/4}}{2^{1/4}}\frac{1}{\lambda_n^k}\frac{1}{(C_{w1}d_n)^k}, 2^{k-1/2}\frac{C_5^{1/2}s_0^{k/2}}{(C_{w1}d_n)^k}\right)\sqrt{P(\mathcal{F}^c)}. \tag{C.15}$$

There are two possibilities in (C.15). First,

$$2C_5^{1/2}s_0^{k/2} \geq \frac{(C_4'/2)^{1/4}}{\lambda_n^k(C_{w_1}d_n)^k}P(\mathcal{F}^c)^{1/2},$$

which is implied by the lower bound on the tuning parameter

$$\lambda_n \geq L_{w41}\frac{P(\mathcal{F}^c)^{1/2k}}{s_0^{1/2}d_n},$$

with $L_{w41} := (C_4'/2)^{1/4k}/(2^{1/k}C_5^{1/2k}C_{w1})$ a positive constant. Then the second possibility is

$$2C_5^{1/2}s_0^{k/2} \geq \frac{2^{k-1/2}C_5^{1/2}s_0^{k/2}}{C_{w1}^k d_n^k}P(\mathcal{F}^c)^{1/2},$$

which is implied by the lower bound on d_n

$$d_n \geq L_{w42}P(\mathcal{F}^c)^{1/4k},$$

with $L_{w42} := 1/(C_{w1}2^{(3/2k)-1})^8$. Next, by the lower bounds on λ_n, d_n in this proof above we have

$$E\|\hat{\beta}_w\|_1^{k} \leq 2C_5^{1/2}s_0^{k/2}.$$

[8]Note that the lower bound on d_n can be as small as $L_{w2}P(\mathcal{F}^c)^{1/2k}$ but that rate will not hold jointly with the upper bound on λ_n in the proof of Theorem 5

Note that by choosing $L_{w5} := 2^{1/k} C_5^{1/2k}$ a positive constant, we have the result. Also the uniformity proof follows through since β_0 is in the upper bound through s_0 definition only.**Q.E.D.**

D Appendix D

This section provides the proofs for conservative Lasso IC(case of $n \to \infty, p \to \infty$), which is explained in Section 5.1. Let "wpa1" denote with probability approaching one. First, we start with l_∞ bound for Lasso estimator. This bound is needed for Conservative Lasso for the proofs of moment bounds. Lemma D1(i) is for sufficiently large n, and then Lemma D1(ii)(iii) are the asymptotic results.

Lemma D.1. *(i). Under Assumptions 1-2, and on $\mathcal{A}_1 \cap \mathcal{A}_2$, with $n \geq N$*

$$\|\hat{\beta} - \beta_0\|_\infty \leq \|\Theta\|_{l_\infty} [\frac{\lambda_n}{2} + t_1 \frac{24\lambda_n s_0}{\phi^2_\Sigma(s_0)} + \lambda_n],$$

with the definition

$$\lambda_{prec} := \|\Theta\|_{l_\infty} [\frac{\lambda_n}{2} + t_1 \frac{24\lambda_n s_0}{\phi^2_\Sigma(s_0)} + \lambda_n].$$

t_1 is defined in (A.18).
(ii). With added Assumptions 4-6,8 to (ii) above, we obtain $\lambda_{prec} = O(s_1 \lambda_n)$, with $d_n = s_1 \lambda_n$.
(iii). The result in (i) holds wpa1 with Assumptions 1-2,4-6,8 and

$$\|\hat{\beta} - \beta_0\|_\infty = O_p(\lambda_{prec}) = O_p(s_1 \lambda_n) = o_p(1).$$

Remark. Also $\|\Theta\|_{l_\infty} = O(s_1)$ allows the row sums of precision matrix to be diverging with n. Hence we relax the restrictive assumption of constant maximum row sum of the precision matrix in Caner and Kock (2018) as well as the one in van de Geer (2016).

Proof of Lemma D.1.

(i). By Lemma 2.5.1 of van de Geer (2014) or (A.25) of Caner and Kock (2018)

$$\|\hat{\beta} - \beta_0\|_\infty \leq \|\Theta\|_{l_\infty} \left[\|\frac{X'u}{n}\|_\infty + \|\hat{\Sigma} - \Sigma\|_\infty \|\hat{\beta} - \beta_0\|_1 + \lambda_n \right].$$

Now on $\mathcal{A}_1 \cap \mathcal{A}_2$ with Lemma A.1-A.2 and (A.18)-(A.19)

$$\|\hat{\beta} - \beta_0\|_\infty \leq \|\Theta\|_{l_\infty} [\frac{\lambda_n}{2} + t_1 \frac{24\lambda_n s_0}{\phi^2_\Sigma(s_0)} + \lambda_n].$$

(ii). So we define

$$\lambda_{prec} := \|\Theta\|_{l_\infty} \lambda_n [\frac{3}{2} + \frac{24 t_1 s_0}{\phi^2_\Sigma(s_0)}],$$

with Assumptions 4-6 added, we have $t_1 = O(\sqrt{lnp/n})$ and so $s_0 \sqrt{lnp/n} = o(1)$, we get via Assumption 8 $\lambda_{prec} = O(s_1 \lambda_n)$, given that $\|\Theta\|_{l_\infty} = O(s_1)$, and $\phi^2_\Sigma(s_0) \geq c > 0$, for $c > 0$ is a positive constant.

(iii). This is true by Assumptions 1-2,4-6,8.**Q.E.D.**

We have the following l_1 norm result, which is Lemma A.1 in Caner and Kock (2018). Their assumptions are slightly stronger, with our new Lemma A.2-A.3 for the sets $\mathcal{A}_1, \mathcal{A}_2$ we can prove under our Assumptions 1- 2,4-6.

Lemma D.2. *Under Assumptions 1-2,4-6,8*

$$\|\hat{\beta}_w - \beta_0\|_1 = O_p(\lambda_n s_0) = o_p(1).$$

We start with the proof of moments for conservative Lasso's moments in asymptotics. The proof of Theorem 4 also allows $n \to \infty$ but here, when $n \to \infty$ special case, conditions simplify for lower bounds in λ_n.

Lemma D.3. *Under Assumptions 1-2,4-6,8 and with the sufficient condition,*

$$\lambda_n s_0^{1/3} s_1^{1/3} P(\mathcal{F}^c)^{-1/6k} \to \infty,$$

we obtain when $n \to \infty$

$$[E\|\hat{\beta}_w - \beta_0\|_1^k]^{1/k} = O(\lambda_n s_0) = o(1).$$

Proof of Lemma D.3. The proof will mirror proof of Theorem 4 above. We show the places that will differ.

Step 1. Using Lemma D.2 above

$$E\|\hat{\beta}_w - \beta_0\|_1^k 1_{\{\mathcal{F}\}} = O(s_0^k \lambda_n^k). \tag{D.1}$$

Step 2. This is exactly the same in Step 2, Theorem 4. It only involved error terms not the penalty.

Step 3. This follows step 3-(C.4) in Theorem 4 proof. Note that when $\max_{1\leq j\leq p} |\hat{\beta}_j| \leq \lambda_{prec}$, $\hat{w}_{min} = 1$. However, when $\max_{1\leq j\leq p} |\hat{\beta}_j| > \lambda_{prec}$, we have with with probability approaching one, given Lemma D.1 and since

$$\hat{w}_{min} = \max_{1\leq j\leq p} |\hat{\beta}_j|/\lambda_{prec},$$

$$\hat{w}_{min}^{-1} \leq \frac{\max_{1\leq j\leq p} |\hat{\beta}_j - \beta_{0,j}| + \max_{1\leq j\leq p} |\beta_{0,j}|}{\lambda_{prec}} \tag{D.2}$$

$$\leq \frac{\lambda_{prec} + C}{\lambda_{prec}} = 1 + \frac{C}{\lambda_{prec}}. \tag{D.3}$$

By Assumption 8 we know $\lambda_{prec} = o(1)$, via Lemma D.1(ii)-(iii) with $d_n = s_1 \lambda_n$

$$\hat{w}_{min}^{-1} = O_p(\lambda_{prec}^{-1}).$$

Regardless of whether $|\hat{\beta}_j|$ is larger than or equal to or less than λ_{prec} we have, with Lemma D.1

$$\frac{1}{\hat{w}_{min}} = O_p(s_1^{-1} \lambda_n^{-1}),$$

since its diverging in n when $|\hat{\beta}_j| > \lambda_{prec}$, and one otherwise. So

$$\left[E\left(\frac{1}{\hat{w}_{min}^{4k}} \right) \right]^{1/2} = O(s_1^{-2k} \lambda_n^{-2k}), \tag{D.4}$$

as well. With (A.29)(D.4)in (C.4)

$$
\begin{aligned}
E\|\hat{\beta}_w - \beta_0\|_1^{2k} &\leq 2^{2k-1}\{\frac{1}{(2\lambda_n)^{2k}}[(E\|u\|_n^2)^{4k}]^{1/2}[E(\hat{w}_{min})^{-4k}]^{1/2} + [2\|\beta_0\|_1]^{2k}E\left[\frac{1}{\hat{w}_{min}}\right]^{2k}\} \\
&= O(\lambda_n^{-2k})O(1)O(\lambda_{prec}^{-2k}) + O(s_0^k)O(\lambda_{prec}^{-2k}) \\
&= O(\lambda_n^{-2k})O(s_1^{-2k}\lambda_n^{-2k}) + O(s_0^k)O(s_1^{-2k}\lambda_n^{-2k}) \qquad \text{(D.5)} \\
&= O(s_1^{-2k}\lambda_n^{-4k}), \qquad \text{(D.6)}
\end{aligned}
$$

where we use $(s_0^k\lambda_n^k)\lambda_n^k \to 0$ by Assumption 5, so to get last equality, the first rate dominated in (D.5).

Step 4. Now merge the rates in (D.1)(D.6)

$$
\begin{aligned}
E\|\hat{\beta}_w - \beta_0\|_1^k &= E\|\hat{\beta}_w - \beta_0\|_1^k 1_{\{\mathcal{F}\}} + E\|\hat{\beta}_w - \beta_0\|_1^k 1_{\{\mathcal{F}^c\}} \\
&\leq O(s_0^k\lambda_n^k) + \sqrt{E\|\hat{\beta}_w - \beta_0\|_1^{2k}}\sqrt{P(\mathcal{F}^c)} \\
&= O(s_0^k\lambda_n^k) + O(s_1^{-k}\lambda_n^{-2k})P(\mathcal{F}^c)^{1/2}. \qquad \text{(D.7)}
\end{aligned}
$$

To establish a rate

$$
s_0^k\lambda_n^k \geq \lambda_n^{-2k}s_1^{-k}P(\mathcal{F}^c)^{1/2}, \qquad \text{(D.8)}
$$

which (D.8) is implied by

$$
\lambda_n s_0^{1/3}s_1^{1/3}P(\mathcal{F}^c)^{-1/6k} \to \infty.
$$

This shows

$$
E\|\hat{\beta}_w - \beta_0\|_1^k = O(s_0^k\lambda_n^k).
$$

Q.E.D

Lemma D.4. *Under Assumptions 1-2,4-6,8 with*

$$
\lambda_n s_0^{1/4}s_1^{1/2}P(\mathcal{F}^c)^{-1/4k} \to \infty,
$$

we obtain with $n \to \infty$

$$
[E\|\hat{\beta}_w\|_1^k]^{1/k} = O(s_0^{1/2}) = o_p(1).
$$

Proof of Lemma D.4. By (C.9)

$$
E\|\hat{\beta}_w\|_1^k \leq E\|\hat{\beta}_w\|_1^k 1_{\{\mathcal{F}\}} + \sqrt{E\|\hat{\beta}_w\|_1^{2k}}\sqrt{P(\mathcal{F}^c)}. \qquad \text{(D.9)}
$$

By Lemma D.2, and (A.29)

$$
\begin{aligned}
\|\hat{\beta}_w\|_1 &\leq \|\hat{\beta}_w - \beta_0\|_1 + \|\beta_0\|_1 \\
&= O_p(\lambda_n s_0) + O(\sqrt{s_0}) = O_p(\sqrt{s_0}), \qquad \text{(D.10)}
\end{aligned}
$$

and the last equality is by Assumption 5, since $s_0\lambda_n \to 0$. Hence by Assumption 6, the upper bound for λ_n in Theorem 5 is satisfied. So

$$
E\|\hat{\beta}_w\|_1^k 1_{\{\mathcal{F}\}} = O(s_0^{k/2}). \qquad \text{(D.11)}
$$

Note that by Assumption 8 and Lemma D.1(ii), $d_n = s_1\lambda_n$. Then to handle the second term on

the right side in (D.9), use (C.14)

$$\begin{aligned}\sqrt{E\|\hat{\beta}_w\|_1^{2k}}\sqrt{P(\mathcal{F}^c)} &= O(max(\frac{1}{\lambda_n^{2k}s_1^k}, \frac{s_0^{k/2}}{s_1^k\lambda_n^k}))\sqrt{P(\mathcal{F}^c)} \\ &= O(\frac{1}{\lambda_n^{2k}s_1^k})\sqrt{P(\mathcal{F}^c)}, \end{aligned} \tag{D.12}$$

where the second rate equality is obtained by Assumption 5, $\frac{s_0^{k/2}/(s_1^k\lambda_n^k)}{1/s_1^k\lambda_n^{2k}} = s_0^{k/2}\lambda_n^k \to 0$. Then merge the rates in (D.9) by (D.11) and (D.12)

$$E\|\hat{\beta}_w\|_1^k = O(s_0^{k/2}),$$

since the first rate dominates by the following lower bound on λ_n, with $d_n := s_1\lambda_n$, with $n \to \infty$

$$\lambda_n s_0^{1/4} s_1^{1/2} P(\mathcal{F}^c)^{-1/4k} \to \infty.$$

Q.E.D.

Proof of Corollary 2. The analysis is the same in Section 4.1 to understand the simplification of the proof of Theorems 3 and 6. That analysis simplifies the upper bounds for λ_n. But the lower bound for λ_n is ($k = 2$) implied by, with $n \to \infty$

$$\frac{\lambda_n}{max\left(\frac{P(\mathcal{F}^c)^{1/8}}{s_0^{1/4}s_1^{1/2}}, \frac{P(\mathcal{F}^c)^{1/12}}{s_0^{1/3}s_1^{1/3}}\right)} \to \infty.$$

Q.E.D.

References

Buhlmann, P. and S. van de Geer (2011). *Statitistics for High-Dimensional Data*. Springer.

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

Caner, M. and A. B. Kock (2019). High dimensional linear gmm. *arXiv:1811.08779*.

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*.

Chetverikov, D., Z. Liao, and V. Chernozhukov (2021). On cross-validated lasso in high dimensions. *Annals of Statistics 49*, 1300–1317.

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.

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*.

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

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.

Hastie, T., R. Tibshirani, and J. Friedman (2011). *The elements of statistical learning*. Springer.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An introduction to statistical learning*. Springer.

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.

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

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

Shaywitz, D. (2020). "the alignment problem" review: When machines miss the point. *The Wall Street Journal*, A25,25 October.

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. (2014). *Statistical Theory for High Dimensional Models*.

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*.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*, 1418–1429.