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

Mehmet Caner† Kfir Eliaz‡

January 25, 2024

Abstract

We consider situations where a user feeds her attributes to a machine learning method that tries to predict her best option based on a random sample of other users. The predictor is incentive-compatible if the user has no incentive to misreport her covariates. Focusing on the popular Lasso estimation technique, we borrow tools from high-dimensional statistics to characterize sufficient conditions that ensure that Lasso is incentive compatible with sufficiently large sample size. We also provide simplification of some conditions for incentive compatibility in the asymptotic case. We extend our results to the Conservative Lasso estimator and provide new moment bounds for this generalized weighted version of Lasso. Our results show that incentive compatibility is achieved if the tuning parameter is kept above some threshold in the case of asymptotics. We present simulations that illustrate how this can be done in practice.

Keywords: Moment oracle inequality, machine learning, overfit

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

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

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

APPENDIX

In the next part, Appendix A considers the proofs when $p > n$, when the model is approximately linear and Appendix B considers the case $p \leq n$, and relaxing Assumption 1(iii). Appendix C covers proofs for GLM Structured Sparsity Estimators. Appendix D covers simulations and tuning parameter choice.

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 Finite Sample IC: Approximately Linear Model

We use an assumption that will provide us maximal inequalities.

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

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

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

for a positive constant $K_1 > 0$. With Assumption A.1 here, Caner and Kock (2019) or

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

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

for a positive constant $K_2 > 0$.

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

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

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

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

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

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

A.2.1 Events

We repeat the events in the main text. Before the assumptions, we need to define events that will be helpful. The first event is:

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

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

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

We define the population adaptive restricted eigenvalue of Σ

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

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

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

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

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

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

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

A.2.2 Proofs of Lemmata

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

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

$$\|\hat{\beta} - \beta_0\|_1 \leq 4\sqrt{2}\sqrt{s_0}\left[\frac{36\lambda_n^2 s_0}{\phi^2_\Sigma(s_0)} + 8c_s^2\right]^{1/2} + \frac{c_s^2}{\lambda_n}$$

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

Remarks. 1. With $c_s = 0$, we obtain the non-asymptotic lasso bound up to constants in Caner and Kock (2018). A result with fixed regressors and normal errors are in Lemma 3.7 of Belloni and Chernozhukov (2009). We have a different and new proof technique here due to random regressors and non-normal errors.

2. We show that in asymptotic case, since we prove $P(\mathcal{F}) \to 1$, the upper bound simplifies and

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

The details are in section A.3.

Proof of Lemma A.1. Let $Y := (Y_1, \cdots, Y_i, \cdots, Y_n)' : n \times 1$ and $u := (u_1, \cdots, u_i, \cdots, u_n)' : n \times 1$, $r := (r_1, \cdots, r_i, \cdots, r_n)' : n \times 1$ vectors. 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 $Y = X\beta_0 + u + r$ on the first left side term as well as the first right side term to simplify the inequality above combining with Holder's Inequality and Cauchy-Schwartz Inequality for the second right side term, and since $c_s^2 := \frac{r'r}{n} = \frac{1}{n}\sum_{i=1}^{n} r_i r_i'$,

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

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

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

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

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

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

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

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

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

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

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

There are two possibilities in (A.9). First

$$\|X(\hat{\beta}-\beta_0)\|_n^2 - 2c_s\|X(\hat{\beta}-\beta_0)\|_n < 0,$$

then clearly we have the upper bound

$$\|X(\hat{\beta}-\beta_0)\|_n \leq 2c_s. \tag{A.10}$$

The more difficult case in (A.9) is:

$$\|X(\hat{\beta}-\beta_0)\|_n^2 - 2c_s\|X(\hat{\beta}-\beta_0)\|_n \geq 0, \tag{A.11}$$

and in that scenario

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

Ignoring the first two terms on the left side of (A.9) (since it is positive or zero) and by (A.11) shows that we satisfy the restricted set condition in empirical adaptive restricted eigenvalue condition, so we have

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

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

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

Then use $3u_1v_1 \leq u_1^2/4 + 9v_1^2$ with $v_1 = \lambda_n\sqrt{s_0}/\hat{\phi}_{\hat{\Sigma}}(s_0)$, $u_1 = \|X(\hat{\beta} - \beta_0)\|_n$ for the first term on the right side of the above inequality and then for the second term on the right side set $u_2 := \|X(\hat{\beta} - \beta_0)\|_n$ and $v_2 := c_s$ and $2u_2v_2 \leq u_2^2/4 + 4v_2^2$ to get

$$\|X(\hat{\beta} - \beta_0)\|_n^2 + \lambda_n \sum_{j \in S_0^c} |\hat{\beta}_j| \leq \frac{\|X(\hat{\beta} - \beta_0)\|_n^2}{4} + \frac{9\lambda_n^2 s_0}{\hat{\phi}_{\hat{\Sigma}}^2(s_0)} + \frac{\|X(\hat{\beta} - \beta_0)\|_n^2}{4} + 4c_s^2.$$

Simplify above, by multiplying all terms by 2

$$\|X(\hat{\beta} - \beta_0)\|_n^2 + 2\lambda_n \sum_{j \in S_0^c} |\hat{\beta}_j| \leq \frac{18\lambda_n^2 s_0}{\hat{\phi}_{\hat{\Sigma}}^2(s_0)} + 8c_s^2.$$

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

$$\|X(\hat{\beta} - \beta_0)\|_n^2 + 2\lambda_n \sum_{j \in S_0^c} |\hat{\beta}_j| \leq \frac{36\lambda_n^2 s_0}{\phi_{\Sigma}^2(s_0)} + 8c_s^2.$$

This implies the oracle inequality by squaring (A.10) bound and comparing with the bound above

$$\|X(\hat{\beta} - \beta_0)\|_n^2 \leq \frac{36\lambda_n^2 s_0}{\phi_{\Sigma}^2(s_0)} + 8c_s^2. \tag{A.12}$$

Now we go back to (A.9) for l_1 bound and provide an upper bound for the term so

$$2c_s\|X(\hat{\beta} - \beta_0)\|_n \leq \|X(\hat{\beta} - \beta_0)\|_n^2 + c_s^2,$$

where we use $2u_3v_3 \leq u_3^2 + v_3^2$ with $u_3 := c_s, v_3 := \|X(\hat{\beta} - \beta_0)\|_n$. So (A.9) is

$$\|X(\hat{\beta} - \beta_0)\|_n^2 + \lambda_n \sum_{j \in S_0^c} |\hat{\beta}_j| \leq 3\lambda_n \sqrt{s_0} \|\hat{\beta}_{S_0} - \beta_{S_0,0}\|_2 + c_s^2 + \|X(\hat{\beta} - \beta_0)\|_n^2. \tag{A.13}$$

To get to the l_1 see the first term in (A.13) on the left and last term on the right cancels and add both sides $\lambda_n \|\hat{\beta}_{S_0} - \beta_{0,S_0}\|_1$ to have

$$\lambda_n \sum_{j \in S_0^c} |\hat{\beta}_j| + \lambda_n \sum_{j \in S_0} |\hat{\beta}_j - \beta_{0,j}| = \lambda_n \|\hat{\beta} - \beta_0\|_1 \leq \lambda_n \|\hat{\beta}_{S_0} - \beta_{0,S_0}\|_1 + 3\lambda_n \sqrt{s_0} \|\hat{\beta}_{S_0} - \beta_{0,S_0}\|_2 + c_s^2,$$

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

$$\lambda_n \|\hat{\beta} - \beta_0\|_1 \leq 4\lambda_n \sqrt{s_0} \|\hat{\beta}_{S_0} - \beta_{0,S_0}\|_2 + c_s^2.$$

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

$$\|\hat{\beta} - \beta_0\|_1 \leq 4\sqrt{s_0} \frac{\|X(\hat{\beta} - \beta_0)\|_n}{\hat{\phi}_{\hat{\Sigma}}(s_0)} + \frac{c_s^2}{\lambda_n}.$$

Use (A.12) and the event $\mathcal{A}_2$ to have

$$\begin{aligned} \|\hat{\beta} - \beta_0\|_1 &\leq 4\sqrt{s_0} \left[\frac{36\lambda_n^2 s_0}{\phi_\Sigma^2(s_0)} + 8c_s^2 \right]^{1/2} \left[\frac{1}{\hat{\phi}_{\hat{\Sigma}}(s_0)} \right] + \frac{c_s^2}{\lambda_n} \\ &\leq 4\sqrt{2}\sqrt{s_0} \left[\frac{36\lambda_n^2 s_0}{\phi_\Sigma^2(s_0)} + 8c_s^2 \right]^{1/2} + \frac{c_s^2}{\lambda_n}. \end{aligned} \tag{A.14}$$

Note that uniformity over $\mathcal{B}_{l_0}(s_0)$ follows since the upper bound in (A.14) depends on β_0 only through s_0. **Q.E.D**

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

$$P(\mathcal{A}_1) \geq 1 - exp(-C_1\kappa_n) - \frac{K_1' E M_1^2}{(n\kappa_n)} = 1 - \frac{1}{p^{C_1}} - \frac{K_1' E M_1^2}{nlnp}$$

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

$$P(\mathcal{A}_1) \geq 1 - exp(-C_1\kappa_n) - K_1' \frac{E M_1^2}{(n\kappa_n)} = 1 - \frac{1}{p^{C_1}} - \frac{K_1' E M_1^2}{nlnp}, \tag{A.15}$$

with a universal positive constant $K_2 > 0$

$$\lambda_n = K_2[\sqrt{\frac{lnp}{n}} + \frac{\sqrt{EM_1^2 lnp}}{n} + \sqrt{\frac{lnp}{n}}]. \tag{A.16}$$

Q.E.D.

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

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

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

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

$$N \geq (32s_0)^2 \left[\frac{K_3\sqrt{lnp^2} + K_3\sqrt{EM_2^2 lnp^2} + K_3\sqrt{lnp}}{\phi_\Sigma^2(s_0)} \right]^2. \tag{A.17}$$

We also show how we derive N expression in the proof of Lemma A.3. If the lower bound in (A.17) is not an integer, N is the smallest integer that is larger than the lower bound.

Proof of Lemma A.3. Start with

$$\begin{aligned} \left| \delta' \frac{X'X}{n} \delta \right| &= \left| \delta' (\frac{X'X}{n} - \Sigma + \Sigma)\delta \right| \\ &\geq |\delta'\Sigma\delta| - |\delta'(\hat{\Sigma} - \Sigma)\delta|. \end{aligned} \tag{A.18}$$

The second term on the right side of (A.18) can be bounded by repeated application of Holders inequality

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

So (A.18) becomes

$$|\delta'\hat{\Sigma}\delta| \geq |\delta'\Sigma\delta| - \|\delta\|_1^2 \|\hat{\Sigma} - \Sigma\|_\infty. \tag{A.19}$$

Now we digress a bit to simplify (A.19). Note that we have the restriction set definition

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

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

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

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

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

Now divide (A.19) by $\|\delta_{S_0}\|_2^2 > 0$ to have

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

Minimize over δ on the both sides

$$\hat{\phi}_{\hat{\Sigma}}^2(s_0) \geq \phi_\Sigma^2(s_0) - 16s_0\|\hat{\Sigma} - \Sigma\|_\infty. \tag{A.20}$$

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

$$t_1 = K_2[\sqrt{\frac{lnp^2}{n}} + \frac{\sqrt{EM_2^2}lnp^2}{n} + \sqrt{\frac{lnp}{n}}]. \tag{A.21}$$

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

$$\begin{aligned}P[16s_0\|\hat{\Sigma} - \Sigma\|_\infty > \epsilon_{n1}] &= P[\|\hat{\Sigma} - \Sigma\|_\infty > t_1] \\ &\leq exp(-C_1lnp) + \frac{K_1'EM_2^2}{(nlnp)}.\end{aligned} \tag{A.22}$$

See that by multiplying the second term on right side of (A.21) by $n^{1/2}$ and hence define another positive sequence

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

If $n \geq N$ using (A.17) we have

$$\epsilon_{n1} \leq \epsilon_{n2} \leq \phi^2_{\Sigma}(s_0)/2, \tag{A.23}$$

since both $\epsilon_{n1}, \epsilon_{n2}$ are positive and $t_1 > 0$. Then with $n \geq N$ and by (A.20)(A.22)

$$\begin{aligned} P[\hat{\phi}^2_{\hat{\Sigma}}(s_0) \geq \phi^2_{\Sigma}(s_0)/2] &\geq 1 - exp(-C_1\kappa_n) - \frac{K_1' E M_2^2}{(n\kappa_n)} \\ &= 1 - \frac{1}{p^{C_1}} - \frac{K_1' E M_2^2}{nlnp} \end{aligned} \tag{A.24}$$

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(i)-2, with $\kappa_n = lnp$, and $n \geq N$*

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

Proof of Lemma A.4.

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

$$\begin{aligned} P(\mathcal{F}^c) &= P(\mathcal{A}_1 \cap \mathcal{A}_2)^c = P(\mathcal{A}_1^c \cup \mathcal{A}_2^c) \leq P(\mathcal{A}_1^c) + P(\mathcal{A}_2^c) \\ &\leq 2exp(-C_1\kappa_n) + \frac{K_1'[EM_1^2 + EM_2^2]}{(n\kappa_n)} \\ &= \frac{2}{p^{C_1}} + \frac{K_1'[EM_1^2 + EM_2^2]}{nlnp}. \end{aligned} \tag{A.25}$$

Q.E.D.

A.2.3 Analysis of Assumption 1(ii)

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, note

that

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

since in the case of s_0 growing with n

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

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

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

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

However,

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

where $\phi_{min}(\Sigma) \geq c > 0$. Hence, if Assumption 1(ii) 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 $C_0 := \frac{\|\beta_0\|_2^2 \phi_{min}(\Sigma)}{\sigma_u^2}$.

In Appendix B, we also consider $\|\beta_0\|_2 = O(\sqrt{s_0})$, where all nonzero coefficients can be large (i.e., none of them are local to zero, as in set D_2 above). In other words, there is no index set D_2 as above, but all nonzero coefficients (their indices) are in the set D_1 above.

A.2.4 New Oracle Inequality Proofs

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

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

Our first result in this section is a k-th moment bound for the l_1 norm of the Lasso bias, with $k \geq 1$.

Theorem A.1 *Suppose Assumptions 1(i)(ii)-2 hold. Let C_4, C_5, C_7, C_8 are some positive constants defined in the proof explicitly. If λ_n is chosen such that*

$$\lambda_n \geq max\left(\frac{C_4 c_s}{s_0^{1/2}}, C_7 \frac{P(\mathcal{F}^c)^{1/4k}}{s_0^{1/2}}, C_8 \frac{P(\mathcal{F}^c)^{1/2k}}{s_0^{1/2}}\right), \tag{A.26}$$

then

$$[E\|\hat{\beta} - \beta_0\|_1^k]^{1/k} \leq (2C_5)^{1/k} s_0 \lambda_n. \tag{A.27}$$

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

Issue of No-Sparsity: Consider the case of a fully dense model with $s_0 = p$, where all regressors are relevant (i.e. all coefficients of β_0 vector are non zero). In this case, the lower

bound (A.26) is easier to achieve compared with $s_0 < p$, but the moment upper bound in Theorem 1 will be larger.

Proof of Theorem A.1. We proceed in four steps.

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

$$E\|\hat{\beta} - \beta_0\|_1^k = E\|\hat{\beta} - \beta_0\|_1^k 1_{\{\mathcal{F}\}} + E\|\hat{\beta} - \beta_0\|_1^k 1_{\{\mathcal{F}^c\}}. \tag{A.28}$$

We want to form rates for the right side terms in (A.28).

Step 1. Note that by Lemma A.1, the first term on the right side of (A.28) is:

$$E\|\hat{\beta} - \beta_0\|_1^k 1_{\{\mathcal{F}\}} \leq 2^{5k/2} s_0^{k/2} [\frac{36}{\phi_\Sigma^2(s_0)} \lambda_n^2 s_0 + 8c_s^2]^{k/2} + \frac{c_s^{2k}}{\lambda_n^k}. \tag{A.29}$$

We simplify the bound for tuning parameter in (A.29). Define a positive constant $C_2 := \frac{\phi_\Sigma(s_0)\sqrt{8}}{6}$, then if we have

$$\lambda_n \geq \frac{C_2 c_s}{s_0^{1/2}}, \tag{A.30}$$

we get

$$\frac{36}{\phi_\Sigma^2(s_0)} \lambda_n^2 s_0 \geq 8c_s^2. \tag{A.31}$$

Given the first term dominates the second term in (A.29) as shown in (A.31), the first term dominates the third term

$$\frac{2^{5k/2} 6^k}{\phi_\Sigma^k(s_0)} \lambda_n^k s_0^k \geq \frac{c_s^{2k}}{\lambda_n^k},$$

if

$$\lambda_n \geq \frac{C_3 c_s}{s_0^{1/2}}, \tag{A.32}$$

where C_3 is a positive constant defined as $C_3 := \frac{\phi_\Sigma^{1/2}(s_0)}{6^{1/2} 2^{5/4}}$.

Combining (A.30)(A.32) with $C_4 := max(C_2, C_3)$

$$\lambda_n \geq C_4 \frac{c_s}{s_0^{1/2}}. \tag{A.33}$$

Under (A.33) (A.29) is upper bounded, with $C_5 := 3(2^{5k/2})6^k/\phi^k_\Sigma(s_0)$ which is a positive constant

$$E\|\hat{\beta} - \beta_0\|_1^k 1_{\{\mathcal{F}\}} \leq C_5 s_0^k \lambda_n^k. \tag{A.34}$$

Now we want to evaluate the second term on the right side of (A.28). But before that we need the following intermediate step.

Step 2. We use Nemirowski's moment inequality, Lemma 14.24 with Lemma 14.14 in Buhlmann and van de Geer (2011), with for all $k \geq 1$, for the first inequality below. Then for the second inequality we use Loeve's c_r inequality, and for the equality we use u_i being iid, also the definition of $\sigma^2 := Eu_i^2$ with $Eu_i^{4k} \leq C' < \infty$, $C' > 0$ a positive constant, $k \geq 1$

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

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

$$|x+y|^{2k} \leq 2^{2k-1}(|x|^{2k} + |y|^{2k}), \tag{A.35}$$

for $k \geq 1$, and x, y being generic scalars, and σ^2 being bounded above by Assumption 2 and using (A.35) by sufficiently large n

$$\begin{aligned} E\left|\frac{1}{n}\sum_{i=1}^n u_i^2\right|^{2k} &= E\left|\frac{1}{n}\sum_{i=1}^n (u_i^2 - \sigma^2) + \sigma^2\right|^{2k} \\ &\leq 2^{2k-1}\left[E\left|\frac{1}{n}\sum_{i=1}^n (u_i^2 - \sigma^2)\right|^{2k} + (\sigma^2)^{2k}\right] \\ &\leq 2^{2k}\sigma^{4k}. \end{aligned} \tag{A.36}$$

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.28). 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 for the first right side term above use the model $Y - X\beta_0 = u + r$, use the triangle inequality for prediction norm

$$\|Y - X\beta_0\| = \|u + r\|_n^2 \leq \|u\|_n^2 + \|r\|_n^2.$$

and noting that $c_s^2 := \|r\|_n^2$

$$\|\hat{\beta}\|_1 \leq \frac{\|u\|_n^2}{2\lambda_n} + \frac{c_s^2}{2\lambda_n} + \|\beta_0\|_1. \tag{A.37}$$

Then use triangle inequality and then the inequality above

$$\|\hat{\beta} - \beta_0\|_1 \leq \|\hat{\beta}\|_1 + \|\beta_0\|_1 \leq \frac{\|u\|_n^2}{2\lambda_n} + \frac{c_s^2}{2\lambda_n} + 2\|\beta_0\|_1. \tag{A.38}$$

We have the following inequality that we use, from (9.63) of Davidson (1994)

$$|x + y + z|^{2k} \leq 3^{2k-1}[|x|^{2k} + |y|^{2k} + |z|^{2k}]. \tag{A.39}$$

Next taking the $2k$ th moment of the sampling error in l_1 norm, by (A.38)-(A.39) and by taking expectations there for the second inequality below

$$E\|\hat{\beta} - \beta_0\|_1^{2k} \leq 3^{2k-1}\{E\left[\frac{\|u\|_n^2}{2\lambda_n}\right]^{2k} + 2\|\beta_0\|_1^{2k} + \left[\frac{c_s^2}{2\lambda_n}\right]^{2k}\} \tag{A.40}$$

We use Assumption 1, $\|\beta_0\|_2 = O(1)$ to have, specifically with C_6 defined here as an upper bound constant, $\|\beta_0\|_2^{2k} \leq C_6$, which C_6 is a positive constant

$$\|\beta_0\|_1^{2k} \leq (\sqrt{s_0}\|\beta_0\|_2)^{2k} \leq C_6 s_0^k. \tag{A.41}$$

Then use the last equation with (A.36) in (A.40) to have

$$E\left[\frac{\|u\|_n^2}{2\lambda_n}\right]^{2k} + 2\|\beta_0\|_1^{2k} + \frac{c_s^{4k}}{(2\lambda_n)^{2k}} \leq 2^{2k}\sigma^{4k}\lambda_n^{-2k} + 2C_6 s_0^k + \frac{c_s^{4k}}{(2\lambda_n)^{2k}}. \tag{A.42}$$

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

We get the rough bound for expectation using (A.42) in (A.40)

$$E\|\hat{\beta} - \beta_0\|_1^{2k} \leq 3^{2k} max \left(2^{2k}\sigma^{4k}\lambda_n^{-2k}, 2C_6 s_0^k, \left(\frac{c_s^2}{2\lambda_n}\right)^{2k} \right). \tag{A.43}$$

Note that rates in (A.29)(A.43) are different and the last rate in this step is a rough bound which will be helpful in the next step.

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

$$\begin{aligned} E\|\hat{\beta} - \beta_0\|_1^k &= E\|\hat{\beta} - \beta_0\|_1^k 1_{\{\mathcal{F}\}} + E\|\hat{\beta} - \beta_0\|_1^k 1_{\{\mathcal{F}^c\}} \\ &\leq C_5 s_0^k \lambda_n^k + \sqrt{E\|\hat{\beta} - \beta_0\|_1^{2k}} \sqrt{E 1_{\{\mathcal{F}^c\}}} \\ &\leq C_5 s_0^k \lambda_n^k + 3^k max(\sqrt{2C_6} s_0^{k/2}, 2^k \sigma^{2k} \lambda_n^{-k}, \frac{c_s^{2k}}{(2\lambda_n)^k}) \sqrt{P(\mathcal{F}^c)} \end{aligned} \tag{A.44}$$

where we use (A.34) and Cauchy-Schwartz inequality for the first inequality, and the second inequality is by (A.43).

We can get the rate with the following condition

$$C_5 s_0^k \lambda_n^k \geq 3^k 2^k \sigma^{2k} \lambda_n^{-k} P(\mathcal{F}^c)^{1/2}. \tag{A.45}$$

We can simplify further (A.45), with the positive constant $C_7 := (3^{1/2} 2^{1/2} \sigma)/C_5^{1/2k}$,

$$\lambda_n \geq C_7 [P(\mathcal{F}^c)^{1/4k} / s_0^{1/2}]. \tag{A.46}$$

If $C_5 s_0^k \lambda_n^k \geq 3^k \sqrt{2C_6} s_0^{k/2} P(\mathcal{F}^c)^{1/2}$ which is possible with definition of a positive constant $C_8 := 3[\sqrt{2C_6}/C_5]^{1/k}$, and by $\lambda_n \geq C_8 \frac{P(\mathcal{F}^c)^{1/2k}}{s_0^{1/2}}$, and we also have $C_5 s_0^k \lambda_n^k \geq 3^k \frac{c_s^{2k}}{(2\lambda_n)^k} P(\mathcal{F}^c)^{1/2}$. [1] So

[1] This is possible by (A.30) since $C_2 \geq \frac{3^{1/2}}{2^{1/2}} \frac{1}{C_5^{1/k}} = \frac{3^{1/2}\phi_\Sigma(s_0)}{3^{1/k}48}$ by C_5 definition before (A.34) in Step 1 of the proof.

$$E\|\hat{\beta} - \beta_0\|_1^k \leq 2C_5 s_0^k \lambda_n^k.$$

The uniformity over $\mathcal{B}_{l_0}(s_0)$ follows since the rates in (A.29)(A.43)-(A.45) depends on β_0 only by s_0. **Q.E.D.**

Remark.

1. 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.48) below

$$\lambda_n \geq \frac{P(\mathcal{F}^c)^{1/4k}}{s_0^{1/4}}, \tag{A.47}$$

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

$$P(\mathcal{F}^c) \leq \frac{2}{(2p)^{\tau^2/2}} \tag{A.48}$$

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

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

Theorem A.2 *Suppose Assumptions 1(i)(ii)-2 hold, with $k \geq 1$. Let C_6, C_9, C_{10} be positive constants that are defined in the proof. If*

$$\frac{C_9}{s_0^{1/2}} \geq \lambda_n \geq max\left(\frac{C_4 c_s}{s_0^{1/2}}, C_7\frac{P(\mathcal{F}^c)^{1/4k}}{s_0^{1/2}}, C_8\frac{P(\mathcal{F}^c)^{1/2k}}{s_0^{1/2}}\right), \tag{A.49}$$

then

$$[E\|\hat{\beta}\|_1^k]^{1/k} \leq (2C_{10})^{1/k} s_0^{1/2}.$$

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

Remarks.

Non-sparse argument: Consider the case of a fully dense model where $s_0 = p$. In this case, the upper bound on λ_n in Theorem A.2 is more difficult to achieve (compared to $s_0 < p$) but the lower bound for λ_n is smaller compared with the sparse case. Note that in the moment upper bound of Theorem A.2, the bound gets larger in the fully dense case compared with the sparse case.

Proof of Theorem A.2.

We start with

$$E\|\hat{\beta}\|_1^k = E\|\hat{\beta}\|_1^k 1_{\{\mathcal{F}\}} + E\|\hat{\beta}\|_1^k 1_{\{\mathcal{F}^c\}} \leq E\|\hat{\beta}\|_1^k 1_{\{\mathcal{F}\}} + \sqrt{E\|\hat{\beta}\|_1^{2k}}\sqrt{P(\mathcal{F}^c)}, \tag{A.50}$$

by using Cauchy-Schwartz inequality. Then use triangle inequality and norm inequality to have for the first right side term on (A.50)

$$\|\hat{\beta}\|_1 \leq \|\hat{\beta} - \beta_0\|_1 + \|\beta_0\|_1 \tag{A.51}$$

$$\leq \|\hat{\beta} - \beta_0\|_1 + \sqrt{s_0}\|\beta_0\|_2 \tag{A.52}$$

by Assumption 2. This last rate in (A.52) shows that with (A.34)(A.41)

$$E\|\hat{\beta}\|_1^k 1_{\{\mathcal{F}\}} \leq C_5 s_0^k \lambda_n^k + C_6^{1/2} s_0^{k/2} \leq 2C_6^{1/2} s_0^{k/2}, \tag{A.53}$$

with $\lambda_n \leq C_9 \frac{1}{s_0^{1/2}}$, with C_9 defined as a positive constant

$$C_9 := [C_6^{1/2}/C_5]^{1/k}. \tag{A.54}$$

To handle the second right side term in (A.50) we start with the second inequality in (A.38) and ignore $\|\beta_0\|_1$ in the middle to have

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

then follow (A.44) to obtain

$$\sqrt{E\|\hat{\beta}\|_1^{2k}}P(\mathcal{F}^c)^{1/2} \leq 3^k \max(C_6^{1/2}s_0^{k/2}, 2^k\sigma^{2k}\lambda_n^{-k}, \frac{c_s^{2k}}{(2\lambda_n)^k})P(\mathcal{F}^c)^{1/2} \tag{A.55}$$

Now use (A.53) with (A.55) in (A.50)

$$E\|\hat{\beta}\|_1^k \leq 2C_6^{1/2}s_0^{k/2} + 3^k \max(C_6^{1/2}s_0^{k/2}, 2^k\sigma^{2k}\lambda_n^{-k}, \frac{c_s^{2k}}{(2\lambda_n)^k})P(\mathcal{F}^c)^{1/2}. \tag{A.56}$$

Clearly the second right side term in the first inequality in (A.53) is larger than or equal to the first term in the same inequality, but this first term in the right side of (A.53) is larger than equal to second term (which is the maximum of 3 terms) on (A.56) by the analysis in (A.45) and below that inequality in the proof of Theorem A.1. So defining $C_{10} := 3^k C_6^{1/2}$

$$E\|\hat{\beta}\|_1^k \leq 2C_{10}s_0^{k/2}. \tag{A.57}$$

The result is uniform over l_0 ball $\mathcal{B}_{l_0}(s_0)$. **Q.E.D.**

A.2.5 Finite Sample Incentive Compatibility Proof

We provide the definition for finite sample IC.

Definition A.1 *The lasso estimator is* ***uniformly incentive-compatible*** *if for every* X_{n+1}*, for every* $R(X_{n+1})$ *and for every* β_0 *that satisfy Assumptions 1-3, and for* $p \geq 2$ *and* $n \geq n_0 > 0$*,*

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

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

Incentive compatibility means that the user is unable to perform better in the mean squared sense by misreporting her personal characteristics. This definition is for sufficiently large n ($n \geq n_0$), and uniform over $\mathcal{B}_{l_0}(s_0)$ ball. We allow p to change with n, but to save notation we do not subscript p.

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

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

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

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

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

$$E[\hat{\beta}' D_{n+1} D_{n+1}' \hat{\beta}] \geq c_n E[\|\hat{\beta}\|_2^2] > 0. \tag{A.58}$$

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

$\hat{\beta}' D_{n+1} = 0$ is shown in the proof of Theorem A.3. This condition very roughly prevents small lies.

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

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

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

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

Theorem A.3 *Suppose all of the following conditions hold: Assumptions 1-3, Condition 1. Then the Lasso estimator is incentive compatible if the tuning parameter satisfies the following:*

$$min\left(\frac{C_9}{s_0^{1/2}}, \frac{c_1^2 c_n}{(4c_2C_5)s_0c_n + (8C_5^{1/2}C_{10}^{1/2})M_3M_4s_0^{3/2}}\right) \geq \lambda_n \geq max\left(\frac{C_4c_s}{s_0^{1/2}}, \frac{C_7P\left(\mathcal{F}^c\right)^{1/8}}{s_0^{1/2}}, \frac{C_8P\left(\mathcal{F}^c\right)^{1/4}}{s_0^{1/2}}\right) \tag{A.59}$$

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

Remarks.

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

2. A natural question that arises is whether the fully dense case of $s_0 = p$ is compatible with the concept of uniform incentive compatibility. In this remark, we exclude the asymptotic case of $n \to \infty, p \to \infty$, which will be discussed in Remark 5-Theorem 1. Clearly, the

condition for the lower bound on λ_n is easier to achieve when $s_0 = p$ compared with sparse case. However, the opposite is true for the upper bound since this bound gets smaller with $s_0 = p$. Nevertheless, it is possible to achieve this bound with large c_1^2 and C_9. This is an important observation as it implies that a fully dense model accommodates uniform incentive compatibility.

3. Condition 1, clearly restricts the choices of λ_n. To see this, c_n appears in both the numerator and the denominator multiplied by positive terms. So a drop in c_n will make the denominator smaller (hence achieving IC easier) but also it will make numerator smaller too (hence achieving IC more difficult). There is a tradeoff, but since s_0 is nondecreasing in n, its effect on denominator may be large in magnitude, so it may be easier to achieve IC. Also this condition plays no role in asymptotics as we see in Theorem 1 in main text.

Proof of Theorem A.3.

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

$$\frac{C_9}{s_0^{1/2}} \geq \lambda_n \geq max(\frac{C_4 c_s}{s_0^{1/2}}, \frac{C_7 P(\mathcal{F}^c)^{1/4k}}{s_0^{1/2}}, \frac{C_8 P(\mathcal{F}^c)^{1/2k}}{s_0^{1/2}}). \tag{A.60}$$

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

$$\begin{aligned} E[R(X_{n+1})'\hat{\beta} - X'_{n+1}\beta_0]^2 &= \{E[R(X_{n+1})'\hat{\beta} - X'_{n+1}\hat{\beta} + X'_{n+1}\hat{\beta} - X'_{n+1}\beta_0]^2\} \\ &= \{E[R(X_{n+1})'\hat{\beta} - X'_{n+1}\hat{\beta}]^2 + E[X'_{n+1}\hat{\beta} - X'_{n+1}\beta_0]^2 \\ &+ E[\hat{\beta}'(R(X_{n+1}) - X_{n+1})X'_{n+1}(\hat{\beta} - \beta_0)] \\ &+ E[(\hat{\beta} - \beta_0)'X_{n+1}(R(X_{n+1})' - X'_{n+1})\hat{\beta}]. \end{aligned} \tag{A.61}$$

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

have

$$\sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \{E[R(X_{n+1})'\hat{\beta} - X'_{n+1}\beta_0]^2 - E[X'_{n+1}\hat{\beta} - X'_{n+1}\beta_0]^2\} = \sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \{E[\hat{\beta}' D_{n+1} D'_{n+1}\hat{\beta}] \tag{A.62}$$

$$+ E[\hat{\beta}' D_{n+1} X'_{n+1}(\hat{\beta} - \beta_0)] \tag{A.63}$$

$$+ E[(\hat{\beta} - \beta_0)' X_{n+1} D'_{n+1}\hat{\beta}]\}. \tag{A.64}$$

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

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

First, in case of $\hat{\beta}' D_{n+1} = 0$, since right side terms, (A.62)-(A.64) are all zero, the Lasso estimator is incentive compatible. $\hat{\beta}' D_{n+1} = 0$ can be zero in three scenarios, either $\hat{\beta} = 0$, or $\hat{\beta} \neq 0_p$, and $D_{n+1} = 0_p$ (case of full truth), or $\hat{\beta} \neq 0_p$, $D_{n+1} \neq 0_p$ but $\hat{\beta}' D_{n+1} = 0$.
In the case of $\hat{\beta}' D_{n+1} \neq 0$, we have the following analysis. We start with (A.62)

$$\begin{aligned} \sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} E[\hat{\beta}' D_{n+1} D'_{n+1}\hat{\beta}] &\geq \inf_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} E[\hat{\beta}' D_{n+1} D'_{n+1}\hat{\beta}] \\ &\geq [\inf_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} E\|\hat{\beta}\|_2^2] c_n, \end{aligned} \tag{A.65}$$

where $c_n > 0$, and can be a local to zero sequence (our proofs go through with $c_n = c$ being a positive constant as well) and we use Condition 1. Then in (A.65) we analyze the first term on the right side. In that respect we start with

$$\begin{aligned} \|\hat{\beta}\|_2^2 &= \|\hat{\beta} - \beta_0 + \beta_0\|_2^2 \\ &= (\hat{\beta} - \beta_0)'(\hat{\beta} - \beta_0) + \beta_0'\beta_0 + 2(\hat{\beta} - \beta_0)'\beta_0 \\ &\geq \|\beta_0\|_2^2 + 2(\hat{\beta} - \beta_0)'\beta_0 \\ &\geq \|\beta_0\|_2^2 - 2\|(\hat{\beta} - \beta_0)\|_1 \|\beta_0\|_\infty, \end{aligned} \tag{A.66}$$

where the first inequality is obtained by observing the first term in the second equality is non-negative, and dropping that first term in the second equality, and the second inequality is observed by Holder's inequality. Use (A.66) in (A.65) to have, by $\sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \|\beta_0\|_\infty \leq c_2 < \infty$, and $c_2 > 0$ is a positive constant

$$\begin{aligned}
\sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} E[\hat{\beta}' D_{n+1} D_{n+1}' \hat{\beta}] &\geq c_1^2 c_n \\
&- 2 \sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} E\|\hat{\beta} - \beta_0\|_1 \sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \|\beta_0\|_\infty c_n \\
&\geq c_n[c_1^2 - (4c_2 C_5) s_0 \lambda_n], \qquad \text{(A.67)}
\end{aligned}$$

by Assumption 1, Theorem A.1, since $s_0 \geq 1$ and the lowest possible value of the l_2 norm of β_0 can be taken as $c_1 > 0$, a positive constant.

Now analyze (A.63), the analysis of (A.64) is the same and thus omitted. See that

$$\begin{aligned}
\hat{\beta}' D_{n+1} X_{n+1}'(\hat{\beta} - \beta_0) &\leq |\hat{\beta}' D_{n+1} X_{n+1}'(\hat{\beta} - \beta_0)| \\
&\leq |\hat{\beta}' D_{n+1}||X_{n+1}'(\hat{\beta} - \beta_0)| \\
&\leq \|\hat{\beta}\|_1 \|D_{n+1}\|_\infty \|X_{n+1}\|_\infty \|\hat{\beta} - \beta_0\|_1, \qquad \text{(A.68)}
\end{aligned}$$

where we use Holder's inequality. Then

$$\begin{aligned}
\sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} E[\hat{\beta}' D_{n+1} X_{n+1}'(\hat{\beta} - \beta_0)] &\leq \|D_{n+1}\|_\infty \|X_{n+1}\|_\infty \sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \|E\left[\|\hat{\beta}\|_1 \|\hat{\beta} - \beta_0\|_1\right] \\
&\leq [M_3][M_4] \sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \left[E\|\hat{\beta}\|_1^2\right]^{1/2} \sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \left[E\|\hat{\beta} - \beta_0\|_1^2\right]^{1/2} \\
&\leq M_3 M_4 [(2C_5^{1/2}) s_0 \lambda_n][(2C_{10}^{1/2}) s_0^{1/2}] \qquad \text{(A.69)}
\end{aligned}$$

where we apply (A.68) for the first inequality and Cauchy-Schwartz Inequality in the second inequality above, and M_3, M_4 definitions. Then we apply Theorems A.1-A.2 with $k = 2$.

By (A.60) (A.67),(A.69) in (A.62)-(A.64)

$$\begin{aligned}\sup_{\beta\in\mathcal{B}_{l_0}(s_0)}\{E[\hat{\beta}'D_{n+1}D'_{n+1}\hat{\beta}]+2E[\hat{\beta}'D_{n+1}X'_{n+1}(\hat{\beta}-\beta_0)\} &\geq \inf_{\beta\in\mathcal{B}_{l_0}(s_0)}\{E[\hat{\beta}'D_{n+1}D'_{n+1}\hat{\beta}]\}\\ &- 2\sup_{\beta\in\mathcal{B}_{l_0}(s_0)}|E[\hat{\beta}'D_{n+1}X'_{n+1}(\hat{\beta}-\beta_0)]|\\ &\geq c_1^2c_n-(4c_2C_5)s_0\lambda_nc_n\\ &- (8C_5^{1/2}C_{10}^{1/2})M_3M_4s_0^{3/2}\lambda_n, \qquad \text{(A.70)}\end{aligned}$$

where c_2, C_5, C_6 are positive constants, and M_3, M_4 are nondecreasing positive sequences in n.

Next, choose

$$\lambda_n \leq \frac{c_1^2c_n}{(4c_2C_5)s_0c_n+(8C_5^{1/2}C_{10}^{1/2})M_3M_4s_0^{3/2}}, \qquad \text{(A.71)}$$

to have (A.70) to be non-negative, hence left side of (A.62)

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

We can see that λ_n has to be in the following bound for incentive compatibility by (A.60)(A.71)

$$min\left(\frac{C_9}{s_0^{1/2}}, \frac{c_1^2c_n}{(4c_2C_5)s_0c_n+(8C_5^{1/2}C_{10}^{1/2})M_3M_4s_0^{3/2}}\right)\geq\lambda_n\geq max\left(\frac{C_4c_s}{s_0^{1/2}}, \frac{C_7P\left(\mathcal{F}^c\right)^{1/8}}{s_0^{1/2}}, \frac{C_8P\left(\mathcal{F}^c\right)^{1/4}}{s_0^{1/2}}\right).$$

Q.E.D.

A.3 Asymptotics

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

Assumption A.2.

$$\frac{\sqrt{EM_F^2}\sqrt{lnp}}{\sqrt{n}}=O(1).$$

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

$$\max_{1\leq j\leq p} |\hat{\mu}_j - \mu_j| = O_p(\sqrt{lnp/n}). \tag{A.72}$$

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

$$\max_{1\leq j\leq p} |\hat{\mu}_j - \mu_j| = O_p(\sqrt{lnn/n}). \tag{A.73}$$

See that $P(\mathcal{F}^c) \to 0$ due to upper bound in the proof of Lemma A.4, (A.25) converging to zero under Assumptions A.1-A.2. Also see that

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

by Assumption 4 with definition of M_1. Then clearly by $c_s = O(\frac{\sqrt{s_0}}{\sqrt{n}})$ and by Assumption 6, $\lambda_n s_0 \to 0$, we have the rate for Lemma A.1

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

since $\lambda_n^2 s_0/c_s^2 \to \infty$.

Proof of Theorem 1. Here we only provide the case of $\hat{\beta}' D_{n+1} \neq 0$ from the proof of Theorem A.3. In that respect since (A.67) is non-negative, there will be no need for Condition 1. We consider the left side of (A.70) when $n \to \infty$

$$\begin{aligned}
\sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \{E[\hat{\beta}' D_{n+1} D_{n+1}' \hat{\beta}] \;&+\; 2E[\hat{\beta}' D_{n+1} X_{n+1}' (\hat{\beta} - \beta_0)\} \\
&\geq \inf_{\beta \in \mathcal{B}_{l_0}(s_0)} \{E[\hat{\beta}' D_{n+1} D_{n+1}' \hat{\beta}]\} - 8C_5^{1/2} C_{10}^{1/2} M_3 M_4 s_0^{3/2} \lambda_n \\
&\geq 0,
\end{aligned}$$

by Assumption 6. Hence incentive compatibility is established, and there is no need for an upper bound for the tuning parameter by Assumption 6. Also the first term (approximation error) in lower bound is satisfied by $\lambda_n/(c_s/s_0^{1/2}) \to \infty$, since $\lambda_n = O(\sqrt{lnp/n})$, $c_s = O(\sqrt{s_0/n})$. Also see that $P(\mathcal{F}^c)^{1/8} \geq P(\mathcal{F}^c)^{1/4}$, hence the lower bound $C_8 P(\mathcal{F}^c)^{1/4}/s_0^{1/2}$ is not binding in asymptotics case. **Q.E.D.**

B Appendix B

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

B.1 When $p \leq n$

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

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

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

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

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

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

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

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

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

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

B.2 Relaxing Assumption 1(ii)

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

Assumption 1(iv).

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

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

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

In terms of discussion A.2.3, this implies $S_0 = D_1$, and D_2 is an empty set. So Assumption 1(iv) can simultaneously allow s_0 increasing with n, and all large nonzero coefficients in S_0. Previously in Assumption 1(ii), there can be only a fixed number of large coefficients, and increasing $(s_0 - f_1)$ number of local to zero (small) coefficients.

C Appendix C: A Nonlinear High Dimensional Model: GLM with Structured Sparsity Estimators

This part consists of three subsections. First notation and terminology is introduced (which can be skipped) and then finite sample IC result with necessary theorems are introduced and ends with asymptotic IC case.

C.1 Notation and Terminology for GLM Structured Sparsity Estimators

This part is a short introduction to weakly decomposable norms and of the definition of events that is used in the proof of Theorem 1 in Caner (2023) which we benefit. Theorem 1 of Caner (2023) is an oracle inequality for GLM structured sparsity estimators. Following is a summary of Section 2 of Caner (2023) or Section 6.4 van de Geer (2016). We divide the index of all regressors $J = \{1, 2, \cdots, p\}$ into two mutually exclusive sets S and S^c. Let $|S|$ represent the cardinality of the index subset S. We define also following norm $\tilde{\Omega}(.)$ on $R^{p-|S|}$. Define β_S as a vector with all entries that correspond to $j \in S$ as kept in the new vector, and all other entries that is not in S, set to zero. β_{S^c} is defined in the same way but S^c entries are kept and the rest is set to zero.

Definition C.1. (weakly decomposable norm).*Fix some index set S. S is known as allowed set if for a norm $\tilde{\Omega}(.)$*

$$\Omega(\beta) \geq \Omega(\beta_S) + \tilde{\Omega}(\beta_{S_c}),$$

where we call $\Omega(.)$ a weakly decomposable norm.

Weakly decomposable norms are generated from convex cones, as shown in Section 6.9 of van de Geer (2016). We also define

$$\underline{\Omega}(\beta) := \Omega(\beta_S) + \tilde{\Omega}(\beta_{S_c}),$$

and hence $\Omega(\beta) \geq \underline{\Omega}(\beta)$ as in (2.6) of Caner (2023) and on p.79 and p.82 of van de Geer (2016).

We introduce the three events that we condition our proofs. To do that we need the following definitions first. Define $\mathcal{B}_{local}$ as a convex subset of the collection

$$\{\tilde{\beta} \in \mathcal{B} : \underline{\Omega}(\tilde{\beta} - \beta) \leq M\},$$

with $M > 0$ a positive constant. $\mathcal{B} \subset \mathbf{R}^p$, and it is a convex subset. $X := n \times p$ regressor matrix.

Define event E_1:

$$E_1 := \left\{ E\rho(y_i, X_i'\tilde{\beta}) - E\rho(y_i, X_i'\beta_0) \geq \frac{\|X(\tilde{\beta} - \beta_0)\|_n^2}{2C_\rho^2} - \frac{M^2 t_1}{2C_\rho} \right\},$$

with $t_1 = O(\sqrt{lnp/n})$ a positive sequence in n, and C_ρ is a positive constant that changes with shape of the loss $\rho(.)$.

This event is peculiar to nonlinear models and it is related to the one margin condition in van de Geer (2016). This condition is shown to hold in Lemma A.2 of Caner (2023) with probability approaching 1. E_1 regulates the loss function behavior around β_0.

Next condition is similar to adaptive restricted eigenvalue condition in linear models, and it is called effective sparsity. Let $L > 0$ a positive constant, and S is an allowed set.

$$\Gamma^2(L, S) := [\min\{E\|X\beta_S - X\beta_{S_c}\|_n^2 : \Omega(\beta_S) = 1, \tilde{\Omega}(\beta_{S_c}) \leq L\}]^{-1},$$

and it is related to more familiar compatibility constant, a positive compatibility constant implies that effective sparsity is finite. To see this see p.81 of van de Geer (2016)or (2.7) of Caner (2023). Also a positive minimum eigenvalue of EX_iX_i' implies a positive compatibility constant. Hence a positive minimum eigenvalue of EX_iX_i' implies effective sparsity being finite.

Sample version of effective sparsity is defined in the following way

$$\hat{\Gamma}^2(L, S) := [\min\{\|X\beta_S - X\beta_{S_c}\|_n^2 : \Omega(\beta_S) = 1, \tilde{\Omega}(\beta_{S_c}) \leq L\}]^{-1},$$

The event E_2 is defined as follows and it is proven to be holding true with wpa1 in Lemma A.3 of Caner (2023), with $L = 2, S = S_0$ (sparse model index set)

$$E_2 := \{2\Gamma^2(2, S_0) \geq \hat{\Gamma}^2(2, S_0)\}.$$

The last event is related to noise condition in linear models.

$$E_3 := \left\{ \sup_{\tilde{\beta} \in \mathcal{B}: \underline{\Omega}(\tilde{\beta} - \beta_0) \leq M} \left| [n^{-1} \sum_{i=1}^n (\rho(y_i, X_i'\tilde{\beta}) - E\rho(y_i, X_i'\tilde{\beta}))] - [n^{-1} \sum_{i=1}^n (\rho(y_i, X_i'\beta_0) - E\rho(y_i, X_i'\beta_0))] \right| \leq \lambda_e M \right\},$$

where $\lambda_n = 16\lambda_e$ as in proof of Theorem 1(ii) in Caner (2023). This last event also holds true with wpa1 as in Lemma A.4 of Caner (2023). Also proof of Theorem 1 of Caner (2023) proves, given $\mathcal{E} = E_1 \cap E_2 \cap E_3$, the complement of that set $P(\mathcal{E}^c) \to 0$ given Assumptions G.1-G.4 here.

C.2 Finite Sample IC for GLM Structured Sparsity Estimator

We formally define IC for GLM structured sparsity estimator.

Definition C.2. The GLM structured sparsity estimator is **uniformly incentive-compatible** if for every X_{n+1}, for every $R(X_{n+1})$ and for every β_0 that satisfy Assumptions G.1-G.6, and for $p \geq 2$ and $n \geq n_0 > 0$,

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

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

In Theorem C.3, we formally prove that GLM structured sparsity estimator is incentive compatible, where we show the need for new upper and lower bounds for λ_n compared to lasso for approximately linear case. We have the counterpart to Theorem A.1 with this new assumption.

Theorem C.1. *Under Assumptions G.1-G.5, C_b, G_1 are positive constants that are defined in the proof, with λ_n is chosen to reflect the following lower bound, for sufficiently large n*

$$\lambda_n \geq G_1 \max \left(\frac{n^{1/4k}}{(lnp)^{1/4k}} \frac{P(\mathcal{E}^c)^{1/4k}}{s_0^{1/2}}, \frac{P(\mathcal{E}^c)^{1/2k}}{s_0} g(s_0) \right),$$

then

$$\left[E\{\underline{\Omega}(\hat{\beta}_G - \beta_0)\}^k \right]^{1/k} \leq (2C_b)^{1/k} s_0 \lambda_n.$$

Proof of Theorem C.1. We proceed in four steps. Set $\mathcal{E} := \{E_1 \cap E_2 \cap E_3\}$. $\mathcal{E}^c$ is the complement.

Step 1. We obtain a moment bound for estimation error under $\mathcal{E}$. By Assumptions G.1-G.4, Theorem 1 of Caner (2023) under $\mathcal{E}$ proves, for sufficiently large n

$$\underline{\Omega}(\hat{\beta}_G - \beta_0) \leq C\lambda_n s_0,$$

uniformly over $\mathcal{B}_{l_0}(s_0)$. Then for a positive constant $C_b > 0$, defined as $C_b \geq C^k$

$$E\underline{\Omega}(\hat{\beta}_G - \beta_0)^k 1_{\{\mathcal{E}\}} \leq C_b \lambda_n^k s_0^k. \tag{C.2}$$

Step 2. We obtain the moment bound for difference between unpenalized loss function and its counterpart at true value of the parameter β_0. Define another positive constant $C_0 > 0$ as the upper bound on the partial derivative of the loss $\rho(,)$ with respect to second argument, in a neighborhood of β_0, where $\bar{\beta} \in (\beta_0, \hat{\beta}_G)$

$$\begin{aligned}
R_n(\hat{\beta}_G) - R_n(\beta_0) &\leq |R_n(\hat{\beta}_G) - R_n(\beta_0)| = |\dot{\rho}(y_i, X_i'\bar{\beta})X_i'(\hat{\beta}_G - \beta_0)| \qquad \text{(C.3)} \\
&\leq C_0|X_i'(\hat{\beta}_G - \beta_0)| \leq C_0[\underline{\Omega}^*(X_i)][\underline{\Omega}(\hat{\beta}_G - \beta_0)] \\
&\leq C_0[\|X_i\|_\infty][\underline{\Omega}(\hat{\beta}_G - \beta_0)] \\
&\leq C_m\|X_i\|_\infty, \qquad \text{(C.4)}
\end{aligned}$$

where we use Assumption G.4 for the second inequality, and for the third inequality we use dual norm inequality in (A.1) of Caner (2023). Then for the fourth inequality we use (A.3) of Caner (2023) and the last result is obtained by $\hat{\beta}_G \in \mathcal{B}_{local}$ by Theorem 1 in Caner (2023) which means uniformly over $B_{l_0}(s_0)$, $\underline{\Omega}(\hat{\beta}_G - \beta_0) \leq M$, where M is a constant under $\mathcal{E}$. Defining a local neighborhood around β_0 and showing the estimator is in this local neighborhood is suggested by p.113, Section 7.6 of van de Geer (2016) in high dimensional

GLM framework with penalties. C_m is a positive constant and defined as multiplication of C_0 and M.

Step 3. As in the proof of Theorem 1, to get a bound for $E[\underline{\Omega}(\hat{\beta}_G - \beta_0)]$ without basing the proof on $\mathcal{E}$ we will get a rough bound for the moment that is mentioned. In that respect by the definition of $\hat{\beta}_G$ as the minimizer of the loss function

$$R_n(\hat{\beta}_G) + \lambda_n \Omega(\hat{\beta}_G) \leq R_n(\beta_0) + \lambda_n \Omega(\beta_0). \tag{C.5}$$

There are two possibilities depending on $R_n(\beta_0) - R_n(\hat{\beta}_G)$. First if $R_n(\beta_0) - R_n(\hat{\beta}_G) \leq 0$ by (C.5)

$$\lambda_n \Omega(\hat{\beta}_G) \leq \lambda_n \Omega(\beta_0). \tag{C.6}$$

Then by $\underline{\Omega}(.) \leq \Omega(.)$ definition and triangle inequality

$$\underline{\Omega}(\hat{\beta}_G - \beta_0) \leq \underline{\Omega}(\hat{\beta}_G) + \underline{\Omega}(\beta_0) \leq \Omega(\hat{\beta}_G) + \Omega(\beta_0) \leq 2\Omega(\beta_0), \tag{C.7}$$

where the last inequality is by (C.6). Under $R_n(\beta_0) - R_n(\hat{\beta}_G) \leq 0$ we have

$$\underline{\Omega}(\hat{\beta}_G - \beta_0) \leq 2\Omega(\beta_0). \tag{C.8}$$

Then we consider $R_n(\beta_0) - R_n(\hat{\beta}_G) > 0$ case. So when $R_n(\beta_0) - R_n(\hat{\beta}_G) > 0$ using (C.3)(C.5) for the first and second inequality below in (C.9) and Step 2

$$\begin{aligned} \Omega(\hat{\beta}_G) &\leq \left\{ \frac{[R_n(\beta_0) - R_n(\hat{\beta})]}{\lambda_n} \right\} + \Omega(\beta_0) \\ &\leq \left\{ \frac{|R_n(\hat{\beta}_G) - R_n(\beta_0)|}{\lambda_n} \right\} + \Omega(\beta_0) \leq \frac{C_m \|X_i\|_\infty}{\lambda_n} + \Omega(\beta_0), \end{aligned} \tag{C.9}$$

and by (C.4) for the last inequality. Then by triangle inequality and $\underline{\Omega}(.) \leq \Omega(.)$ by (C.9)

$$\begin{aligned} \underline{\Omega}(\hat{\beta}_G - \beta_0) &\leq \underline{\Omega}(\hat{\beta}_G) + \underline{\Omega}(\beta_0) \\ &\leq \Omega(\hat{\beta}_G) + \Omega(\beta_0) \\ &\leq \frac{C_m \|X_i\|_\infty}{\lambda_n} + 2\Omega(\beta_0). \end{aligned} \tag{C.10}$$

Clearly (C.10) is larger than (C.8). So we have

$$\underline{\Omega}(\hat{\beta}_G - \beta_0) \leq \frac{C_m \|X_i\|_\infty}{\lambda_n} + 2\Omega(\beta_0). \tag{C.11}$$

Then take the $2k$ th moment and use (A.35) with the definition of the positive constant $G_0 := \max(2^{k-1/2} C_m^k, 2^{2k-1/2})$

$$\begin{aligned} E\underline{\Omega}(\hat{\beta}_G - \beta_0)^{2k} &\leq E\left[\frac{C_m \|X_i\|_\infty}{\lambda_n} + 2\Omega(\beta_0)\right]^{2k} \\ &\leq 2^{2k-1}\left\{E\left[\frac{C_m \|X_i\|_\infty}{\lambda_n}\right]^{2k} + 2^{2k}\Omega(\beta_0)^{2k}\right\} \\ &= G_0^2[\frac{E\|X_i\|_\infty^{2k}}{\lambda_n^{2k}} + \Omega(\beta_0)^{2k}]. \end{aligned} \tag{C.12}$$

Step 4. Now we merge the bounds in (C.2)(C.12).

$$E\underline{\Omega}(\hat{\beta} - \beta_0)^k = E\underline{\Omega}(\hat{\beta} - \beta_0)^k 1_{\{\mathcal{E}\}} + E\underline{\Omega}(\hat{\beta} - \beta_0)^k 1_{\{\mathcal{E}^c\}}$$

Next by Assumption G.2, Cauchy-Schwartz Inequality, and M_2 definition $EM_2^{2k} := E\|X_i\|_\infty^{2k}$, and by Assumption G.5

$$\begin{aligned} E\underline{\Omega}(\hat{\beta}_G - \beta_0)^k &\leq C_b \lambda_n^k s_0^k + \sqrt{E[\underline{\Omega}(\hat{\beta}_G - \beta_0)^{2k}]} P(\mathcal{E}^c)^{1/2} \\ &\leq C_b \lambda_n^k s_0^k + G_0 \max\left(\frac{n^{1/2}}{(lnp)^{1/2}\lambda_n^k}, g(s_0)^k\right) P(\mathcal{E}^c)^{1/2}. \end{aligned} \tag{C.13}$$

We show now that $C_b \lambda_n s_0^k$ is larger than equal to other two possibilities under a lower bound condition on λ_n. Let

$$C_b \lambda_n^k s_0^k \geq \frac{G_0 n^{1/2}}{(lnp)^{1/2}\lambda_n^k} P(\mathcal{E}^c)^{1/2}.$$

This bound is possible by choosing

$$\lambda_n \geq [G_0/C_b]^{1/2k} \frac{n^{1/4k}}{(lnp)^{1/4k}} \frac{P(\mathcal{E}^c)^{1/4k}}{s_0^{1/2}}. \tag{C.14}$$

For the second possibility, we need

$$C_b \lambda_n^k s_0^k \geq G_0 g(s_0)^k P(\mathcal{E}^c)^{1/2},$$

which is possible under the following lower bound,

$$\lambda_n \geq [G_0/C_b]^{1/k} \frac{P(\mathcal{E}^c)^{1/2k}}{s_0} g(s_0). \tag{C.15}$$

So if we define a positive constant $G_1 := max((\frac{G_0}{C_b})^{1/2k}, (\frac{G_0}{C_b})^{1/k})$

$$\lambda_n \geq G_1 \max\left(\frac{n^{1/4k}}{(lnp)^{1/4k}} \frac{P(\mathcal{E}^c)^{1/4k}}{s_0^{1/2}}, \frac{P(\mathcal{E}^c)^{1/2k}}{s_0} g(s_0) \right),$$

then

$$E\underline{\Omega}(\hat{\beta}_G - \beta_0)^k \leq 2C_b \lambda_n^k s_0^k,$$

and the uniformity over $\mathcal{B}_{l_0}(s_0)$ still holds since the right side depends on β_0 through s_0 only. **Q.E.D.**

The following Theorem C.2 is the moment bound for GLM structured sparsity estimator, and this is a new result in the literature and will be used in finite sample IC proof in Theorem C.3.

Theorem C.2. *Under Assumptions G.1-G.5 with the following bounds for λ_n,*

$$\frac{g(s_0)}{Cs_0} \geq \lambda_n \geq G_1 [\frac{n}{lnp}]^{1/4k} \frac{P(\mathcal{E}^c)^{1/4k}}{s_0^{1/2}},$$

we have

$$[E\underline{\Omega}(\hat{\beta}_G)^k]^{1/k} \leq 2^{1+1/k} g(s_0).$$

Proof of Theorem C.2. We start with

$$E\underline{\Omega}(\hat{\beta}_G)^k = E\underline{\Omega}(\hat{\beta})^k 1_{\{\mathcal{E}\}} + E\underline{\Omega}(\hat{\beta}_G)^k 1_{\{\mathcal{E}^c\}}. \tag{C.16}$$

We first consider the first term on the right side of (C.16). By triangle inequality

$$\underline{\Omega}(\hat{\beta}_G) \leq \underline{\Omega}(\hat{\beta}_G - \beta_0) + \underline{\Omega}(\beta_0).$$

By Theorem 1 of Caner (2023) and Assumption G.5 and $\underline{\Omega}(.) \leq \Omega(.)$

$$\underline{\Omega}(\hat{\beta}_G) \leq C\lambda_n s_0 + g(s_0). \tag{C.17}$$

Then by (C.17) and (A.35) under $\mathcal{E}$

$$E\underline{\Omega}(\hat{\beta}_G)^k \leq E[C\lambda_n s_0 + g(s_0)]^k \leq 2^{k-1}[(C\lambda_n s_0)^k + g(s_0)^k]. \tag{C.18}$$

Then with

$$C\lambda_n \leq \frac{g(s_0)}{s_0}. \tag{C.19}$$

we obtain

$$E\underline{\Omega}(\hat{\beta}_G)^k 1_{\{\mathcal{E}\}} \leq 2^k g(s_0)^k. \tag{C.20}$$

Now we consider the second term on the right side of (C.16). Use (C.9) and $\underline{\Omega}(.) \leq \Omega(.)$ and (A.35), and also define a positive constant $G_2 := max(2^{k-1/2}C_m^k, 2^{k-1/2})$

$$\begin{aligned} E\underline{\Omega}(\hat{\beta}_G)^{2k} &\leq E\left[\frac{C_m\|X_i\|_\infty}{\lambda_n} + \Omega(\beta_0)\right]^{2k} \\ &\leq 2^{2k-1}[\frac{C_m^{2k}}{\lambda_n^{2k}}E\|X_i\|_\infty^{2k} + \Omega(\beta_0)^{2k}] \\ &\leq G_2^2[\frac{E\|X_i\|_\infty^{2k}}{\lambda_n^{2k}} + g(s_0)^{2k}], \end{aligned} \tag{C.21}$$

where the last inequality is by Assumption G.5. Combine (C.20)(C.21) in (C.16)

$$\begin{aligned} E\underline{\Omega}(\hat{\beta})^k &= E\underline{\Omega}(\hat{\beta}_G)^k 1_{\{\mathcal{E}\}} + E\underline{\Omega}(\hat{\beta}_G)^k 1_{\{\mathcal{E}^c\}} \\ &\leq E\underline{\Omega}(\hat{\beta}_G)^k 1_{\{\mathcal{E}\}} + \sqrt{E\underline{\Omega}(\hat{\beta}_G)^{2k}}P(\mathcal{E}^c)^{1/2} \\ &\leq 2^k g(s_0)^k + G_2 max\left(\frac{\sqrt{n}}{\sqrt{lnp}}\lambda_n^{-k}, g(s_0)^k\right)P(\mathcal{E}^c)^{1/2}, \end{aligned} \tag{C.22}$$

where we use Assumption G.2 for the last inequality, $E\|X_i\|_\infty^{2k} = O(n/lnp)$. Note that the second term in max(,) will not contribute to the rate since the first term on the right side of (C.22) can be always be bigger than the second term multiplied by G_2 in max(,) with choice of $2^k \geq G_2$ which is possible by small C_m in G_2 definition above (C.21). Then note that by (C.19), we have that $g(s_0)^k \geq C^k\lambda_n^k s_0^k$ but $C^k\lambda_n^k s_0^k$ dominates the first term multiplied by G_2 in max(,) in (C.22) and this can be seen by (C.13)(C.14), and set $C_b = C^k$ in step 1 of the proof of Theorem 1, and since $G_0 \geq G_2$ by definition. So by the proof of

Theorem C.1, if

$$\lambda_n \geq G_1[\frac{n}{lnp}]^{1/4k}\frac{P(\mathcal{E}^c)^{1/4k}}{s_0^{1/2}},$$

we get

$$E\underline{\Omega}(\hat{\beta}_G)^k \leq 2^{k+1}g(s_0)^k.$$

Uniformity over $\mathcal{B}_{l_0}(s_0)$ follows through since the bound depends on β_0 through s_0 only. **Q.E.D.**

We provide two extra assumptions so that we can have finite sample IC. In asymptotic IC, these are not needed. For some $\delta > 0$, we have the following lower bound for the partial derivative with respect to second argument of GLM loss (unpenalized)

Assumption G.7.

$$\inf_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \inf_{|\bar{a}-a_0|\leq\delta} \inf_{y_{n+1}\subset\mathcal{Y}} \dot{\rho}(y_{n+1}, \bar{a})^2 \geq c > 0.$$

We impose $\bar{a} = \bar{R}'\hat{\beta}_G$, where $\bar{R} \in [R(X_{n+1}), X_{n+1}]$ in the following proof of Theorem C.3. Also we need the following lower and upper bounds on β_0, and it is a very mild restriction.

Assumption G.8. We need at least one nonzero (i.e not local to zero and not zero) coefficient in β_0, and the largest possible absolute coefficient in β_0 is a constant.

Theorem C.3. *Incentive compatibility is achieved by Assumptions G.1-G.5, G.7-G.8 and Condition 1 implemented with $\hat{\beta}_G$, with sufficiently large n, and with the following bounds on λ_n, with G_1, G_3, G_4 are positive constants that are defined in the proof,*

$$\min\left(\frac{g(s_0)}{Cs_0}, \frac{cc_n}{(G_3)c_n s_0 + (G_4)M_3M_4 s_0 g(s_0)}\right) \geq \lambda_n \geq G_1 \max\left(\frac{n^{1/8}}{(lnp)^{1/8}}\frac{P(\mathcal{E}^c)^{1/8}}{s_0^{1/2}}, P(\mathcal{E}^c)^{1/4}\frac{g(s_0)}{s_0}\right).$$

Proof of Theorem C.3 We start with the definition of incentive compatibility

$$E[\rho(y_{n+1}, R(X_{n+1})'\hat{\beta}_G) - \rho(y_{n+1}, X'_{n+1}\beta_0)]^2 - \{E[\rho(y_{n+1}, X'_{n+1}\hat{\beta}_G) - \rho(y_{n+1}, X'_{n+1}\beta_0)]^2\}. \tag{C.23}$$

Add and subtract $\rho(y_{n+1}, X_{n+1}'\hat{\beta}_G)$ on the first expected term in (C.23) and simplify to have

$$\begin{aligned} & E[\rho(y_{n+1}, R(X_{n+1})'\hat{\beta}_G) - \rho(y_{n+1}, X_{n+1}'\hat{\beta}_G)]^2 \\ + \quad & 2E\{[\rho(y_{n+1}, R(X_{n+1})'\hat{\beta}_G) - \rho(y_{n+1}, X_{n+1}'\hat{\beta}_G)][\rho(y_{n+1}, X_{n+1}'\hat{\beta}_G) - \rho(y_{n+1}, X_{n+1}'\beta_0)]\}. \end{aligned} \tag{C.24}$$

We simplify the first term in (C.24) via a mean value expansion

$$\rho(y_{n+1}, R(X_{n+1})'\hat{\beta}_G) - \rho(y_{n+1}, X_{n+1}'\hat{\beta}_G) = \dot{\rho}(y_{n+1}, \bar{R}\hat{\beta}_G)[R(X_{n+1}) - X_{n+1}]'\hat{\beta}_G, \tag{C.25}$$

with $\bar{R} \in [R(X_{n+1}), X_{n+1}]$. Also to analyze one of the terms in (C.24) we have the mean value expansion

$$\rho(y_{n+1}, X_{n+1}'\hat{\beta}_G) - \rho(y_{n+1}, X_{n+1}'\beta_0) = \dot{\rho}(y_{n+1}, X_{n+1}'\bar{\beta})X_{n+1}'(\hat{\beta}_G - \beta_0), \tag{C.26}$$

where $\bar{\beta} \in (\beta_0, \hat{\beta}_G)$. Use (C.25)-(C.26) in (C.23)(C.24) with $D_{n+1} := R(X_{n+1}) - X_{n+1}$

$$\begin{aligned} E[\rho(y_{n+1}, R(X_{n+1})'\hat{\beta}_G) \quad - \quad & \rho(y_{n+1}, X_{n+1}'\beta_0)]^2 - \{E[\rho(y_{n+1}, X_{n+1}'\hat{\beta}_G) - \rho(y_{n+1}, X_{n+1}'\beta_0)]^2\} \\ = \quad & E[\dot{\rho}(y_{n+1}, \bar{R}\hat{\beta})D_{n+1}'\hat{\beta}_G]^2 \\ + \quad & 2E\{[\dot{\rho}(y_{n+1}, \bar{R}\hat{\beta}_G)D_{n+1}'\hat{\beta}]'[\dot{\rho}(y_{n+1}, X_{n+1}'\bar{\beta})X_{n+1}'(\hat{\beta}_G - \beta_0)]\}. \end{aligned} \tag{C.27}$$

Consider the first term on the right side of (C.26) by Condition 1 with $\hat{\beta}_G$ and Assumption G.7

$$E[\dot{\rho}(y_{n+1}, \bar{R}\hat{\beta}_G)D_{n+1}'\hat{\beta}]^2 \geq cc_n E\|\hat{\beta}_G\|_2^2 > 0. \tag{C.28}$$

Next, using the analysis in first inequality in (A.66)

$$\|\hat{\beta}_G\|_2^2 \geq \|\beta_0\|_2^2 + 2(\hat{\beta}_G - \beta_0)'\beta_0 \geq \|\beta_0\|_2^2 - 2|(\hat{\beta}_G - \beta_0)'\beta_0|. \tag{C.29}$$

But by (A.1)-(A.2) of Caner (2023)

$$|(\hat{\beta}_G - \beta_0)'\beta_0| \leq \underline{\Omega}(\hat{\beta}_G - \beta_0)\underline{\Omega}^*(\beta_0) \leq \underline{\Omega}(\hat{\beta}_G - \beta_0)\|\beta_0\|_\infty. \tag{C.30}$$

Use (C.30) in (C.29)

$$\|\hat{\beta}_G\|_2^2 \geq \|\beta_0\|_2^2 - 2\underline{\Omega}(\hat{\beta}_G - \beta_0)\|\beta_0\|_\infty. \tag{C.31}$$

Use (C.31) in (C.28) by Assumption G.8, where we absorb all small coefficients as $c > 0$,

$$\begin{aligned} E[\dot{\rho}(y_{n+1}, \bar{R}\hat{\beta}_G)D'_{n+1}\hat{\beta}_G]^2 &\geq cc_n \\ &- 2cc_n[\sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} E\underline{\Omega}(\hat{\beta}_G - \beta_0)] \sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \|\beta_0\|_\infty. \quad \text{(C.32)} \end{aligned}$$

We consider the second term on the right side of (C.27), let $C_0 > 0$ be a positive constant

$$\begin{aligned} [\dot{\rho}(y_{n+1}, \bar{R}'\hat{\beta}_G)D'_{n+1}\hat{\beta}_G]'[\dot{\rho}(y_{n+1}, X'_{n+1}\bar{\beta})X'_{n+1}(\hat{\beta}_G - \beta_0)] &\leq |\dot{\rho}(y_{n+1}, \bar{R}'\hat{\beta}_G||\dot{\rho}(y_{n+1}, X'_{n+1}\bar{\beta})| \\ &\times |\hat{\beta}'D_{n+1}X'_{n+1}(\hat{\beta}_G - \beta_0)| \\ &\leq C_0^2|\hat{\beta}'D_{n+1}||X'_{n+1}(\hat{\beta}_G - \beta_0)| \\ &\leq C_0^2\underline{\Omega}(\hat{\beta}_G)\underline{\Omega}^*(D_{n+1}) \\ &\times \underline{\Omega}^*(X_{n+1})\underline{\Omega}(\hat{\beta}_G - \beta_0) \\ &\leq C_0^2\underline{\Omega}(\hat{\beta}_G)\|D_{n+1}\|_\infty \\ &\times \|X_{n+1}\|_\infty\underline{\Omega}(\hat{\beta}_G - \beta_0), \quad \text{(C.33)} \end{aligned}$$

where the second inequality is by Assumption G.4 and the third inequality is dual norm inequality in Lemma A.1 of Caner (2023) or (6.2) of van de Geer (2016) and the last inequality is by Lemma A.1(iii) of Caner (2023). Use (C.33) in the second term on the right side of (C.27), with M_3, M_4 definitions (they are non-random terms)

$$\begin{aligned} &E[\dot{\rho}(y_{n+1}, \bar{R}'\hat{\beta}_G)\hat{\beta}'_G D_{n+1}X'_{n+1}(\hat{\beta}_G - \beta_0)\dot{\rho}(y_{n+1}, X'_{n+1}\bar{\beta})] \\ \leq\ & C_0^2 M_3 M_4 \sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} [E\underline{\Omega}(\hat{\beta}_G)^2]^{1/2} \sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \sqrt{E\underline{\Omega}(\hat{\beta}_G - \beta_0)^2}. \quad \text{(C.34)} \end{aligned}$$

Combine (C.32)(C.34) in (C.27), with definitions of positive constants, and $C > 0$ is a positive constant $G_3 := 2(2C_b)^{1/2}C, G_4 := 2C_0^2 2^{3/2}(2C_b)^{1/2} = 8C_0^2 C_b^{1/2}$ to have with $k = 2$ in Theorems C.1-C.2, for $y_{n+1} \subset \mathcal{Y}$

$$\begin{aligned} &\{E[\rho(y_{n+1}, R(X_{n+1})'\hat{\beta}_G) - \rho(y_{n+1}, X'_{n+1}\hat{\beta}_G)]^2 - E[\rho(y_{n+1}, X'_{n+1}\hat{\beta}_G) - \rho(y_{n+1}, X'_{n+1}\beta_0)]^2\} \\ \geq\ & c_n c - 2c_n[\sup_{\beta \in \mathcal{B}_{l_0}(s_0)} E\underline{\Omega}(\hat{\beta}_G - \beta_0)] \sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} \|\beta_0\|_\infty \\ -\ & 2C_0^2 M_3 M_4[\sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} (E\underline{\Omega}(\hat{\beta}_G - \beta_0)^2)^{1/2}][\sup_{\beta_0 \in \mathcal{B}_{l_0}(s_0)} (E\underline{\Omega}(\hat{\beta}_G)^2)^{1/2}] \\ \geq\ & cc_n - G_3 c_n s_0 \lambda_n - G_4 M_3 M_4 s_0 \lambda_n g(s_0) \geq 0. \quad \text{(C.35)} \end{aligned}$$

with choice of

$$\frac{cc_n}{G_3 c_n s_0 + G_4 M_3 M_4 s_0 g(s_0)} \geq \lambda_n,$$

to have incentive compatibility with sufficiently large n. The result is obtained since the integral over the distribution for y_{n+1} can be obtained by seeing (C.28)(C.33) with Assumptions G.4 and G.7 and (C.35) is true for all $y_{n+1} \subset \mathcal{Y}$. The uniformity over $\mathcal{B}_{l_0}(s_0)$ is achieved since all the bounds depend on β_0 only through s_0.**Q.E.D.**

C.3 Asymptotic IC Proof: Proof of Theorem 2

Proof of Theorem 2. The proof of asymptotics are easily derived from the finite sample proof in Theorem C.3 proof above. So we only use the parts that are different here. First of all, the first term in (C.24) (squared expectations) is always nonnegative and hence poses no issues from asymptotic point of view. So there will be no extra conditions attached to IC because of that term in asymptotic proof here. The second term in (C.24) has to converge to zero so that we can have asymptotic IC. To analyze that term, we consider (C.34). In order to consider conditions in Theorems C.1-C.2 we need

$$M_3 M_4 s_0 g(s_0) \lambda_n \to 0. \tag{C.36}$$

With (C.36) the second term in (C.24) goes to zero. Now we consider the conditions for Theorems C.1-C.2, and see whether there are simplifications in asymptotic case. Since Theorems C.1-C.2 are the main ingredients for finite sample IC proof, we need to check these lower and upper bound conditions for λ_n. Starting with Theorem C.1 conditions with $k = 2$

$$\lambda_n (lnp)^{1/8} n^{-1/8} P(\mathcal{E}^c)^{-1/8} s_0^{1/2} \to \infty. \tag{C.37}$$

and

$$\lambda_n P(\mathcal{E}^c)^{-1/4} s_0 (g(s_0))^{-1} \to \infty. \tag{C.38}$$

Then the upper bound condition in Theorem C.2 will not be needed in asymptotic case, $\lambda_n s_0 g(s_0) \to 0$ by (C.36) above since M_3, M_4 are nondecreasing in n, so $\lambda_n s_0 / g(s_0) \to 0$ since $g(s_0)$ is nondecreasing in n as in Assumption G.5. So (C.37)-(C.38) are the conditions for asymptotic IC.**Q.E.D.**

D Appendix D

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 Theorems 1-2), incentive compatibility is analyzed through the lens of "small" and "large" lies. Finally, we show that incentive compatibility is not vacuous, it is possible for a new user to gain from lying.

In choosing the tuning parameter we use the asymptotic case since this involves only a lower bound, and will be shown to perform well even with $p = 100$ and $n = 100$. We provide two setups. The first one is a linear model (no approximation error) which simplifies our nonlinear but approximately linear model in Section 2. The linear setup gives us a good benchmark. We use lasso estimation in linear setup. The second setup is nonlinear, and we use logistical lasso with weighted group norm. Also by choosing these setups we consider a continuous outcome variable in the first setup, and a discrete outcome in the second setup.

Our first simulation setup is a sparse linear model approximating nonlinearity. The approximation error, hence $r_i = 0.5 * u_i$, for all $i = 1, \cdots, n$ where the approximation error is correlated with idiosyncratic error u_i. Let

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

where $\beta_0 = (1, 0'_{p-s_0}, 1'_{s_0-1})'$, 0_{p-s_0} is a $p - s_0$ column vector of all zero elements, and 1_{s_0-1} is a $s_0 - 1$ dimensional column vector of all ones. The term s_0 represent the sparsity of

the above model and we set $s_0 = 5$. The error term u_i has t distribution with 5 degrees of freedom.

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

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

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

In the second setup, we choose logistical loss with weighted group lasso penalty as in (4.3) of Caner (2023). Let y_i be binary outcomes (0 or 1) and it is iid across $i = 1, \cdots, n$, and X_i is iid across i too. There are m disjoint groups, and group indices are G_j, for $j = 1, \cdots, m$. For example G_2 represent the indices that belong to second group. Also see that $\cup_{j=1}^{m} G_j = \{1, \cdots, p\}$. The cardinality of each group is represented by $|G_j|$. Let β_{G_j} represent the entries in β that correspond to G_j the group. Denote

$$\beta = (\beta_{G_1}, \cdots, \beta_{G_j}, \cdots, \beta_{G_m})',$$

which is a $p \times 1$ vector, with $\beta_{G_j} : p_j \times 1$ and $p_j = |G_j|$. So corresponding to each group there are p_j of these coefficients. Let

$$\hat{\beta}_G := (\hat{\beta}_{G_1}, \cdots, \hat{\beta}_{G_j}, \cdots, \hat{\beta}_{G_m})'.$$

Also denote X_{i,G_j} as the predictors in G_j th group at i th observation, $i = 1, \cdots, n, j =$

$1, \cdots, m$. Our norm is weighted group lasso norm

$$\Omega(\beta) = \sum_{j=1}^{m} \sqrt{|G_j|} \|\beta_{G_j}\|_2.$$

We define the estimator as

$$\hat{\beta}_G := argmin_{\beta \in \mathcal{B} \subset \mathbf{R}^p} \left[\frac{1}{n} \sum_{i=1}^{n} \left\{ -y_i \sum_{j=1}^{m} X'_{i,G_j} \beta_{G_j} + ln[1 + exp(\sum_{j=1}^{m} X'_{i,G_j} \beta_{G_j}) \right\} + 2\lambda_n \sum_{j=1}^{m} \sqrt{|G_j|} \|\beta_{G_j}\|_2. \right]$$

We use Design 1 in Supplement of Caner (2023). There are five groups, $m = 5$, and of equal size, and there is one intercept, which is not included in the groups. Our parameter vector takes the following values, with $0'_{p/5}$ representing a row vector of zeros of dimension $p/5$, if $p = 100$, the vector has 20 zeros, $0.5_{p/5}$ representing a row vector of 0.5 at each cell of dimension $p/5$,

$$\beta_0 := (\beta_{0,1} = 1, \beta_{0,G_1} = 0'_{p/5}, \beta_{0,G_2} = 0.5'_{p/5}, \beta_{0,G_3} = 0'_{p/5}, \beta_{0,G_4} = 0'_{p/5}, \beta_{0,G_5} = 0'_{p/5})'.$$

Groups 1 and 2 and Groups 4 and 5 are correlated with each other at 0.2 correlation level. Regressors in each group are correlated among themselves , across i_l, $l = 1, 2, 3, 4, 5$, the regressors are iid, with multivariate normal with zero mean, $X_{i_l} \sim N(0, \Sigma)$, where (j, k) th cell of Σ is

$$\Sigma_{j,k} = \rho^{|k-j|},$$

with $\rho = 0.75$.

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 the report is 0.2 across all attributes) respectively in the linear setup. Tables 3-4 consider the GLM Structured Sparsity for the nonlinear setup. The number of iterations is 1000.

D.1 Tuning Parameter Selection: Lasso

For Lasso, we aim to demonstrate that with a "large" tuning parameter as in Theorem 1, incentive compatibility can be achieved when the sample size n is large enough. In case s_0 is increasing in n, and $s_0 \geq 1$, the lower bound IC condition is:

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

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

$$P(\mathcal{F}^c) \leq \frac{2}{p^{C_1}} + \frac{K_1'[EM_1^2 + EM_2^2]}{nlnp} \leq \frac{2}{p^{C_1}} + \frac{C_\lambda}{(lnp)^2}, \tag{D.1}$$

by observing that for $l = 1, 2$

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

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

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

The tuning parameter is as follows

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

and to have the probability $P(\mathcal{F}^c)$ close to zero, even for $p = 2$, we need a large C_1, and a small C_λ. Note that (D.2) satisfies the lower bound condition.

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

Note that the criterion for choosing the tuning parameter should take incentive compatibility into account. Hence, we choose only C_λ with GIC, but the structure of our tuning parameter is determined by our characterization of incentive compatibility. Therefore, our choice of λ_n is *above* a lower bound, which prevents overfitting (this is the novel insight of Theorem 1). Note that in the literature for inference in high dimensional parameters, the tuning parameter is either selected by cross-validation or information criterion. Consistency of the Lasso estimators is the key in these types of selection, hence it can force the researcher to select a low value for tuning parameter. For example a standard package as glmnet in R as a default tries 100 lambda choices, from a large set of values near zero to a larger ones. But a low tuning parameter can create overfit which may violate incentive compatibility. Our choice in (D.2) prevents this problem.

Algorithm for Tuning Parameter Selection in Lasso

Step 1. Given p, C_1, C_λ we start with λ_n as in (D.2)

Step 2. In (D.2) set $C_1 = 6$, and form a grid for C_λ choices $[0.0, 0.05, 0.1, 0.2]$.

Step 3. Run lasso, $\hat{\beta}$ for each choice of λ_n and record this, and also record each sparsity estimate (no of nonzero coefficients) for each λ_n.

Step 4. By choosing C_λ, and hence forming a grid of $\Lambda = [\lambda_{n,1}, \lambda_{n,2}, \lambda_{n,3}, \lambda_{n,4}]$ corre-

sponding to each C_λ across the grid, choose the optimal tuning parameter according to Generalized information Criterion (GIC)

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

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.

D.2 Tuning Parameter Selection: Generalized Structured Sparsity Estimators

In this scenario, there are two lower bounds and we have to choose λ_n in a way that is higher than these bounds asymptotically. We can simplify these lower bounds. In one of the lower bounds we have the ratio of $s_0/g(s_0)$, as $g(s_0)$ is the signal coming from penalty, i.e. for l_1 norm, $g(s_0) = O(s_0^{1/2})$ if $\|\beta_0\|_0 = O(1)$ (bounded signal). To get that $g(s_0) = \|\beta_0\|_1 \leq \sqrt{s_0}\|\beta_0\|_2 = s_0^{1/2} O(1) = O(s_0^{1/2})$. So we can assume $s_0/g(s_0) \to \infty$ also for other possible norms with more structure than l_1 norm. This simplifies our tuning parameter choice with this mild assumption. The issue becomes what is $P(\mathcal{E}^c)$? Theorem 1(ii) in Caner (2023) shows that

$$P(\mathcal{E}^c) \leq \frac{3}{p^{2C_1}} + \frac{1}{p^{C_1}} + \frac{7}{4} \frac{C_\lambda}{(lnp)^2}.$$

Note again that we can take $C_1 = 6$, to make $P(\mathcal{E}^c)$ as small as possible. As in lasso choose $C_\lambda := [10, 5, 2.5, 1]$ which is larger than lasso constants since the nonlinear case demands larger tuning parameter compared with linear case as discussed after Theorem 2, Remarks. We also tried smaller C_λ choices, and our qualitative results stayed the same. Since $n \geq lnp$

by Assumption we can choose

$$\lambda_n^* = \left\{ n \left[\frac{3}{p^{2C_1}} + \frac{1}{p^{C_1}} + \frac{7}{4} \frac{C_\lambda}{(lnp)^2} \right] \right\}^{1/8}. \tag{D.3}$$

Note that this will satisfy the lower bound conditions in Theorem 2 given $s_0/g(s_0) \to \infty$, and since $P(\mathcal{E}^c) \to 0$.

Algorithm for Tuning Parameter Selection in GLM Structured Sparsity Estimators

Step 1. Given p, C_1, C_λ we start with λ_n as in (D.3).

Step 2. In (D.3) set $C_1 = 6$, and form a grid for C_λ choices $[10, 5, 2.5, 1]$.

Step 3. Run GLM structured sparsity estimator, $\hat{\beta}_G$ for each choice of λ_n and record this for each λ_n. Record the cardinality of active set for each $\hat{\beta}_G$.

Step 4. By choosing C_λ, and hence forming a grid of $\Lambda = [\lambda_{n,1}, \lambda_{n,2}, \lambda_{n,3}, \lambda_{n,4}]$ corresponding to each C_λ across the grid, choose the optimal tuning parameter according to Generalized information Criterion (GIC) in Caner (2023) Supplement Appendix

$$\lambda_{n,IC}^* := argmin_{\lambda_n \in \Lambda} \left[\sum_{i=1}^{n} \frac{\rho_{\lambda_n}(y_i, X_i' \hat{\beta}_G)}{n} + \frac{\hat{s}(\lambda_n)}{n} ln(n) ln(ln(p)) \right],$$

where $\hat{s}(\lambda_n)$ is the cardinality of the active set in the GLM structured sparsity estimator, given a choice of λ_n in a grid Λ, and $\rho_{\lambda_n}(y_i, X_i' \hat{\beta}_G)$ is the unpenalized GLM loss regression, given a choice of λ_n in a grid Λ.

D.3 Results

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

The numbers in each cell of the tables correspond to the disutility of the user (i.e., the mean square difference between the statistician's estimate and the optimal action). Hence, smaller numbers correspond to higher payoffs. Let us compare the tables when $p = 300$ and $n = 200$. In Table 1, which corresponds to a large magnitude of a lie, the user's disutility from reporting the truth is 0.69, while the disutility from lying is 42.16. Hence, the $n + 1$ user prefers to be truthful. In Table 2, for a small lie, truth-telling induces a disutility of 0.69, while lying induces a lower disutility of 0.23. 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 Tables 1-2. 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 similar pattern for GLM Structured Sparsity Estimation. With larger p it is more difficult catch a lie in a nonlinear system like GLM. To give an example, for GLM Structured Sparsity estimators with $p = 100$ and $n = 100$, in Table 3 the new user prefers to tell the truth with a mean squared error of 6.02 from truth compared to 62.04 from lying. However, with a larger $p = 200, 300$, it is profitable to lie.

Table 1: Lasso-Incentive Compatibility:

Difference 2	$n = 100$		$n = 200$		$n = 300$	
Dimension	Truth	Report	Truth	Report	Truth	Report
$p = 100$	0.80	38.77	0.63	41.12	0.52	42.10
$p = 200$	0.60	51.44	0.27	55.51	0.20	57.02
$p = 300$	1.22	36.54	0.69	42.16	0.59	44.01

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 errors are desired.

Table 2:Lasso-Incentive Compatibility:

Difference 0.2	$n = 100$		$n = 200$		$n = 300$	
Dimension	Truth	Report	Truth	Report	Truth	Report
$p = 100$	0.80	0.35	0.63	0.16	0.52	0.10
$p = 200$	0.60	1.58	0.27	1.21	0.20	1.17
$p = 300$	1.22	0.61	0.69	0.23	0.59	0.14

Note: See Table 1 note.

Table 3: GLM Structured Sparsity-Incentive Compatibility:

Difference 2	$n = 100$		$n = 200$		$n = 300$	
Dimension	Truth	Report	Truth	Report	Truth	Report
$p = 100$	6.02	62.04	5.57	121.12	6.06	171.42
$p = 200$	5.41	0.05	4.16	0.01	2.99	0.01
$p = 300$	7.09	0.24	3.67	0.01	3.09	0.01

Note: "Truth" refers to $E[\rho(y_{n+1}, X'_{n+1}\hat{\beta}_G) - \rho(y_{n+1}, X'_{n+1}\beta_0)]^2$, and "Report" refers to $E[\rho(y_{n+1}, R(X_{n+1})'\hat{\beta}_G) - \rho(y_{n+1}, X'_{n+1}\beta_0)]^2$ in Incentive Compatibility Definition.

Table 4: GLM Structured Sparsity-Incentive Compatibility:

Difference 0.2	$n = 100$		$n = 200$		$n = 300$	
Dimension	Truth	Report	Truth	Report	Truth	Report
$p = 100$	6.02	8.12	5.57	8.48	6.06	9.88
$p = 200$	5.41	3.78	4.16	2.20	2.99	1.35
$p = 300$	7.09	5.24	3.67	1.92	3.95	1.26

Note: See Table 3 note.

References

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

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

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

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

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.

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

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