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

Mehmet Caner† Kfir Eliaz‡

August 30, 2021

Abstract

We consider situations where a user feeds her attributes to a machine learning method that tries to predict her best option based on a random sample of other users. The predictor is incentive-compatible if the user has no incentive to misreport her covariates. Focusing on the popular Lasso estimation technique, we borrow tools from high-dimensional statistics to characterize sufficient conditions that ensure that Lasso is incentive compatible in large samples. 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. We present simulations that illustrate how this can be done in practice.

*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 Eccles School of Business, the University of Utah. Email: kfire@tauex.tau.ac.il.

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 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 user may not know the values of the coefficients $\beta_{\mathbf{0}}$, in which

[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.

case she would have some (possibly degenerate) prior beliefs over them. 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 he 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 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, and for *any* prior belief on β_0 : I.e., if for every possible value of β_0 and X_{n+1}, the expected value of $(X'_{n+1}\beta_0 - R(X_{n+1})'\hat{\beta})^2$ is minimized at the truth $R(X_{n+1}) = X_{n+1}$, where the expectation is taken with respect to the statistician's sample.

Intuition suggests that an individual cannot benefit from lying to a procedure that is meant to predict the best outcome for her. To counter this intuition, Eliaz and Spiegler (2019), and Eliaz and Spiegler (2020) use the above framework to illustrate that a user may have a strict incentive to lie about her attributes when the prediction is based on a linear regression that penalizes non-zero estimated coefficients. The rough intuition is that the user believes that despite the statistician's good intentions, these estimation techniques lead to distortions, which she tries to undo by lying. For instance, given the user's beliefs about the true model parameters, she may be concerned that the estimator will admit too many irrelevant attributes, and hence, she reports a zero value for these attributes (see Eliaz and Spiegler (2019), and Eliaz and Spiegler (2020) for more details). However, these papers focus on particular examples in which attributes are *binary*, the statistician has the *same* (fixed) finite number of observations on each possible combination of attribute values, and the penalty parameter is *fixed* and does *not* adjust to the sample size. That is, these papers only raise the problem of incentive compatibility but do not provide an econometric solution. Hence, they leave open the following important question: For a general environment, are there conditions ensuring that a penalized regression model is incentive compatible in large samples?

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

This paper addresses the above open question by first focusing on the most popular form of penalized regressions - the *Lasso* estimator.[2] Borrowing tools from high-dimensional statistics, we establish sufficient conditions for incentive compatibility of the Lasso estimator in large samples. 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 2 in Section 4). 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

[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.

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 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 in large samples. 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 resuls when we relax our assumption on the signal to noise ratio. Finally, Appendix C contains the proofs for the general weighted Lasso.

2 The model

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

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., $\|A\|_{l_1} = \max_{1 \le k \le l} \sum_{i=1}^m |A_{ik}|$ which is also called the induced l_1 norm of A. Let $\|A\|_{l_\infty} := \max_{1 \le i \le m} \sum_{k=1}^l |A_{ik}|$ 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

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

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. A user may or may not know the value of the coefficients $(\beta_{0,1}, ..., \beta_{0,p})$. In the latter case, she has some (possibly degenerate) prior beliefs over their values.

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]

[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 a couple of examples for such data, which may be obtained

The X_i' s are also i.i.d. across i and exogenous, and will be discussed in detail in Assumption 1 in the next subsection. β_0 is a $p \times 1$ vector, representing the true parameters in f. 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). s_0 is a nondecreasing function of n, and we assume $s_0 \geq 1$. These facts are known to an "oracle" but not to the statistician (and possibly not to a user).

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

$$\hat{\beta} = argmin_{\beta \in R^p} \frac{\sum_{i=1}^{n} (y_i - X_i'\beta)^2}{n} + 2\lambda_n \|\beta\|_1, \tag{2}$$

where $\lambda_n > 0$ is the penalty (also called tuning parameter) that decreases with the number of observations at the rate of $\lambda_n = O(\sqrt{lnp/n})$ (an explicit expression for the sequence λ_n is given in equation (A.14) in Appendix A).[4]

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}.

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 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, she takes into account that she does not observe the statistician's sample, and hence, does

from a third party, or from marketing surveys.

[4]We established this rate in Lemma A.2 in Appendix A.

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 minimize her expected loss $E_{\beta_0,\hat{\beta}}(R(X_{n+1})'\hat{\beta} - X'_{n+1}\beta_0)^2$, 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 $\hat{\beta}$. Hence, the new user may decide to lie and report $R(X_{n+1}) \neq X_{n+1}$. In particular, she may decide to "opt out" and submit a vector of zeros.[5] Our objective is to understand under what conditions it is in the user's best interest to be truthful regardless of her prior beliefs on β_0.

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

Definition 1. *An estimator is* ***incentive-compatible*** *if for every* X_{n+1}*, for every* $R(X_{n+1})$ *and for every every* β_0*,*

$$E[R(X_{n+1})'\hat{\beta} - X'_{n+1}\beta_0]^2 \geq E[X'_{n+1}\hat{\beta} - X'_{n+1}\beta_0]^2. \tag{3}$$

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 by misreporting her personal characteristics, *regardless* of her beliefs over the true model's parameters in mean squared sense.[6] How should we interpret this requirement, given that we do not

[5] 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.

[6] If we were to relax the requirement that truth-telling is preferred for *every* prior belief over the true

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.

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.

To see that our definition of incentive-compatibility is not vacuous, simply add and

model's parameters, we would need to make some assumptions on the user's prior beliefs (see, e.g. Eliaz and Spiegler (2020)). Thus, our incentive-compatibility has the merit of being robust to *any* specification of prior beliefs.

subtract the term $X'_{n+1}\hat{\beta}$ inside the squared brackets on the left side term of (3), such that

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

Canceling common terms reduces incentive-compatibility to the following inequality:

$$
E\left[\|(R(X_{n+1}) - X_{n+1})'\hat{\beta}\|_2^2\right] \geq -2E[\hat{\beta}'(R(X_{n+1}) - X_{n+1})X'_{n+1}(\hat{\beta} - \beta_0)]. \tag{4}
$$

Note that this inequality 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 $\tilde{X}_{n+1} - X_{n+1}$ 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. For instance, with one explanatory variable, a very small lie can lead to a very small positive number on the left hand side (due to a small lie being squared), while the right hand side may be positive and slightly larger. For our main result, we analyze the asymptotic version of incentive compatibility.

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:

$$
\int E[(R(X_{n+1})'\hat{\beta} - X'_{n+1}\beta_0)^2]dP_{X_{n+1}} \geq \int E[(X'_{n+1}\hat{\beta} - X'_{n+1}\beta_0)^2]dP_{X_{n+1}},
$$

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.

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). [7] 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

[7] 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).

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 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.

2.4 The Statistician's Data

In this subsection, we introduce a number of restrictions on the statistician's data. To describe these restrictions, we shall make use of the following notations. Define an l_0 ball $\mathcal{B}_{l_0}(s_0) = \{\|\beta_0\|_{l_0} \leq s_0\}$. 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. Our first requirement extends the sub-Gaussian data assumption used in statistics:

Assumption 1. *(i).*$E(u_i|X_i) = 0$, X_i, u_i *are identical and independent across* $i = 1, \cdots, n$, *and for some positive constant* C,

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

$$E|u_i|^l \leq C < \infty$$

where $l = max(2k, 4)$ *for all* $k \geq 1$,

(ii). $\phi_{min}(\Sigma) \geq c > 0$, *where c is a positive constant.*

Our second set of restrictions applies to the first and second moments. These will guarantee the consistency of the Lasso estimator, but will not ensure incentive compatibility (sufficient conditions for incentive compatibility will be introduced in Section 4). We start by defining the maximal value of certain cross products, which will be related to the behavior of moments in high dimensions in our next assumption.

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

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

Note 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. 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. [8]

Assumption 2. *(i).*

$$\frac{\sqrt{lnp}}{\sqrt{n}}[max((EM_1^2)^{1/2}, (EM_2^2)^{1/2})] \to 0.$$

(ii). $s_0(\frac{lnp}{n})^{1/2} \to 0.$

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

Assumption 2(i) and 2(ii) are standard in high dimensional econometrics. In particular, 2(i) is used in Chernozhukov et al. (2017) allowing them to apply a concentration inequality, and 2(ii) is a standard sparsity condition. Note that with Assumption 2(ii), Lasso prevents

[8] Alternatively, we could provide primitives on Assumption 2 using boundedness of individual moments of X, u.

underfitting since letting $\lambda = O(\sqrt{\frac{lnp}{n}})$ implies that $s_0\lambda_n \to 0$, which ensures that λ_n cannot be large enough to generate underfitting. This allows us to establish the consistency of Lasso in Lemmas A.1-A.3 in the Appendix.

Assumption 2(iii) ensures that the signal to noise ratio is bounded (see p.2343 of Jankova and van de Geer (2018)). 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 1). Hence, when $E(u_i|X_i) = 0$, which is imposed in Assumption 1,

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

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 2(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 $C_0 := \frac{\|\beta_0\|_2^2\phi_{min}(\Sigma)}{\sigma_u^2}$.

The empirical implication of this is that only a fixed number of nonzero coefficients can be constants, and the other nonzero coefficients have to be local to zero. To see this implication, 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

$$\sqrt{\sum_{j\in S_0} \beta_{0,j}^2} = \sqrt{\sum_{j\in F_1} \beta_{0,j}^2 + \sum_{j\in S_0-F_1} \beta_{0,j}^2} \leq \sqrt{f_1C^2 + (s_0 - f_1)\frac{C^2}{s_0 - f_1}} = O(1),$$

where $F_1 := \{j : |\beta_{0,j}| = C\}$ with $|F_1| = f_1$ being a fixed number, C is a generic positive constant and $F_2 := \{j : |\beta_{0,j}| = \frac{C}{\sqrt{s_0-f_1}}\}$ with $|F_2| = s_0 - f_1$. For ease of exposition, we set all coefficients in F_1 and F_2 to be the same constants, C and $C/\sqrt{s_0 - f_1}$, respectively. F_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 2(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 F_2 above). In other words, there is no index set F_2 as above, but all nonzero coefficients (their indices) are in the set F_1 above.

As p and n grow large, the total number of nonzero coefficients s_0 (also known as the *sparsity index*) can grow arbitrarily large. To guarantee consistency and unbiasedness, it is typically assumed that the product of the sparsity index and the tuning parameter should go to zero. However, this standard condition does not guarantee the incentive compatibility of the Lasso estimator as can be seen in the proof of Theorem 3 below.

3 New Oracle Inequalities for Lasso

Oracle inequalities in high dimensional statistics are upper bounds on prediction and estimation errors. For our main result, we require moment bounds on the Lasso estimator's error in l_1 norm. By taking the sample size to be large, we can show that the upper bound on the mean of higher-order moments of Lasso estimation errors tend to zero. We then use this asymptotic result to establish the incentive compatibility of the Lasso estimator in large samples. To illustrate this, we note that from the proof of Theorem 3 in Appendix A.2.4, the incentive compatibility constraint is tied to the following expression

$$E[R(X_{n+1})'\hat{\beta} - X'_{n+1}\beta_0]^2 - E[X'_{n+1}\hat{\beta} - X'_{n+1}\beta_0]^2 = E[\hat{\beta}'(R(X_{n+1}) - X_{n+1})(R(X_{n+1}) - X_{n+1})'\hat{\beta}] \quad (5)$$

$$+ E[\hat{\beta}'(R(X_{n+1}) - X_{n+1})X'_{n+1}(\hat{\beta} - \beta_0)] \quad (6)$$

$$+ E[(\hat{\beta} - \beta_0)'X_{n+1}(R(X_{n+1})' - X'_{n+1})\hat{\beta}]. \quad (7)$$

For incentive compatibility to hold in large samples, we need the sum of the right-hand side terms to be greater than or equal to zero. The first term on the right-hand side (5) is always non-negative. Hence, if we prove that (6) and (7) converge to zero, we establish asymptotic incentive compatibility. However, the size of terms in (6) and (7) will depend on the mean of higher-order estimation errors of Lasso.

To bound these errors, we prove 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 also consider 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. 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.6) and (A.9), 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. *Under Assumptions 1-2, if n is sufficiently large and $\lambda_n \geq \frac{P(\mathcal{F}^c)^{1/4k}}{s_0^{1/2}}$, then*

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

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

If we set $k = 1$ we can learn whether the Lasso estimator is unbiased. By the above Theorem, Assumption 2 and (A.15) imply $s_0 \lambda_n \to 0$. Hence, in large samples, we have unbiasedness in the large λ_n case. Next, we provide the k-th moment bound for l_1 norm for the Lasso estimator.

Theorem 2. *Under Assumptions 1-2, if n is sufficiently large n and $\lambda_n \geq P(\mathcal{F}^c)^{1/2k}/s_0^{1/2}$, then*

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

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

This is a new result and a simple extension of Theorem 1 above. The rate in Theorem diverges to infinity if $s_0 \to \infty$ as $n \to \infty$.

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. In other words, we establish conditions such that when $n \to \infty$,

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

for all X'_{n+1} and $R(X_{n+1})'$ and for every β_0, where the expectation is taken with respect to the statistician's realized sample (since the reporting user does not observe this sample).

The proof of this result makes use of the following notation.

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

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

Note that M_4 is the absolute magnitude of the potential misreport on a given variable j by the $n+1$ user. Since we deal with ex-post incentive compatibility, M_3 and M_4 are deterministic but can grow with n. Hence, we allow M_3 and M_4 to be nondecreasing in n.

Theorem 3. *Under Assumptions 1 and 2, the Lasso estimator is incentive compatible in large samples ($n \to \infty$) if the following conditions hold:*

$$\lambda_n \geq P(\mathcal{F}^c)^{1/8}/s_0^{1/2} \tag{8}$$

and

$$s_0^{3/2}\sqrt{\frac{lnp}{n}}[M_3][M_4] \to 0. \tag{9}$$

Furthermore, incentive compatibility is valid uniformly over $\mathcal{B}_{l_0}(s_0) = \{\|\beta_0\|_{l_0} \leq s_0\}$.

Remarks.

1. Theorem 3 establishes that a *sufficient* condition for incentive compatibility is that the tuning parameter λ_n needs to be large "enough". A simple way to choose λ_n to satisfy (8) is to use the upper bound of the exception probability

$$\lambda_n := \textit{upperbound}(P(\mathcal{F}^c)^{1/8}),$$

in Lemma A.4. The simulations in Section 6 address the issue of whether such a bound is feasible.

2. 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 (as Lemma A.1 in Appendix A shows, this rate is $s_0\lambda_n \to 0$). 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 the point with a simple example. Suppose we take a value for λ_n below the upper bound in Theorem 1 (i.e., $\lambda_n < P(\mathcal{F}^c)^{1/4k}/s_0^{1/2}$). In particular, take $\lambda_n = P(\mathcal{F}^c)^{1/2k}/s_0^{1/2}$. Then from the proof of Theorem 1-(A.34) we obtain that

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

But given our choice of λ_n,

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

Hence, even though there is consistency ($s_0\lambda_n \to 0$) under this λ_n choice (see Remark 5), the moment bound estimation error is *diverging.*

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

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.

3. The second sufficient condition (9) allows the distance between the user's report $R(X_{n+1})$ and the truth X_{n+1} to be of any magnitude since $M_4 \equiv \|R(X_{n+1})-X_{n+1}\|_\infty$ can be arbitrarily large. Since the above conditions are sufficient but not necessary, it remains an open question whether incentive compatibility can be achieved with a tuning parameter that is lower than the threshold in (8) without restricting the magnitude of the deviation between the user's reported and true attributes.

4. Note that (9) requires stricter sparsity than Assumption 2. If $M_3 = O(1)$ and $M_4 = O(1)$, then condition (9) amounts to $s_0^{3/2}\sqrt{\frac{ln(p)}{n}} \to 0$, which is a sparsity requirement still stronger than Assumption 2(ii). In addition, if we let $M_4 = O(ln\,(n))$ and $M_3 = O(ln\,(n))$, then $s_0^{3/2}\sqrt{\frac{ln(p)}{n}}(ln\,(n))^2 \leq 4s_0^{3/2}\sqrt{\frac{ln(p)}{n}} \to 0$ is needed to get incentive compatibility with $n \leq p$.

5. A natural question that arises is whether condition (8) is compatible with the l_1 norm consistency of Lasso. In other words, consistency requires a small λ_n, but incentive compatibility requires a large λ_n, 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 $s_0\lambda_n \to 0$. To verify whether this is possible, we can take the lower bound on the tuning parameter in (8) and see whether we can achieve consistency. Note that

$$s_0\lambda_n = s_0\frac{P(\mathcal{F}^c)^{1/8}}{s_0^{1/2}} = s_0^{1/2}P(\mathcal{F}^c)^{1/8}, \tag{10}$$

From (A.22) in the Appendix, an upper bound on this exception probability is:

$$P(\mathcal{F}^c) \leq \frac{2}{p^{C_1}} + \frac{K[EM_1^2 + EM_2^2]}{nlnp}, \tag{11}$$

where C_1 and K are positive constants. With $l = 1, 2$, it therefore follows from (10) and (11) that we need

$$s_0^4/p^{C_1} \to 0, \quad s_0^4 \max_l EM_l^2/nlnp \to 0,$$

to have consistency. These two conditions are not unreasonable in the sense that they are consistent with (n,p) increasing to infinity. Also they are compatible with moments satisfying condition (9) in Theorem 3.

6. Finally, note that $\lambda_n = O(\sqrt{\frac{ln(p)}{n}})$ represents an upper bound in terms of rates for λ_n, whereas (8) represents a lower bound. We can then take for a positive constant $C > 0$

$$C\frac{\sqrt{ln(p)}}{\sqrt{n}} \geq \lambda_n \geq \frac{P(\mathcal{F}^c)^{1/8}}{s_0^{1/2}}.$$

The question is, are there suitable combinations of n and p that satisfy these inequalities? By using algebra and the upper bound for exception probability (A.22), we obtain the requirement that,

$$Cs_0^{1/2} \geq \left[\frac{2n}{p^{C_1}} + \frac{K[EM_1^2 + EM_2^2]}{nlnp}\right]^{1/8} \frac{\sqrt{n}}{\sqrt{lnp}},$$

which is plausible for $p > n$ and large n since the left hand side may diverge and the right side may go to zero. This may be the case for example when p is exponential in n, or large C.

7. When we relax Assumption 2(iii) 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. Remarks 5-6 above still apply but with slightly stronger sparsity conditions.

8. While Theorem 3 provides sufficient conditions for incentive-compatibility, we can also derive a necessary condition. As we show in Appendix A, incentive-compatibility implies equation (A.52). By Markov's inequality, we obtain that if the Lasso estimator is incentive-compatibile in large samples, then

$$\left[\sum_{j=1}^{p}(\hat{\beta}_j - \beta_{0,j})X_{n+1,j}\right]\left[\sum_{j=1}^{p}(R(X_{n+1,j}) - X_{n+1,j})\hat{\beta}_j\right] \xrightarrow{p} 0,$$

which implies a weighted consistency condition for the Lasso estimator. Note that a condition like (8) is not involved, and hence, there is still a gap between our sufficient and necessary conditions. It remains a challenging open question whether there exist conditions for asymptotic incentive compatibility of Lasso that are both necessary and sufficient.

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 C.1 in Appendix C.

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. 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 assumption:

Assumption 3. *Define the precision matrix,* $\Theta := \Sigma^{-1}$. *Then*

(i).

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

with s_1 *a nondecreasing positive sequence in* n.

(ii). $0 < c \leq \max_{1\leq j\leq p} |\beta_{0,j}| \leq C < \infty$.

(iii). $s_1\lambda_n = o(1)$.

Assumption 3(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 3(ii) prevents the maximum absolute coefficient from being a sequence that is local to zero. A local to zero sequence for the maximum coefficient is unrealistic, and furthermore, it implies that all the coefficients in the model converge to zero, which renders the model useless to begin with. Assumption 3(iii) 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 3(i).

Theorem 4. *Under Assumptions 1-3, with sufficiently large* n, *and with*

$$\lambda_n \geq \frac{P(\mathcal{F}^c)^{1/6k}}{s_0^{1/3} s_1^{1/3}}$$

then we obtain

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

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

To the best of our knowledge, this is an entirely 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 proofs are not trivial and involve finding the rate for minimal estimated weight.

Note that the lower bound for the tuning parameter in Theorem 4 may be weakly higher than the one in Theorem 1: if $s_1 \leq \sqrt{s_0}$

$$\frac{P(\mathcal{F}^c)^{1/6k}}{s_0^{1/3} s_1^{1/3}} \geq \frac{P(\mathcal{F}^c)^{1/4k}}{s_0^{1/2}},$$

since $s_0 \geq 1$. If, however, $s_1 > \sqrt{s_0}$, then it is not clear which bound will be higher.

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

Theorem 5. *Under Assumptions 1-3, with sufficiently large n, and with*

$$\lambda_n \geq \frac{P(\mathcal{F}^c)^{1/4k}}{s_0^{1/4} s_1^{1/2}}$$

then we obtain

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

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

We are now ready to characterize the sufficient conditions for incentive-compatibility in large samples of the Conservative Lasso.

Theorem 6. *Under Assumptions 1-3, with sufficiently large n, and with*

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

and

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

then Conservative Lasso is Incentive Compatible in large samples. The result is valid uniformly over $\mathcal{B}_{l_0}(s_0)$.

Remarks.

1. If our Assumption 3 imposed $s_1 = O(1)$, as in Caner and Kock (2018), then we would need a larger bound for the tuning parameter of the Conservative Lasso compared with Lasso. This follows from observing that under this restriction on s_1,

$$\frac{P(\mathcal{F}^c)^{1/8}}{s_0^{1/2}} \leq \frac{P(\mathcal{F}^c)^{1/8}}{s_0^{1/4} s_1^{1/2}}.$$

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

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

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

3. A natural question that arises is whether a lower bound for λ_n is compatible with consistency as in Remark 5 of Theorem 3. Applying the lower bound in Theorem 6, let

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

Using (11) 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 Theorem 6 results in weaker conditions. These conditions are

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

4. Note that in a high dimensional penalized regression, the tuning parameter λ_n is an upper bound on the noise as defined by $\mathcal{A}_1$ in (A.6). As in the case of Lasso (see Remark 6 following Theorem 3), we verify whether the lower bound for incentive-compatibility is compatible with the upper bound for noise reduction. Namely, we check whether

$$C \frac{\sqrt{ln\,(p)}}{\sqrt{n}} \geq \lambda_n \geq \frac{P(\mathcal{F}^c)^{1/8}}{s_0^{1/4} s_1^{1/2}}.$$

The question is, are there suitable combinations of n and p that satisfy these inequalities? By using algebra and the upper bound for the exception probability (A.22), we obtain the requirement that,

$$Cs_0^{1/4}s_1^{1/2} \geq \left[\frac{2n}{p^{C_1}} + \frac{K[EM_1^2 + EM_2^2]}{nlnp}\right]^{1/8} \frac{\sqrt{n}}{\sqrt{lnp}},$$

which is plausible for $p > n$ since the left-hand side may diverge and the right-hand side may go to zero. For example, this may be the case when p is exponential in n. If, instead, we were to use

$$C\frac{\sqrt{lnp}}{\sqrt{n}} \geq \lambda_n \geq \frac{P(\mathcal{F}^c)^{1/12}}{s_0^{1/3}s_1^{1/3}}.$$

we would obtain the requirement that,

$$Cs_0^{1/3}s_1^{1/3} \geq \left[\frac{2n}{p^{C_1}} + \frac{K[EM_1^2 + EM_2^2]}{nlnp}\right]^{1/12} \frac{\sqrt{n}}{\sqrt{lnp}},$$

As before, this is plausible for $p > n$ since the left hand side may diverge and the right side may go to zero (e.g., when p is exponential in n). Note that given that the upper bounds for λ_n is the same, (see Lemma A.4 of Caner and Kock (2018)), the lower bound for Conservative Lasso is weakly higher than that of Lasso. Thus, the range of (n, p) values that satisfy both bounds is smaller in Conservative Lasso.

5. We can also relax Assumption 2(iii) to $\|\beta_0\|_2 = O(\sqrt{s_0})$. The analysis will be similar to that of Lasso (see Appendix B).

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 Theorem 3), incentive compatibility is analyzed through the lens of a small lie and larger lie. Finally, we show that incentive compatibility is not vacuous, it is possible to have new users lying to the machines and benefit from that.

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. Let s_0 represent the sparsity of the above model and set $s_0 = 5$.

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. That new user's draw is deterministic-non-random. 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). 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 Theorem 3, 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 Theorem 3 is the upper bound on the exception probability,

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

The issue is to make the exception probability, $P(\mathcal{F}^c)$ operational and usable. Note that an upper bound on this probability is (with positive constants $C_1 > 0, C_2 > 0, K > 0$)

$$P(\mathcal{F}^c) \leq \frac{2}{p^{C_1}} + \frac{K[EM_1^2 + EM_2^2]}{nlnp} \leq \frac{2}{p^{C_1}} + \frac{C_2}{(lnp)^2}, \tag{12}$$

by observing that for $l = 1, 2$

$$\begin{aligned} \frac{K \max_l EM_l^2}{nlnp} &= \left[\frac{K^{1/2}\sqrt{\max_l EM_l^2}}{\sqrt{n}\sqrt{lnp}} \right]^2 \\ &= \left[\frac{K^{1/2}\sqrt{\max_l EM_l^2}\sqrt{lnp}}{\sqrt{n}} \right]^2 (\frac{1}{lnp})^2 \\ &\leq \frac{C_2}{(lnp)^2}, \end{aligned}$$

where we use Assumption 2(i). Hence, we can write the upper bound of the exception probability by using $p \geq 1$

$$\frac{2}{p^{C_1}} + \frac{C_2}{(lnp)^2} \leq 2 + \frac{C_2}{(lnp)^2}.$$

The tuning parameter is as follows

$$\lambda_n := \left[2 + \frac{C_2}{(lnp)^2} \right]^{1/8}, \tag{13}$$

where C_2 can start from a small positive value and stop at a large positive value. We select the values for C_2 and λ_n according to the Generalized Information Criterion (GIC) as in Caner and Kock (2018), which gives consistent model selection with weighted Lasso choices in the least squares framework (the choice of tuning parameter with GIC in least squares with Lasso and Conservative Lasso is shown to be 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_2 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 Theorem 3). On the other hand, to prevent a very large λ_n and ensure consistency of Lasso, the lower bound inversely depends on p.

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_2 in a grid of values $[2 + \frac{C_2}{(lnp)^2}]$ as in (13). Let $C_2 := [0.01, 0.1, 0.5, 1, 2, 10, 100]$, so Λ is the grid of values of λ_n depending on C_2. The number of iterations is 1,000.

For the Conservative Lasso, the same type of tuning parameter analysis is used, but with Theorem 6, instead of Theorem 3. Hence, the tuning parameter choice for conservative lasso is:

$$\lambda_n := \left[2 + \frac{C_2}{(lnp)^2} \right]^{1/12}, \tag{14}$$

The Choice of C_2 is done in the same way as in lasso above. The "Report" column 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 3.08, while the disutility from lying is 3.86. Hence, the $n+1$ user prefers to be truthful. In Table 2, for a small lie, truth-telling induces a disutility of 3.08, while lying induces a higher disutility of 2.04. 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 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

Table 1: Lasso-Incentive Compatibility:

Difference 2	$n = 100$		$n = 200$		$n = 300$	
Dimension	Truth	Report	Truth	Report	Truth	Report
$p = 100$	2.71	4.25	2.85	3.54	2.72	3.71
$p = 200$	0.99	18.01	0.76	17.95	0.68	17.93
$p = 300$	3.35	4.22	3.08	3.86	3.05	3.61

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
$p = 100$	2.71	1.80	2.85	1.90	2.71	1.76
$p = 200$	0.99	1.64	0.76	1.40	0.68	1.32
$p = 300$	3.35	2.32	3.08	2.04	3.05	2.00

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.

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 1.55 from lying compared to 2.49 from truth-telling. However, with a larger lie, the mean squared error from lying increases to 5.45 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
$p = 100$	2.49	5.45	2.56	4.78	2.43	4.85
$p = 200$	0.89	19.78	0.70	19.48	0.63	19.47
$p = 300$	3.06	5.27	2.79	4.94	2.78	4.65

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
$p = 100$	2.49	1.55	2.56	1.58	2.43	1.47
$p = 200$	0.89	1.57	0.70	1.36	0.63	1.29
$p = 300$	3.06	2.03	2.79	1.75	2.78	1.72

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 previously made by people raises the question of how people would interact with these automated systems. In particular, would people 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.

In the next part, Appendix A considers the proofs when $p > n$, and Appendix B considers the case $p \leq n$, and relaxing Assumption 2(iii). Appendix C covers Conservative Lasso proofs.

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 two assumptions 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 $C > 0$.

Assumption A.2. Assume

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

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))

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

for a constant $K > 0$. With Assumptions A.1-A.2 here, Caner and Kock (2019) or Lemma

E.1 of Chernozhukov et al. (2017) provides

$$E \max_{1 \le j \le p} |\hat{\mu}_j - \mu_j| \le K[\frac{\sqrt{lnp}}{\sqrt{n}} + \frac{\sqrt{EM_F^2 lnp}}{n}]$$
$$= O(\frac{\sqrt{lnp}}{\sqrt{n}}). \quad \text{(A.2)}$$

Define the sequence $\kappa_n = lnp$. Set $t = t_n = (n\kappa_n)^{1/2}$ to have (A.1) as

$$P\left[\max_{1 \le j \le p} |\hat{\mu}_j - \mu_j| \ge 2E \max_{1 \le j \le p} |\hat{\mu}_j - \mu_j| + \frac{\sqrt{\kappa_n}}{\sqrt{n}}\right] \le exp(-C_1\kappa_n) + K\frac{EM_F^2}{n\kappa_n}$$
$$= \frac{1}{p^{C_1}} + \frac{KEM_F^2}{nlnp} \quad \text{(A.3)}$$

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

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

$$P(\max_{1 \le j \le p} |\hat{\mu}_j - \mu_j| \ge 2K[\frac{\sqrt{lnp}}{\sqrt{n}} + \frac{(EM_F^2)^{1/2} lnp}{n}] + \frac{\sqrt{lnp}}{\sqrt{n}})$$
$$\le \frac{1}{p^{C_1}} + \frac{KEM_F^2}{n(lnp)} = o(1), \quad \text{(A.4)}$$

by Assumptions A1-A.2. This shows also that, since EM_F^2 is nondecreasing in n

$$\max_{1 \le j \le p} |\hat{\mu}_j - \mu_j| = O_p(\sqrt{lnp}/\sqrt{n}). \quad \text{(A.5)}$$

A.2.1 Events

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

$$\mathcal{A}_1 = \left\{2\left\|\frac{u'X}{n}\right\|_\infty \le \lambda_n\right\}, \quad \text{(A.6)}$$

which controls the noise. This is the maximal correlation between regressors and errors. We want this to be bounded with probability approaching one, and this upper bound, λ_n, itself is converging to zero in our proofs. We show that in Lemma A.2. So in large samples, this proof technique amounts to verification of exogeneity of regressors. This is standard in high dimensional econometrics, for a recent analysis see Lemma A.4 of Caner and Kock (2018).

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

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

Note that if $\Sigma = EX_iX_i'$ has full rank, the population adaptive restricted eigenvalue being positive is satisfied by Assumption 1. Also instead of minimizing all over R^p, we minimize vectors that satisfy $\|\delta_{S^c}\|_1 \leq 3\|\delta_S\|_1$. 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 1:

$$\hat{\phi}^2_{\hat{\Sigma}}(s) = \min\left\{ \frac{\delta'\hat{\Sigma}\delta}{\|\delta_S\|_2^2} : \delta \in R^p - \{0\}, \|\delta_{S_c}\|_1 \leq 3\sqrt{s}\|\delta_S\|_2, |S| \leq s \right\}. \tag{A.8}$$

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:

$$\mathcal{A}_2 = \left\{ \hat{\phi}^2_{\hat{\Sigma}}(s_0) \geq \phi^2_{\Sigma}(s_0)/2 \right\}. \tag{A.9}$$

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. Empirical adaptive eigenvalue is over a restricted set which we prove to be positive, with probability approaching one, in Lemma A.3. This is also standard in high dimensional econometrics, see Lemma A.6 of Caner and Kock (2018). 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

$$\|X(\hat\beta - \beta_0)\|_n^2 + \lambda_n \sum_{j\in S_0^c} |\hat\beta_j| \leq 3\lambda_n \sqrt{s_0}\|\hat\beta_{S_0} - \beta_{0,S_0}\|_2. \tag{A.10}$$

Ignoring the first term on the left of (A.10), (A.10) 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.8) in (A.10)

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

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

$$\|X(\hat{\beta}-\beta_0)\|_n^2 \leq \frac{18\lambda_n^2 s_0}{\phi_\Sigma^2(s_0)}. \tag{A.11}$$

To get to the l_1 bound ignore the first term in (A.10) 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.11) and the event $\mathcal{A}_2$ to have

$$\begin{aligned}\|\hat{\beta}-\beta_0\|_1 &\leq 4\sqrt{s_0}\left[\frac{3\sqrt{2}\lambda_n\sqrt{s_0}}{\phi_\Sigma(s_0)}\right]\left[\frac{1}{\hat{\phi}_{\hat{\Sigma}}(s_0)}\right] \\ &\leq \frac{24\lambda_n s_0}{\phi_\Sigma^2(s_0)}. \end{aligned} \tag{A.12}$$

Note that uniformity over $\mathcal{B}_{l_0}(s_0)$ follows since the upper bound in (A.12) depends on β_0 only through s_0. **Q.E.D**

Lemma A.2. *(i). Under Assumption 1, and since* $\kappa_n = lnp$

$$P(\mathcal{A}_1) \geq 1 - exp(-C_1\kappa_n) - \frac{KEM_1^2}{(n\kappa_n)} = 1 - \frac{1}{p^{C_1}} - \frac{KEM_1^2}{nlnp}$$

(ii). Under added Assumption 2 to Assumption 1, $P(\mathcal{A}_1) \to 1$.

(iii). Under added Assumption 2 to Assumption 1, $\lambda_n = O(\sqrt{lnp/n})$.

Proof of Lemma A.2. (i). Establish the probability bound on $\mathcal{A}_1$ via Assumption 1, using (A.3)(A.4) with $F_i = X_i u_i$ there and $\kappa_n = lnp$, we have

$$P(\mathcal{A}_1) \geq 1 - exp(-C_1\kappa_n) - K\frac{EM_1^2}{(n\kappa_n)} = 1 - \frac{1}{p^{C_1}} - \frac{KEM_1^2}{nlnp}, \tag{A.13}$$

with

$$\lambda_n = K[\sqrt{\frac{lnp}{n}} + \frac{\sqrt{EM_1^2}lnp}{n}] + \sqrt{\frac{lnp}{n}}. \tag{A.14}$$

(ii). By Assumption 2, we have the proof.

(iii). By Assumption 2, we have

$$\lambda_n = O(\sqrt{lnp/n}). \tag{A.15}$$

Q.E.D.

Lemma A.3. *Under Assumptions 1, 2,* $\kappa_n = lnp$

$$P(\mathcal{A}_2) \geq 1 - exp(-C_1\kappa_n) - \frac{KEM_2^2}{(n\kappa_n)} = 1 - \frac{1}{p^{C_1}} - \frac{KEM_2^2}{nlnp} = 1 - o(1).$$

Proof of Lemma A.3. Start with

$$\left|\delta'\frac{X'X}{n}\delta\right| = \left|\delta'(\frac{X'X}{n} - \Sigma + \Sigma)\delta\right|$$
$$\geq |\delta'\Sigma\delta| - |\delta'(\hat{\Sigma} - \Sigma)\delta|. \tag{A.16}$$

The second term on the right side of (A.16) can be bounded by repeated application of

Holders inequality

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

So (A.16) becomes

$$|\delta'\hat{\Sigma}\delta| \geq |\delta'\Sigma\delta| - \|\delta\|_1^2 \|\hat{\Sigma} - \Sigma\|_\infty. \tag{A.17}$$

Now we digress a bit to simplify (A.17). 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

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

Now divide (A.17) by $\|\delta_{S_0}\|_2^2 > 0$ to have

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

Minimize over δ on the both sides

$$\hat{\phi}_{\hat{\Sigma}}^2(s_0) \geq \phi_\Sigma^2(s_0) - 16 s_0 \|\hat{\Sigma} - \Sigma\|_\infty. \tag{A.18}$$

So if we can prove that with probability approaching one, $16 s_0 \|\hat{\Sigma} - \Sigma\|_\infty \leq \phi_\Sigma^2(s_0)/2$, that will imply of $\hat{\phi}_{\hat{\Sigma}}^2(s_0) \geq \phi_\Sigma^2(s_0)/2$ with probability approaching one. Define $\epsilon_n = 16 s_0 t_1$, where

$$t_1 = K[\sqrt{\frac{lnp^2}{n}} + \frac{\sqrt{EM_2^2}lnp^2}{n}] + \sqrt{\frac{lnp}{n}}. \tag{A.19}$$

By (A.3)(A.4), via Assumption 1

$$P[16s_0\|\hat{\Sigma}-\Sigma\|_\infty > \epsilon_n] = P[\|\hat{\Sigma}-\Sigma\|_\infty > t_1]$$
$$\leq exp(-C_1 lnp) + \frac{KEM_2^2}{(nlnp)}$$
$$\to 0, \tag{A.20}$$

where we use Assumption 2 for the probability tail converging to zero. Also see that by Assumption 2, $\epsilon_n \to 0$ since $s_0\sqrt{lnp/n} \to 0$. So we get, with probability approaching one, $16s_0\|\hat{\Sigma}-\Sigma\|_\infty \leq \epsilon_n \leq \phi_\Sigma^2(s_0)/2$, since left side of that inequality converges to zero in probability, and the right side is constant. Then by (A.18)(A.20)

$$P[\hat{\phi}_{\hat{\Sigma}}^2(s_0) \geq \phi_\Sigma^2(s_0)/2] \geq 1 - exp(-C_1\kappa_n) - \frac{KEM_2^2}{(n\kappa_n)}$$
$$= 1 - \frac{1}{p^{C_1}} - \frac{KEM_2^2}{nlnp}$$
$$= 1 - o(1). \tag{A.21}$$

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* $\kappa_n = lnp$

$$P(\mathcal{F}^c) \leq 2exp(-C_1\kappa_n) + \frac{K[EM_1^2 + EM_2^2]}{(n\kappa_n)}$$
$$= \frac{2}{p^{C_1}} + \frac{K(EM_1^2 + EM_2^2)}{nlnp} = o(1).$$

Proof of Lemma A.4.

Now we provide an upper bound for the probability $P(\mathcal{F}^c)$ in our case under Assumptions

1, 2, by using Lemmata A.2-A.3

$$\begin{aligned}
P(\mathcal{F}^c) &= P(\mathcal{A}_1 \cap \mathcal{A}_2)^c = P(\mathcal{A}_1^c \cup \mathcal{A}_2^c) \leq P(\mathcal{A}_1^c) + P(\mathcal{A}_2^c) \\
&\leq 2exp(-C_1\kappa_n) + \frac{K[EM_1^2 + EM_2^2]}{(n\kappa_n)} \\
&= \frac{2}{p^{C_1}} + \frac{K[EM_1^2 + EM_2^2]}{nlnp} \\
&\to 0. \qquad \text{(A.22)}
\end{aligned}$$

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 several steps.

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

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

We want to form rates for the right side terms in (A.23).

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

$$E\|\hat{\beta} - \beta_0\|_1^k 1_{\{\mathcal{F}\}} = O(s_0^k \lambda_n^k). \qquad \text{(A.24)}$$

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. Use Nemirowski's moment inequality, Lemma 14.24 in Buhlmann and van de Geer (2011), with for all $k \geq 1$, for the first inequality, and for the second inequality by Loeve's

c_r inequality, and for the equality we use u_i being iid, also the definition of $\sigma^2 := Eu_i^2$,

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

by Assumption 1. Before the next result we provide the inequality,

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

for $k \geq 1$, and x, y being generic scalars, and σ^2 being bounded above by Assumption 1 and using (A.25)

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

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 k th moment of the sampling error in l_1 norm, and using (A.25) by taking expectations there for the second inequality below

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

We use the assumption $\|\beta_0\|_2 = O(1)$ to have

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

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

$$E\left[\frac{\|u\|_n^2}{2\lambda_n}\right]^k + 2\|\beta_0\|_1^k = O(\lambda_n^{-k}) + O(s_0^{k/2}) = O(\max(s_0^{k/2}, \lambda_n^{-k})). \tag{A.30}$$

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

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

$$E\|\hat{\beta} - \beta_0\|_1^k = O(\max(s_0^{k/2}, \lambda_n^{-k})). \tag{A.31}$$

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. The rate in (A.31) is diverging to infinity.

We can simplify the rate further, since we are aiming for an asymptotic result for incentive compatibility, with sufficiently large n, by Assumption 2, $s_0^{k/2}\lambda_n^k = \frac{s_0^k \lambda_n^k}{s_0^{k/2}} \leq 1$. This last inequality implies that with sufficiently large n, the rate in (A.31) is

$$E\|\hat{\beta} - \beta_0\|_1^k = O(\lambda_n^{-k}). \tag{A.32}$$

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

$$\begin{aligned} E\|\hat{\beta} - \beta_0\|_1^k &= E\|\hat{\beta} - \beta_0\|_1^k 1_{\{\mathcal{F}\}} + E\|\hat{\beta} - \beta_0\|_1^k 1_{\{\mathcal{F}^c\}} \\ &\leq O(s_0^k \lambda_n^k) + \sqrt{E\|\hat{\beta} - \beta_0\|_1^{2k}}\sqrt{E 1_{\{\mathcal{F}^c\}}} \\ &= O(s_0^k \lambda_n^k) + O(\lambda_n^{-k})\sqrt{P(\mathcal{F}^c)} \end{aligned} \tag{A.33}$$

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

We can get the rate:

$$s_0^k \lambda_n^k \geq \lambda_n^{-k} P(\mathcal{F}^c)^{1/2}. \tag{A.34}$$

By (A.33)(A.34)

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

We can simplify further (A.34),

$$\lambda_n \geq P(\mathcal{F}^c)^{1/4k} / s_0^{1/2}. \tag{A.35}$$

So if $\lambda_n \geq P(\mathcal{F}^c)^{1/4k} / s_0^{1/2}$ then

$$E\|\hat{\beta} - \beta_0\|_1^k = O(s_0^k \lambda_n^k). \tag{A.36}$$

The uniformity over $\mathcal{B}_{l_0}(s_0)$ follows since the rates in (A.24)(A.31)-(A.34) depends on β_0 only by s_0. **Q.E.D.**

Remark. 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.38) below

$$\lambda_n \geq \frac{P(\mathcal{F}^c)^{1/4k}}{s_0^{1/4}}, \tag{A.37}$$

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

$$P(\mathcal{F}^c) \leq \frac{2}{(2p)^{\tau^2/2}} \tag{A.38}$$

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.38) is different than our Lemma A.4.

Proof of Theorem 2.

We start with

$$E\|\hat{\beta}\|_1^k = E\|\hat{\beta}\|_1^k 1_{\{\mathcal{F}\}} + E\|\hat{\beta}\|_1^k 1_{\{\mathcal{F}^c\}} \leq E\|\hat{\beta}\|_1^k 1_{\{\mathcal{F}\}} + \sqrt{E\|\hat{\beta}\|_1^{2k}}\sqrt{P(\mathcal{F}^c)}, \tag{A.39}$$

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

$$\begin{aligned}
\|\hat{\beta}\|_1 &\leq \|\hat{\beta} - \beta_0\|_1 + \|\beta_0\|_1 \\
&\leq \frac{24\lambda_n s_0}{\phi_\Sigma^2(s_0)} + \sqrt{s_0}\|\beta_0\|_2 \\
&= O_p(\sqrt{s_0}),
\end{aligned}$$

by Assumptions 1, 2. This last rate shows that

$$E\|\hat{\beta}\|_1^k 1_{\{\mathcal{F}\}} = O(s_0^{k/2}). \tag{A.40}$$

To handle the second right side term in (A.39) 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.30) to get

$$\begin{aligned}\sqrt{E\|\hat{\beta}\|_1^{2k}}P(\mathcal{F}^c)^{1/2} &= O(\max(s_0^{k/2}, \lambda_n^{-k}))P(\mathcal{F}^c)^{1/2} \\ &= O(\lambda_n^{-k} P(\mathcal{F}^c)^{1/2}), \end{aligned} \tag{A.41}$$

and to get the second equality by Assumption 2(ii) $(s_0\lambda_n)^k / s_0^{k/2} \leq 1$ since the ratio on the left converges to zero, so this means $s_0^{k/2} \leq \lambda_n^{-k}$ with sufficiently large n.

Now use (A.40) with (A.41) in (A.39)

$$E\|\hat{\beta}\|_1^k = O(s_0^{k/2}) + O(\lambda_n^{-k} P(\mathcal{F}^c)^{1/2}). \tag{A.42}$$

If $\lambda_n \geq P(\mathcal{F}^c)^{1/2k}/s_0^{1/2}$ it is clear that

$$s_0^{k/2} \geq \lambda_n^{-k} P(\mathcal{F}^c)^{1/2}, \tag{A.43}$$

So by (A.43) in (A.42) we have the desired result. **Q.E.D.**

Q.E.D.

A.2.4 Main Theorem Proof: Incentive Compatibility

Proof of Theorem 3.

By Theorem 1 and 2 we can choose the larger of λ_n in those theorems, with $s_0 \geq 1$, and since it is nondecreasing with n,

$$\lambda_n \geq \frac{P(\mathcal{F}^c)^{1/4k}}{s_0^{1/2}} \geq \frac{P(\mathcal{F}^c)^{1/2k}}{s_0^{1/2}} \tag{A.44}$$

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

$$\begin{aligned} E[R(X_{n+1})'\hat{\beta} - X'_{n+1}\beta_0]^2 &= E[R(X_{n+1})'\hat{\beta} - X'_{n+1}\hat{\beta} + X'_{n+1}\hat{\beta} - X'_{n+1}\beta_0]^2 \\ &= E[R(X_{n+1})'\hat{\beta} - X'_{n+1}\hat{\beta}]^2 + E[X'_{n+1}\hat{\beta} - X'_{n+1}\beta_0]^2 \\ &+ E[\hat{\beta}'(R(X_{n+1}) - X_{n+1})X'_{n+1}(\hat{\beta} - \beta_0)] \\ &+ E[(\hat{\beta} - \beta_0)'X_{n+1}(R(X_{n+1})' - X'_{n+1})\hat{\beta}]. \end{aligned} \tag{A.45}$$

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

$$E[R(X_{n+1})'\hat{\beta} - X'_{n+1}\beta_0]^2 - E[X'_{n+1}\hat{\beta} - X'_{n+1}\beta_0]^2 = E[\hat{\beta}'D_{n+1}D'_{n+1}\hat{\beta}] \tag{A.46}$$

$$+ E[\hat{\beta}'D_{n+1}X'_{n+1}(\hat{\beta} - \beta_0)] \tag{A.47}$$

$$+ E[(\hat{\beta} - \beta_0)'X_{n+1}D'_{n+1}\hat{\beta}]. \tag{A.48}$$

Now analyze (A.47), the analysis of (A.48) is the same and thus omitted. See that

$$\begin{aligned} \hat{\beta}'D_{n+1}X'_{n+1}(\hat{\beta} - \beta_0) &\leq |\hat{\beta}'D_{n+1}X'_{n+1}(\hat{\beta} - \beta_0)| \\ &\leq |\hat{\beta}'D_{n+1}||X'_{n+1}(\hat{\beta} - \beta_0)| \\ &\leq \|\hat{\beta}\|_1\|D_{n+1}\|_\infty\|X_{n+1}\|_\infty\|\hat{\beta} - \beta_0\|_1, \end{aligned} \tag{A.49}$$

where we use Holder's inequality. Then

$$E[\hat{\beta}' D_{n+1} X_{n+1}' (\hat{\beta} - \beta_0)] \leq \|D_{n+1}\|_\infty \|X_{n+1}\|_\infty \| E \left[\|\hat{\beta}\|_1 \|\hat{\beta} - \beta_0\|_1 \right] \tag{A.50}$$

$$\leq \left[\|D_{n+1}\|_\infty \right] \left[\|X_{n+1}\|_\infty \right] \left[E\|\hat{\beta}\|_1^2 \right]^{1/2} \left[E\|\hat{\beta} - \beta_0\|_1^2 \right]^{1/2} \tag{A.51}$$

$$= [M_4][M_3] \left[E\|\hat{\beta}\|_1^2 \right]^{1/2} \left[E\|\hat{\beta} - \beta_0\|_1^2 \right]^{1/2} \tag{A.52}$$

where we apply (A.49) for the first inequality and Holder's Inequality in the second inequality above, and the last equality comes from M_3, M_4 definitions. Then we apply Theorems 1-2 with $k = 2$. We assume $\lambda_n \geq P(\mathcal{F}^c)^{1/8}/s_0^{1/2}$ and if

$$s_0^{3/2} \sqrt{\frac{lnp}{n}} [M_3][M_4] \to 0, \tag{A.53}$$

we see that (A.52) goes to zero, by Theorems 1-2, and $\lambda_n = O(\sqrt{\frac{lnp}{n}})$.

So looking at incentive compatibility definition and (A.46)-(A.48)

$$E[R(X_{n+1})'\hat{\beta} - X_{n+1}'\beta_0]^2 - E[X_{n+1}'\hat{\beta} - X_{n+1}'\beta_0]^2 = E[\hat{\beta}' D_{n+1} D_{n+1}' \hat{\beta}] + o(1), \tag{A.54}$$

where the first right side term in (A.54) is nonnegative and the other terms are negligible in large samples by (A.53).

The uniformity over $\mathcal{B}_{l_0}(s_0)$ goes through since Theorems 1, 2 depend on β_0 only through s_0, and they are the main ingredient in the proof.

Q.E.D.

B Appendix B

Here we consider results when $p \leq n$, and relaxing Assumption 2(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 $\kappa_n = lnn$. Change Assumption 2(ii) so that

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

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

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

by Assumptions A1-A.2. To see this point

$$\frac{EM_F^2}{nlnn} = \left[\left(\frac{(EM_F^2)^{1/2}\sqrt{lnp}}{\sqrt{n}}\right)\frac{1}{\sqrt{lnn}\sqrt{lnp}}\right]^2 = o(1). \tag{B.2}$$

This shows also that

$$\max_{1\leq j\leq p} |\hat{\mu}_j - \mu_j| = O_p(\sqrt{lnn}/\sqrt{n}). \tag{B.3}$$

Lemma A.1 will be the same. Lemma A.2(i) lower bound probability has $\kappa_n = lnn$ now. Lemma A.2(ii) is the same. Lemma A.2(iii) will change to $\lambda_n = O(\sqrt{lnn}/\sqrt{n})$. Lemma A.3 use $\kappa_n = lnn$, so (A.19) becomes

$$t_1 = K[\frac{\sqrt{lnp^2}}{\sqrt{n}} + \frac{\sqrt{EM_2^2} lnp^2}{n}] + \frac{\sqrt{lnn}}{n}.$$

Lemma A.4 is the same with $\kappa_n = lnn$.

Given these results, the proof of Theorem 1 is the same with $\lambda_n = O(\sqrt{\frac{lnn}{n}})$. Theorem 2 does not change. Theorem 3 condition will be changing to

$$s_0^{3/2}\sqrt{\frac{lnn}{n}}[M_3][M_4] \to 0,$$

B.2 Relaxing Assumption 2(iii)

In this subsection we relax Assumption 2(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 2(iv).

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

Assumption 2(iii) which is suggested by Jankova and van de Geer (2018) and simplifies their paper in semiparametric efficient estimators. Our Assumption 2(iv) here generalizes that assumption and in the case of s_0 being constant becomes Assumption 2(iii). The implication of Assumption 2(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 Section 2 discussion after Assumption 2, this implies $S_0 = F_1$, and F_2 is an empty set. So Assumption 2(iv) can simultaneously allow s_0 increasing with n, and all large nonzero coefficients in S_0. Previously in Assumption 2(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 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 2(iii) there. The first change comes in step 3 of Theorem 1 proof. First (A.29) changes to $\|\beta_0\|_1^k = O(s_0^k)$ under Assumption 2(iv) instead of Assumption 2(iii). Then (A.30) becomes

$$E\left[\frac{\|u\|_n^2}{2\lambda_n}\right]^k + 2\|\beta_0\|_1^k = O(\max(s_0^k, \lambda_n^{-k})). \tag{B.4}$$

Then (A.33) changes to following

$$\begin{aligned} E\|\hat{\beta} - \beta_0\|_1^k &= O(s_0^k \lambda_n^k) + O(\max(s_0^k, \lambda_n^{-k})\sqrt{P(\mathcal{F}^c)} \\ &= O(s_0^k \lambda_n^k) + O(\lambda_n^{-k}\sqrt{P(\mathcal{F}^c)}), \end{aligned} \tag{B.5}$$

where we use Assumption 2 with $s_0^k \lambda_n^k \leq 1$ and sufficiently large n to show the last equality. Instead of (A.34) we have the following conditions, to establish the rate for the oracle inequality (i.e. mean l_1 norm bound to k th order)

$$s_0^k \lambda_n^k \geq \lambda_n^{-k} P(\mathcal{F}^c)^{1/2}. \tag{B.6}$$

Using (B.5)-(B.6)

$$E\|\hat{\beta} - \beta_0\|_1^k = O(s_0^k \lambda_n^k). \tag{B.7}$$

The condition (B.6) can be written as

$$\lambda_n \geq P(\mathcal{F}^c)^{1/4k} / s_0^{1/2}, \tag{B.8}$$

where the tuning parameter choice under Assumption 2(iv) which is (B.8) is the same. The discussion after this in step 4 is the same, given Assumption 2(i)-(ii). So we have the following result:

Corollary B.1. *Under Assumptions 1, 2(i)(ii)(iv), with sufficiently large n*

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

we have

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

The result is also uniform over l_0 ball $\mathcal{B}_{l_0}$

Now we modify the proof of Theorem 2. In that respect, by Assumption 2(iv) the rate after (A.39) becomes

$$\|\hat{\beta}\|_1 = O_p(s_0). \tag{B.9}$$

Then (A.42) changes to

$$E\|\hat{\beta}\|_1^k = O(s_0^k) + O(\lambda_n^{-k} P(\mathcal{F}^c)^{1/2}). \tag{B.10}$$

We can show that

$$s_0^k \geq \lambda_n^{-k} P(\mathcal{F}^c)^{1/2}, \tag{B.11}$$

if we have

$$\lambda_n \geq P(\mathcal{F}^c)^{1/2k} / s_0. \tag{B.12}$$

Then given (B.12), using (B.11) in (B.10) we have

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

So we established the following Corollary to Theorem 2. The result is different from Theorem 2 and the k th moment of l_1 error grows faster here in Corollary B.2 if s_0 increases with n. So relaxed assumption comes with a cost that will affect main incentive compatibility condition.

Corollary B.2. *Under Assumptions 1, 2(i)(ii)(iv), with sufficiently large n*

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

we have

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

The result is also uniform over l_0 ball $\mathcal{B}_{l_0}$

Now we follow the proof of Theorem 3 and substitute Assumption 2(iv) instead of Assumption 2(iii). Note that our λ_n choice must choose the maximum of the ones in Corollary B.1 and B.2. Clearly Corollary B.1 tuning parameter is larger than the one in Corollary B.2. The only place we have to change there is (A.53). Given

$$\lambda_n \geq max\left(\frac{P(\mathcal{F}^c)^{1/4}}{s_0}, \frac{P(\mathcal{F}^c)^{1/8}}{s_0^{1/2}}\right) = \frac{P(\mathcal{F}^c)^{1/8}}{s_0^{1/2}},$$

since $s_0 \geq 1$ we need

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

to have Incentive Compatibility in large samples. So we have the following counterpart to Theorem 3.

Corollary B.3.*Under Assumptions 1, 2(i)(ii)(iv) and with sufficiently large n*

$$\lambda_n \geq \frac{P(\mathcal{F}^c)^{1/8}}{s_0^{1/2}}$$

and

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

lasso is Incentive Compatible. The result is also uniform over l_0 ball $\mathcal{B}_{l_0}$.

Clearly, there is a difference between Theorem 3 and Corollary B.3 here. Incentive compatibility of lasso is more difficult to achieve, due to sparsity, s_0, having exponent of 2 here instead of 3/2 in Theorem 3.

C Appendix C

This section provides the proofs for conservative lasso IC, which is explained in Section 5. 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 C.1. *(i). Under Assumption 1, and on $\mathcal{A}_1 \cap \mathcal{A}_2$*

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

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.19).

(ii). With added Assumption 2 to (ii), and assuming $\|\Theta\|_{l_\infty} = O(s_1)$, $\lambda_{prec} = O(s_1\lambda_n)$.

(iii). The result in (i) holds wpa1 (i.e. with probability approaching one) with Assumptions 1-3, and

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

Remark. Result (iii) holds without the need to be in $\mathcal{A}_1 \cap \mathcal{A}_2$. This is Lemma A.7 of Caner and Kock (2018), where we prove it under Assumptions 1-2 which are weaker moment conditions than the one in Caner and Kock (2018) due to usage of new maximal inequalities in Section A.2. 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 C.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 and (A.19)-(A.20)

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

(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 Assumption 2 added, and since $t_1 = O(\sqrt{lnp/n})$ and so $s_0\sqrt{lnp/n} = o(1)$ by Assumption 2(ii) we get $\lambda_{prec} = O(s_1\lambda_n)$, 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 Lemma A.2-Lemma A.3 given (ii) and Assumption 3.**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, part (ii) of Lemma C.2. Part (i) below is from their paper.

Lemma C.2. *Let $0 < a_n \leq 1$, where a_n is a deterministic-positive sequence in n, then*

(i). on the set $\mathcal{A}_1 \cap \mathcal{A}_2$

$$\|\hat{\beta}_w - \beta_0\|_1 \leq 4(a_n+1)(2a_n+1)\frac{\lambda_n s_0}{\phi^2_\Sigma(s_0)}.$$

(ii). with Assumptions 1-2

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

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.2 above

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

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.

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

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

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

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

Thus since $\max_{1 \le j \le p} \hat{w}_j \le 1$, and define $\hat{w}_{min} := \min_{1 \le j \le p} \hat{w}_j$ we can rewrite (C.2)

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

Use triangle inequality

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

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

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

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

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

Then with with probability approaching one, given Lemma C.1

$$
\begin{aligned}
\hat{w}_{min}^{-1} &\leq \frac{\max_{1\leq j\leq p}|\hat{\beta}_j - \beta_{0,j}| + \max_{1\leq j\leq p}|\beta_{0,j}|}{\lambda_{prec}} \qquad \text{(C.5)} \\
&\leq \frac{\lambda_{prec} + C}{\lambda_{prec}} = 1 + \frac{C}{\lambda_{prec}}. \qquad \text{(C.6)}
\end{aligned}
$$

By Assumption 3 we know $\lambda_{prec} = o(1)$, via Lemma C.1

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

Regardless of whether $|\hat{\beta}_j|$ is larger than or equal to or less than λ_{prec} we have

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

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

$$E\left(\frac{1}{\hat{w}_{min}^k}\right) = O(s_1^{-k}\lambda_n^{-k}).$$

Also

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

as well. With these two rates and by (A.26)(A.29) in (C.4)

$$\begin{aligned}
&\frac{1}{(2\lambda_n)^k}[E\|u\|_n^{4k}]^{1/2}[E(\hat{w}_{min})^{-2k}]^{1/2} + [2\|\beta_0\|_1]^k E\left[\frac{1}{\hat{w}_{min}}\right]^k \\
&= O(\lambda_n^{-k})O(1)O(\lambda_{prec}^{-k}) + O(s_0^{k/2})O(\lambda_{prec}^{-k}) \\
&= O(\lambda_n^{-k})O(s_1^{-k}\lambda_n^{-k}) + O(s_0^{k/2})O(s_1^{-k}\lambda_n^{-k}) \qquad (C.7)\\
&= O(s_1^{-k}\lambda_n^{-2k}), \qquad (C.8)
\end{aligned}$$

where we use $s_0^{k/2}\lambda_n^k = (s_0\lambda_n)^k/s_0^{k/2} \leq 1$ by Assumption 2 with sufficiently large n, so to get last equality, the first rate dominated in (C.7).

So using (C.8) in (C.4)

$$E\|\hat{\beta}_w - \beta_0\|_1^k = O(s_1^{-k}\lambda_n^{-2k}).$$

This is a key finding and entirely new, shows that general weight function in conservative lasso made the error larger compared with lasso ,where weights are all one in lasso, since we have an extra $s_1^{-k}\lambda_n^{-k}$ term extra compared with lasso in (A.32). Extending lasso to general weights as conservative lasso made the moment estimation worse due to weights being very small. Conservative lasso came with better selection properties then lasso but here it lacks in estimating moments.

Step 4. Now merge the rates in (C.1)(C.8)

$$
\begin{aligned}
E\|\hat{\beta}_w - \beta_0\|_1^k &= E\|\hat{\beta}_w - \beta_0\|_1^k 1_{\{\mathcal{F}\}} + E\|\hat{\beta}_w - \beta_0\|_1^k 1_{\{\mathcal{F}^c\}} \\
&\leq O(s_0^k \lambda_n^k) + \sqrt{E\|\hat{\beta}_w - \beta_0\|_1^{2k}} \sqrt{P(\mathcal{F}^c)} \\
&= O(s_0^k \lambda_n^k) + O(s_1^{-k} \lambda_n^{-2k}) P(\mathcal{F}^c)^{1/2}. \qquad \text{(C.9)}
\end{aligned}
$$

To establish a rate

$$
s_0^k \lambda_n^k \geq \lambda_n^{-2k} s_1^{-k} P(\mathcal{F}^c)^{1/2}, \qquad \text{(C.10)}
$$

which (C.10) is implied by

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

This shows

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

Q.E.D

Proof of Theorem 5. By (A.43)

$$
E\|\hat{\beta}_w\|_1^k \leq E\|\hat{\beta}_w\|_1^k 1_{\{\mathcal{F}\}} + \sqrt{E\|\hat{\beta}_w\|_1^{2k}} \sqrt{P(\mathcal{F}^c)}. \qquad \text{(C.11)}
$$

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

$$
\begin{aligned}
\|\hat{\beta}_w\|_1 &\leq \|\hat{\beta}_w - \beta_0\|_1 + \|\beta_0\|_1 \\
&= O_p(\lambda_n s_0) + O(\sqrt{s_0}) = O_p(\sqrt{s_0}), \qquad \text{(C.12)}
\end{aligned}
$$

and the last equality is by Assumption 2, since $s_0 \lambda_n \to 0$. So

$$
E\|\hat{\beta}_w\|_1^k 1_{\{\mathcal{F}\}} = O(s_0^{k/2}). \qquad \text{(C.13)}
$$

Then to handle the second term on the right side in (C.11), use (C.3)

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

Next repeat exactly (C.3)-(C.8) to have

$$\sqrt{E\|\hat{\beta}_w\|_1^{2k}}\sqrt{P(\mathcal{F}^c)} = O(\lambda_n^{-2k} s_1^{-k})\sqrt{P(\mathcal{F}^c)}, \tag{C.14}$$

since $(s_0\lambda_n)^k/s_0^{k/2} \leq 1$ for sufficiently large n. Use (C.13)(C.14) in (C.11) to have

$$E\|\hat{\beta}_w\|_1^k = O(s_0^{k/2}) + O(\lambda_n^{-2k} s_1^{-k} P(\mathcal{F}^c)^{1/2}). \tag{C.15}$$

Note that compared to lasso, our second rate is different by $\lambda_n^{-k} s_1^k$, this is due to usage of weights, namely minimum weight estimate being at rate of λ_n, and we use inverse of that estimate in the bounds. To get a rate for k th moment of conservative lasso (in l_1 norm) with

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

first rate in (C.15) above dominates the second one, which gets us

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

Q.E.D.

Proof of Theorem 6. Given Theorems 4-5 proof here follows exactly from the proof of Theorem 3. But the lower bound for λ_n is: $(k = 2)$

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

.**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*.

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.