Single Point Transductive Prediction

Nilesh Tripuraneni [1] **Lester Mackey** [2]

Abstract

Standard methods in supervised learning separate training and prediction: the model is fit independently of any test points it may encounter. However, can knowledge of the next test point $\mathbf{x}_\star$ be exploited to improve prediction accuracy? We address this question in the context of linear prediction, showing how techniques from semiparametric inference can be used transductively to combat regularization bias. We first lower bound the $\mathbf{x}_\star$ prediction error of ridge regression and the Lasso, showing that they must incur significant bias in certain test directions. We then provide non-asymptotic upper bounds on the $\mathbf{x}_\star$ prediction error of two transductive prediction rules. We conclude by showing the efficacy of our methods on both synthetic and real data, highlighting the improvements single point transductive prediction can provide in settings with distribution shift.

1. Introduction

We consider the task of prediction given independent datapoints $((y_i, \mathbf{x}_i))_{i=1}^n$ from a linear model,

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta}_0 + \epsilon_i, \quad \mathbb{E}[\epsilon_i] = 0, \quad \epsilon_i \perp\!\!\!\perp \mathbf{x}_i \qquad (1)$$

in which our observed targets $\mathbf{y} = (y_1, \dots, y_n) \in \mathbb{R}^n$ and covariates $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^\top \in \mathbb{R}^{n \times p}$ are related by an unobserved parameter vector $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ and noise vector $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n) \in \mathbb{R}^n$.

Most approaches to linear model prediction are *inductive*, divorcing the steps of training and prediction; for example, regularized least squares methods like ridge regression (Hoerl & Kennard, 1970) and the Lasso (Tibshirani, 1996) are fit independently of any knowledge of the next target test point $\mathbf{x}_\star$. This suggests a tantalizing *transductive* question:

[1]Department of EECS, University of California, Berkeley [2]Microsoft Research New England, Cambridge, MA, USA. Correspondence to: Nilesh Tripuraneni <nilesh_tripuraneni@berkeley.edu>.

can knowledge of a single test point $\mathbf{x}_\star$ be leveraged to improve prediction for $\mathbf{x}_\star$? In the random design linear model setting (1), we answer this question in the affirmative.

Specifically, in Section 2 we establish *out-of-sample* prediction lower bounds for the popular ridge and Lasso estimators, highlighting the significant dimension-dependent bias introduced by regularization. In Section 3 we demonstrate how this bias can be mitigated by presenting two classes of transductive estimators that exploit explicit knowledge of the test point $\mathbf{x}_\star$. We provide non-asymptotic risk bounds for these estimators in the random design setting, proving that they achieve dimension-free $O(\frac{1}{n})$ $\mathbf{x}_\star$-prediction risk for n sufficiently large. In Section 4, we first validate our theory in simulation, demonstrating that transduction improves the prediction accuracy of the Lasso with fixed regularization even when $\mathbf{x}_\star$ is drawn from the training distribution. We then demonstrate that under distribution shift, our transductive methods outperform even the popular cross-validated Lasso, cross-validated ridge, and cross-validated elastic net estimators (which attempt to find an optimal data-dependent trade-off between bias and variance) on both synthetic data and a suite of five real datasets.

1.1. Related Work

Our work is inspired by two approaches to semiparametric inference: the debiased Lasso approach introduced by (Zhang & Zhang, 2014; Van de Geer et al., 2014; Javanmard & Montanari, 2014) and the orthogonal machine learning approach of (Chernozhukov et al., 2017). The works (Zhang & Zhang, 2014; Van de Geer et al., 2014; Javanmard & Montanari, 2014) obtain small-width and asymptotically-valid confidence intervals (CIs) for individual model parameters $(\boldsymbol{\beta}_0)_j = \langle \boldsymbol{\beta}_0, \mathbf{e}_j \rangle$ by debiasing an initial Lasso estimator (Tibshirani, 1996). The works (Chao et al., 2014; Cai & Guo, 2017; Athey et al., 2018) each consider a more closely related problem of obtaining prediction confidence intervals using a generalization of the debiased Lasso estimator of (Javanmard & Montanari, 2014). The work of Chernozhukov et al. (2017) describes a general-purpose procedure for extracting $\sqrt{n}$-consistent and asymptotically normal target parameter estimates in the presence of nuisance parameters. Specifically, Chernozhukov et al. (2017) construct a two-stage estimator where one initially fits first-stage estimates of nuisance parameters using arbitrary ML

estimators on a first-stage data sample. In the second-stage, these first-stage estimators are used to provide estimates of the relevant model parameters using an orthogonalized method-of-moments. Wager et al. (2016) also uses generic ML procedures as regression adjustments to form efficient confidence intervals for treatment effects.

These pioneering works all focus on improved confidence interval construction. Here we show that the semiparametric techniques developed for hypothesis testing can be adapted to provide practical improvements in mean-squared prediction error. Our resulting mean-squared error bounds complement the in-probability bounds of the aforementioned literature by controlling prediction performance across all events.

Our approach to prediction also bears some resemblance to semi-supervised learning (SSL) – transferring predictive power between labelled and unlabelled examples (Zhu, 2005). In contrast with SSL, the goal of transductive learning is to predict solely the labels of the observed unlabeled features. Alquier & Hebiri (2012) formulate a transductive version of the Lasso and Dantzig selector estimators in the fixed design setting focused only on predicting a subset of points. Bellec et al. (2018) also prove risk bounds for transductive and semi-supervised ℓ_1-regularized estimators in the high-dimensional setting. A principal difference between our approaches is that we make no distributional assumptions about the sequence of test points $\mathbf{x}_\star$ and do not assume simultaneous access to a large pool of test data. Rather our procedures receive access to only a single arbitrary test point $\mathbf{x}_\star$, and our aim is accurate prediction for that point. Conventionally, SSL benefits from access to a large pool of test points; we are unaware of other results that benefit from access to single test point $\mathbf{x}_\star$. Moreover, existing SSL methods in the sparse linear regression setting do not remove regularization bias to achieve dimension-independent $O(\frac{1}{n})$ rates but rather provide dimension-dependent bounds (Bellec et al., 2018).

1.2. Problem Setup

Our principal aim in this work is to understand the $\mathbf{x}_\star$ *prediction risk*,

$$\mathcal{R}(\mathbf{x}_\star, \hat{y}) = \mathbb{E}[(y_\star - \hat{y})^2] - \sigma_\epsilon^2 = \mathbb{E}[(\hat{y} - \langle \mathbf{x}_\star, \boldsymbol{\beta}_0 \rangle)^2] \quad (2)$$

of an estimator $\hat{y}$ of the unobserved test response $y_\star = \mathbf{x}_\star^\top \boldsymbol{\beta}_0 + \epsilon_\star$. Here, $\epsilon_\star$ is independent of $\mathbf{x}_\star$ with variance σ_ϵ^2. We exclude the additive noise σ_ϵ^2 from our risk definition, as it is irreducible for any estimator. Importantly, to accommodate non-stationary learning settings, we consider $\mathbf{x}_\star$ to be fixed and arbitrary; in particular, $\mathbf{x}_\star$ need not be drawn from the training distribution. Hereafter, we will make use of several assumptions which are standard in the random design linear regression literature.

Assumption 1 (Well-specified Model). *The data* $(\mathbf{X}, \mathbf{y})$ *is generated from the model* (1).

Assumption 2 (Bounded Covariance). *The covariate vectors have common covariance* $\boldsymbol{\Sigma} = \mathbb{E}[\mathbf{x}_i \mathbf{x}_i^\top]$ *with* $\boldsymbol{\Sigma}_{ii} \leq 1/2$, $\sigma_{\max}(\boldsymbol{\Sigma}) \leq C_{\max}$ *and* $\sigma_{\min}(\boldsymbol{\Sigma}) \geq C_{\min}$. *We further define the precision matrix* $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ *and condition number* $C_{cond} = C_{\max}/C_{\min}$.

Assumption 3 (Sub-Gaussian Design). *Each covariate vector* $\boldsymbol{\Sigma}^{-1/2}\mathbf{x}_i$ *is sub-Gaussian with parameter* $\kappa \geq 1$, *in the sense that,* $\mathbb{E}[\exp(\mathbf{v}^\top \mathbf{x}_i)] \leq \exp(\kappa^2 \|\boldsymbol{\Sigma}^{1/2}\mathbf{v}\|^2/2)$.

Assumption 4 (Sub-Gaussian Noise). *The noise* ϵ_i *is sub-Gaussian with variance parameter* σ_ϵ^2.

Throughout, we will use bold lower-case letters (e.g., $\mathbf{x}$) to refer to vectors and bold upper-case letters to refer to matrices (e.g., $\mathbf{X}$). We use $[p]$ for the set $\{1, \ldots, p\}$. Vectors or matrices subscripted with an index set S indicate the subvector or submatrix supported on S. The expression $s_{\boldsymbol{\beta}_0}$ indicates the number of non-zero elements in $\boldsymbol{\beta}_0$ and $\text{supp}(\boldsymbol{\beta}_0) = \{j : (\boldsymbol{\beta}_0)_j \neq 0\}$. We will use $\gtrsim, \lesssim$, and $\asymp$ to denote greater than, less than, and equal to up to a constant that is independent of p and n.

2. Lower Bounds for Regularized Prediction

We begin by providing lower bounds on the $\mathbf{x}_\star$ prediction risk of Lasso and ridge regression; the corresponding predictions take the form $\hat{y} = \langle \mathbf{x}_\star, \hat{\boldsymbol{\beta}} \rangle$ for a regularized estimate $\hat{\boldsymbol{\beta}}$ of the unknown vector $\boldsymbol{\beta}_0$.

2.1. Lower Bounds for Ridge Regression Prediction

We first consider the $\mathbf{x}_\star$ prediction risk of the ridge estimator $\hat{\boldsymbol{\beta}}_R(\lambda) \triangleq \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2$ with regularization parameter $\lambda > 0$. In the asymptotic high-dimensional limit (with $n, p \to \infty$) and assuming the training distribution equals the test distribution, Dobriban et al. (2018) compute the predictive risk of the ridge estimator in a dense random effects model. By contrast, we provide a non-asymptotic lower bound which does not impose any distributional assumptions on $\mathbf{x}_\star$ or on the underlying parameter vector $\boldsymbol{\beta}_0$. Theorem 1, proved in Appendix B.1, isolates the error in the ridge estimator due to bias for any choice of regularizer λ.

Theorem 1. *Under Assumption 1, suppose* $\mathbf{x}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \mathbf{I}_p)$ *with independent noise* $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}_n \sigma_\epsilon^2)$. *If* $n \geq p \geq 20$,

$$\mathbb{E}[\langle \mathbf{x}_\star, \hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0 \rangle^2] \geq$$
$$\frac{\|\boldsymbol{\beta}_0\|_2^2}{\sigma_\epsilon^2} \cdot \frac{n}{4} \left(\frac{\lambda/n}{\lambda/n+7}\right)^2 \cdot \|\mathbf{x}_\star\|_2^2 \cdot \frac{\sigma_\epsilon^2}{n} \cdot \cos(\mathbf{x}_\star, \boldsymbol{\beta}_0)^2.$$

Notably, the dimension-free term $\|\mathbf{x}_\star\|_2^2 \cdot \frac{\sigma_\epsilon^2}{n}$ in this bound coincides with the $\mathbf{x}_\star$ risk of the ordinary least squares (OLS)

estimator in this setting. The remaining multiplicative factor indicates that the ridge risk can be substantially larger if the regularization strength λ is too large. In fact, our next result shows that, surprisingly, over-regularization can result even when λ is tuned to minimize held-out prediction error over the training population. The same undesirable outcome results when λ is selected to minimize ℓ_2 estimation error; the proof can be found in Appendix B.2.

Corollary 1. *Under the conditions of Theorem 1, if $\tilde{\mathbf{x}} \overset{d}{=} \mathbf{x}_1$ and $\tilde{\mathbf{x}}$ is independent of $(\mathbf{X}, \mathbf{y})$, then for* $\mathrm{SNR} \triangleq \|\boldsymbol{\beta}_0\|_2^2 / \sigma_\epsilon^2$,

$$\lambda_* \triangleq \operatorname{argmin}_\lambda \mathbb{E}[\langle \tilde{\mathbf{x}}, \hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0 \rangle^2] =$$

$$\operatorname{argmin}_\lambda \mathbb{E}[\|\hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\|_2^2] = \frac{p}{\mathrm{SNR}}, \text{ and, for } n \geq \frac{1}{6}\frac{p}{\mathrm{SNR}},$$

$$\mathbb{E}[\langle \mathbf{x}_\star, \hat{\boldsymbol{\beta}}_R(\lambda_*) - \boldsymbol{\beta}_0 \rangle^2] \geq \frac{p^2}{n\,\mathrm{SNR}} \cdot \|\mathbf{x}_\star\|_2^2 \cdot \frac{\sigma_\epsilon^2}{n} \cdot \frac{\cos(\mathbf{x}_\star, \boldsymbol{\beta}_0)^2}{784}.$$

Several insights can be gathered from the previous results. First, the expression $\mathbb{E}[\langle \tilde{\mathbf{x}}, \hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0 \rangle^2]$ minimized in Corollary 1 is the expected prediction risk $\mathbb{E}[(\tilde{y} - \tilde{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}_R(\lambda))^2] - \sigma_\epsilon^2$ for a new datapoint $(\tilde{\mathbf{x}}, \tilde{y})$ drawn from the training distribution. This is the population analog of held-out validation error or cross-validation error that is often minimized to select λ in practice. Second, in the setting of Corollary 1, taking $\mathrm{SNR} = \frac{1}{6}\frac{p}{n}$ yields

$$\mathbb{E}[\langle \mathbf{x}_\star, \hat{\boldsymbol{\beta}}_R(\lambda_*) - \boldsymbol{\beta}_0 \rangle^2] \geq p \cdot \|\mathbf{x}_\star\|_2^2 \cdot \frac{\sigma_\epsilon^2}{n} \cdot \frac{3\cos(\mathbf{x}_\star, \boldsymbol{\beta}_0)^2}{392}.$$

More generally, if we take $\cos(\mathbf{x}_\star, \boldsymbol{\beta}_0)^2 = \Theta(1)$, $\mathrm{SNR} = o(\frac{p^2}{n})$ and $\mathrm{SNR} \geq \frac{1}{6}\frac{p}{n}$ then,

$$\mathbb{E}[\langle \mathbf{x}_\star, \hat{\boldsymbol{\beta}}_R(\lambda_*) - \boldsymbol{\beta}_0 \rangle^2] \geq \omega(\|\mathbf{x}_\star\|_2^2 \cdot \frac{\sigma_\epsilon^2}{n}).$$

If λ is optimized for estimation error or for prediction error with respect to the training distribution, the ridge estimator must incur much larger test error then the OLS estimator in some test directions. Such behavior can be viewed as a symptom of over-regularization – the choice λ_* is optimized for the training distribution and cannot be targeted to provide uniformly good performance over all $\mathbf{x}_\star$. In Section 3 we show how transductive techniques can improve prediction in this regime.

The chief difficulty in lower-bounding the $\mathbf{x}_\star$ prediction risk in Theorem 1 lies in controlling the expectation over the design $\mathbf{X}$, which enters nonlinearly into the prediction risk. Our proof circumvents this difficulty in two steps. First, the isotropy and independence properties of Wishart matrices are used to reduce the computation to that of a 1-dimensional expectation with respect to the unordered eigenvalues of $\mathbf{X}$. Second, in the regime $n \geq p$, the sharp concentration of Gaussian random matrices in spectral norm is exploited to essentially approximate $\frac{1}{n}\mathbf{X}^\top\mathbf{X} \approx \mathbf{I}_p$.

2.2. Lower Bounds for Lasso Prediction

We next provide a strong lower bound on the out-of-sample prediction error of the Lasso estimator $\hat{\boldsymbol{\beta}}_L(\lambda) \triangleq \operatorname{argmin}_{\boldsymbol{\beta}} \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$ with regularization parameter $\lambda > 0$. There has been extensive work (see, e.g., Raskutti et al., 2011) establishing minimax lower bounds for the in-sample prediction error and parameter estimation error of any procedure given data from a sparse linear model. However, our focus is on out-of-sample prediction risk for a specific procedure, the Lasso. The point $\mathbf{x}_\star$ need not be one of the training points (in-sample) nor even be drawn from the same distribution as the covariates. Theorem 2, proved in Appendix C.1, establishes that a well-regularized Lasso program suffers significant biases even in a simple problem setting with i.i.d. Gaussian covariates and noise.[1]

Theorem 2. *Under Assumption 1, fix any $s \geq 0$, and let $\mathbf{x}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \mathbf{I}_p)$ with independent noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}_n\sigma_\epsilon^2)$. If $\lambda \geq (8 + 2\sqrt{2})\sigma_\epsilon\sqrt{\log(2ep)/n}$ and $p \geq 20$, then there exist universal constants c_1, c_2, c_3 such that for all $n \geq c_1 s^2 \log(2ep)$,*

$$c_3\lambda^2\|\mathbf{x}_\star\|_{(s)}^2 \geq \sup_{\boldsymbol{\beta}_0 \in \mathbb{B}_0(s)} \mathbb{E}[\langle \mathbf{x}_\star, \hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0 \rangle^2]$$

$$\geq \sup_{\boldsymbol{\beta}_0 \in \mathbb{B}_0(s), \|\boldsymbol{\beta}_0\|_\infty \leq \lambda} \mathbb{E}[\langle \mathbf{x}_\star, \hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0 \rangle^2] \geq c_2\lambda^2\|\mathbf{x}_\star\|_{(s)}^2$$

where the trimmed norm $\|\mathbf{x}_\star\|_{(s)}$ is the sum of the magnitudes of the s largest magnitude entries of $\mathbf{x}_\star$.

In practice we will always be interested in a known $\mathbf{x}_\star$ direction, but the next result clarifies the dependence of our Lasso lower bound on sparsity for worst-case test directions $\mathbf{x}_\star$ (see Appendix C.2 for the proof):

Corollary 2. *In the setting of Theorem 2, for $q \in [1, \infty]$,*

$$\sup_{\|\mathbf{x}_\star\|_q = 1} \sup_{\boldsymbol{\beta}_0 \in \mathbb{B}_0(s)} \mathbb{E}[\langle \mathbf{x}_\star, \hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0 \rangle^2] \geq c_2\lambda^2 s^{2-2/q}.$$

We make several comments regarding these results. First, together Theorem 2 yields an $\mathbf{x}_\star$-specific lower bound – showing that given any potential direction $\mathbf{x}_\star$ there will exist an underlying s-sparse parameter $\boldsymbol{\beta}_0$ for which the Lasso performs poorly. Morever, the magnitude of error suffered by the Lasso scales both with the regularization strength λ and the norm of $\mathbf{x}_\star$ along its top s coordinates. Second, the constraint on the regularization parameter in Theorem 2, $\lambda \gtrsim \sigma_\epsilon\sqrt{\log p/n}$, is a sufficient and standard choice to obtain consistent estimates with the Lasso (see Wainwright, Ch. 7 for example). Third, simplifying to the case of $q = 2$, we see that Corollary 2 implies the Lasso must incur worst-case $\mathbf{x}_\star$ prediction error $\gtrsim \frac{\sigma_\epsilon^2 s \log p}{n}$, matching upper bounds

---

[1] A yet tighter lower bound is available if, instead of being fixed, $\mathbf{x}_\star$ follows an arbitrary distribution, and the expectation is taken over $\mathbf{x}_\star$ as well. See the proof for details.

for Lasso prediction error. In particular such a bound is not dimension-free, possessing a dependence on $s \log p$, even though the Lasso is only required to predict well along a *single* direction.

The proof of Theorem 2 uses two key ideas. First, in this benign setting, we can show that $\hat{\boldsymbol{\beta}}_L(\lambda)$ has support strictly contained in the support of $\boldsymbol{\beta}_0$ with at least constant probability. We then adapt ideas from the study of debiased lasso estimation in (Javanmard & Montanari, 2014) to sharply characterize the coordinate-wise bias of the Lasso estimator along the support of $\boldsymbol{\beta}_0$; in particular we show that a worst-case $\boldsymbol{\beta}_0$ can match the signs of the s largest elements of $\mathbf{x}_\star$ and have magnitude λ on each non-zero coordinate. Thus the bias induced by regularization can coherently sum across the s coordinates in the support of $\boldsymbol{\beta}_0$. A similar lower bound follows by choosing $\boldsymbol{\beta}_0$ to match the signs of $\mathbf{x}_\star$ on any subset of size s. This sign alignment between $\mathbf{x}_\star$ and $\boldsymbol{\beta}_0$ is also explored in the independent and concurrent work of (Bellec & Zhang, 2019, Thm. 2.2).

3. Upper Bounds for Transductive Prediction

Having established that regularization can lead to excessive prediction bias, we now introduce two classes of estimators which can mitigate this bias using knowledge of the single test direction $\mathbf{x}_\star$. While our presentation focuses on the prediction risk (2), which features an expectation over $\hat{y}$, our proofs in the appendix also provide identical high probability upper bounds on $(\hat{y} - \langle \mathbf{x}_\star, \boldsymbol{\beta}_0 \rangle)^2$.

3.1. Javanmard-Montanari (JM)-style Estimator

Our first approach to single point transductive prediction is inspired by the debiased Lasso estimator of Javanmard & Montanari (2014) which was to designed to construct confidence intervals for individual model parameters $(\boldsymbol{\beta}_0)_j$. For prediction in the $\mathbf{x}_\star$ direction, we will consider the following generalization of the Javanmard-Montanari (JM) debiasing construction[2]:

$$\hat{y}_{\text{JM}} = \langle \mathbf{x}_\star, \hat{\boldsymbol{\beta}} \rangle + \frac{1}{n} \mathbf{w}^\top \mathbf{X}^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \quad \text{for} \quad (3)$$

$$\mathbf{w} = \operatorname{argmin}_{\tilde{\mathbf{w}}} \tilde{\mathbf{w}}^\top \boldsymbol{\Sigma}_n \tilde{\mathbf{w}} \text{ s.t. } \|\boldsymbol{\Sigma}_n \tilde{\mathbf{w}} - \mathbf{x}_\star\|_\infty \leq \lambda_{\mathbf{w}}. \quad (4)$$

Here, $\hat{\boldsymbol{\beta}}$ is any (ideally ℓ_1-consistent) initial pilot estimate of $\boldsymbol{\beta}_0$, like the estimate $\hat{\boldsymbol{\beta}}_L(\lambda)$ returned by the Lasso. When $\mathbf{x}_\star = \mathbf{e}_j$ the estimator (3) reduces exactly to the program in (Javanmard & Montanari, 2014), and equivalent generalizations have been used in (Chao et al., 2014; Athey et al., 2018; Cai & Guo, 2017) to construct prediction intervals and to estimate treatment effects. Intuitively, $\mathbf{w}$ approximately inverts the population covariance matrix along the direction defined by $\mathbf{x}_\star$ (i.e., $\mathbf{w} \approx \boldsymbol{\Omega} \mathbf{x}_\star$). The second term

---

[2]In the event the constraints are not feasible we define $\mathbf{w} = 0$.

in (3) can be thought of as a high-dimensional one-step correction designed to remove bias from the initial prediction $\langle \mathbf{x}_\star, \hat{\boldsymbol{\beta}} \rangle$; see (Javanmard & Montanari, 2014) for more intuition on this construction. We can now state our primary guarantee for the *JM-style estimator* (3); the proof is given in Appendix D.1.

Theorem 3. *Suppose Assumptions 1, 2, 3 and 4 hold and that the transductive estimator $\hat{y}_{\text{JM}}$ of (3) is fit with regularization parameter $\lambda_{\mathbf{w}} = 8a\sqrt{C_{cond}}\kappa^2 \|\mathbf{x}_\star\|_2 \sqrt{\frac{\log(p \vee n)}{n}}$ for some $a > 0$. Then there is a universal constant c_1 such that if $n \geq c_1 a^2 \log(2e(p \vee n))$,*

$$\mathbb{E}[(\hat{y}_{\text{JM}} - \langle \boldsymbol{\beta}_0, \mathbf{x}_\star \rangle)^2] \leq \quad (5)$$
$$O\left(\frac{\sigma_\epsilon^2 \mathbf{x}_\star \boldsymbol{\Omega} \mathbf{x}_\star}{n} + r_{\boldsymbol{\beta},1}^2 (\lambda_{\mathbf{w}}^2 + \|\mathbf{x}_\star\|_\infty^2 \frac{1}{(n \vee p)^{c_3}}) \right).$$

for $c_3 = \frac{a^2}{4} - \frac{1}{2}$ and $r_{\boldsymbol{\beta},1} = (\mathbb{E}[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1^4])^{1/4}$, the ℓ_1 error of the initial estimate. Moreover, if $\lambda_{\mathbf{w}} \geq \|\mathbf{x}_\star\|_\infty$, then $\mathbb{E}[(\hat{y}_{\text{JM}} - \langle \boldsymbol{\beta}_0, \mathbf{x}_\star \rangle)^2] = \mathbb{E}[\langle \mathbf{x}_\star, \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_0 \rangle^2]$. Here the $O(\cdot)$ masks constants depending only on $\kappa, C_{\min}, C_{\max}, C_{cond}$.

Intuitively, the first term in our bound (5) can be viewed as the variance of the estimator's prediction along the direction of $\mathbf{x}_\star$ while the second term can be thought of as the (reduced) bias of the estimator. We consider the third term to be of higher order since a (and in turn c_3) can be chosen as a large constant. Finally, when $\lambda_{\mathbf{w}} \geq \|\mathbf{x}_\star\|_\infty$ the error of the transductive procedure reduces to that of the pilot regression procedure. When the Lasso is used as the pilot regression procedure we can derive the following corollary to Theorem 3, also proved in Appendix D.2.

Corollary 3. *Recall $s_{\boldsymbol{\beta}_0} = \|\boldsymbol{\beta}_0\|_0$. Under the conditions of Theorem 3, consider the JM-style estimator (3) with pilot estimate $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_L(\lambda)$ with $\lambda \geq 80\sigma_\epsilon \sqrt{\frac{\log(2ep/s_{\boldsymbol{\beta}_0})}{n}}$. If $p \geq 20$, then there exist universal constants c_1, c_2 such that if $\|\boldsymbol{\beta}_0\|_\infty / \sigma_\epsilon = o(e^{c_1 n})$ and $n \geq c_2 \max\{\frac{s_{\boldsymbol{\beta}_0} \kappa^4}{C_{\min}}, a^2\} \log(2e(p \vee n))$,*

$$\mathbb{E}[(\hat{y}_{\text{JM}} - \langle \boldsymbol{\beta}_0, \mathbf{x}_\star \rangle)^2] \leq O\left(\frac{\sigma_\epsilon^2 \mathbf{x}_\star \boldsymbol{\Omega} \mathbf{x}_\star}{n} \right.$$
$$\left. + \lambda^2 s_{\boldsymbol{\beta}_0}^2 (\lambda_{\mathbf{w}}^2 + \|\mathbf{x}_\star\|_\infty^2 \frac{1}{(n \vee p)^{c_3}}) \right)$$

Here the $O(\cdot)$ masks constants depending only on $\kappa, C_{\min}, C_{\max}, C_{cond}$.

We make several remarks to further interpret this result. First, to simplify the presentation of the results (and match the lower bound setting of Theorem 2) consider the setting in Corollary 3 with $a \asymp 1$, $\lambda \asymp \sigma_\epsilon \sqrt{\log p / n}$, and $n \gtrsim s_{\boldsymbol{\beta}_0}^2 \log p \log(p \vee n)$. Then the upper bound in Theorem 3 can be succinctly stated as $O(\frac{\sigma_\epsilon^2 \|\mathbf{x}_\star\|_2^2}{n})$. In short, the transductive estimator attains a dimension-free rate for sufficiently large n. Under the same conditions the Lasso estima-

tor suffers a prediction error of $\Omega(\|\mathbf{x}_\star\|_{(s)}^2 \frac{\sigma_\epsilon^2 \log p}{n})$ as Theorem 2 and Corollary 2 establish. Thus transduction guarantees improvement over the Lasso lower bound whenever $\mathbf{x}_\star$ satisfies the soft sparsity condition $\frac{\|\mathbf{x}_\star\|_2}{\|\mathbf{x}_\star\|_{(s)}} \lesssim \sqrt{\log p}$. Since $\mathbf{x}_\star$ is observable, one can selectively deploy transduction based on the soft sparsity level $\frac{\|\mathbf{x}_\star\|_2}{\|\mathbf{x}_\star\|_{(s)}}$ or on bounds thereof. Second, the estimator described in (4) and (3) is transductive in that it is tailored to an individual test-point $\mathbf{x}_\star$. The corresponding guarantees in Theorem 3 and Corollary 3 embody a computational-statistical tradeoff. In our setting, the detrimental effects of regularization can be mitigated at the cost of extra computation: the convex program in (4) must be solved for each new $\mathbf{x}_\star$. Third, the condition $\|\boldsymbol{\beta}_0\|_\infty / \sigma_\epsilon = o(e^{c_1 n})$ is not used for our high-probability error bound and is only used to control prediction risk (2) on the low-probability event that the (random) design matrix $\mathbf{X}$ does not satisfy a restricted eigenvalue-like condition. For comparison, note that our Theorem 2 lower bound establishes substantial excess Lasso bias even when $\|\boldsymbol{\beta}_0\|_\infty = \lambda = o(1)$.

Finally, we highlight that Cai & Guo (2017) have shown that the JM-style estimator with a scaled lasso base procedure and $\lambda_\mathbf{w} \asymp \sqrt{\frac{\log p}{n}}$ produce CIs for $\mathbf{x}_\star^\top \boldsymbol{\beta}_0$ with minimax rate optimal length when $\mathbf{x}_\star$ is sparsely loaded. Although our primary focus is in improving the mean-square prediction risk (2), we conclude this section by showing that a different setting of $\lambda_\mathbf{w}$ yields minimax rate optimal CIs for dense $\mathbf{x}_\star$ and simultaneously minimax rate optimal CIs for sparse and dense $\mathbf{x}_\star$ when $\boldsymbol{\beta}_0$ is sufficiently sparse.

Proposition 4. *Under the conditions of Theorem 3 with $\sigma_\epsilon = 1$, consider the JM-style estimator (3) with pilot estimate $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_L(\lambda)$ and $\lambda = 80\sqrt{\frac{\log(2p)}{n}}$. Fix any $C_1, C_2, C_3 > 0$, and instate the assumptions of Cai & Guo (2017), namely that the vector $\mathbf{x}_\star$ satisfies $\frac{\max_j |(\mathbf{x}_\star)_j|}{\min_j |(\mathbf{x}_\star)_j|} \leq C_1$ and $s_{\boldsymbol{\beta}_0} \asymp p^\gamma$ for $0 \leq \gamma < \frac{1}{2}$. Then for $n \gtrsim s_{\boldsymbol{\beta}_0} \log p$ the estimator $\hat{y}_{\mathrm{JM}}$ (3) with $\lambda_\mathbf{w} = 8\sqrt{C_{cond}}\kappa^2 \frac{1}{s_{\boldsymbol{\beta}_0} \sqrt{\log p}}\|\mathbf{x}_\star\|_2$ yields (minimax rate optimal) $1 - \alpha$ confidence intervals for $\mathbf{x}_\star^\top \boldsymbol{\beta}_0$ of expected length*

- *$O(\|\mathbf{x}_\star\|_\infty \cdot s_{\boldsymbol{\beta}_0} \sqrt{\frac{\log p}{n}})$ in the dense $\mathbf{x}_\star$ regime where $\|\mathbf{x}_\star\|_0 = C_3 p^{\gamma_q}$ with $2\gamma < \gamma_q < 1$ (matching the result of (Cai & Guo, 2017, Thm. 4)).*
- *$O(\|\mathbf{x}_\star\|_2 \cdot \frac{1}{\sqrt{n}})$ in the sparse $\mathbf{x}_\star$ regime of (Cai & Guo, 2017, Thm. 1) where $\|\mathbf{x}_\star\|_0 \leq C_2 s_{\boldsymbol{\beta}_0}$ if $n \gtrsim s_{\boldsymbol{\beta}_0}^2 (\log p)^2$.*

Here the $O(\cdot)$ masks constants depending only on $\kappa, C_1, C_2, C_3, C_{\min}, C_{\max}, C_{cond}$.

3.2. Orthogonal Moment (OM) Estimators

Our second approach to single point transductive prediction is inspired by orthogonal moment (OM) estimation (Chernozhukov et al., 2017). OM estimators are commonly used

to estimate single parameters of interest (like a treatment effect) in the presence of high-dimensional or nonparametric nuisance. To connect our problem to this semiparametric world, we first frame the task of prediction in the $\mathbf{x}_\star$ direction as one of estimating a single parameter, $\theta_0 = \mathbf{x}_\star^\top \boldsymbol{\beta}_0$. Consider the linear model equation (1)

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta}_0 + \epsilon_i = ((\mathbf{U}^{-1})^\top \mathbf{x}_i)^\top \mathbf{U}\boldsymbol{\beta}_0 + \epsilon_i$$

with a data reparametrization defined by the matrix $\mathbf{U} = \|\mathbf{x}_\star\|_2 \cdot \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{R} \end{bmatrix}$ for $\frac{\mathbf{x}_\star}{\|\mathbf{x}_\star\|_2} = \mathbf{u}_1$ so that $\mathbf{e}_1^\top \mathbf{U}\boldsymbol{\beta}_0 = \mathbf{x}_\star^\top \boldsymbol{\beta}_0 = \theta_0$. Here, the matrix $\mathbf{R} \in \mathbb{R}^{(p-1)\times p}$ has orthonormal rows which span the subspace orthogonal to $\mathbf{u}_1$ – these are obtained as the non-$\mathbf{u}_1$ eigenvectors of the projector matrix $\mathbf{I}_p - \mathbf{u}_1 \mathbf{u}_1^\top$. This induces the data reparametrization $\mathbf{x}' = [t, \mathbf{z}] = (\mathbf{U}^{-1})^\top \mathbf{x}$. In the reparametrized basis, the linear model becomes,

$$y_i = \theta_0 t_i + \mathbf{z}_i^\top \mathbf{f}_0 + \epsilon_i, \qquad t_i = \mathbf{g}_0(\mathbf{z}_i) + \eta_i,$$
$$\mathbf{q}_0(\mathbf{z}_i) \triangleq \theta_0 \mathbf{g}_0(\mathbf{z}_i) + \mathbf{z}_i^\top \mathbf{f}_0 \tag{6}$$

where we have introduced convenient auxiliary equations in terms of $\mathbf{g}_0(\mathbf{z}_i) \triangleq \mathbb{E}[t_i \mid \mathbf{z}_i]$.

To estimate $\theta_0 = \mathbf{x}_\star^\top \boldsymbol{\beta}_0$ in the presence of the unknown nuisance parameters $\mathbf{f}_0, \mathbf{g}_0, \mathbf{q}_0$, we introduce a thresholded-variant of the two-stage method of moments estimator proposed in (Chernozhukov et al., 2017). The method of moments takes as input a moment function m of both data and parameters that uniquely identifies the target parameter of interest. Our reparameterized model form (6) gives us access to two different *Neyman orthogonal* moment functions described (Chernozhukov et al., 2017):

f moments: $m(t_i, y_i, \theta, \mathbf{z}_i^\top \mathbf{f}, \mathbf{g}(\mathbf{z}_i)) =$
$$(y_i - t_i\theta - \mathbf{z}_i^\top \mathbf{f})(t_i - \mathbf{g}(\mathbf{z}_i)) \tag{7}$$
q moments: $m(t_i, y_i, \theta, \mathbf{q}(\mathbf{z}_i), \mathbf{g}(\mathbf{z}_i)) =$
$$(y_i - \mathbf{q}(\mathbf{z}_i) - \theta(t_i - \mathbf{g}(\mathbf{z}_i)))(t_i - \mathbf{g}(\mathbf{z}_i)).$$

These orthogonal moment equations enable the accurate estimation of a target parameter θ_0 in the presence of high-dimensional or nonparametric nuisance parameters (in this case $\mathbf{f}_0$ and $\mathbf{g}_0$). We focus our theoretical analysis and present description on the set of **f** moments since the analysis is similar for the **q**, although we investigate the practical utility of both in Section 4.

Our OM proposal to estimate θ_0 now proceeds as follows. We first split our original dataset of n points into two[3] disjoint, equal-sized folds $(\mathbf{X}^{(1)}, \mathbf{y}^{(1)}) = \{(\mathbf{x}_i, y_i) : i \in \{1, \ldots, \frac{n}{2}\}\}$ and $(\mathbf{X}^{(2)}, \mathbf{y}^{(2)}) = \{(\mathbf{x}_i, y_i) : i \in \{\frac{n}{2} + 1, \ldots, n\}\}$. Then,

---

[3]In practice, we use K-fold cross-fitting to increase the sample-efficiency of the scheme as in (Chernozhukov et al., 2017); for simplicity of presentation, we defer the description of this slight modification to Appendix G.4.

- The first fold $(\mathbf{X}^{(1)}, \mathbf{y}^{(1)})$ is used to run two *first-stage* regressions. We estimate $\boldsymbol{\beta}_0$ by linearly regressing $\mathbf{y}^{(1)}$ onto $\mathbf{X}^{(1)}$ to produce $\hat{\boldsymbol{\beta}}$; this provides an estimator of $\mathbf{f}_0$ as $\mathbf{e}_{-1}^\top \mathbf{U} \hat{\boldsymbol{\beta}} = \hat{\mathbf{f}}$. Second we estimate $\mathbf{g}_0$ by regressing $\mathbf{t}^{(1)}$ onto $\mathbf{z}^{(1)}$ to produce a regression model $\hat{\mathbf{g}}(\cdot) : \mathbb{R}^{p-1} \to \mathbb{R}$. Any arbitrary linear or non-linear regression procedure can be used to fit $\hat{\mathbf{g}}(\cdot)$.

- Then, we estimate $\mathbb{E}[\eta_1^2]$ as $\mu_2 = \frac{1}{n/2} \sum_{i=\frac{n}{2}+1}^{n} t_i(t_i - \hat{\mathbf{g}}(\mathbf{z}_i))$ where the sum is taken over the second fold of data in $(\mathbf{X}^{(2)}, \mathbf{y}^{(2)})$; crucially $(t_i, \mathbf{z}_i)$ are independent of $\hat{\mathbf{g}}(\cdot)$ in this expression.

- If $\mu_2 \leq \tau$ for a threshold τ we simply output $\hat{y}_{\mathrm{OM}} = \mathbf{x}_\star^\top \hat{\boldsymbol{\beta}}$. If $\mu_2 \geq \tau$ we estimate θ_0 by solving the empirical moment equation:

$$\sum_{i=\frac{n}{2}+1}^{n} m(t_i, y_i, \hat{y}_{\mathrm{OM}}, \mathbf{z}_i^\top \hat{\mathbf{f}}, \hat{\mathbf{g}}(\mathbf{z}_i)) = 0 \implies$$
$$\hat{y}_{\mathrm{OM}} = \frac{\frac{1}{n/2} \sum_{i=\frac{n}{2}+1}^{n} (y_i - \mathbf{z}_i^\top \hat{\mathbf{f}})(t_i - \hat{\mathbf{g}}(\mathbf{z}_i))}{\mu_2} \quad (8)$$

  where the sum is taken over the second fold of data in $(\mathbf{X}^{(2)}, \mathbf{y}^{(2)})$ and m is defined in (7).

If we had oracle access to the underlying $\mathbf{f}_0$ and $\mathbf{g}_0$, solving the population moment condition $\mathbb{E}_{t_1, y_1, \mathbf{z}_1}[m(t_1, y_1, \theta, \mathbf{z}_1^\top \mathbf{f}_0, \mathbf{g}_0(\mathbf{z}_1))] = 0$ for θ would exactly yield $\theta_0 = \mathbf{x}_\star^\top \boldsymbol{\beta}_0$. In practice, we first construct estimates $\hat{\mathbf{f}}$ and $\hat{\mathbf{g}}$ of the unknown nuisance parameters to serve as surrogates for $\mathbf{f}_0$ and $\mathbf{g}_0$ and then solve an empirical version of the aforementioned moment condition to extract $\hat{y}_{\mathrm{OM}}$. A key property of the moments in (7) is their Neyman orthogonality: they satisfy $\mathbb{E}[\nabla_{\mathbf{z}_1^\top \mathbf{f}} m(t_1, y_1, \theta_0, \mathbf{z}_1^\top \mathbf{f}_0, \mathbf{g}_0(\mathbf{z}_1))] = 0$ and $\mathbb{E}[\nabla_{\mathbf{g}(\mathbf{z}_1)} m(t_1, y_1, \theta_0, \mathbf{z}_1^\top \mathbf{f}_0, \mathbf{g}_0(\mathbf{z}_1))] = 0$. Thus the solution of the empirical moment equations is first-order insensitive to errors arising from using $\hat{\mathbf{f}}, \hat{\mathbf{g}}$ in place of $\mathbf{f}_0$ and $\mathbf{g}_0$. Data splitting is further used to create independence across the two stages of the procedure. In the context of testing linearly-constrained hypotheses of the parameter $\boldsymbol{\beta}_0$, Zhu & Bradic (2018) propose a two-stage OM test statistic based on the transformed f moments introduced above; they do not use cross-fitting and specifically employ adaptive Dantzig-like selectors to estimate $\mathbf{f}_0$ and $\mathbf{g}_0$. Finally, the thresholding step allows us to control the variance increase that might arise from μ_2 being too small and thereby enables our non-asymptotic prediction risk bounds. Before presenting the analysis of the OM estimator (8) we introduce another condition[4]:

Assumption 5. *The noise η_i is independent of $\mathbf{z}_i$.*

---

[4]This assumption is not essential to our result and could be replaced by assuming η_i satisfies $\mathbb{E}[\eta_i | \mathbf{z}_i] = 0$ and is almost surely (w.r.t. to $\mathbf{z}_i$) sub-Gaussian with a uniformly (w.r.t. to $\mathbf{z}_i$) bounded variance parameter.

Recall $\hat{\mathbf{g}}$ is evaluated on the (independent) second fold data $\mathbf{z}$. We now obtain our central guarantee for the OM estimator (proved in Appendix E.1).

Theorem 5. *Let Assumptions 1, 2, 3, 4 and 5 hold, and assume that $\mathbf{g}_0(\mathbf{z}_i) = \mathbf{g}_0^\top \mathbf{z}_i$ in (6) for $\mathbf{g}_0 = \arg\min_{\mathbf{g}} \mathbb{E}[(t_1 - \mathbf{z}_1^\top \mathbf{g})^2]$. Then the thresholded orthogonal ML estimator $\hat{y}_{\mathrm{OM}}$ of (8) with $\tau = \frac{1}{4} \sigma_\eta^2$ satisfies*

$$\mathbb{E}[(\hat{y}_{\mathrm{OM}} - \mathbf{x}_\star^\top \boldsymbol{\beta}_0)^2] \leq$$
$$\|\mathbf{x}_\star\|_2^2 \left[O(\frac{\sigma_\epsilon^2}{\sigma_\eta^2 n}) + O(\frac{r_{\boldsymbol{\beta},2}^2 r_{\mathbf{g},2}^2}{(\sigma_\eta^2)^2}) + O(\frac{r_{\boldsymbol{\beta},2}^2 \sigma_\eta^2 + r_{\mathbf{g},2}^2 \sigma_\epsilon^2}{(\sigma_\eta^2)^2 n}) \right] \quad (9)$$

where $r_{\boldsymbol{\beta},2} = (\mathbb{E}[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^4])^{1/4}$ and $r_{\mathbf{g},2} = (\mathbb{E}[(\hat{\mathbf{g}}(\mathbf{z}_n) - \mathbf{g}_0(\mathbf{z}_n))^4])^{1/4}$ denote the expected prediction errors of the first-stage estimators, and the $O(\cdot)$ masks constants depending only on $C_{\min}, C_{\max}, C_{cond}, \kappa$.

Since we are interested in the case where $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{g}}(\cdot)$ have small error (i.e., $r_{\boldsymbol{\beta},2} = r_{\mathbf{g},2} = o(1)$), the first term in (9) can be interpreted as the variance of the estimator's prediction along the direction of $\mathbf{x}_\star$, while the remaining terms represent the reduced bias of the estimator. We first instantiate this result in the setting where both $\boldsymbol{\beta}_0$ and $\mathbf{g}_0$ are estimated using ridge regression (see Appendix E.2 for the corresponding proof).

Corollary 4 (OM Ridge). *Assume $\|\boldsymbol{\beta}_0\|_\infty / \sigma_\epsilon = O(1)$. In the setting of Theorem 5, suppose $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{g}}(\mathbf{z}_i) = \hat{\mathbf{g}}^\top \mathbf{z}_i$ are fit with the ridge estimator with regularization parameters $\lambda_{\boldsymbol{\beta}}$ and $\lambda_{\mathbf{g}}$ respectively. Then there exist universal constants c_1, c_2, c_3, c_4, c_5 such that if $p \geq 20$, $c_1 \frac{n^2 C_{\min}}{p C_{cond}} e^{-nc_2/\kappa^4 C_{cond}^2} \leq \lambda_{\boldsymbol{\beta}} \leq c_3 (C_{cond} C_{\max} n)^{1/3}$, and $c_4 \frac{n^2 C_{\min}}{p C_{cond}} e^{-nc_2/\kappa^4 C_{cond}^2} \leq \lambda_{\mathbf{g}} \leq p \left(\frac{C_{\max} \|\mathbf{x}_\star\|_2^2}{C_{cond}} \frac{n}{p} \sigma_\eta^4 \right)^{1/3}$ for $n \geq c_5 \kappa^4 C_{cond}^2 p$,*

$$\mathbb{E}[(\hat{y}_{\mathrm{OM}} - \mathbf{x}_\star^\top \boldsymbol{\beta}_0)^2]$$
$$\leq \|\mathbf{x}_\star\|_2^2 \left[O(\frac{\sigma_\epsilon^2}{\sigma_\eta^2 n}) + O(\frac{p^2}{(\sigma_\eta^2)^2 n^2}) + O(\frac{p(\sigma_\eta^2 + \sigma_\epsilon^2)}{(\sigma_\eta^2)^2 n^2}) \right]$$

where the $O(\cdot)$ masks constants depending only on $C_{\min}, C_{\max}, C_{cond}, \kappa$.

Similarly, when $\boldsymbol{\beta}_0$ and $\mathbf{g}_0$ are estimated using the Lasso we conclude the following (proved in Appendix E.2).

Corollary 5 (OM Lasso). *In the setting of Theorem 5, suppose $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{g}}(\mathbf{z}_i) = \hat{\mathbf{g}}^\top \mathbf{z}_i$ are fit with the Lasso with regularization parameters $\lambda_{\boldsymbol{\beta}} \geq 80 \sigma_\epsilon \sqrt{\log(2ep/s_{\boldsymbol{\beta}_0})/n}$ and $\lambda_{\mathbf{g}} \geq 80 \sigma_\eta \sqrt{\log(2ep/s_{\mathbf{g}})/n}$ respectively. If $p \geq 20$, $s_{\boldsymbol{\beta}_0} = \|\boldsymbol{\beta}_0\|_0$, and $s_{\mathbf{g}_0} = \|\mathbf{g}_0\|_0$, then there exist universal constants c_1, c_2 such that if $\|\boldsymbol{\beta}_0\|_\infty / \sigma_\epsilon = o(e^{c_1 n})$, then for $n \geq \frac{c_1 \kappa^4}{C_{\min}} \max\{s_{\boldsymbol{\beta}_0}, s_{\mathbf{g}}\} \log(2ep)$,*

$$\mathbb{E}[(\hat{y}_{\mathrm{OM}} - \mathbf{x}_\star^\top \boldsymbol{\beta}_0)^2] \leq$$
$$\|\mathbf{x}_\star\|_2^2 \left[O(\frac{\sigma_\epsilon^2}{\sigma_\eta^2 n}) + O(\frac{\lambda_{\boldsymbol{\beta}}^2 \lambda_{\mathbf{g}}^2 s_{\boldsymbol{\beta}_0} s_{\mathbf{g}_0}}{(\sigma_\eta^2)^2}) + O(\frac{\lambda_{\boldsymbol{\beta}}^2 s_{\boldsymbol{\beta}_0} \sigma_\eta^2 + \lambda_{\mathbf{g}}^2 s_{\mathbf{g}_0} \sigma_\epsilon^2}{(\sigma_\eta^2)^2 n}) \right]$$

where the $O(\cdot)$ masks constants depending only on $C_{\min}, C_{\max}, C_{cond}, \kappa$.

We make several comments regarding the aforementioned results. First, Theorem 5 possesses a double-robustness property. In order for the dominant bias term $O(r_{\boldsymbol{\beta},2}^2 r_{\mathbf{g},2}^2)$ to be small, it is sufficient for *either* $\boldsymbol{\beta}_0$ or $\mathbf{g}_0$ to be estimated at a fast rate or *both* to be estimated at a slow rate. As before, the estimator is transductive and adapted to predicting along the direction $\mathbf{x}_\star$. Second, in the case of ridge regression, to match the lower bound of Corollary 1, consider the setting where $n = \Omega(p^2)$, SNR $= o(\frac{p^2}{n})$, $\cos(\mathbf{x}_\star, \boldsymbol{\beta}_0)^2 = \Theta(1)$ and SNR $\gtrsim \frac{p}{n}$. Then, the upper bound[5] can be simplified to $O(\|\mathbf{x}_\star\|_2^2 \frac{\sigma_\epsilon^2}{n})$. By contrast, Corollary 1 shows the error of the optimally-tuned ridge estimator is lower bounded by $\omega(\|\mathbf{x}_\star\|_2^2 \frac{\sigma_\epsilon^2}{n})$; for example, the error is $\Omega(p\|\mathbf{x}_\star\|_2^2 \frac{\sigma_\epsilon^2}{n})$ when SNR $= \frac{1}{6}\frac{p}{n}$. Hence, the performance of the ridge estimator can be significantly worse then its transductive counterpart. Third, if we consider the setting of Corollary 5 where $n \gtrsim s_{\boldsymbol{\beta}_0} s_{\mathbf{g}_0} (\log p)^2$ while we take $\lambda_{\boldsymbol{\beta}} \asymp \sigma_\epsilon \sqrt{\log p/n}$ and $\lambda_{\mathbf{g}} \asymp \sigma_\eta \sqrt{\log p/n}$, the error of the OML estimator attains the fast, dimension-free $O(\|\mathbf{x}_\star\|_2^2 \frac{\sigma_\epsilon^2}{n})$ rate. On the other hand, Corollary 2 shows the Lasso suffers prediction error $\Omega(\|\mathbf{x}_\star\|_{(s)}^2 \frac{\sigma_\epsilon^2 \log p}{n})$, and hence again strict improvement is possible over the baseline when $\frac{\|\mathbf{x}_\star\|_2}{\|\mathbf{x}_\star\|_{(s)}} \lesssim \sqrt{\log p}$. Finally, although Theorem 5 makes stronger assumptions on the design of $\mathbf{X}$ than the JM-style estimator introduced in (4) and (3), one of the primary benefits of the OM framework is its flexibility. All that is required for the algorithm are "black-box" estimates of $\mathbf{g}_0$ and $\boldsymbol{\beta}_0$ which can be obtained from more general ML procedures than the Lasso.

4. Experiments

We complement our theoretical analysis with a series of numerical experiments highlighting the failure modes of standard inductive prediction. In Sections 4.1 and 4.2, error bars represent ± 1 standard error of the mean computed over 20 independent problem instances. We refer the reader to Appendix G for further details on the experimental set-up in each section. Python code to reproduce our experiments can be found at https://github.com/nileshtrip/SPTransducPred.

4.1. Excess Lasso Bias without Distribution Shift

We construct problem instances for Lasso estimation by independently generating $\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{I}_p)$, $\epsilon_i \sim \mathcal{N}(0, 1)$, and $(\boldsymbol{\beta}_0)_j \sim \mathcal{N}(0, 1)$ for j less then the desired sparsity level $s_{\boldsymbol{\beta}_0}$ while $(\boldsymbol{\beta}_0)_j = 0$ otherwise. We fit the Lasso estimator,

---

[5] Note that in this regime, $\sqrt{\text{SNR}} = \|\boldsymbol{\beta}_0\|_2/\sigma_\epsilon = o(1)$ and hence the condition $\|\boldsymbol{\beta}_0\|_\infty/\sigma_\epsilon = O(1)$ in Corollary 4 is satisfied.

JM-style estimator with Lasso pilot, and the OM f-moment estimator with Lasso first-stage estimators. We set all hyperparameters to their theoretically-motivated values. As
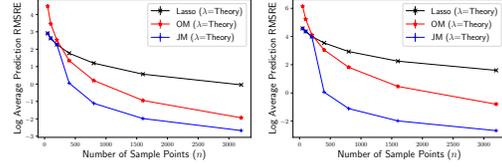

Figure 1. Lasso vs. transductive Lasso prediction without distribution shift. Hyperparameters are set according to theory (see Section 4.1). Left: $p = 200$, $s_{\boldsymbol{\beta}_0} = 20$. Right: $p = 200$, $s_{\boldsymbol{\beta}_0} = 100$.

Figure 1 demonstrates, both transductive methods significantly reduce the prediction risk of the Lasso estimator when the hyperparameters are calibrated to their theoretical values, even for a dense $\boldsymbol{\beta}_0$ (where $\frac{p}{s_{\boldsymbol{\beta}_0}} = 2$).

4.2. Benefits of Transduction under Distribution Shift

The no distribution shift simulations of Section 4.1 corroborate the theoretical results of Corollaries 3 and 5. However, since our transductive estimators are tailored to each individual test point $\mathbf{x}_\star$, we expect these methods to provide an even greater gain when the test distribution deviates from the training distribution.

In Figure 2, we consider two cases where the test distribution is either mean-shifted or covariance-shifted from the training distribution and evaluate the ridge estimator with the optimal regularization parameter for the training distribution, $\lambda_* = \frac{p\sigma_\epsilon^2}{\|\boldsymbol{\beta}_0\|_2^2}$. We independently generated $\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{I}_p)$, $\epsilon_i \sim \mathcal{N}(0, 1)$, and $\boldsymbol{\beta}_0 \sim \mathcal{N}(0, \frac{1}{\sqrt{p}}\mathbf{I}_p)$. In the case with a mean-shifted test distribution, we generated $\mathbf{x}_\star \sim \mathcal{N}(10\boldsymbol{\beta}_0, \mathbf{I}_p)$ for each problem instance while the covariance-shifted test distribution was generated by taking $\mathbf{x}_\star \sim \mathcal{N}(0, 100\boldsymbol{\beta}_0\boldsymbol{\beta}_0^\top)$. The first and second plots in Figure 2 show the OM estimator with λ_*-ridge pilot provides significant gains over the baseline λ_*-ridge estimator.
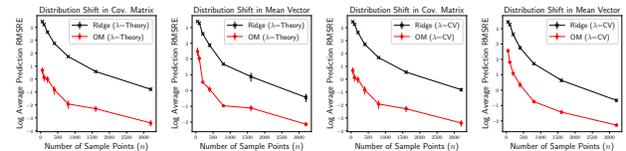

Figure 2. Ridge vs. transductive ridge prediction ($p = 200$) under train-test distribution shift. Hyperparameters are set according to theory (1st and 2nd plots) or via CV (3rd and 4th plots).

In Figure 3 we also consider two cases where the test distribution is shifted for Lasso estimation but otherwise identical to the previous set-up in Section 4.1. For covariance shifting, we generated $(\mathbf{x}_\star)_i \overset{\text{indep}}{\sim} \mathcal{N}(0, 100)$ for $i \in \text{supp}(\boldsymbol{\beta}_0)$ and $(\mathbf{x}_\star)_i = 0$ otherwise for each problem instance. For mean

Table 1. Test set RMSE of OLS, CV-tuned ridge, Lasso, and elastic net, and transductive CV-tuned ridge, Lasso, and elastic net on five real-world datasets. All hyperparameters are set via CV. Error bars represent ± 1 standard error of the mean computed over the test set.

| Method	Wine	Parkinson	Fire	Fertility	Triazines (no shift)
OLS	1.0118±0.0156	12.7916±0.1486	82.7147±35.5141	0.3988±0.0657	0.1716±0.037
Ridge	0.9936±0.0155	12.5267±0.1448	82.3462±35.5955	0.399±0.0665	0.1469±0.0285
OM f (Ridge)	0.9883±0.0154	12.4686±0.1439	82.3522±35.5519	0.3987±0.0655	**0.1446±0.029**
OM q (Ridge)	**0.7696±0.0145**	**12.0891±0.1366**	**81.9794±35.7872**	**0.3977±0.0653**	0.1507±0.0242
Lasso	0.9812±0.0155	12.2535±0.1356	82.0656±36.0321	0.4092±0.0716	0.1482±0.0237
JM (Lasso)	1.0118±0.0156	12.7916±0.1486	82.7147±35.5141	0.3988±0.0657	0.173±0.0367
OM f (Lasso)	0.9473±0.0152	**11.869±0.1339**	**81.794±35.5699**	0.398±0.0665	**0.1444±0.0239**
OM q (Lasso)	**0.7691±0.0144**	11.8692±0.1339	81.811±35.5637	**0.3976±0.0656**	0.1479±0.0226
Elastic	0.9652±0.0154	12.2535±0.1356	81.8428±35.8333	.0492±0.0716	0.1495±0.0238
OM f (Elastic)	0.9507±0.0152	**11.8369±0.1338**	**81.7719±35.6166**	0.398±0.0655	**0.1445±0.024**
OM q (Elastic)	**0.7693±0.0145**	11.8658±0.1341	81.803±35.6485	**0.3976±0.0657**	0.147±0.0228
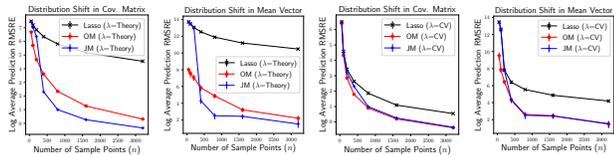
Figure 3. Lasso vs. transductive Lasso prediction ($p = 200$) under mean ($s_{\beta_0} = 100$) or covariance ($s_{\beta_0} = 20$) train-test distribution shifts. Hyperparameters are set according to theory (1st and 2nd plots) or via CV (3rd and 4th plots).

shifting, we generated $\mathbf{x}_\star \sim \mathcal{N}(10\beta_0, \mathbf{I}_p)$ for each problem instance. The first and second plots in Figure 3 show the transductive effect of the OM and JM estimators improves prediction risk with respect to the Lasso when the regularization hyperparameters are selected via theory. We also note that comparing CV-tuned ridge or Lasso to OM and JM with CV-tuned base procedures, shows the the benefit of transduction in this practical setting where regularization hyperparameters are chosen by CV.

As the third and fourth in Figure 2 show, selecting λ via CV leads to over-regularization of the ridge estimator, and the transductive methods provide substantial gains over the base ridge estimator. In the case of the Lasso, the third and fourth plots in Figure 3 show the residual bias of the CV Lasso also causes it to incur significant error in its test predictions, while the transductive methods provide substantial gains by adapting to each $\mathbf{x}_\star$.

4.3. Improving Cross-validated Prediction

Motivated by our findings on synthetic data, we conclude by reporting the performance of our methods on real data with and without distribution shift. We also include the popular elastic net estimator as a base regression procedure alongside the ridge and Lasso estimators. All hyperparameters are selected by CV.

For the OM estimators we exploited the flexibility of the framework by including a suite of methods for the aux-

iliary $\mathbf{g}$ regressions: Lasso estimation, random forest regression, and a $\mathbf{g} = 0$ baseline. Amongst these, we select the method with the least estimated asymptotic variance, which can be done in a data-dependent way *without* introducing any extra hyperparameters into the implementation; see Appendix G for further details on the methodology and regression datasets from the UCI dataset repository (Dua & Graff, 2017).

The $\mathbf{f}$ and $\mathbf{q}$ regressions were always fit with Lasso, ridge, or elastic net estimation. In Table 1 we can see the orthogonal moment estimators generically provide gains over the CV Lasso, CV ridge regression, and CV elastic net on datasets with intrinsic distribution shift and perform comparably on a dataset without explicit distribution shift. On Wine, we see a substantial performance gain from 0.96-0.99 RMSE without transduction to 0.77 with OM q transduction. The gains on other datasets are smaller but notable as they represent consistent improvements over the de facto standard of CV prediction.

We also report the performance of ordinary least squares (OLS) which produces an unbiased estimate of the entire parameter vector β_0. OLS fares worse than most methods on each dataset due to a increase in variance. In contrast, our proposed transductive procedures limit the variance introduced by targeting a single parameter of interest, $\langle \mathbf{x}_\star, \beta_0 \rangle$.

5. Discussion and Future Work

We have highlighted the detrimental effects of bias in linear prediction and presented two transductive estimators that amelioriate this bias. We provided theoretical guarantees for these estimators and demonstrated their practical utility–especially under distribution shift–on both synthetic/real data. Promising directions for future work include improving the quality of our OM debiasing techniques using higher-order orthogonal moments (Mackey et al., 2017) and exploring the utility of these debiasing techniques for other regularizers (e.g., group Lasso (Yuan & Lin, 2006) penalties) and

alternative models such as generalized linear models/kernel machines.

6. Acknowledgements

References

Alquier, P. and Hebiri, M. Transductive versions of the lasso and the dantzig selector. *Journal of Statistical Planning and Inference*, 142(9):2485–2500, 2012.

Athey, S., Imbens, G. W., and Wager, S. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4): 597–623, 2018.

Bellec, P. C. and Zhang, C.-H. De-biasing the lasso with degrees-of-freedom adjustment. *arXiv preprint arXiv:1902.08885*, 2019.

Bellec, P. C., Lecué, G., and Tsybakov, A. B. Slope meets lasso: improved oracle bounds and optimality. *arXiv preprint arXiv:1605.08651*, 2016.

Bellec, P. C., Dalalyan, A. S., Grappin, E., Paris, Q., et al. On the prediction loss of the lasso in the partially labeled setting. *Electronic Journal of Statistics*, 12(2):3443–3472, 2018.

Bickel, P. J., Ritov, Y., Tsybakov, A. B., et al. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Bishop, A. N., Moral, P. D., and Niclas, A. An introduction to wishart matrix moments. *Foundations and TrendsÂ® in Machine Learning*, 11(2):97–218, 2018. ISSN 1935-8237. doi: 10.1561/2200000072. URL http://dx.doi.org/10.1561/2200000072.

Cai, T. T. and Guo, Z. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of statistics*, 45(2):615–646, 2017.

Chao, S.-K., Ning, Y., and Liu, H. On high dimensional post-regularization prediction intervals, 2014.

Chen, X., Monfort, M., Liu, A., and Ziebart, B. D. Robust covariate shift regression. In *Artificial Intelligence and Statistics*, pp. 1270–1279, 2016.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., et al. Double/debiased machine learning for treatment and causal parameters. Technical report, 2017.

Diamond, S. and Boyd, S. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Dobriban, E., Wager, S., et al. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Hoerl, A. E. and Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Hsu, D., Kakade, S. M., and Zhang, T. Random design analysis of ridge regression. In *Conference on learning theory*, pp. 9–1, 2012.

Javanmard, A. and Montanari, A. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

Mackey, L., Syrgkanis, V., and Zadik, I. Orthogonal machine learning: Power and limitations. *arXiv preprint arXiv:1711.00342*, 2017.

Moritz, P., Nishihara, R., Wang, S., Tumanov, A., Liaw, R., Liang, E., Elibol, M., Yang, Z., Paul, W., Jordan, M. I., et al. Ray: A distributed framework for emerging {AI} applications. In *13th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 18)*, pp. 561–577, 2018.

Raskutti, G., Wainwright, M. J., and Yu, B. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, 57 (10):6976–6994, 2011.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

van de Geer, S. Statistical theory for high-dimensional models, 2014a.

van de Geer, S. Statistical theory for high-dimensional models. *arXiv preprint arXiv:1409.8557*, 2014b.

Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., et al. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

Wager, S., Du, W., Taylor, J., and Tibshirani, R. J. High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(45):12673–12678, 2016.

Wainwright, M. J. High-dimensional statistics: A non-asymptotic viewpoint.

Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68 (1):49–67, 2006.

Zhang, C.-H. and Zhang, S. S. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

Zhu, X. J. Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2005.

Zhu, Y. and Bradic, J. Linear hypothesis testing in dense high-dimensional linear models. *Journal of the American Statistical Association*, 113(524):1583–1600, 2018.

A. Notation

We first establish several useful pieces of notation used throughout the Appendices. We will say that a mean-zero random variable x is sub-gaussian, $x \sim \mathrm{sG}(\kappa)$, if $\mathbb{E}[\exp(\lambda x))] \leq \exp\left(\frac{\kappa^2 \lambda^2}{2}\right)$ for all λ. We will say that a mean-zero random variable x is sub-exponential, $x \sim \mathrm{sE}(\nu, \alpha)$, if $\mathbb{E}[\exp(\lambda x)] \leq \exp\left(\frac{\nu^2 \lambda^2}{2}\right)$ for all $|\lambda| \leq \frac{1}{\alpha}$. We will say that a mean-zero random vector is sub-gaussian, $\mathbf{x} \sim \mathrm{sG}(\kappa)$, if $\forall \mathbf{v} \in \mathbb{R}^p$, $\mathbb{E}[\exp(\mathbf{v}^\top \mathbf{x})] \leq \exp\left(\frac{\kappa^2 \|\mathbf{v}\|_2^2}{2}\right)$. Moreover a standard Chernoff argument shows if $x \sim \mathrm{sE}(\nu, \alpha)$ then $\Pr[|x| \geq t] \leq 2 \exp\left(-\frac{1}{2} \min(\frac{t^2}{\nu^2}, \frac{t}{\alpha})\right)$.

B. Proofs for Section 2.1: Lower Bounds for Prediction with Ridge Regression

Here we provide lower bounds on the prediction risk of the ridge regression estimator. To do so, we show that under Gaussian design and independent Gaussian noise ϵ the ridge regression estimator can perform poorly.

Recall we define the ridge estimator as $\hat{\boldsymbol{\beta}}_R(\lambda) = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2\right)$ which implies $\hat{\boldsymbol{\beta}}_R(\lambda) = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{y}$. For convenience we further define $\hat{\boldsymbol{\Sigma}} = \frac{\mathbf{X}^\top \mathbf{X}}{n}$, $\hat{\boldsymbol{\Sigma}}_\lambda = \frac{\mathbf{X}^\top \mathbf{X}}{n} + \frac{\lambda}{n}\mathbf{I}_p$ and $\boldsymbol{\Pi}_\lambda = \mathbf{I}_p - (\hat{\boldsymbol{\Sigma}}_\lambda)^{-1}\hat{\boldsymbol{\Sigma}}$. Note that under Assumption 1, $\hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0 = -\boldsymbol{\Pi}_\lambda \boldsymbol{\beta}_0 + \hat{\boldsymbol{\Sigma}}_\lambda^{-1} \mathbf{X}^\top \boldsymbol{\epsilon}/n$, which can be thought of as a standard bias-variance decomposition for the ridge estimator. We begin by stating a standard fact about Wishart matrices we will repeatedly use throughout this section.

Proposition 6. *Let $\mathbf{x}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \mathbf{I}_p)$ for $i \in [n]$. Then the eigendecomposition of the sample covariance $\hat{\boldsymbol{\Sigma}}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top = \mathbf{V}^\top \mathbf{D} \mathbf{V}$ satisfies the following properties:*

- *The orthonormal matrix $\mathbf{V}$ is uniformly distributed (with respect to the Haar measure) over the orthogonal group $O(p)$.*

- *The matrices $\mathbf{V}$ and $\mathbf{D}$ are independent. Moreover, by isotropy, $\mathbf{D}$ is equivalent in distribution to the random matrix $z\mathbf{I}_p$ where z is an unordered eigenvalue of $\hat{\boldsymbol{\Sigma}}_n$.*

Proof. Statements and proofs of these standard facts about Wishart matrices can be found in Bishop et al. (2018).   □

B.1. Theorem 1

We now provide the proof of our primary lower bound on the prediction risk of the ridge estimator,

Proof of Theorem 1. The first statement follows by using Lemma 1 and taking the expectation over $\mathbf{X}$,

$$\mathbb{E}[\langle \mathbf{x}_\star, \hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0 \rangle^2] = \mathbb{E}[(\mathbf{x}_\star^\top \boldsymbol{\Pi}_\lambda \boldsymbol{\beta}_0)^2] + \frac{\sigma_\epsilon^2}{n}\mathbf{x}_\star^\top \mathbb{E}[(\hat{\boldsymbol{\Sigma}}_\lambda)^{-1}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Sigma}}_\lambda)^{-1}]\mathbf{x}_\star = \mathbb{E}[(\mathbf{x}_\star^\top \boldsymbol{\Pi}_\lambda \boldsymbol{\beta}_0)^2] + \sigma_\epsilon^2 \frac{\|\mathbf{x}_\star\|_2^2}{n}\mathbb{E}[\frac{z}{(z+\lambda/n)^2}]$$

The computation of the variance term uses the eigendecomposition of $\boldsymbol{\Sigma}$ and Proposition 6,

$$\mathbb{E}[(\hat{\boldsymbol{\Sigma}}_\lambda)^{-1}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Sigma}}_\lambda)^{-1}] = \mathbb{E}[\mathbf{V}^\top \mathbb{E}[(\mathbf{D} + \lambda/n\mathbf{I}_p)^{-2}\mathbf{D}]\mathbf{V}] = \mathbb{E}[\frac{z}{(z+\lambda/n)^2}]\mathbf{I}_p.$$

We now lower bound the bias. Again by Proposition 6 and the eigendecomposition of $\hat{\boldsymbol{\Sigma}}_n$, $\mathbb{E}[\boldsymbol{\Pi}_\lambda] = \mathbb{E}[\frac{\lambda/n}{z+\lambda/n}]\mathbf{I}_p$. Using Jensen's inequality,

$$\mathbb{E}[(\boldsymbol{\beta}_0^\top \boldsymbol{\Pi}_\lambda \mathbf{x}_\star)^2] \geq (\boldsymbol{\beta}_0^\top \mathbb{E}[\boldsymbol{\Pi}_\lambda]\mathbf{x}_\star)^2 = \|\mathbf{x}_\star\|_2^2 \|\boldsymbol{\beta}_0\|_2^2 \cos(\mathbf{x}_\star, \boldsymbol{\beta}_0)^2 \mathbb{E}[(\frac{\lambda/n}{z+\lambda/n})]^2.$$

The final expectation over the unordered eigenvalue distribution can be controlled using the sharp concentration of Gaussian random matrices. Namely for $n \geq p$, $\left\|\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}\right\|_2 \leq 2\epsilon + \epsilon^2$ for $\epsilon = \sqrt{\frac{p}{n}} + \delta$ with probability at least $1 - 2\epsilon^{-n\delta^2/2}$ (Wainwright, Theorem 6.1, Example 6.2). Taking $\delta = 1/2\sqrt{p/n}$ and assuming that $p \geq 20$ we conclude that $\left\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_2 \leq 6\sqrt{\frac{p}{n}}$ with probability at least $\frac{1}{2}$ – let $\mathcal{E}$ denote this event. Note that by the Weyl inequalities, on the

event $\mathcal{E}$, all of the eigenvalues of $\hat{\Sigma}$ are uniformly close to the eigenvalues of Σ. Hence if $n \geq p$, on $\mathcal{E}$ we must have that $\hat{\Sigma}_n \preceq 7\mathbf{I}_p$, and hence the unordered eigenvalue $z \leq 7$ as well. Thus it follows that $(\mathbb{E}[\frac{1}{\lambda/n+z}])^2 \geq (\mathbb{E}[\frac{1}{\lambda/n+z}\mathbf{I}[\mathcal{E}]])^2 \geq (\mathbb{E}[\frac{1}{\lambda/n+7}\mathbf{I}[\mathcal{E}]])^2 \geq \frac{1}{4}\frac{1}{(\lambda/n+7)^2}$. Combining the expressions yields the conclusion. $\qquad\square$

B.2. Corollary 1

We now prove Corollary 1.

Proof of Corollary 1. The expression for $\lambda_* = \arg\min_\lambda \mathbb{E}[\|\hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\|_2^2]$ can be computed using Lemma 2. Since, $\arg\min_\lambda \mathbb{E}[\|\tilde{y} - \tilde{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}_R(\lambda)\|_2^2 = \mathbb{E}[\|\hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\|_2^2] + \sigma_\epsilon^2$, equality of the minimizers follows for both expressions.

Define $\mathrm{SNR} = \frac{\|\boldsymbol{\beta}_0\|_2^2}{\sigma_\epsilon^2}$ and $a = \sqrt{\frac{4C}{\mathrm{SNR}}}$. If, in addition, $n \geq a^2$ and $\lambda \geq \frac{7an}{\sqrt{n}-a}$, we claim,

$$\mathbb{E}[\langle \mathbf{x}_\star, \hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\rangle^2] \geq C\cos(\mathbf{x}_\star, \boldsymbol{\beta}_0)^2 \cdot \|\mathbf{x}_\star\|_2^2 \cdot \frac{\sigma_\epsilon^2}{n}.$$

This lower bound follows by simply rearranging the lower bound from Theorem 1 – some algebraic manipulation give the conditions that $\frac{\lambda/n}{\lambda/n+7} \geq \frac{a}{\sqrt{n}} \implies \lambda \geq \frac{a}{\sqrt{n}}(\lambda + 7n) \implies \lambda(1 - \frac{a}{\sqrt{n}}) \geq 7a\sqrt{n} \implies \lambda \geq \frac{7a\sqrt{n}}{1 - \frac{a}{\sqrt{n}}} \implies \lambda \geq \frac{7an}{\sqrt{n}-a}$.

After defining $\lambda_* = p/\mathrm{SNR} = b$ the previous inequality over λ_* to achieve the desired conclusion can be rearranged to $b(\sqrt{n}-a) \geq 7an \implies n - \frac{b}{7a}\sqrt{n} + \frac{b}{7} \implies n - \frac{b}{7a}\sqrt{n} + \frac{b}{7} \leq 0$. The corresponding quadratic equation in $\sqrt{n}$ has roots $r_+ = \frac{1}{14}\left(\frac{b}{a} + \frac{\sqrt{b^2 - 28a^2b}}{a}\right), r_- = \frac{1}{14}\left(\frac{b}{a} - \frac{\sqrt{b^2 - 28a^2b}}{a}\right)$. In order to ensure both roots are real we must have $b \geq 28a^2 \implies p \geq 120C$. The condition that $r_- \leq \sqrt{n} \leq r_+$ can be equivalently expressed as,

$$\left|\sqrt{n} - \frac{1}{14}\frac{b}{a}\right| \leq \frac{\sqrt{b^2 - 28a^2b}}{a} \iff$$

$$\left|\sqrt{n} - \frac{1}{14}\frac{p}{\sqrt{4C\mathrm{SNR}}}\right| \leq \sqrt{\frac{p^2}{4C\mathrm{SNR}} - 28\frac{p}{\mathrm{SNR}}}.$$

Defining C such that $\sqrt{n} - \frac{1}{14}\frac{p}{\sqrt{4C\mathrm{SNR}}} = 0 \implies C = \frac{p^2}{784n\mathrm{SNR}}$. The remaining condition simplifies as, $\sqrt{\frac{p^2}{4C\mathrm{SNR}} - 28\frac{p}{\mathrm{SNR}}} \geq 0 \implies 196n - 28\frac{p}{\mathrm{SNR}} \geq 0 \implies n \geq \frac{1}{7}\frac{p}{\mathrm{SNR}}$. The condition $p \geq 120C \implies n \geq \frac{1}{6}\frac{p}{\mathrm{SNR}}$. Accordingly, under these conditions,

$$\mathbb{E}[\langle \mathbf{x}_\star, \hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\rangle^2] \geq \frac{\cos(\mathbf{x}_\star, \boldsymbol{\beta}_0)^2}{784}\frac{p^2}{n\mathrm{SNR}}\|\mathbf{x}_\star\|_2^2 \cdot \frac{\sigma_\epsilon^2}{n}$$

$\qquad\square$

We first compute the (conditional on $\mathbf{X}$) prediction risk of this estimator alongst $\mathbf{x}_\star$ as,

Lemma 1. *Let the independent noise distribution be Gaussian, $\epsilon \sim \mathcal{N}(0, \mathbf{I}_n\sigma_\epsilon^2)$, and Assumption 1 hold. Then,*

$$\mathbb{E}[\langle \mathbf{x}_\star, \hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\rangle^2 | \mathbf{X}] = (\mathbf{x}_\star^\top \boldsymbol{\Pi}_\lambda \boldsymbol{\beta}_0)^2 + \sigma_\epsilon^2 \mathbf{x}_\star^\top (\hat{\boldsymbol{\Sigma}}_\lambda)^{-1}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Sigma}}_\lambda)^{-1}\mathbf{x}_\star/n$$

Proof. Using the standard bias-variance decomposition $\hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0 = -\boldsymbol{\Pi}_\lambda \boldsymbol{\beta}_0 + \hat{\boldsymbol{\Sigma}}_\lambda^{-1}\mathbf{X}^\top \epsilon/n$, squaring and taking the expectation over ϵ (which is mean-zero) gives the result. $\qquad\square$

We now calculate the optimal choice of the ridge parameter λ to minimize the parameter error in the ℓ_2 distance.

Lemma 2. *Under Assumption 1, let $\mathbf{x}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \mathbf{I}_p)$ with independent noise $\epsilon \sim \mathcal{N}(0, \mathbf{I}_n\sigma_\epsilon^2)$. Then,*

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\right\|_2^2\right] = \|\boldsymbol{\beta}_0\|_2^2 \mathbb{E}[(\frac{\lambda/n}{z + \lambda/n})^2] + \frac{\sigma_\epsilon^2 p}{n}\mathbb{E}[\frac{z}{(z + \lambda/n)^2}]$$

and the optimal $\lambda_* = \arg\min_\lambda \mathbb{E}\left[\left\|\hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\right\|_2^2\right]$, is $\lambda_*/p = \frac{\sigma_\epsilon^2}{\|\boldsymbol{\beta}_0\|_2^2}$.

Proof. We first compute the (expected) mean-squared error. Using Lemma 1, summing over $\mathbf{x}_\star = \mathbf{e}_i$, and taking a further expectation over $\mathbf{X}$ we have that,

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\right\|_2^2\right] = \mathbb{E}\left[\sum_{i=1}^p (\mathbf{e}_i^\top \boldsymbol{\Pi}_\lambda \boldsymbol{\beta}_0)^2\right] + \frac{\sigma_\epsilon^2}{n}\mathbb{E}\left[\sum_{i=1}^p \mathbf{e}_i^\top (\hat{\boldsymbol{\Sigma}}_\lambda)^{-1}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Sigma}}_\lambda)^{-1}\mathbf{e}_i\right]$$

The computation of both the bias and variance terms exploits Proposition 6 along with the eigendecomposition of $\hat{\boldsymbol{\Sigma}}_n$. For the bias term,

$$\mathbb{E}[\sum_{i=1}^p (\mathbf{e}_i^\top \boldsymbol{\Pi}_\lambda \boldsymbol{\beta}_0)^2] = \mathbb{E}[\boldsymbol{\beta}_0^\top \boldsymbol{\Pi}_\lambda^2 \boldsymbol{\beta}_0] = \mathbb{E}[\boldsymbol{\beta}_0^\top \mathbf{V}^\top (\mathbb{E}[\mathbf{I}_p - 2(\mathbf{D} + \lambda\mathbf{I}_p)^{-1}\mathbf{D} + (\mathbf{D} + \lambda\mathbf{I}_p)^{-2}\mathbf{D}^2])\mathbf{V}\boldsymbol{\beta}_0] = \|\boldsymbol{\beta}_0\|_2^2 v$$

where $v = \mathbb{E}[(\frac{\lambda/n}{\lambda/n+z})^2]$. Similarly for the variance term,

$$\frac{\sigma_\epsilon^2}{n}\mathbb{E}[\sum_{i=1}^p \mathbf{e}_i^\top (\hat{\boldsymbol{\Sigma}}_\lambda)^{-1}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Sigma}}_\lambda)^{-1}\mathbf{e}_i] = \frac{\sigma_\epsilon^2}{n}\mathbb{E}[\text{Tr}\left[(\hat{\boldsymbol{\Sigma}}_\lambda)^{-1}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Sigma}}_\lambda)^{-1}\right]] = \frac{\sigma_\epsilon^2}{n}\mathbb{E}[\text{Tr}\left[\mathbf{V}\mathbb{E}[w]\mathbf{I}_p\mathbf{V}^\top\right]] = \frac{\sigma_\epsilon^2 p}{n}\mathbb{E}[w]$$

where $\mathbb{E}[w] = \mathbb{E}[\frac{z}{(z+\lambda/n)^2}]$. Combining we have that,

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\right\|_2^2\right] = \|\boldsymbol{\beta}_0\|_2^2\mathbb{E}[(\frac{\lambda/n}{z + \lambda/n})^2] + \frac{\sigma_\epsilon^2 p}{n}\mathbb{E}[\frac{z}{(z + \lambda/n)^2}].$$

In general this expression is a complicated function of λ, however conveniently,

$$\frac{d}{d\lambda}\mathbb{E}\left[\left\|\hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\right\|_2^2\right] = 2\lambda n\|\boldsymbol{\beta}_0\|_2^2\mathbb{E}[\frac{z}{(z + \lambda n)^3}] - 2n^2\frac{\sigma_\epsilon^2 p}{n}\mathbb{E}[\frac{z}{(\lambda n + z)^3}] \implies \lambda_*/p = \frac{\sigma_\epsilon^2}{\|\boldsymbol{\beta}_0\|_2^2}.$$

$\square$

C. Proofs for Section 2.2: Lower Bounds for Prediction with the Lasso

Here we provide lower bounds on the prediction risk of the Lasso estimator. In order to do so we will exhibit a benign instance of the design matrix for which for the Lasso performs poorly.

C.1. Theorem 2

We begin by stating a more general version of Theorem 2 and provide its proof

Theorem 7. *Under Assumption 1, fix any $s \geq 0$, and let $\mathbf{x}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \mathbf{I}_p)$ with independent noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}_n\sigma_\epsilon^2)$. Then, if $\hat{\boldsymbol{\beta}}_L(\lambda)$ denotes the solution of the Lasso program, with regularization parameter chosen as $\lambda \geq (8 + 2\sqrt{2})\sigma_\epsilon\sqrt{\log(2ep)/n}$, and $p \geq 20$, there exist universal constants c_1, c_2, c_3 such that for all $n \geq c_1 s^2 \log(2ep)$ and for fixed $\mathbf{x}_\star \sim \mathcal{P}^\star$ independently of $\mathbf{X}, \boldsymbol{\epsilon}$,*

$$\sup_{\boldsymbol{\beta}_0 \in \mathbb{B}_0(s)} \mathbb{E}[\langle \mathbf{x}_\star, \hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\rangle^2] \geq \sup_{\substack{\boldsymbol{\beta}_0 \in \mathbb{B}_0(s), \\ \|\boldsymbol{\beta}_0\|_\infty \leq \lambda}} \mathbb{E}[\langle \mathbf{x}_\star, \hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\rangle^2] \geq c_2\lambda^2\Lambda_s[\mathbb{E}[\mathbf{x}_\star\mathbf{x}_\star^\top]] \geq c_2\lambda^2\|\mathbb{E}[\mathbf{x}_\star]\|_{(s)}^2$$

where the trimmed norm $\|\mathbf{x}_\star\|_{(s)}$ *is the sum of the magnitudes of the s largest magnitude entries of $\mathbf{x}_\star$ and $\Lambda_s[\mathbb{E}[\mathbf{x}_\star\mathbf{x}_\star^\top]]$ is the maximum s-sparse eigenvalue of $\mathbb{E}[\mathbf{x}_\star\mathbf{x}_\star^\top]$. Moreover, for deterministic $\mathbf{x}_\star$,*

$$\sup_{\boldsymbol{\beta}_0 \in \mathbb{B}_0(s)} \mathbb{E}[\langle \mathbf{x}_\star, \hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\rangle^2] \leq c_3\lambda^2\|\mathbb{E}[\mathbf{x}_\star]\|_{(s)}^2$$

Proof of Theorem 2 and Theorem 7. Let $\mathbf{v}_\star$ denote the maximum s-sparse eigenvector of $\mathbb{E}[\mathbf{x}_\star\mathbf{x}_\star^\top]$ (which is normalized as have $\|\mathbf{v}\|_2 = \|\mathbf{v}\|_{(s)} = 1$) and $\Lambda_s[\mathbb{E}[\mathbf{x}_\star\mathbf{x}_\star^\top]]$ its corresponding eigenvalue. We begin by restricting $\boldsymbol{\beta}_0$ to have support

on these s coordinates of $\mathbf{v}_\star$, denoted by S; we subsequently choose the magnitude of the elements $\boldsymbol{\beta}_0$. Now under the conditions of the Proposition, we can guarantee support recovery of the Lasso solution, $S_{\hat{\boldsymbol{\beta}}_L} \subseteq S_{\boldsymbol{\beta}_0} \equiv S$, with probability at least $\frac{1}{2}$ by Proposition 9. Denote this event by $\mathcal{S}$.

Thus, for this choice of $\boldsymbol{\beta}_0$,

$$\mathbb{E}[\langle \mathbf{x}_\star, \hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0 \rangle^2] \geq \mathbb{E}[\langle (\mathbf{x}_\star)_S, (\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0)_S \rangle^2 \mathbf{I}[\mathcal{S}]] = \mathbb{E}[\langle (\mathbf{x}_\star)_S, \mathbf{I}[\mathcal{S}](\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0)_S \rangle^2]$$
$$\geq \langle \mathbb{E}[\mathbf{I}[\mathcal{S}](\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0)_S], \mathbb{E}[\mathbf{x}_\star \mathbf{x}_\star^\top]_S \mathbb{E}[\mathbf{I}[\mathcal{S}](\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0)_S \rangle \tag{10}$$

using Jensen's inequality and independence of $\mathbf{x}_\star$ and $\hat{\boldsymbol{\beta}}_L(\lambda)$ in the inequality.

We now focus on characterizing the bias of the Lasso solution $\hat{\boldsymbol{\beta}}_L(\lambda)$ on the coordinates contained in S (in fact using properties of the debiased Lasso estimator). Consider a single coordinate $i \in S$, and without loss of generality assume that $(\mathbf{x}_\star)_i > 0$, in which case we choose $(\boldsymbol{\beta}_0)_i > 0$. We will argue that the magnitude of $(\boldsymbol{\beta}_0)_i$ can be chosen so that $\mathbb{E}[(\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0)_i] < c < 0$ for appropriate c under the conditions of the theorem. Note that under our assumptions $\kappa = C_{\max} = C_{\min} = 1$ for the following.

Recall, since $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$, from the KKT conditions applied to the Lasso objective we have that,

$$\frac{1}{n}\mathbf{X}^\top(\mathbf{X}^\top \hat{\boldsymbol{\beta}}_L(\lambda) - \mathbf{y}) + \lambda \mathbf{v} = 0, \quad \mathbf{v} \in \partial\left(\|\hat{\boldsymbol{\beta}}_L(\lambda)\|_1\right) \implies$$

$$\underbrace{(\mathbf{I} - \hat{\boldsymbol{\Sigma}}_n)(\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0)}_{\boldsymbol{\Delta}} + \underbrace{\frac{1}{n}\mathbf{X}^\top \boldsymbol{\epsilon}}_{\mathbf{Z}} - \lambda \mathbf{v} = \hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0$$

We can now use this relation to control the coordinate-wise Lasso bias,

$$\mathbb{E}[\mathbf{I}[\mathcal{S}](\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0)_i] = \mathbb{E}[(\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0)_i \mathbf{I}[\mathcal{S} \cap \{(\hat{\boldsymbol{\beta}}_L(\lambda))_i > 0\}]] + \mathbb{E}[(\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0)_i \mathbf{I}[\mathcal{S} \cap \{(\hat{\boldsymbol{\beta}}_L(\lambda))_i \leq 0\}]] =$$
$$\mathbb{E}[(\mathbf{Z} + \boldsymbol{\Delta} - \lambda \mathbf{v})_i \mathbf{I}[\mathcal{S} \cap \{(\hat{\boldsymbol{\beta}}_L(\lambda))_i > 0\}]] + \mathbb{E}[(\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0)_i \mathbf{I}[\mathcal{S} \cap \{(\hat{\boldsymbol{\beta}}_L(\lambda))_i \leq 0\}]] \leq$$
$$\mathbb{E}[|\mathbf{Z}_i| + |\boldsymbol{\Delta}_i|] - \lambda\mathbb{E}[\mathbf{I}[\mathcal{S} \cap \{(\hat{\boldsymbol{\beta}}_L(\lambda))_i > 0\}]] - (\boldsymbol{\beta}_0)_i \mathbb{E}[\mathbf{I}[\mathcal{S} \cap \{(\hat{\boldsymbol{\beta}}_L(\lambda))_i \leq 0\}]] \leq$$
$$\mathbb{E}[|\mathbf{Z}_i| + |\boldsymbol{\Delta}_i|] - \min(\lambda, (\boldsymbol{\beta}_0)_i) \underbrace{\Pr[\mathcal{S}]}_{\geq 1/2}.$$

At this point we fix the magnitude of $(\boldsymbol{\beta}_0)_i = \lambda$ for $i \in S$. We can now bound the expectations of our first two terms. For the first term $\mathbf{Z}_i = \frac{1}{n}\mathbf{e}_i^\top \mathbf{X}^\top \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_\epsilon^2 \mathbf{I}_n)$ and $\mathbf{v} = \mathbf{X}\mathbf{e}_i \sim \mathcal{N}(0, \sigma_\epsilon^2 \mathbf{I}_n)$ independently of $\boldsymbol{\epsilon}$. Thus,

$$\mathbb{E}[|\mathbf{Z}_i|] \leq \frac{1}{n}\sqrt{\mathbb{E}[(\mathbf{v}^\top \boldsymbol{\epsilon})^2]} = \frac{\sigma}{\sqrt{n}}.$$

For the second term,

$$\mathbb{E}[|\boldsymbol{\Delta}_i|] \leq \sqrt{\mathbb{E}[\|(\hat{\boldsymbol{\Sigma}}_n - \mathbf{I})\mathbf{e}_i\|_\infty^2]}\sqrt{\mathbb{E}[\|\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0\|_1^2]}$$

From the proof of Lemma 5, with $\mathbf{x}_\star = \mathbf{e}_i$ and $\boldsymbol{\Omega} = \mathbf{I}$, we have that $\Pr\left[\|(\hat{\boldsymbol{\Sigma}}_n - \mathbf{I})\mathbf{e}_i\|_\infty \geq t\right] \leq 2p \cdot \exp\left(-\frac{n}{2}\min((\frac{t}{\kappa'})^2, \frac{t}{\kappa'})\right)$ where $\kappa' = 8$. Note for $n \geq (a/\kappa')^2 \log p$, $a\sqrt{\frac{\log p}{n}} \leq \kappa'$. Defining $A = \|(\hat{\boldsymbol{\Sigma}}_n - \mathbf{I})\mathbf{e}_i\|_\infty$,

$$\mathbb{E}[A^2] = \int_0^\infty 2t \Pr[A > t] \leq 4\left[\int_0^{a\sqrt{\log p/n}} t \cdot 1 + \int_{a\sqrt{\log p/n}}^{\kappa'} p \cdot t \exp\left(-\frac{n}{2}\left(\frac{t}{\kappa'}\right)^2\right) + \int_{\kappa'}^\infty p \cdot t \exp\left(-\frac{n}{2}\frac{t}{\kappa'}\right)\right]$$

$$\leq 4\left[\frac{a^2}{2}\frac{\log p}{n} + \frac{\kappa'^2 p^{1-\frac{a^2}{2\kappa'^2}}}{n} + \frac{2\kappa'^2 e^{-n/2}(2+n)p}{n^2}\right] \leq \left(8\kappa'^2 + \frac{20\kappa'^2}{p\log p \cdot n}\right)\frac{\log p}{n} \leq 9\kappa'^2\frac{\log p}{n}.$$

where the last sequence of inequalities follows by choosing $a = 2\kappa'$, assuming $n \geq \max\{4 \log p, 2\}$, and then assuming $p \geq 20$. Using Lemma 10 and 15 we have that,

$$\mathbb{E}[\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_1^2] \leq \left(\frac{49\lambda s_{\boldsymbol{\beta}_0}}{4}\right)^2 + \left(\frac{49}{8}\frac{(8 + 2\sqrt{2})\sigma_\epsilon}{\sqrt{n}}\right)^2 + \left(\frac{\sigma_\epsilon^4}{\lambda_*^2} + 2^4 s_{\boldsymbol{\beta}_0}^2 \lambda^2\right)\left(2e^{-c/2 \cdot n}\right)$$

using our choice of $|(\boldsymbol{\beta}_0)_i| = \lambda$ for each of the s non-zero coordinates in $\boldsymbol{\beta}_0$ (so $\|\boldsymbol{\beta}_0\|_1 \leq s_{\boldsymbol{\beta}_0}\lambda$). Here λ_* is the lower bound on λ from the Theorem statement. Under the assumption that $n \geq c_1 s_{\boldsymbol{\beta}_0}^2 \log(2ep)$ and $p \geq 20$, there exists c_1 such that $\left(\frac{\sigma_\epsilon^4}{\lambda_*^2} + 2^4 s_{\boldsymbol{\beta}_0}^2 \lambda^2\right)\left(2e^{-c_2/2 \cdot n}\right) \leq (8 + 2\sqrt{2})^2 \sigma_\epsilon^2/n + 2^5 \lambda^2 s_{\boldsymbol{\beta}_0}^2$. Once again using $p \geq 20$ and that $\lambda \geq \lambda_*$ we have that,

$$\mathbb{E}[\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_1^2] \leq 300\lambda^2 s_{\boldsymbol{\beta}_0}^2.$$

Assembling, we conclude that,

$$\mathbb{E}[\mathbf{I}[\mathcal{S}](\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0)_i] \leq \mathbb{E}[|\mathbf{Z}_i| + |\boldsymbol{\Delta}_i|] - \min(\lambda, (\boldsymbol{\beta}_0)_i)\underbrace{\Pr[\mathcal{S}]}_{\geq 1/2} \leq \frac{\sigma_\epsilon}{\sqrt{n}} + 300\lambda s_{\boldsymbol{\beta}_0}\sqrt{\frac{\log(2ep)}{n}} - \frac{1}{2}\lambda \leq -\frac{2}{5}\lambda.$$

The last inequality holds using that $\lambda \geq \lambda_*$ and $n \geq c_1 s_{\boldsymbol{\beta}_0}^2 \log(2ep)$ for sufficiently large c_1.

This allows us to conclude that $(\mathbf{v}_\star^\top \mathbb{E}[\mathbf{I}[\mathcal{S}](\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0)_S])^2 \geq c^2\lambda^2\|\mathbf{v}_\star\|_{(s)}^2 \geq c^2\lambda^2$. Finally if we consider a spectral decomposition of $\mathbb{E}[\mathbf{x}_\star \mathbf{x}_\star^\top]_S$ we can conclude that, $\langle \mathbb{E}[\mathbf{I}[\mathcal{S}](\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0)_S], \mathbb{E}[\mathbf{x}_\star \mathbf{x}_\star^\top]_S \mathbb{E}[\mathbf{I}[\mathcal{S}](\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0)_S\rangle \geq \Lambda_s[\mathbb{E}[\mathbf{x}_\star \mathbf{x}_\star^\top]](\mathbf{v}_\star^\top(\mathbb{E}[\mathbf{I}[\mathcal{S}](\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0)_S]))^2$, which yields the desired conclusion after combining with (10). The final inequality in the display, $\Lambda_s[\mathbb{E}[\mathbf{x}_\star \mathbf{x}_\star^\top]] \geq \|\mathbb{E}[\mathbf{x}_\star]\|_{(s)}^2$ follows by Jensen's inequality and the variational characterization of the s-sparse eigenvalues. The claim for fixed deterministic $\mathbf{x}_\star$ follows immediately from this result.

To show tightness of the upper bound for deterministic $\mathbf{x}_\star$, we first apply the Holder inequality on the top-s norm and its dual (see Proposition 10) to see that,

$$\mathbb{E}[\langle \mathbf{x}_\star, \hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\rangle^2] \leq \|\mathbf{x}_\star\|_{(s)}^2 \mathbb{E}\left[\max\left(\frac{\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_1}{s_{\boldsymbol{\beta}_0}}, \|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_\infty\right)^2\right]$$

Since for $a, b \geq 0$, $\max(a, b)^2 \leq 2(a^2 + b^2)$ it suffices to bound the expectation of each term individually. From the previous computations we recall that $\mathbb{E}[\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_1^2] \leq 300\lambda^2 s_{\boldsymbol{\beta}_0}^2$. Finally by appealing to Lemma 4 and similar computations to before, we have that,

$$\mathbb{E}[\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_\infty^2] \leq 30\left(\mathbb{E}[(\|\mathbf{X}^\top \boldsymbol{\epsilon}\|_\infty/n)^2] + \sqrt{\mathbb{E}[\|\boldsymbol{\Sigma}_n - \mathbf{I}_d\|_\infty^4]}\sqrt{\mathbb{E}[\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_1^4]} + (\frac{\lambda}{2})^2\right) \leq$$

$$O(\lambda_*^2) + O((\sqrt{\log(2ep)/n} \cdot \lambda s_{\boldsymbol{\beta}_0})^2) + O(\lambda^2) \leq O(\lambda^2),$$

using once again that $\lambda \geq \lambda_*$ and that $n \geq c_1 s_{\boldsymbol{\beta}_0}^2 \log(2ep)$ for sufficiently large c_1. Recall we define $\mathbf{Z}_i = \frac{1}{n}\mathbf{e}_i^\top \mathbf{X}^\top \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_\epsilon^2 \mathbf{I}_n)$ and $\mathbf{v} = \mathbf{X}\mathbf{e}_i \sim \mathcal{N}(0, \sigma_\epsilon^2 \mathbf{I}_n)$ independently of $\boldsymbol{\epsilon}$. Hence appealing to Lemma 8 and using a union bound,

$$\Pr\left[\max_i |\mathbf{Z}|_i \geq t\right] \leq 2p\exp\left(-\frac{n}{2}\min((t/\kappa)^2, t/\kappa)\right) \implies \mathbb{E}[(\|\mathbf{X}^\top \boldsymbol{\epsilon}\|_\infty/n)^2] \leq O\left(\left(\sigma_\epsilon\sqrt{\frac{\log p}{n}}\right)^2\right) \leq O(\lambda^2)$$

for $\kappa = 8\sigma_\epsilon^2$ by integrating the tail bound using similar computations to before when $n \geq c_1 \log p$ for large-enough constant c_1. Combining these results shows that,

$$\mathbb{E}[\langle \mathbf{x}_\star, \hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\rangle^2] \leq c_3\|\mathbf{x}_\star\|_{(s)}^2 \lambda^2$$

for some large-enough c_3. $\square$

C.2. Corollary 2 and Supporting Lemmas

We now provide a short proof of the supporting corollary.

Proof of Corollary 2. This follows from Theorem 2 since for a fixed $\mathbf{x}_\star$ we have that $\mathbb{E}[\mathbf{x}_\star] = \mathbf{x}_\star$ and $\sup_{\|\mathbf{x}_\star\|_q = 1} \|\mathbf{x}_\star\|_{(s)}^2 \geq s^{2-2/q}$. □

The construction of this lower bound utilizes a support recovery result which requires the following conditions on the sample design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$,

Condition 1. (Lower Eigenvalue on Support). The smallest eigenvalue of the sample covariance sub-matrix indexed by S is bounded below:

$$\sigma_{\min}\left(\frac{\mathbf{X}_S^\top \mathbf{X}_S}{n}\right) \geq c_{\min} > 0$$

Condition 2. (Mutual Incoherence). There exists some $\alpha \in [0, 1)$ such that

$$\max_{j \in S^c} \left\| (\mathbf{X}_S^\top \mathbf{X}_S)^{-1} \mathbf{X}_S^\top \mathbf{X} \mathbf{e}_j \right\|_1 \leq \alpha$$

Condition 3. (Column Normalization). There exists some C such that

$$\max_{j=1,\ldots,p} \|\mathbf{X}\mathbf{e}_j\|_2 / \sqrt{n} \leq C$$

Importantly all of these conditions can be verified w.h.p when $n \gtrsim s_{\boldsymbol{\beta}_0} \log p$ for covariates $\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{I}_p)$ using standard matrix concentration arguments. To state our first lower bound it is also convenient to define $\Pi_{S^\perp}(\mathbf{X}) = \mathbf{I}_n - \mathbf{X}_S(\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\mathbf{X}_S^\top$, which is a type of orthogonal projection matrix.

Given these conditions we can state a conditional (on $\mathbf{X}$) support recovery result,

Proposition 8. *Let Conditions (1), (2) and (3) hold for the sample covariance matrix $\mathbf{X}$, the independent noise distribution be Gaussian, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}_n \sigma_\epsilon^2)$, and Assumption 1 hold (with $s_{\boldsymbol{\beta}_0}$-sparse underlying parameter $\boldsymbol{\beta}_0$). Then, for any choice of regularization parameter $\lambda = \frac{2C\sigma}{1-\alpha}\sqrt{\frac{2\log(p-s_{\boldsymbol{\beta}_0})}{n}} + \delta$ for $\delta > 0$, the support of $\hat{\boldsymbol{\beta}}_L(\lambda)$ is strictly contained in the support of $\boldsymbol{\beta}_0$:*

$$S_{\hat{\boldsymbol{\beta}}_L(\lambda)} \subseteq S_{\boldsymbol{\beta}_0}$$

with probability at least $1 - 4e^{-n\delta^2/2}$.

Proof. Conditions (1) and (2), and the fact that $\lambda \geq \frac{2}{1-\alpha}\left\|\mathbf{X}_{S^c}^\top \Pi_{S^\perp}(\mathbf{X})\frac{\boldsymbol{\epsilon}}{n}\right\|_\infty$ are sufficient show a support recovery result. Under these conditions, for all s-sparse $\boldsymbol{\beta}_0$, there is a unique optimal solution to the Lagrangian Lasso program $\hat{\boldsymbol{\beta}}_L(\lambda)$ and the support of $\hat{\boldsymbol{\beta}}_L(\lambda)$, $S_{\hat{\boldsymbol{\beta}}_L(\lambda)}$, is contained within the support $S_{\boldsymbol{\beta}_0}$ (no false inclusion property) (Wainwright, Theorem 7.21). We can simplify the condition on the regularization parameter from Proposition (Wainwright, Theorem 7.21) using a standard union bound/Gaussian tail bound argument (using Assumption 4) along with the column normalization condition (Condition (3)) to show that $\lambda = \frac{2C\sigma}{1-\alpha}\sqrt{\frac{2\log(p-s_{\boldsymbol{\beta}_0})}{n}} + \delta$ satisfies $\lambda \geq \frac{2}{1-\alpha}\left\|\mathbf{X}_{S^c}^\top \Pi_{S^\perp}(\mathbf{X})\frac{\boldsymbol{\epsilon}}{n}\right\|_\infty$ with probability at least $1 - 4e^{-n\delta^2/2}$ (over the randomness in ϵ) (Wainwright, Corollary 7.22). Combining yields the desired conclusion. □

The aforementioned result holds conditional on $\mathbf{X}$. However, we can verify that Conditions, (1), (2), (3) hold true w.h.p. even if we sample $\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{I}_p)$ (see Lemma 3). Thus, we can show a Lasso prediction error bound that holds in expectation over all the randomness in the training data $(\mathbf{X}, \boldsymbol{\epsilon})$.

To do so we introduce the following standard result showing Conditions (1), (2), (3) can be verified w.h.p. for i.i.d. covariates from $\mathcal{N}(0, \mathbf{I}_p)$.

Lemma 3. *Let* $\mathbf{x}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \mathbf{I}_p)$ *for* $i \in [n]$. *Then there exists a universal constant* c_2, *such that for* $n \geq c_2 s_{\boldsymbol{\beta}_0} \log p$ *and* $p \geq 20$, *Conditions 1, 2, 3 each hold with probability at least* $\frac{99}{100}$.

Proof. The proofs of these follow by standard matrix concentration arguments. Condition (3) can be verified w.h.p. for $C = 1$ (as a function of n) identically to Lemma 9 for $n \gtrsim \log p$. Condition (2) can also be verified w.h.p. for $\alpha = \frac{1}{2}$ for $n \gtrsim s_{\boldsymbol{\beta}_0} \log(p - s_{\boldsymbol{\beta}_0})$, see for example (Wainwright, Ch.7, p.221, Exercise 19). While finally, Condition (1) can also be verified w.h.p. for $c_{\min} = \frac{1}{2}$ when $n \gtrsim s_{\boldsymbol{\beta}_0}$ using standard operator norm bounds for Gaussian ensembles (see for example, (Wainwright, Theorem 6.1, Example 6.3)). $\qquad\square$

Combining Lemma 3 and Proposition 8 yields the desired conclusion which we formalize below.

Proposition 9. *Under Assumption 1, suppose* $\mathbf{x}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \mathbf{I}_p)$ *with independent noise* $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}_n \sigma_\epsilon^2)$. *Then, if* $\hat{\beta}_L(\lambda)$ *denotes the solution of the Lasso program, with regularization parameter chosen as* $\lambda \geq 8\sigma_\epsilon \sqrt{\log p / n}$, *there exists a universal constant* c_1 *such that for all* $n \geq c_1 s_{\boldsymbol{\beta}_0} \log p$,

$$S_{\hat{\boldsymbol{\beta}}_L(\lambda)} \subseteq S_{\boldsymbol{\beta}_0}$$

with probability at least $\frac{1}{2}$.

Proof. The proof follows using the independence of $\boldsymbol{\epsilon}$ and $\mathbf{X}$, by combining the results of Proposition 8 and Lemma 3 with a union bound (and taking n sufficiently large). $\qquad\square$

We next state a useful supremum norm bound applicable to the Lasso under random design from van de Geer (2014a),

Lemma 4 (Lemma 2.5.1 in van de Geer (2014a)). *Under Assumption 1, if* $\hat{\beta}_L(\lambda)$ *denotes the solution of the Lasso program, with regularization parameter chosen as* λ,

$$\left\| \hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0 \right\|_\infty \leq \left\| \boldsymbol{\Omega} \mathbf{X}^\top \boldsymbol{\epsilon} \right\|_\infty / n + \|\boldsymbol{\Omega}\|_1 \left(\left\| \boldsymbol{\Sigma}_n - \mathbf{I}_d \right\|_\infty \left\| \hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0 \right\|_1 + \frac{\lambda}{2} \right)$$

Finally, we state a useful (and standard fact) from convex analysis.

Proposition 10. *If* $\|\mathbf{x}\|_{(k)}$ *denotes the top-k norm, the sum of the magnitudes of the s largest magnitude entries of* $\mathbf{x}$, *then its dual norm is* $\|\mathbf{x}\|_{(k),*} = \max(\|\mathbf{x}\|_1 / k, \|\mathbf{x}\|_\infty)$.

D. Proofs for Section 3.1: Javanmard-Montanari (JM)-style Estimator

In this section we provide the proof of the prediction risk bounds for the JM-style estimator.

D.1. Theorem 3

We provide the proof of Theorem 3.

Proof of Theorem 3. Recall that we will use $r_{\boldsymbol{\beta},1} = (\mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}}[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1^4])^{1/4}$. This estimator admits the error decomposition,

$$\hat{y}_{\text{JM}} - \langle \mathbf{x}_\star, \boldsymbol{\beta}_0 \rangle = \frac{1}{n} \mathbf{w}^\top \mathbf{X}^\top \boldsymbol{\epsilon} + \langle \mathbf{x}_\star - \boldsymbol{\Sigma}_n \mathbf{w}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \rangle$$

and hence,

$$\mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}}[(\hat{y}_{\text{JM}} - \langle \mathbf{x}_\star, \boldsymbol{\beta}_0 \rangle)^2] \leq 2 \left(\mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}}[(\frac{1}{n} \mathbf{w}^\top \mathbf{X}^\top \boldsymbol{\epsilon})^2] + \mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}}[\langle \mathbf{x}_\star - \boldsymbol{\Sigma}_n \mathbf{w}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \rangle^2] \right)$$

The first term can be thought of as the variance contribution while the second is the contribution due to bias. For the variance term, we begin by evaluating the expectation over $\boldsymbol{\epsilon}$. Using independence (w.r.t. to $\mathbf{X}$) and sub-gaussianity of $\boldsymbol{\epsilon}$,

$$\mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}}[(\frac{1}{n} \mathbf{w}^\top \mathbf{X}^\top \boldsymbol{\epsilon})^2] = \frac{1}{n} \mathbb{E}_{\mathbf{X}} \mathbb{E}_{\boldsymbol{\epsilon}}[(\sum_{i=1}^n \mathbf{w}^\top \mathbf{x}_i \epsilon_i)^2 | \mathbf{X}] = \frac{\sigma_\epsilon^2}{n} \mathbb{E}_{\mathbf{X}}[\mathbf{w}^\top \boldsymbol{\Sigma}_n \mathbf{w}]$$

Now using Corollary 6 and defining $\kappa_1' = 8\kappa^2/C_{\min}\|\mathbf{x}_\star\|_2^2$ we have that,

$$\mathbb{E}_{\mathbf{X}}[\mathbf{w}^\top \mathbf{\Sigma}_n \mathbf{w}] \leq \mathbf{x}_\star^\top \mathbf{\Omega} \mathbf{x}_\star + \frac{3\kappa_1'}{\sqrt{n}}.$$

using the condition $n \geq 2$. Turning to the bias term, the Holder and Cauchy-Schwarz inequalities give, $\mathbb{E}_{\mathbf{X},\epsilon}[\langle \mathbf{x}_\star - \hat{\mathbf{\Sigma}}_n \mathbf{w}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \rangle^2] \leq \mathbb{E}_{\mathbf{X},\epsilon}[\left\|\mathbf{x}_\star - \hat{\mathbf{\Sigma}}_n \mathbf{w}\right\|_\infty^2 \left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right\|_1^2] \leq \sqrt{\mathbb{E}_{\mathbf{X}}[\left\|\mathbf{x}_\star - \hat{\mathbf{\Sigma}}_n \mathbf{w}\right\|_\infty^4] \mathbb{E}_{\mathbf{X},\epsilon}[\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right\|_1^4]}.$

We begin by evaluating the first expectation $\mathbb{E}_{\mathbf{X}}[\left\|\mathbf{x}_\star - \hat{\mathbf{\Sigma}}_n \mathbf{w}\right\|_\infty^4]$ which follows from Corollary 6,

$$\sqrt{\mathbb{E}_{\mathbf{X}}[\left\|\mathbf{x}_\star - \hat{\mathbf{\Sigma}}_n \mathbf{w}\right\|_\infty^4]} \leq \lambda_{\mathbf{w}}^2 + \sqrt{2}\|\mathbf{x}_\star\|_\infty^2 (p \vee n)^{-c_3}$$

for $n \geq a^2 \log(p \vee n)$ and $c_3 = a^2/4 - \frac{1}{2}$ with $\kappa_2' = 8\kappa^2 \sqrt{C_{\text{cond}}} \|\mathbf{x}_\star\|_2$. By definition of the base estimation procedure we can assemble to obtain the desired error is bounded by,

$$\leq O(\frac{\sigma_\epsilon^2 \mathbf{x}_\star \mathbf{\Omega} \mathbf{x}_\star}{n} + \frac{\sigma_\epsilon^2 \kappa_1'}{n^{3/2}} + r_{\beta,1}^2 ((\lambda_{\mathbf{w}}^2 + \|\mathbf{x}_\star\|_\infty^2 (p \vee n)^{-c_3})))$$

where $\lambda_{\mathbf{w}} = ak^2 \kappa_2' \sqrt{\frac{\log(p\vee n)}{n}}$.

For the second claim note by Corollary 6, that $\mathbf{w} = 0$ and hence we can write the error of the estimator as,

$$\hat{y}_{\text{JM}} - \langle \mathbf{x}_\star, \boldsymbol{\beta}_0 \rangle = \langle \mathbf{x}_\star, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \rangle \implies \mathbb{E}_{\mathbf{X},\epsilon}[(\hat{y}_{\text{JM}} - \langle \mathbf{x}_\star, \boldsymbol{\beta}_0 \rangle)^2] = \mathbb{E}_{\mathbf{X},\epsilon}[\langle \mathbf{x}_\star, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \rangle^2].$$

$\square$

We can now instantiate the result of the previous theorem in the setting where the Lasso estimator is used as the base-regression procedure.

D.2. Corollary 3 and Supporting Lemmas

We provide the proof of Corollary 3.

Proof of Corollary 3. The second expectation $\mathbb{E}_{\mathbf{X},\epsilon}[\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_1^4]$ can be evaluated using Lemmas 13 and 15 from which we find,

$$r_{\beta,1}^2 = \sqrt{\mathbb{E}_{\mathbf{X},\epsilon}[\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_1^4]} \leq O\left(\frac{\lambda_{\boldsymbol{\beta}_0} s_{\boldsymbol{\beta}_0}}{C_{\min}}\right)^2 + O\left(\frac{\sigma_\epsilon}{\sqrt{n}}\right)^2 + O\left(\left(\frac{\sigma_\epsilon^4}{\lambda_{\boldsymbol{\beta}}^2} + \|\boldsymbol{\beta}_0\|_1^2\right)\left(e^{-\frac{c}{4\kappa^4} n}\right)\right)$$

Assuming $p \geq 20$ and $n \geq c_2 \frac{\kappa^4}{C_{\min}} s \log(2ep)$, there exists sufficiently large c_2 such that $\left(\frac{\sigma_\epsilon^4}{\lambda_{\boldsymbol{\beta}}^2} + \|\boldsymbol{\beta}_0\|_1^2\right)\left(\sqrt{2}e^{-n\frac{c}{4\kappa^4}}\right) \leq O(\frac{\sigma_\epsilon^2}{n} + \|\boldsymbol{\beta}_0\|_1^2 e^{-nc/(4\kappa^4)}) \leq O(\frac{\sigma_\epsilon^2}{n})$ since $\|\boldsymbol{\beta}_0\|_\infty/\sigma_\epsilon = o(e^{c_1 n})$ for some sufficiently small c_1. Thus we have $r_{\beta,1}^2 \leq O\left(\frac{\lambda_{\boldsymbol{\beta}}^2 s_{\boldsymbol{\beta}_0}^2}{C_{\min}^2} + \frac{\sigma_\epsilon^2}{n}\right) = O(\frac{\lambda_{\boldsymbol{\beta}}^2 s_{\boldsymbol{\beta}_0}^2}{C_{\min}^2})$ due to the lower bound on $\lambda_{\boldsymbol{\beta}}$. Combining with Theorem 3 gives the result,

$$O\left(\frac{\sigma_\epsilon^2 \mathbf{x}_\star \mathbf{\Omega} \mathbf{x}_\star}{n} + \left(\frac{\lambda_{\boldsymbol{\beta}}^2 s_{\boldsymbol{\beta}_0}^2}{C_{\min}^2}\right)(\lambda_{\mathbf{w}}^2 + \|\mathbf{x}_\star\|_\infty^2 (p \vee n)^{-c_3})\right)$$

$\square$

We now connect our results to the problem of constructing CIs in sparse linear regression – namely the results in Cai & Guo (2017). We first define formally what it means for a set S to be a $1 - \alpha$ CI in this context – namely that for all $\boldsymbol{\beta}_0$, $\liminf_{n,p\to\infty} \Pr_{\boldsymbol{\beta}_0}[\mathbf{x}_\star^\top \boldsymbol{\beta}_0 \in S] \geq 1 - \alpha$.

Proof of Proposition 4. Before beginning, we first recall the tail bound in Bellec et al. (2016, Theorem 4.2), which provides that,

$$\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_q \leq \frac{49}{8}\left(\frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))} \vee \frac{1}{\phi_0^2}\right)\lambda s^{1/q}$$

with probability at least $1 - \delta_0/2$, where $\delta(\lambda) = \exp\left(-\left(\frac{\lambda\sqrt{n}}{(8+2\sqrt{2})\sigma}\right)\right)$ for all design matrices in $\mathbf{X} \in \mathcal{E}_n(s, 7)$ where $\phi_0^2 = \phi_{SRE}^2(s, 7)$. Note by Theorem 11 we have that under our design assumptions $\mathbf{X} \in \mathcal{E}_n(s, 7)$ with probability at least $1 - 3\exp\left(-cn/\kappa^4\right)$ for $n \gtrsim s\log p$. Hence taking $q = 1$ and $\lambda \asymp \sqrt{\log p/n}$, $s\log(1/\delta(\lambda)) \asymp s\exp\left(-c\sqrt{\log p}\right) \asymp \exp\left(\gamma\log p - c\sqrt{\log p}\right)$ for $0 \leq \gamma < \frac{1}{2}$. Hence, for $\delta_0 \asymp p^{-\gamma/2}$, $\frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))} \to 0$. Accordingly, for sufficiently large p, we

$$\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_1 \leq K_1 s\sqrt{\frac{\log p}{n}}$$

with probability at least $1 - O(\exp(-cn)) - O(p^{-\gamma/2})$. Define the set $S_1 = [\mathbf{x}_\star^\top\hat{\boldsymbol{\beta}}_L(\lambda) + \|\mathbf{x}_\star\|_\infty K s_{\boldsymbol{\beta}_0}\sqrt{\frac{\log p}{n}}, \mathbf{x}_\star^\top\hat{\boldsymbol{\beta}}_L(\lambda) - \|\mathbf{x}_\star\|_\infty K s_{\boldsymbol{\beta}_0}\sqrt{\frac{\log p}{n}}]$ for future reference.

In the case of the dense loading regime we have that $\frac{\|\mathbf{x}_\star\|_\infty}{\|\mathbf{x}_\star\|_2} \asymp p^{-\gamma_q/2}$, and take $\lambda_\mathbf{w} = 8\sqrt{C_{\text{cond}}}\kappa^2\frac{1}{s\sqrt{\log p}}\|\mathbf{x}_\star\|_2$. This choice of satisfies $\lambda_\mathbf{w} \asymp \underbrace{\frac{p^{\gamma_q/2-\gamma}}{\sqrt{\log p}}}_{\to\infty}\|\mathbf{x}_\star\|_\infty$. Hence by the definition of the JM program, for sufficiently large p, its minimizer is $\mathbf{w} = 0$ almost surely as argued in the proof of Theorem 3 – in which case $\hat{y}_{\text{JM}} = \mathbf{x}_\star^\top\hat{\boldsymbol{\beta}}_L(\lambda)$ almost surely. Hence in this regime, $S_1 = [\hat{y}_{\text{JM}} + \|\mathbf{x}_\star\|_\infty K_1 s_{\boldsymbol{\beta}_0}\sqrt{\frac{\log p}{n}}, \hat{y}_{\text{JM}} - \|\mathbf{x}_\star\|_\infty K_1 s_{\boldsymbol{\beta}_0}\sqrt{\frac{\log p}{n}}]$ provides valid coverage by the previous arguments.

To show the second claim consider the set $S_2 = [\hat{y}_{\text{JM}} + 1.01/\sqrt{n}z_{\alpha/2}\|\mathbf{x}_\star\|_2\sqrt{\mathbf{w}^\top\boldsymbol{\Sigma}_n\mathbf{w}} + \sqrt{n}, \hat{y}_{\text{JM}} + 1.01/\sqrt{n}z_{\alpha/2}\|\mathbf{x}_\star\|_2\sqrt{\mathbf{w}^\top\boldsymbol{\Sigma}_n\mathbf{w}} + K_2/\sqrt{n}]$, and note fom the proof of Theorem 3 we can see

$$|\hat{y}_{\text{JM}} - \langle\mathbf{x}_\star, \boldsymbol{\beta}_0\rangle| = \left|\frac{1}{n}\mathbf{w}^\top\mathbf{X}^\top\boldsymbol{\epsilon} + \langle\mathbf{x}_\star - \boldsymbol{\Sigma}_n\mathbf{w}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\rangle\right|$$

using the results in therein that $\left|\langle\mathbf{x}_\star - \boldsymbol{\Sigma}_n\mathbf{w}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\rangle\right| \leq \left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right\|_1\|\mathbf{x}_\star - \boldsymbol{\Sigma}_n\mathbf{w}\|_\infty \lesssim s\sqrt{\log p/n}\cdot\|\mathbf{x}_\star\|_2\frac{1}{s\sqrt{\log p}} \leq K_2\|\mathbf{x}_\star\|_2\frac{1}{\sqrt{n}}$ with probability at least $1 - O(\exp(-cn) - O(p^{-1})$ with the aforementioned choice of a in the regime $n \gtrsim s^2(\log p)^2$ (which implies $a \gtrsim 1$). Conditionally on $\mathbf{X}$ we then have that $\frac{1}{n}\mathbf{w}^\top\mathbf{X}^\top\boldsymbol{\epsilon} \sim \mathcal{N}(0, \frac{1}{n}\mathbf{w}^\top\boldsymbol{\Sigma}_n\mathbf{w})$. Combining these results with a union bound show thats $\liminf_{n,p}\Pr\left[\mathbf{x}_\star^\top\boldsymbol{\beta}_0 \in S_2\right] \to 1 - \alpha$ with as $n, p \to \infty$. Finally, since by Lemma 6 we have that $\sqrt{\mathbf{w}^\top\boldsymbol{\Sigma}_n\mathbf{w}} \leq 1.01\sqrt{\mathbf{x}_\star^\top\boldsymbol{\Omega}\mathbf{x}_\star}$ with probability at least $1 - \exp(-cn)$, and Corollary 6, we $\mathbb{E}[\mathbf{w}^\top\boldsymbol{\Sigma}_n\mathbf{w}] \leq \mathbf{x}_\star\boldsymbol{\Omega}\mathbf{x}_\star + O(\frac{1}{\sqrt{n}})$ we can see that in the regime $n \gtrsim s^2(\log p)^2$ the interval S_2 indeed has expected length $O(\frac{\|\mathbf{x}_\star\|_2}{\sqrt{n}})$ which is optimal in this regime.

$\square$

Here we collect several useful lemmas which follow from standard concentration arguments useful both in the analysis of the upper bound on the JM estimator and in the Lasso lower bound.

To begin we show the convex program defining the JM estimator is feasible with high probability. For convenience we define the event $\mathcal{F}(a)$ to be the event that the convex program defining the JM estimator in (4) with choice of regularization parameter $\lambda_\mathbf{w} = a\sqrt{\log p/n}$ possesses $\mathbf{w}_0 = \boldsymbol{\Omega}\mathbf{x}_\star$ as a feasible point.

Lemma 5. *Let Assumption 2 and 3 hold for the design $\mathbf{X}$ and assume $n \geq a^2\log(p \vee n)$ with $\kappa_2' = 8\kappa^2\sqrt{C_{cond}}\|\mathbf{x}_\star\|_2$. If $\mathbf{x}_\star \in \mathbb{R}^p$ then for $\mathbf{w}_0 = \boldsymbol{\Omega}\mathbf{x}_\star$,*

$$\Pr\left[\left\|\hat{\boldsymbol{\Sigma}}_n\mathbf{w}_0 - \mathbf{x}_\star\right\|_\infty \geq a\kappa_2'\sqrt{\log(p \vee n)/n}\right] \leq 2(p \vee n)^{-c_2}$$

for $c_2 = \frac{a^2}{2} - 1$. Hence the convex program in (4) with regularization parameter $\lambda_\mathbf{w} = a\kappa_2'\sqrt{\frac{\log(p\vee n)}{n}}$ admits $\mathbf{w}_0$ as a feasible point with probability at least $1 - 2(p \vee n)^{-c_2}$.

Proof. This follows from a standard concentration argument for sub-exponential random variables. Throughout we will use $\tilde{\mathbf{x}}_\ell = \mathbf{\Omega}^{1/2}\mathbf{x}_\ell$. Consider some $j \in [p]$ and define $z_\ell^j = \mathbf{e}_j^\top \mathbf{\Omega}^{1/2}\tilde{\mathbf{x}}_\ell \cdot \tilde{\mathbf{x}}_\ell^\top \mathbf{\Sigma}^{1/2}\mathbf{x}_\star - \mathbf{e}_j^\top\mathbf{x}_\star$ which satisfies $\mathbb{E}[z_\ell^j] = 0$, is independent over $\ell \in [n]$, and for which $\mathbf{e}_j^\top(\hat{\mathbf{\Sigma}}_n\mathbf{w}_0 - \mathbf{x}_\star) = \frac{1}{n}\sum_{j=1}^n z_j^\ell$. Since $\mathbf{e}_j^\top\mathbf{\Omega}^{1/2}\tilde{\mathbf{x}}_\ell \sim \mathrm{sG}(\kappa\|\mathbf{e}_j^\top\mathbf{\Omega}^{1/2}\|_2) \sim \mathrm{sG}(\kappa/\sqrt{C_{\min}})$, and $(\mathbf{x}_\star)^\top\mathbf{\Sigma}^{1/2}\tilde{\mathbf{x}}_\ell \sim \mathrm{sG}(\kappa\|\mathbf{\Sigma}^{1/2}\mathbf{x}_\star\|_2) \sim \mathrm{sG}(\kappa\sqrt{C_{\max}}\|\mathbf{x}_\star\|_2)$, z_ℓ^j is a mean-zero $\mathrm{sE}(8\kappa^2\sqrt{C_{\mathrm{cond}}}\|\mathbf{x}_\star\|_2, 8\kappa^2\sqrt{C_{\mathrm{cond}}}\|\mathbf{x}_\star\|_2)$ r.v. by Lemma 8. Defining $\kappa_2' = 8\kappa^2\sqrt{C_{\max}/C_{\min}}\|\mathbf{x}_\star\|_2$, applying the tail bound for sub-exponential random variables, and taking a union bound over the p coordinates implies that,

$$\Pr[\|\mathbf{\Sigma}_n\mathbf{w}_0 - \mathbf{x}_\star\|_\infty \geq t] \leq \Pr[\|\mathbf{\Sigma}_n\mathbf{w}_0 - \mathbf{x}_\star\|_\infty \geq t] \leq 2p\exp\left[-\frac{n}{2}\min((t/\kappa_2')^2, t/\kappa_2'))\right].$$

Choosing $t = a\kappa_2'\sqrt{\log(p \vee n)/n}$, assuming $n \geq a^2\log(p \vee n)$, gives the conclusion

$$\Pr\left[\|\mathbf{\Sigma}_n\mathbf{w}_0 - \mathbf{x}_\star\|_\infty \geq a\kappa_2'\sqrt{\log(p \vee n)/n}\right] \leq 2(p \vee n)^{-a^2/2+1}$$

and the conclusion follows. $\square$

We can now provide a similar concentration argument to bound the objective of the JM program.

Lemma 6. *Let Assumption 2 and 3 hold for the design* $\mathbf{X}$. *Let* $\mathbf{w}$ *be the solution of the convex program in* (4) *with regularization parameter set as* $\lambda_\mathbf{w}$. *If* $\mathbf{x}_\star \in \mathbb{R}^p$, *then,*

$$\Pr\left[\mathbf{w}^\top\mathbf{\Sigma}_n\mathbf{w} \geq \mathbf{x}_\star\mathbf{\Omega}\mathbf{x}_\star + t\right] \leq 2\exp\left[-n/2\min((t/\kappa_1')^2, t/\kappa_1')\right]$$

for $\kappa_1'^2 = 8\kappa^2/C_{\min}\|\mathbf{x}_\star\|_2^2$.

Proof. The argument once again follows from a standard concentration argument for sub-exponential random variables. Considering,

$$(\mathbf{x}_\star\mathbf{\Omega})^\top\mathbf{\Sigma}_n\mathbf{\Omega}\mathbf{x}_\star = [(\mathbf{x}_\star\mathbf{\Omega})^\top\mathbf{\Sigma}_n\mathbf{\Omega}\mathbf{x}_\star - \mathbf{x}_\star\mathbf{\Omega}\mathbf{x}_\star] + \mathbf{x}_\star\mathbf{\Omega}\mathbf{x}_\star = \frac{1}{n}\sum_{j=1}^n (z_j^2 - \mathbf{x}_\star\mathbf{\Omega}\mathbf{x}_\star) + \mathbf{x}_\star\mathbf{\Omega}\mathbf{x}_\star$$

where $z_j = \mathbf{x}_\star^\top\mathbf{\Omega}\mathbf{x}_j$ is mean-zero with $s_j \sim \mathrm{sG}(\kappa\|\mathbf{\Omega}^{1/2}\mathbf{x}_\star\|_2) \sim \mathrm{sG}(\kappa/\sqrt{C_{\min}}\|\mathbf{x}_\star\|_2)$. Since $\mathbb{E}[z_j^2] = \mathbf{x}_\star\mathbf{\Omega}\mathbf{x}_\star$, Lemma 8 implies $z_j^2 - \mathbf{x}_\star\mathbf{\Omega}\mathbf{x}_\star \sim \mathrm{sE}(8\kappa^2/C_{\min}\|\mathbf{x}_\star\|_2^2, 8\kappa^2/C_{\min}\|\mathbf{x}_\star\|_2^2)$ and is mean-zero. The sub-exponential tail bound gives,

$$\Pr\left[\frac{1}{n}\sum_{\ell=j}^n z_j^2 \geq \mathbf{x}_\star\mathbf{\Omega}\mathbf{x}_\star + t\right] \leq \exp\left[-n/2\min((t/\kappa_1')^2, t/\kappa_1')\right]$$

where $\kappa_1' = 8\kappa^2/C_{\min}\|\mathbf{x}_\star\|_2^2$. Hence, since on the event $\mathcal{F}(a)$, we have that $\mathbf{w}^\top\mathbf{\Sigma}_n\mathbf{w} \leq (\mathbf{x}_\star\mathbf{\Omega})^\top\mathbf{\Sigma}_n\mathbf{\Omega}\mathbf{x}_\star$ (recall $\mathbf{w}_0 = \mathbf{\Omega}\mathbf{x}_\star$ is feasible on $\mathcal{F}(a)$),

$$\Pr\left[\mathbf{w}^\top\mathbf{\Sigma}_n\mathbf{w} \geq \mathbf{x}_\star\mathbf{\Omega}\mathbf{x}_\star + t\right] \leq \Pr\left[\{\mathbf{w}^\top\mathbf{\Sigma}_n\mathbf{w} \geq \mathbf{x}_\star\mathbf{\Omega}\mathbf{x}_\star + t\} \cap \mathcal{F}(a)\right] + \Pr\left[\{\mathbf{w}^\top\mathbf{\Sigma}_n\mathbf{w} \geq \mathbf{x}_\star\mathbf{\Omega}\mathbf{x}_\star + t\} \cap \mathcal{F}(a)^c\right]$$

$$\leq \Pr\left[\frac{1}{n}\sum_{\ell=j}^n z_j^2 \geq \mathbf{x}_\star\mathbf{\Omega}\mathbf{x}_\star + t\right] + 0 \leq \exp\left[-n/2\min((t/\kappa_1')^2, t/\kappa_1'))\right],$$

since by definition on the event $\mathcal{F}(a)^c$ the convex program outputs $\mathbf{w} = \mathbf{0}$ and $\mathbf{x}_\star\mathbf{\Omega}\mathbf{x}_\star \geq 1/C_{\max} > 0$. $\square$

Finally we can easily convert these tail bounds into moment bounds,

Corollary 6. *Let Assumption 2 and 3 hold for the design* $\mathbf{X}$. *Let* $\mathbf{w}$ *be the solution of the convex program in* (4) *with regularization parameter set as* $\lambda_\mathbf{w} = a\kappa_2'\sqrt{\frac{\log(p\vee n)}{n}}$. *If* $\mathbf{x}_\star \in \mathbb{R}^p$, *then,*

$$\mathbb{E}[\mathbf{w}^\top\hat{\mathbf{\Sigma}}_n\mathbf{w}] \leq \mathbf{x}_\star\mathbf{\Omega}\mathbf{x}_\star + \frac{3\kappa_1'}{\sqrt{n}}$$

for $\kappa_1' = 8\kappa^2/C_{\min}\|\mathbf{x}_\star\|_2^2$ *and assuming* $n \geq a^2\log(p \vee n)$,

$$\sqrt{\mathbb{E}[\left\|\hat{\boldsymbol{\Sigma}}_n\mathbf{w} - \mathbf{x}_\star\right\|_\infty^4]} \leq \lambda_\mathbf{w}^2 + \sqrt{2}\|\mathbf{x}_\star\|_\infty^2(p \vee n)^{-c_2}$$

for $c_2 = a^2/4 - 1/2$ *with* $\kappa_2' = 8\kappa^2\sqrt{C_{cond}}\|\mathbf{x}_\star\|_2$. *Moreover if* $\lambda_\mathbf{w} \geq \|\mathbf{x}_\star\|_\infty$ *then* $\mathbf{w} = 0$ *almost surely.*

Proof. Using Lemma 6 we have that,

$$\mathbb{E}[\mathbf{w}^\top\hat{\boldsymbol{\Sigma}}_n\mathbf{w}] = \mathbf{x}_\star\boldsymbol{\Omega}\mathbf{x}_\star + \int_0^\infty \Pr[\mathbf{w}^\top\boldsymbol{\Sigma}_n\mathbf{w} \geq \mathbf{x}_\star\boldsymbol{\Omega}\mathbf{x}_\star + t]$$

$$\leq \mathbf{x}_\star\boldsymbol{\Omega}\mathbf{x}_\star + \int_0^{\kappa_1'}[\exp[-n/2(t/\kappa_1')^2]dt + \int_{\kappa_1'}^\infty[\exp[-n/2(t/\kappa_1')]dt$$

$$\leq \mathbf{x}_\star\boldsymbol{\Omega}\mathbf{x}_\star + \frac{2\kappa_1'}{\sqrt{n}} + \frac{2\kappa_1'e^{-n/2}}{n} \leq \mathbf{x}_\star\boldsymbol{\Omega}\mathbf{x}_\star + \frac{3\kappa_1'}{\sqrt{n}}$$

which holds for $n \geq 2$.

Similarly, directly applying Lemma 5 we obtain,

$$\mathbb{E}[\left\|\hat{\boldsymbol{\Sigma}}_n\mathbf{w} - \mathbf{x}_\star\right\|_\infty^4] = \mathbb{E}[\left\|\hat{\boldsymbol{\Sigma}}_n\mathbf{w} - \mathbf{x}_\star\right\|_\infty^4\mathbb{1}[\mathcal{F}(a)]] + \mathbb{E}[\left\|\hat{\boldsymbol{\Sigma}}_n\mathbf{w} - \mathbf{x}_\star\right\|_\infty^4\mathbb{1}[\mathcal{F}^c(a)]] \leq$$

$$\lambda_\mathbf{w}^4 + \|\mathbf{x}_\star\|_\infty^4\Pr[\mathcal{F}^c(a)] \leq \lambda_\mathbf{w}^4 + 2\|\mathbf{x}_\star\|_\infty^4(p \vee n)^{-c_2}$$

$c_2 = a^2/2 - 1$, since the convex program outputs $\mathbf{w} = 0$ on the event $\mathcal{F}^c(a)$. The first conclusion follows using subadditivity of $\sqrt{\cdot}$.

For the second statement note the convex program in (4) always admits $\mathbf{w} = 0$ as a feasible point under the condition $\lambda_\mathbf{w} \geq \|\mathbf{x}_\star\|_\infty$, in which case $\mathbf{w} = 0$ is a global minima of the objective since $\hat{\boldsymbol{\Sigma}}_n$ is p.s.d. $\qquad\square$

E. Proofs for Section 3.2: Orthogonal Moment Estimators

We begin by providing the consistency proofs for the orthogonal moment estimators introduced in Section 3.2. However, first we make a remark which relates the assumptions on the design we make to the properties of the noise variable η.

Remark 1. Under the random design assumption on $\mathbf{x}$, if we consider $\mathbf{x}' = [t, \mathbf{z}] = (\mathbf{U}^{-1})^\top\mathbf{x}$, then by Assumption 3, $\mathbf{g}_0 = \arg\min_g \mathbb{E}_\mathbf{X}[(t - \mathbf{z}^\top\mathbf{g}_0)^2]$ can be thought of as the best linear approximator interpreted in the regression framework. Hence it can also be related to the precision matrix and residual variance as:

$$\boldsymbol{\Omega}_{t,\cdot} = \frac{(1, -\mathbf{g}_0)}{\sigma_\eta^2}.$$

In this setting, we have that $\mathbb{E}[\eta^2] = \boldsymbol{\Sigma}_{tt} - \mathbf{g}_0^\top\boldsymbol{\Sigma}_{zz}\mathbf{g}_0 \geq 0$. Moreover from the variational characterization of the minimum eigenvalue we also have that $\mathbb{E}[\eta^2] \geq C_{\min}/\|\mathbf{x}_\star\|_2$. Thus $\|\mathbf{g}_0\|_2^2 \leq \frac{\boldsymbol{\Sigma}_{tt}}{C_{\min}} \leq C_{cond}/\|\mathbf{x}_\star\|_2$ and $\mathbb{E}[\eta^2] \leq \boldsymbol{\Sigma}_{tt} \leq C_{\max}/\|\mathbf{x}_\star\|_2$. Moreover, the treatment noise η is also a sub-Gaussian random variable, since $\eta = t - \mathbf{z}^\top\mathbf{g}_0 = (1, -\mathbf{g}_0)^\top\mathbf{x}'$. Recall by Assumption 2 that $\mathbb{E}[(\mathbf{x}^\top\mathbf{v})^p] \leq \kappa^{2p}\|\boldsymbol{\Sigma}^{1/2}\mathbf{v}\|_2^{2p}$ while $\eta = (1, -\mathbf{g}_0)^\top\mathbf{x}'$. Thus we have that $\mathbb{E}[\eta^{2p}] = \kappa^{2p}C_{\max}^p(1 + \|\mathbf{g}_0\|_2^2)^p/\|\mathbf{x}_\star\|_2^{2p} \leq O((\kappa^2 C_{cond}C_{\max}/\|\mathbf{x}_\star\|_2^2)^p)$. Similarly $\mathbb{E}[(\mathbf{z}^\top\mathbf{g}_0)^{2p}] \leq (\kappa^2\|\mathbf{g}_0\|_2^2 C_{\max}/\|\mathbf{x}_\star\|_2^2)^p$.

E.1. Theorem 5

We now present the Proof of Theorem 5.

Proof of Theorem 5. To begin we rescale the $\mathbf{x}_\star$ such that is has unit-norm (and restore the scaling in the final statement of the proof). In order to calculate the mean-squared error of our prediction $\mathbb{E}[(\hat{y}_{\text{OM}} - \mathbf{x}_\star^\top\boldsymbol{\beta}_0)^2]$, it is convenient to organize the calculation in an error expansion in terms of the moment function m. For convenience we define the following (held-out) prediction errors $\boldsymbol{\Delta}_f(\mathbf{z}_i) = \mathbf{z}_i^\top(\hat{\mathbf{f}} - \mathbf{f}_0)$, and $\boldsymbol{\Delta}_g(\mathbf{z}_i) = \hat{\mathbf{g}}(\mathbf{z}_i) - \mathbf{g}_0(\mathbf{z}_i)$ of $\mathbf{f}$ and $\mathbf{g}(\cdot)$ which are trained on first-stage

data but evaluated against the second-stage data. Note that as assumed in the Theorem, $\mathbf{g}_0(\mathbf{z}) = \mathbf{z}^\top \mathbf{g}_0$. Also note the moment equations only depend on $\mathbf{f}$ and $\mathbf{g}(\cdot)$ implicitly through the evaluations $\mathbf{z}^\top \mathbf{f}$ and $\mathbf{g}(\mathbf{z})$, so derivatives of the moment expressions with respect to $\mathbf{z}^\top \mathbf{f}$ and $\mathbf{g}(\mathbf{z})$, refer to derivatives with respect to scalar. Recall the sums of the empirical moment equation here only range over the second fold of data, while $\hat{\mathbf{f}}$ and $\hat{\mathbf{g}}$ are fit on the first fold. The empirical moment equations can be expanded (exactly) as,

$$\underbrace{\frac{1}{n/2}\sum_{i=1}^{n/2} \nabla_\theta m(t_i, y_i, \theta_0, \mathbf{z}_i^\top \hat{\mathbf{f}}, \hat{\mathbf{g}}(\mathbf{z}_i))(\theta_0 - \hat{y}_{\text{OM}})}_{J} = \frac{1}{n/2}\sum_{i=1}^{n/2} m(t_i, y_i, \theta_0, \mathbf{z}_i^\top \hat{\mathbf{f}}, \hat{\mathbf{g}}(\mathbf{z}_i))$$

since by definition $\frac{1}{n/2}\sum_{i=1}^{n/2} m(t_i, y_i, \hat{y}_{\text{OM}}, \mathbf{z}_i^\top \hat{\mathbf{f}}, \hat{\mathbf{g}}(\mathbf{z}_i)) = 0$. Then we further have that,

$$\frac{1}{n/2}\sum_{i=1}^{n/2} m(t_i, y_i, \theta_0, \mathbf{z}_i^\top \hat{\mathbf{f}}, \hat{\mathbf{g}}(\mathbf{z}_i)) = \underbrace{\frac{1}{n/2}\sum_{i=1}^{n/2} \nabla m(t_i, y_i, \theta_0, \mathbf{z}_i^\top \mathbf{f}_0, \mathbf{g}_0(\mathbf{z}_i))}_{A} +$$

$$\underbrace{\frac{1}{n/2}\sum_{i=1}^{n/2} \nabla_{\mathbf{z}^\top \mathbf{f}} m(t_i, y_i, \theta_0, \mathbf{z}_i^\top \mathbf{f}_0, \mathbf{g}_0(\mathbf{z}_i))^\top (\mathbf{\Delta}_f)}_{B_1} +$$

$$\underbrace{\frac{1}{n/2}\sum_{i=1}^{n/2} \nabla_{\mathbf{g}(\mathbf{z})} m(t_i, y_i, \theta_0, \mathbf{z}_i^\top \mathbf{f}_0, \mathbf{g}_0(\mathbf{z}_i))^\top (\mathbf{\Delta}_g)}_{B_2} +$$

$$\underbrace{\frac{1}{n/2}\sum_{i=1}^{n/2} \nabla_{\mathbf{z}^\top \mathbf{f}, \mathbf{g}(\mathbf{z})} m(t_i, y_i, \theta_0, \mathbf{z}_i^\top \mathbf{f}_0, \mathbf{g}_0(\mathbf{z}_i))[\mathbf{\Delta}_f, \mathbf{\Delta}_g]}_{C}$$

We first turn to controlling the moments of A, B_1, B_2, C. We use as convenient shorthand $\zeta = \kappa^2 C_{\max}$. Similarly we also use $r_{f,2} = (\mathbb{E}[\mathbf{\Delta}_f(\mathbf{z})^4])^{1/4}$.

1. For $A = \frac{1}{n/2}\sum_{i=1}^{n/2} \eta_i \epsilon_i$, note that $\mathbb{E}[m(t_i, y_i, \theta_0, \mathbf{z}_i^\top \mathbf{f}_0, \mathbf{g}_0(\mathbf{z}_i))|\mathbf{z}_i] = 0$ so it follows that,

$$\mathbb{E}[A^2] = O(\frac{1}{n}\mathbb{E}[\eta^2 \epsilon^2]) = \frac{1}{n}\sigma_\epsilon^2 \sigma_\eta^2$$

2. For $B_1 = \frac{1}{n/2}\sum_{i=1}^{n/2} \mathbf{\Delta}_f(\mathbf{z}_i)\eta_i$. Note $\mathbb{E}[\nabla_{\mathbf{z}^\top \mathbf{f}} m(t_i, y_i, \theta_0, \mathbf{z}_i^\top \mathbf{f}_0, \mathbf{g}_0(\mathbf{z}_i))|\mathbf{z}_i] = 0$ since $\mathbb{E}[\eta_i|\mathbf{z}_i] = 0$. So we have using sub-gaussianity of the random vector $\mathbf{z}$, sub-gaussianity of η and independence that,

$$\mathbb{E}[B_1^2] = O(\frac{1}{n}\mathbb{E}[(\mathbf{\Delta}_f(\mathbf{z}))^2 \eta^2]) \leq O(\frac{1}{n} r_{f,2}^2 \sigma_\eta^2)$$

3. For $B_2 = \sum_{i=1}^{n} \mathbf{\Delta}_g(\mathbf{z}_i)\epsilon_i$. Note $\mathbb{E}[\nabla_{\mathbf{g}(\mathbf{z})} m(t_i, y_i, \theta_0, \mathbf{z}_i^\top \mathbf{f}_0, \mathbf{g}_0(\mathbf{z}_i))|\mathbf{z}_i] = 0$ using independence of ϵ_i and the fact $\mathbb{E}[\epsilon_i] = 0$. Once again using independence,

$$\mathbb{E}[B_2^2] = \frac{1}{n}\mathbb{E}[\epsilon^2(\mathbf{\Delta}_g(\mathbf{z}))^2] \leq O(\frac{1}{n}\sigma_\epsilon^2 r_g^2)$$

4. For $C = \frac{1}{n}\sum_{i=1}^{n} \mathbf{\Delta}_g(\mathbf{z}_i)\mathbf{\Delta}_f(\mathbf{z}_i)$. Note that in general for the remainder term $\mathbb{E}[\nabla_{\mathbf{z}^\top \mathbf{f}, \mathbf{g}(\mathbf{z})} m(t_i, y_i, \theta_0, \mathbf{z}_i^\top \mathbf{f}_0, \mathbf{g}_0(\mathbf{z}_i))|\mathbf{z}_i] \neq 0$; however in some cases we can exploit unless we can exploit *unconditional* orthogonality: $\mathbb{E}[\nabla_{\mathbf{z}^\top \mathbf{f}, \mathbf{g}(\mathbf{z})} m(t_i, y_i, \theta_0, \mathbf{z}_i^\top \mathbf{f}_0, \mathbf{g}_0(\mathbf{z}_i))] = 0$ to obtain an improved rate although this is not mentioned in the main text.

- In the absence of unconditional orthogonality, we have by the Cauchy-Schwarz inequality that,

$$\mathbb{E}[C^2] \leq O(\sqrt{\mathbb{E}[(\mathbf{\Delta}_g(\mathbf{z}))^4]}\sqrt{\mathbb{E}[(\mathbf{\Delta}_f(\mathbf{z}))^4]}) \leq O(r_{f,2}^2 r_{\mathbf{g},2}^2)$$

- In the presence of unconditional orthogonality we have that,

$$\mathbb{E}[C^2] = \frac{1}{n} r_{f,2}^2 r_{\mathbf{g},2}^2$$

as before using Cauchy-Schwarz but cancelling the cross-terms.

Now we can amalgamate our results. Before doing so note that $r_{f,2} \leq \zeta r_{\boldsymbol{\beta},2}$ since in the description of the algorithm the estimator is defined by rotating an estimate of $\boldsymbol{\beta}_0$ in the base regression procedure (and consistency of the (held-out) prediction error is preserved under orthogonal rotations).

First define the event $\mathcal{J} = \{J \leq \frac{1}{4}\sigma_\eta^2\}$. For the orthogonal estimator defined in the algorithm, on the event $\mathcal{J}$ the estimator will output the estimate from the first-stage base regression using $\hat{y}_{\text{OM}} = \mathbf{x}_\star^\top \boldsymbol{\beta}_0$. So introducing the indicator of this event, and using Cauchy-Schwarz, we have that,

$$\mathbb{E}[(\hat{y}_{\text{OM}} - \theta_0)^2] = \left[\mathbb{E}[(\hat{y}_{\text{OM}} - \theta_0)^2 \mathbb{1}(\mathcal{J})] + \sqrt{\mathbb{E}[\|\mathbf{\Delta}_\beta(\mathbf{x}_\star)\|_2^4]}\sqrt{\Pr[\bar{\mathcal{J}}]}\right]$$

$$\leq \left[O(\frac{\mathbb{E}[A^2 + B_1^2 + B_2^2 + C^2]}{(\sigma_\eta^2)^2}) + O(r_{\boldsymbol{\beta},2}^2)\sqrt{O(((\frac{\xi}{\sigma_\eta^2})^4 + \frac{\xi^2}{(\sigma_\eta^2)^4}r_g^4)\cdot\frac{1}{n^2})}\right]$$

$$\leq \|\mathbf{x}_\star\|_2^2 \left[O(\frac{\sigma_\eta^2\sigma_\epsilon^2 + \zeta r_{\boldsymbol{\beta},2}^2\sigma_\eta^2 + r_g^2\sigma_\epsilon^2}{(\sigma_\eta^2)^2 n}) + O((((\frac{\xi}{\sigma_\eta^2})^2 + \frac{\xi}{(\sigma_\eta^2)^2}r_{\mathbf{g},2}^2)\cdot\frac{1}{n})\cdot r_{\boldsymbol{\beta},2}^2) + \right.$$

$$\left\{\begin{array}{ll} O(\frac{\zeta^2 r_{\boldsymbol{\beta},2}^2 r_{\mathbf{g},2}^2}{(\sigma_\eta^2)^2 n}) & \text{with unconditional orthogonality} \\ O(\frac{\zeta^2 r_{\boldsymbol{\beta},2}^2 r_{\mathbf{g},2}^2}{(\sigma_\eta^2)^2}) & \text{without unconditional orthogonality} \end{array}\right]$$

where $\Pr[\bar{\mathcal{J}}]$ is computed using Lemma 7. If we consider the case without unconditional orthogonality, and assume since $C_{\max} \geq \sigma_\eta^2 \geq C_{\min}$, the above results simplifies (ignoring conditioning-dependent factors) to the theorem statement,

$$\|\mathbf{x}_\star\|_2^2 \left[O(\frac{\sigma_\epsilon^2}{\sigma_\eta^2 n}) + O((\frac{r_{\boldsymbol{\beta},2}^2 r_{\mathbf{g},2}^2}{(\sigma_\eta^2)})^2) + O(\frac{r_{\boldsymbol{\beta},2}^2\sigma_\eta^2 + r_g^2\sigma_\epsilon^2}{(\sigma_\eta^2)^2 n})\right]$$

$\square$

Lemma 7. *Let Assumptions 2, 3, and 5 hold and suppose* $\mathbf{g}_0(\mathbf{z}) = \mathbf{z}^\top \mathbf{g}_0$ *in* (6). *Defining* $J = \frac{1}{n}\sum_{i=1}^n J_i = \frac{1}{n}\sum_{i=1}^n t_i(t_i - \hat{\mathbf{g}}(\mathbf{z}_i))$ *as in the description of first-order OM estimator with* $\tau \leq \frac{1}{4}\mathbb{E}[\eta^2]$, *then,*

$$\Pr\left[\frac{1}{n}\sum_{i=1}^n J_i \leq \tau\right] \leq O(((\frac{\xi}{\sigma_\eta^2})^4 + \frac{\xi^2}{(\sigma_\eta^2)^4}r_{\mathbf{g},2}^4)\cdot\frac{1}{n^2})$$

where $\xi = C_{cond}C_{\max}\kappa^2$ *and* $\zeta = \kappa^2 C_{\max}$ *and* $r_{\mathbf{g},2} = (\mathbb{E}[\|\mathbf{\Delta}_g(\mathbf{z})\|_2^4])^{1/4}$.

Proof. To begin we rescale the $\mathbf{x}_\star$ such that is has unit-norm (and restore the scaling in the final statement of the proof). We begin by establishing concentration of the J term which justifies the thresholding step in the estimator using a 4th-moment Markov inequality. We have that $J = \frac{1}{n}\sum_{i=1}^n \nabla_\theta m(t_i, y_i, \theta_0, \mathbf{z}_i^\top \hat{\mathbf{f}}, \hat{\mathbf{g}}(\mathbf{z}_i)) = \frac{1}{n}\sum_{i=1}^n t_i(t_i - \hat{\mathbf{g}}(\mathbf{z}_i)) = \frac{1}{n}\sum_{i=1}^n J_i$. Note that we assume $t_i = \mathbf{z}_i^\top \mathbf{g}_0 + \eta_i$. The, for an individual term we have that,

$$J_i = (\mathbf{z}_i^\top \mathbf{g}_0 + \eta_i)(\mathbf{\Delta}_g(\mathbf{z}_i) + \eta_i) = \underbrace{\eta_i^2 + \mathbf{z}_i^\top \mathbf{g}_0 \eta_i}_{a_i} + \underbrace{\mathbf{z}_i^\top \mathbf{g}_0(\mathbf{\Delta}_g(\mathbf{z}_i))}_{b_i}$$

Recall by Remark 1, that $\eta = (1, -\mathbf{g}_0)^\top \mathbf{x}'$, and that $\|\mathbf{g}_0\|_2^2 = O(C_{\text{cond}})$. Using sub-gaussianity of $\mathbf{x}'$ we have that $\eta_i \sim \text{sE}(8C_{\text{cond}}\kappa^2 C_{\max}, 8C_{\text{cond}}\kappa^2 C_{\max})$ by Lemma 8. Similarly, $\mathbf{z}_i^\top \mathbf{g}_0 \eta_i \sim \text{sE}(8C_{\text{cond}}\kappa^2 C_{\max}, 8C_{\text{cond}}\kappa^2 C_{\max})$ since $\mathbf{z}_i^\top \mathbf{g}_0 \sim \text{sG}(C_{\max}\kappa^2 C_{\text{cond}})$. We introduce $\xi = C_{\text{cond}}C_{\max}\kappa^2$ and $\zeta = \kappa^2 C_{\max}$.

Analyzing each term, we have that,

- For the first terms, $\mathbb{E}[\eta_i^2] = \mathbb{E}[\eta^2]$. Similarly for the second term, note $\mathbb{E}[b_i] = 0$ since η_i is conditionally (on $\mathbf{z}$) mean-zero. Hence we have that each a_i is mean-zero and $a_i \sim \mathrm{sE}(16\xi, 16\xi)$.

- For the final term, note $\mathbb{E}[(\mathbf{z}_i^\top \mathbf{g}_0(\boldsymbol{\Delta}(\mathbf{z}_i)))^4] \leq O(\xi^2 r_g^4)$ by Cauchy-Schwarz.

Since, $J = \frac{1}{n}\sum_{i=1}^n a_i + b_i + c_i$, if $\left|\frac{1}{n}\sum_{i=1}^n b_i + c_i\right| \leq \epsilon'$ and $\frac{1}{n}\sum_{i=1}^n a_i \geq \epsilon' + \tau$ then $\frac{1}{n}\sum_{i=1}^n J_i > \tau$. So a union bound gives,

$$\Pr\left[\frac{1}{n}\sum_{i=1}^n J_i \leq \tau\right] \leq \Pr\left[\frac{1}{n}\sum_{i=1}^n a_i < +\epsilon' + \tau\right] + \Pr\left[\left|\frac{1}{n}\sum_{i=1}^n b_i + c_i\right| \geq \epsilon'\right]$$

Using a sub-exponential tail bound for the first term and the 4th-moment Marcinkiewicz–Zygmund inequality for the second we obtain,

- For the first term

$$\Pr\left[\frac{1}{n}\sum_{i=1}^n a_i - \mathbb{E}[\eta^2] \leq -\underbrace{(-\epsilon' - \tau + \mathbb{E}[\eta^2])}_{t}\right] \leq O(\exp\left(-cn\min(\frac{t^2}{\xi^2}, \frac{t}{\xi})\right)$$

  for some universal constant c (that may change line to line).

- For the second term

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^n b_i\right| \geq +\epsilon'\right] \leq O(\frac{\xi^2 r_g^4}{(\epsilon')^4 n^2})$$

Taking $\epsilon' = \frac{1}{8}\sigma_\eta^2$ and $\tau \leq \frac{1}{4}\sigma_\eta^2$ it follows that $t \geq \frac{1}{2}\sigma_\eta^2$. Hence the second term can be simplified to $O(\exp\left(-cn\min(\frac{t^2}{\xi^2}, \frac{t}{\xi})\right) = O(\max(\frac{\xi^2}{\sigma_\eta^2}, \frac{\xi}{\sigma_\eta})^2 \frac{1}{n^2})$ Hence the desired bound becomes, $\Pr\left[\frac{1}{n}\sum_{i=1}^n J_i \leq \epsilon\right] \leq O(\frac{\xi^2 r_g^4}{(\sigma_\eta^2)^4 n^2}) + O(\exp\left(-cn\min(\frac{t^2}{\xi^2}, \frac{t}{\xi})\right) = O(((\frac{\xi}{\sigma_\eta^2})^4 + \frac{\xi^2}{(\sigma_\eta^2)^4}r_g^4) \cdot \frac{1}{n^2})$. $\qquad\square$

E.2. Corollaries 4 and 5

We conclude the section by presenting the proofs of Corollary 4 and Corollary 5 which instantiate the OM estimators when both first-stage regressions are estimated with the Lasso.

First we prove Corollary 4.

Proof of Corollary 4. It suffices to compute $r_{\boldsymbol{\beta},2}$ and $r_{\mathbf{g},2}$. By using Lemma 19,

$$r_{\boldsymbol{\beta},2} \leq O\left(\frac{\sigma_\epsilon^2 p}{n}\right)$$

by utilizing condition on λ_β in the theorem statement and that $\|\boldsymbol{\beta}_0\|_\infty/\sigma_\epsilon \leq O(1)$ and $n \geq \Omega(\kappa^4 C_{\text{cond}}^2 p)$. Similarly, for the case of $r_{\mathbf{g},2}$ in the case the estimator is parametric Lasso estimator it follows that $r_{\mathbf{g},2} = (\mathbb{E}[(\mathbf{z}^\top(\mathbf{g}_0 - \mathbf{g}))^4])^{1/4} \leq O(\sqrt{\zeta}\mathbb{E}[(\|\mathbf{g}_0 - \mathbf{g}\|_2^4])^{1/4})$ where $\zeta = \kappa^2 C_{\max}$. Similar to above we obtain that,

$$r_{\mathbf{g},2} \leq O(\frac{\sigma_\eta^2 p}{n})$$

since we can verify that the conditions of Lemma 19 also hold when t is regressed against $\mathbf{z}$ under the hypotheses of the result. In particular, note since the regression for $\mathbf{g}$ is performed between t and $\mathbf{z}$ (which up to an orthogonal rotation is a subvector of the original covariate $\mathbf{x}$ itself), the minimum eigenvalue for this regression is lower-bounded by the minimum eigenvalue of $\mathbf{X}$. Moreover by Remark 1, $\|\mathbf{g}_0\|_2 \leq \sqrt{C_{\text{cond}}}$. $\qquad\square$

Proof of Corollary 5. It suffices to compute $r_{\boldsymbol{\beta},2}$ and $r_{\mathbf{g},2}$. The computation for $r_{\boldsymbol{\beta},2}$ is similar to the one for $r_{\boldsymbol{\beta},1}$. By combining Lemma 13 and Lemma 15, and assuming $p \geq 20$ and $n \geq \frac{c_2 \kappa^4}{C_{\min}} s \log(2ep)$, there exists sufficiently large c such that,

$$
\begin{aligned}
r_{\boldsymbol{\beta},2} &\leq O\left(\left(\frac{\lambda_\beta \sqrt{s_\beta}}{C_{\min}^2}\right) + O\left(\left(\frac{\sigma_\epsilon}{\sqrt{ns_\beta}}\right) + \left(\frac{1}{n\lambda_\beta \sqrt{s_\beta}}\right)\right) + O\left(\frac{\sigma_\epsilon}{\sqrt{n}} + \|\boldsymbol{\beta}_0\|_1 e^{-nc/(8\kappa^4)}\right) = \\
&O\left(\frac{\lambda_\beta \sqrt{s_{\beta_0}}}{C_{\min}^2}\right) + O(\|\boldsymbol{\beta}_0\|_1 e^{-nc/(8\kappa^4)}) \leq O\left(\frac{\lambda_\beta \sqrt{s_{\beta_0}}}{C_{\min}^2}\right)
\end{aligned}
$$

using the lower bound on λ_β in the theorem statement and that $\|\boldsymbol{\beta}_0\|_\infty / \sigma_\epsilon = o(e^{c_1 n})$ for some sufficiently small c_1. Similarly, for the case of $r_{\mathbf{g},2}$ in the case the estimator is parametric Lasso estimator it follows that $r_{\mathbf{g},2} = (\mathbb{E}[(\mathbf{z}^\top(\mathbf{g}_0 - \mathbf{g}))^4])^{1/4} \leq O(\sqrt{\zeta}\mathbb{E}[(\|\mathbf{g}_0 - \mathbf{g}\|_2^4])^{1/4})$ where $\zeta = \kappa^2 C_{\max}$. Similar to above we obtain that,

$$
r_{\mathbf{g},2} \leq O\left(\frac{\lambda_g \sqrt{s_{\mathbf{g}_0}}}{C_{\min}^2}\right) + O(\|\mathbf{g}_0\|_1 e^{-nc/(8\kappa^4)}) \leq O\left(\frac{\lambda_{\mathbf{g}} \sqrt{s_{\mathbf{g}_0}}}{C_{\min}^2}\right)
$$

since we can verify that the conditions of Lemma 13 and Lemma 15 also hold when t is regressed against $\mathbf{z}$ under the hypotheses of the result. In particular, note since the regression for $\mathbf{g}$ is performed between t and $\mathbf{z}$ (which up to an orthogonal rotation is a subvector of the original covariate $\mathbf{x}$ itself), the strong-restricted eigenvalue for this regression is lower-bounded by the strong-restricted eigenvalue of $\mathbf{X}$. Moreover by Remark 1, $\|\mathbf{g}_0\|_1 \leq s_{\mathbf{g}_0}\|\mathbf{g}_0\|_2 \leq s_{\mathbf{g}_0}\sqrt{C_{\text{cond}}}$. $\qquad\square$

F. Auxiliary Lemmas

We now introduce a standard concentration result we will repeatedly use throughout,

Lemma 8. *Let x, y be mean-zero random variables that are both sub-Gaussian with parameters κ_1 and κ_2 respectively. Then $z = xy - \mathbb{E}[xy] \sim sE(8\kappa_1\kappa_2, 8\kappa_1\kappa_2)$.*

Proof. Using the dominated convergence theorem,

$$
\begin{aligned}
\mathbb{E}[e^{\lambda z}] &= 1 + \sum_{k=2}^\infty \frac{\lambda^k \mathbb{E}[(xy - \mathbb{E}[xy])^k]}{k!} \\
&\leq 1 + \sum_{k=2}^\infty \frac{\lambda^k 2^{k-1}(\mathbb{E}[|xy|^k] + \mathbb{E}[|xy|]^k)}{k!} \\
&\leq 1 + \sum_{k=2}^\infty \frac{\lambda^k 2^k \sqrt{\mathbb{E}[x^{2k}]\mathbb{E}[y^{2k}]}}{k!} \\
&\leq 1 + \sum_{k=2}^\infty \frac{\lambda^k 2^k (2\kappa_1\kappa_2)^k (2k)\Gamma(k)}{k!} = 1 + 2(4\lambda\kappa_1\kappa_2)^2 \sum_{k=0}^\infty (4\lambda\kappa_1\kappa_2)^k \\
&\leq 1 + 4(4\lambda\kappa_1\kappa_2)^2 = 1 + 64\lambda^2\kappa_1^2\kappa_2^2 \quad \text{for } |\lambda| \leq \frac{1}{8\kappa_1\kappa_2} \\
&\leq e^{(\lambda \cdot 8\kappa_1\kappa_2)^2} \leq e^{(\lambda \cdot 8\kappa_1\kappa_2)^2/2}
\end{aligned}
$$

$\qquad\square$

where we have used the fact a sub-Gaussian random variable x with parameter κ satisfies $\mathbb{E}[|x|^k] \leq (2\kappa^2)^{k/2} k\Gamma(k/2)$ (which itself follows from integrating the sub-gaussian tail bound), along with the Cauchy-Schwarz and Jensen inequalities.

F.1. Random Design Matrices and Lasso Consistency

Here we collect several useful results we use to show consistency of the Lasso estimator in the random design setting.

Note Assumption 2 ensures the population covariance for the design $\mathbf{X}$ satisfies $\boldsymbol{\Sigma}_{ii} \leq 1/2$, and a standard sub-exponential concentration argument establishes the result for a random design matrix under Assumption 3. Accordingly, we introduce,

Definition 1. The design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ if satisfies the 1-column normalization condition if

$$\max_{i \in [p]} \|\mathbf{X}\mathbf{e}_j\|_2^2/n = \hat{\mathbf{\Sigma}}_{ii} \leq 1$$

and we have that,

Lemma 9. *Let $\kappa' = 8\sqrt{2}\kappa$. If Assumptions 2 and 3 hold, then*

$$\Pr\left[\max_{i \in [p]}[(\hat{\mathbf{\Sigma}}_n)_{ii} - \mathbf{\Sigma}_{ii}] \geq t\right] \leq p \exp\left(-\frac{n}{2}\min(\frac{t^2}{\kappa'^2}, \frac{t}{\kappa'})\right)$$

and if $n \geq 2a \max(\kappa'^2, \kappa') \log p$, then with probability at least $1 - p^{-a}$

$$\max_{i \in [p]}(\hat{\mathbf{\Sigma}}_n)_{ii} \leq 1.$$

Proof. Note that $\mathbf{x}_i = \mathbf{x}^\top \mathbf{e}_i$ satisfies $\mathbb{E}[\exp(\lambda \mathbf{x}_i)] \leq \exp(\lambda^2 \kappa^2 \mathbf{\Sigma}_{ii}/2)$. For fixed i we have that $(\hat{\mathbf{\Sigma}}_n)_{ii} = \frac{1}{n}\sum_{i=1}^n(\mathbf{x}_i^2 - \mathbf{\Sigma}_{ii})$. Since $\mathbf{x}_i \sim sG(\kappa\sqrt{\mathbf{\Sigma}_{ii}})$, using Lemma 8 along with a sub-exponential tail bound we have that,

$$\Pr\left[(\hat{\mathbf{\Sigma}}_n)_{ii} \geq \mathbf{\Sigma}_{ii} + t\right] \leq \exp\left(-\frac{n}{2}\min(\frac{t^2}{\kappa'^2}, \frac{t}{\kappa'})\right)$$

defining $\kappa' = 8\kappa\sqrt{\mathbf{\Sigma}_{ii}} \leq 4\sqrt{2}\kappa$. Since $\mathbf{\Sigma}_{ii} \leq \frac{1}{2}$ using a union bound over the p coordinates we have that $\max_{i \in [p]}(\hat{\mathbf{\Sigma}}_n)_{ii} \geq 1$, with probability less than $p \exp\left(-\frac{n}{2}\min(\frac{t^2}{\kappa'^2}, \frac{t}{\kappa'})\right)$. If $t = \frac{1}{2}$ and $n \geq 2a \max(\frac{\kappa'^2}{t^2}, \frac{\kappa'}{t}) \log p$ the stated conclusion holds. $\qquad\square$

Similarly, although the sample covariance will not be invertible for $p > n$ we require it to be nonsingular along a restricted set of directions. To this end we introduce the strong restricted eigenvalue condition (or SRE condition) defined in (Bellec et al., 2016, Equation 4.2) which is most convenient for our purposes.

Definition 2. Given a symmetric covariance matrix $\mathbf{Q} \in \mathbb{R}^{p \times p}$ satisfying $\max_{i \in [p]} \mathbf{Q}_{ii} \leq 1$, an integer s, and parameter L, the strong restricted eigenvalue of Q is,

$$\phi_{SRE}^2(Q, s, L) \equiv \min_\theta \left\{ \frac{\langle \theta, \mathbf{Q}\theta \rangle}{\|\theta_S\|_2^2} : \theta \in \mathbb{R}^p, \|\theta\|_1 \leq (1 + L)\sqrt{s}\|\theta\|_2 \right\}.$$

In general the cone to which θ belongs in Definition 2 is more constraining then the cone associated with the standard restricted eigenvalue condition of Bickel et al. (2009). Interestingly, due to the inclusion of the 1-column normalization constraint in Definition 2, up to absolute constants, the SRE condition is equivalent to the standard RE condition (with the 1-column normalization constraint also included in its definition) (Bellec et al., 2016, Proposition 8.1).

Importantly, using further equivalence with s-sparse eigenvalue condition, (Bellec et al., 2016, Theorem 8.3) establishes the SRE condition holds with high probability under the sub-gaussian design assumption.

Theorem 11. *Bellec et al. (2016, Theorem 8.3). Let Assumptions 2 and 3 hold. Then there exist absolute constants $c_1, c_2 > 0$ such that for $L \geq 0$, if $n \geq \frac{c_1 \kappa^4 (2+L)^2}{C_{\min}} s \log(2ep/s)$, then with probability at least $1 - 3\exp(-c_2 n/\kappa^4)$, we have that*

$$\max_{i \in [p]}(\hat{\mathbf{\Sigma}}_n)_{ii} \leq 1$$

and

$$\phi_{SRE}^2(\hat{\mathbf{\Sigma}}_n, s, L) \geq \frac{C_{\min}}{2}$$

This result follows from (Bellec et al., 2016, Theorem 8.3), the stated implication therein that the weighted restricted eigenvalue condition implies the strong restricted eigenvalue condition with adjusted constants, along with the fact that $\phi^2_{SRE}(\mathbf{\Sigma}, s, L) \geq C_{\min}$.

We define the sequence of sets,

$$\mathcal{E}_n(s, L) = \{\mathbf{X} \in \mathbb{R}^{n \times p} : \phi^2_{SRE}(\hat{\mathbf{\Sigma}}_n, s, L) \geq \frac{C_{\min}}{2}, \max_{i \in [p]} \hat{\mathbf{\Sigma}}_{ii} \leq 1, \hat{\mathbf{\Sigma}} = \mathbf{X}^\top \mathbf{X}/n\}$$

characterizing the class of design matrices satisfying both Definitions 1 and 2.

There are many classical results on ℓ_1/ℓ_2 consistency of the Lasso program,

$$\hat{\boldsymbol{\beta}}_L = \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

for sparse regression (see for example (van de Geer, 2014b, Ch. 6)) when the model is specified as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$ for ϵ_i i.i.d. that are sub-Gaussian with variance parameter σ^2. Such classical results have the confidence level of the non-asymptotic error tied directed directly to the tuning parameter. However, recently (Bellec et al., 2016), through a more refined analysis, has obtained optimal rates for the Lasso estimator over varying confidence levels for a *fixed* regularization parameter. These results allow us to provide clean upper bounds on the Lasso parameter error in expectation.

Lemma 10. *Let $s \in [p]$, assume that the deterministic design matrix $\mathbf{X} \in \mathcal{E}_n(s, 7)$, and let Assumption 4 hold with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. If $\hat{\boldsymbol{\beta}}_L(\lambda)$ denotes the Lasso estimator with $\lambda \geq (8 + 2\sqrt{2})\sigma\sqrt{\frac{\log(2ep/s)}{n}}$, $1 \leq q \leq 2$, and $\|\boldsymbol{\beta}_0\|_0 \leq s$ then letting $\phi_0^2 = \phi^2_{SRE}(s, 7)$,*

$$\mathbb{E}[\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_q^k] \leq \left(\frac{49\lambda s^{1/q}}{8\phi_0^2}\right)^k + \left(\frac{49}{8} \frac{(8 + 2\sqrt{2})\sigma}{s^{1-1/q}\sqrt{n}}\right)^k \frac{k(k-1)}{2}$$

Proof. The proof follows easily by integrating the tail bound in Bellec et al. (2016, Theorem 4.2), which provides that,

$$\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_q \leq \frac{49}{8} \left(\frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))} \vee \frac{1}{\phi_0^2}\right) \lambda s^{1/q}$$

with probability at least $1 - \delta_0/2$, where $\delta(\lambda) = \exp\left(-\left(\frac{\lambda\sqrt{n}}{(8+2\sqrt{2})\sigma}\right)\right)$, which satisfies $\delta(\lambda) \leq \frac{s}{2ep}$. Now, define δ_0^* as the smallest $\delta_0 \in (0, 1)$ for which $\frac{1}{\phi_0^2} = \frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))}$, in which case $\delta_0^* = (\delta(\lambda))^{\frac{s}{\phi_0^2}}$.

Then, $Z_q = \frac{8s\log(1/\delta(\lambda))}{49\lambda s^{1/q}} \leq \log(1/\delta_0)$ with probability at least $1 - \delta_0/2$, for all $\delta_0 \in (0, \delta_0^*]$. Equivalently, $\Pr[Z_q > t] \leq \frac{e^{-t}}{2}$ for all $t \geq T = \log(1/\delta_0^*) = \frac{s}{\phi_0^2}\log(1/\delta(\lambda))$. Thus,

$$\mathbb{E}[Z_q^k] = \int_0^\infty kt^{k-1}\Pr[Z_q > t]dt = \int_0^T kt^{k-1} + \int_T^\infty kt^{k-1}\frac{e^{-t}}{2} \leq$$
$$T^k + \int_0^\infty kt^{k-1}\frac{e^{-t}}{2} \leq T^k + \frac{k(k-1)}{2}.$$

which implies the conclusion,

$$\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_q^k \leq \left(\frac{49}{8} \frac{T\lambda s^{1/q}}{s\log(1/\delta(\lambda))}\right)^k + \left(\frac{49}{8} \frac{\lambda s^{1/q}}{s\log(1/\delta(\lambda))}\right)^k \frac{k(k-1)}{2} \leq$$
$$\left(\frac{49\lambda s^{1/q}}{8\phi_0^2}\right)^k + \left(\frac{49}{8} \frac{(8 + 2\sqrt{2})\sigma}{s^{1-1/q}\sqrt{n}}\right)^k \frac{k(k-1)}{2}$$

where $\lambda \geq (8 + 2\sqrt{2})\sigma\sqrt{\frac{\log(2ep/s)}{n}}$. $\qquad\square$

Although the main results of Bellec et al. (2016) are stated for Gaussian noise distributions, Bellec et al. (2016, Theorem 9.1) also provides a complementary high-probability upper bound for the empirical process $\frac{1}{n}\boldsymbol{\epsilon}^\top \mathbf{X}\mathbf{u}$ when $\boldsymbol{\epsilon}$ is sub-gaussian:

Lemma 11. *Bellec et al. (2016, Theorem 9.1) Let $\delta_0 \in (0,1)$, and let Assumption 4 hold (with variance parameter renamed to σ^2) and assume the deterministic design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ satisfies $\max_{i\in[p]} \|\mathbf{X}\mathbf{e}_i\|_2/\sqrt{n} \leq 1$. Then with probability at least $1 - \delta_0$, for all $u \in \mathbb{R}^p$,*

$$\frac{1}{n}\boldsymbol{\epsilon}^\top \mathbf{X}\mathbf{u} \leq 40\sigma \max \left(\sum_{j=1}^p u_j^\sharp \sqrt{\frac{\log(2p/j)}{n}}, \frac{\|\mathbf{X}\mathbf{u}\|_2}{\sqrt{n}} \frac{\sqrt{\pi/2} + \sqrt{2\log(1/\delta_0)}}{\sqrt{n}} \right)$$

The upper bound contains an additional, additive $\frac{\sqrt{\pi/2}}{\sqrt{n}}$ correction along with a change in absolute constants with respect to Bellec et al. (2016, Theorem 4.1). Hence we trace through the proof of Bellec et al. (2016, Theorem 4.2) to derive a corresponding statement of Bellec et al. (2016, Theorem 4.2) for sub-gaussian distributions.

Lemma 12. *Let $s \in [p]$, $\gamma \in (0,1)$ and $\tau \in (0, 1-\gamma]$ and assume the SRE(s, c_0) condition holds $c_0(\gamma, \tau) = \frac{1+\gamma+\tau}{1-\gamma-\tau}$. Let $\lambda \geq \frac{40\sigma}{\gamma}\sqrt{\frac{\log(2ep/s)}{n}}$. Then on the event in Lemma 11, for $1 \leq q \leq 2$,*

$$\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_q \leq \left(\frac{C_{\gamma,\tau}(s,\lambda,\delta_0)}{\tau}\lambda s + \frac{\pi(1+\tau+\gamma)^2}{\gamma^2 \tau n\lambda} \right)^{2/q-1} \left(3\left(\frac{C_{\gamma,0}(s,\lambda,\delta_0)}{1+\gamma}\lambda\sqrt{s} + \frac{\pi(1+\gamma)}{\gamma^2\lambda\sqrt{2sn}}\right) \right)^{2-2/q}$$

where $C_{\gamma,\tau} = (1+\gamma+\tau)^2 \left(\frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))} \vee \frac{1}{\phi_0^2(s,c_0(\gamma,\tau))} \right)$.

Proof. The argument simply requires tracing through the proof of Bellec et al. (2016, Theorem 4.2) to accommodate the additional $O(\frac{1}{\sqrt{n}})$ term (and is nearly identical to Bellec et al. (2016, Theorem 4.2)), so we only highlight the important modifications.

Following the proof of Bellec et al. (2016, Theorem 4.2) we have,

$$2\tau\lambda\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_1 + 2\|\mathbf{X}(\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0)\|_2^2/n \leq \Delta^* \tag{11}$$

where $\Delta^* = 2\tau\lambda\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_1 + \frac{2}{n}\boldsymbol{\epsilon}^\top \mathbf{X}(\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0) + 2\lambda\|\boldsymbol{\beta}_0\|_1 - 2\lambda\|\hat{\boldsymbol{\beta}}_L(\lambda)\|_1$ Letting $\mathbf{u} = \hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0$, we obtain

$$\Delta^* \leq 2\lambda \left((1+\tau)\sqrt{s}\|\mathbf{u}\|_2 - (1-\tau)\sum_{j=s+1}^p u_j^\sharp \right) + 2\max(F(\mathbf{u}), G(\mathbf{u}))$$

where $F(\mathbf{u}) = \gamma\lambda\left(\sqrt{s}\|\mathbf{u}\|_2 + \sum_{j=s+1}^p u_j^\sharp \right)$ and $G(\mathbf{u}) = 40\sigma(\frac{\|\mathbf{X}\mathbf{u}\|_2}{\sqrt{n}} \frac{\sqrt{\pi/2}+\sqrt{2\log(1/\delta_0)}}{\sqrt{n}})$. By definition of $\delta(\lambda) = \exp\left(-(\frac{\gamma\lambda\sqrt{n}}{40\sigma})^2\right)$ we have equivalently that, $G(\mathbf{u}) = \left(\lambda\sqrt{s}\gamma\sqrt{\log(1/\delta_0)/(s\log(1/\delta(\lambda)))} + \frac{40\sqrt{\pi/2}\sigma}{\sqrt{n}} \right) \|\mathbf{X}\mathbf{u}\|_2/\sqrt{n}$.

We now consider two cases

1. $G(\mathbf{u}) > F(\mathbf{u})$. Then,

$$\|\mathbf{u}\|_2 \leq \left(\sqrt{\frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))}} + \frac{40\sqrt{\pi/2}\sigma}{\lambda\sqrt{s}\gamma\sqrt{n}} \right) \|\mathbf{X}\mathbf{u}\|_2/\sqrt{n} \tag{12}$$

Thus,

$$\Delta^* \leq 2\lambda(1+\tau)\sqrt{s}\|\mathbf{u}\|_2 + 2G(\mathbf{u})$$

$$2\lambda\sqrt{s}(1+\tau+\gamma)\left(\sqrt{\frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))}} + \frac{40\sqrt{\pi/2}\sigma}{\lambda\sqrt{s}\gamma\sqrt{n}} \right) \|\mathbf{X}\mathbf{u}\|_2/\sqrt{n} \leq$$

$$2\lambda^2 s(1+\tau+\gamma)^2 \left(\frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))} + \frac{800\pi\sigma^2}{\lambda^2 s\gamma^2 n} \right) + \|\mathbf{Xu}\|_2^2/n =$$

$$2\lambda^2 s(1+\tau+\gamma)^2 (\frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))}) + \|\mathbf{Xu}\|_2^2/n + \frac{1600\pi\sigma^2(1+\tau+\gamma)^2}{\gamma^2 n} \tag{13}$$

2. $G(\mathbf{u}) \leq F(\mathbf{u})$. In this case,

$$\Delta^* \leq 2\lambda \left((1+\gamma+\tau)\sqrt{s}\|\mathbf{u}\|_2 - (1-\gamma-\tau) \sum_{j=s+1}^p u_j^\sharp \right) = \Delta \tag{14}$$

Since $\Delta > 0$, $\mathbf{u}$ belongs to the $SRE(s, c_0)$ cone and hence $\phi_0(s, c_0)\|\mathbf{u}\|_2 \leq \|\mathbf{Xu}\|_2$. So,

$$\Delta^* \leq \Delta \leq \frac{2(1+\gamma+\tau)\lambda\sqrt{s}}{\phi_0(s, c_0)}\|\mathbf{Xu}\|/\sqrt{n} \leq \left(\frac{(1+\gamma+\tau)\lambda\sqrt{s}}{\phi_0(s, c_0)} \right)^2 + \|\mathbf{Xu}\|^2/n \tag{15}$$

Assembling the two cases we conclude that,

$$2\tau\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_1 \leq 2C_{\gamma,\tau}(s, \lambda, \delta_0)\lambda s + \frac{1600\pi\sigma^2(1+\tau+\gamma)^2}{\gamma^2 n\lambda}$$

where $C_{\gamma,\tau}(s, \lambda, \delta_0) = (1+\gamma+\tau)^2 \left(\frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))} \vee \frac{1}{\phi_0^2(s, c_0(\gamma,\tau))} \right)$.

Turning to upper bounding $\mathbf{u}$ in the ℓ_2 norm, we specialize to $\tau = 0$ and consider cases 1 and 2 from before.

1. $G(\mathbf{u}) > F(\mathbf{u})$, then using Equations 11 and 13 we have,

$$\|\mathbf{Xu}\|_2^2/n \leq 2\lambda^2 s(1+\gamma)^2 (\frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))}) + \frac{1600\pi\sigma^2(1+\gamma)^2}{\gamma^2 n}$$

Combining the previous display with (12) we have,

$$\|\mathbf{u}\|_2 \leq \left(\sqrt{2\lambda^2 s(1+\gamma)^2 (\frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))})} + \sqrt{\frac{1600\pi\sigma^2(1+\gamma)^2}{\gamma^2 n}} \right) \left(\sqrt{\frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))}} + \frac{40\sqrt{\pi/2}\sigma}{\lambda\sqrt{s}\gamma\sqrt{n}} \right)$$

$$= \sqrt{2s}(1+\gamma)\lambda \left(\frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))} \right) + \frac{1600\pi(1+\gamma)\sigma^2}{\gamma^2\lambda\sqrt{2s}n} + \sqrt{\sqrt{2s}(1+\gamma)\lambda\frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))} \cdot \frac{1600\pi\sigma^2(1+\gamma)}{\gamma^2\lambda\sqrt{2s}n}}$$

$$\leq \frac{3}{2} \left(\sqrt{2s}(1+\gamma)\lambda \left(\frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))} \right) + \frac{1600\pi\sigma^2(1+\gamma)}{\gamma^2\lambda\sqrt{2s}n} \right)$$

using subadditivity of $\sqrt{\cdot}$.

2. $G(\mathbf{u}) \leq F(\mathbf{u})$. Equations 11 and 14 implies that $\Delta \geq \Delta^* \geq 0$ a.s. Hence $\mathbf{u}$ is contained in $SRE(s, \frac{1+\gamma}{1-\gamma})$, and

$$\|\mathbf{u}\|_2 \leq \frac{\|\mathbf{Xu}\|_2}{n\phi_0(s, \frac{1+\gamma}{1-\gamma})} \leq \frac{(1+\gamma)\lambda\sqrt{s}}{\phi_0^2(s, \frac{1+\gamma}{1-\gamma})}$$

using (11) and (15), and recalling we set $\tau = 0$. Assembling these two cases we conclude,

$$(1+\gamma)\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_2 \leq 3 \left(C_{\gamma,0}(s, \lambda, \delta_0)\lambda\sqrt{s} + \frac{1600\pi\sigma^2(1+\gamma)^2}{\gamma^2\lambda\sqrt{2s}n} \right)$$

So using the norm interpolation inequality $\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_q \leq \|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_1^{2/q-1}\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_2^{2-2/q}$,

$$\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_q \leq \left(\frac{C_{\gamma,\tau}(s, \lambda, \delta_0)}{\tau}\lambda s + \frac{1600\pi\sigma^2(1+\tau+\gamma)^2}{\gamma^2\tau n\lambda} \right)^{2/q-1} \left(3(\frac{C_{\gamma,0}(s, \lambda, \delta_0)}{1+\gamma}\lambda\sqrt{s} + \frac{1600\pi\sigma^2(1+\gamma)}{\gamma^2\lambda\sqrt{2s}n}) \right)^{2-2/q}$$

$$\square$$

We can now derive a corresponding moment bound for error as before[6],

Lemma 13. *Let $s \in [p]$, assume that the deterministic design matrix $\mathbf{X} \in \mathcal{E}_n(s,7)$, and let Assumption 4 hold (with variance parameter renamed to σ^2). If $\hat{\boldsymbol{\beta}}_L(\lambda)$ denotes the Lasso estimator with $\lambda \geq 80\sigma\sqrt{\frac{\log(2ep/s)}{n}}$, $1 \leq q \leq 2$, and $\|\boldsymbol{\beta}_0\|_0 \leq s$ then letting $\phi_0^2 = \phi_{SRE}^2(s,7)$,*

$$\mathbb{E}[\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_1^k] \leq 2^{k-1}\left(\left(13\frac{\lambda s}{\phi_0^2}\right)^k + \left(13\frac{40\sigma}{\sqrt{n}}\right)^k \frac{k(k-1)}{2} + (\frac{250000}{n\lambda})^k\right)$$

$$\mathbb{E}[\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_2^k] \leq 2^{k-1}\left(\left(5\frac{\lambda\sqrt{s}}{\phi_0^2}\right)^k + \left(13\frac{40\sigma}{\sqrt{ns}}\right)^k \frac{k(k-1)}{2} + (\frac{25000}{n\lambda\sqrt{s}})^k\right)$$

Proof. We instantiate the result of Lemma 13 with $\gamma = 1/2$ and $\tau = 1/4$ in which case $c_0 = 7$, $(1+\gamma+\tau)^2 = 49/16$, $\frac{1+\gamma}{1-\gamma} = 3$, $1+\gamma = 3/2$. Defining $D(\delta_0, \lambda, s) = \left(\frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))} \vee \frac{1}{\phi_0^2}\right)$ and $\phi_0^2 = \phi_0^2(s,7)$ we have,

$$\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_1 \leq 13D(\delta_0, \lambda, s)\lambda s + \frac{250000\sigma^2}{n\lambda}$$

$$\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_2 \leq 5D(\delta_0, \lambda, s)\lambda\sqrt{s} + \frac{25000\sigma^2}{\lambda\sqrt{s}n}$$

with probability $1-\delta_0$ where $\delta(\lambda) = \exp\left(-(\frac{\lambda\sqrt{n}}{80\sigma})\right)$. Now, define δ_0^* as the smallest $\delta_0 \in (0,1)$ for which $\frac{1}{\phi_0^2} = \frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))}$, in which case $\delta_0^* = (\delta(\lambda))^{\frac{s}{\phi_0^2}}$.

Then, $Z_1 = \frac{(\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_1 - \frac{250000\sigma^2}{n\lambda})s\log(1/\delta(\lambda))}{13\lambda s} \leq \log(1/\delta_0)$ and $Z_2 = \frac{(\|\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0\|_2 - \frac{25000\sigma^2}{n\lambda\sqrt{s}})s\log(1/\delta(\lambda))}{5\lambda\sqrt{s}}$ with probability at least $1 - \delta_0$, for all $\delta_0 \in (0, \delta_0^*]$. Equivalently, $\Pr[Z_q > t] \leq e^{-t}$ for all $t \geq T = \log(1/\delta_0^*) = \frac{s}{\phi_0^2}\log(1/\delta(\lambda))$ for $q \in \{1, 2\}$. As before,

$$\mathbb{E}[Z_q^k] \leq T^k + k(k-1).$$

Since $\mathbb{E}[\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_q^k] = \mathbb{E}[(\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_q - c + c)^k] \leq 2^{k-1}\left(\mathbb{E}[(\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_q - c)^k] + c^k\right)$, we conclude,

$$\mathbb{E}[\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_1^k] \leq 2^{k-1}\left(\left(13\frac{T\lambda s}{s\log(1/\delta(\lambda))}\right)^k + \left(13\frac{\lambda s}{s\log(1/\delta(\lambda))}\right)^k \frac{k(k-1)}{2} + (\frac{250000\sigma^2}{n\lambda})^k\right) \leq$$

$$2^{k-1}\left(\left(13\frac{\lambda s}{\phi_0^2}\right)^k + \left(13\frac{40\sigma}{\sqrt{n}}\right)^k \frac{k(k-1)}{2} + (\frac{250000\sigma^2}{n\lambda})^k\right)$$

and

$$\mathbb{E}[\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_2^k] \leq 2^{k-1}\left(\left(5\frac{T\lambda\sqrt{s}}{s\log(1/\delta(\lambda))}\right)^k + \left(5\frac{\lambda\sqrt{s}}{s\log(1/\delta(\lambda))}\right)^k \frac{k(k-1)}{2} + (\frac{25000\sigma^2}{n\lambda\sqrt{s}})^k\right) \leq$$

$$2^{k-1}\left(\left(5\frac{\lambda\sqrt{s}}{\phi_0^2}\right)^k + \left(5\frac{40\sigma}{\sqrt{ns}}\right)^k \frac{k(k-1)}{2} + (\frac{25000\sigma^2}{n\lambda\sqrt{s}})^k\right)$$

where $\lambda \geq 80\sigma\sqrt{\frac{\log(2ep/s)}{n}}$. □

The aforementioned results establish Lasso consistency (in expectation) conditioned on the event $\mathbf{X} \in \mathcal{E}_n(s,7)$. Generalizing these results to an unconditional statement (on $\mathbf{X}$) requires the following deterministic lemma to control the norm of the error vector $\left\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\right\|_1$ on the "bad" events $\mathbf{X} \notin \mathcal{E}_n(s,7)$ where we cannot guarantee a "fast" rate for the Lasso.

---

[6]for convenience we only state for the ℓ_1 and ℓ_2 norms an analogous result to Lemma 10 can be derived with more computation.

Lemma 14. *Let $\hat{\boldsymbol{\beta}}_L(\lambda)$ be the solution of the Lagrangian lasso, then*

$$\left\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\right\|_1 \leq \frac{1}{2n}\|\boldsymbol{\epsilon}\|_2^2/\lambda + 2\|\boldsymbol{\beta}_0\|_1.$$

Proof. By definition we have that,

$$\frac{1}{2n}\left\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_L(\lambda)\right\|_2^2 + \lambda\left\|\hat{\boldsymbol{\beta}}_L\right\|_1 \leq \frac{1}{2n}\|\boldsymbol{\epsilon}\|_2^2 + \lambda\|\boldsymbol{\beta}_0\|_1 \implies \left\|\hat{\boldsymbol{\beta}}_L(\lambda)\right\|_1 \leq \frac{1}{2n}\|\boldsymbol{\epsilon}\|_2^2/\lambda + \|\boldsymbol{\beta}_0\|_1$$

So by the triangle inequality we obtain that,

$$\left\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\right\|_1 \leq \frac{1}{2n}\|\boldsymbol{\epsilon}\|_2^2/\lambda + 2\|\boldsymbol{\beta}_0\|_1.$$

$\square$

With this result in hand we can combine our previous results to provide our final desired consistency result for the Lasso.

Lemma 15. *Let Assumptions 1, 2, 3, 4 hold (with variance parameter renamed to σ^2). Then there exist absolute constants $c_1, c_2 > 0$ such that if $n \geq \frac{c_1(k)\kappa^4}{C_{\min}} s \log(2ep/s)$, and $\hat{\boldsymbol{\beta}}_L(\lambda)$ is a solution of the Lagrangian Lasso then for $q \in 1, 2$*

$$\mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}}\left[\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_q^k\right] \leq \mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}}\left[\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_q^k \mathbb{1}[\mathbf{X} \in \mathcal{E}_n(s, 7)]\right] + \left(\frac{\sigma^{2k}}{\lambda^k} + 2^{2k}\|\boldsymbol{\beta}_0\|_1^k\right)\left(2e^{-\frac{c_2}{2}n}\right)$$

where the first term can be bounded exactly as the conclusion of either Lemmas 10 or 13 with appropriate choice of regularization parameter $\lambda_{\boldsymbol{\beta}}$.

Proof. Consider the event $\{\mathbf{X} \notin \mathcal{E}_n(s, 7)\}$. For $q \in 1, 2$, we can split the desired expectation over the corresponding indicator r.v. giving,

$$\mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}}[\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_q^k] = \mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}}\left[\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_q^k \mathbb{1}[\mathbf{X} \in \mathcal{E}_n(s, 7)]\right] + \mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}}\left[\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_q^k \mathbb{1}[\mathbf{X} \notin \mathcal{E}_n(s, 7)]\right] \quad (16)$$

The first term can be bounded using independence of $\mathbf{X}$ and $\boldsymbol{\epsilon}$ to integrate over $\boldsymbol{\epsilon}$ restricted to the set $\{\mathbf{X} \notin \mathcal{E}_n(s, 7)\}$ (by applying Lemmas 10 and 13). The second term can be bounded using Cauchy-Schwarz and Lemma 14 which provides a coarse bound on the Lasso performance which always holds,

$$\mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}}\left[\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_q^k \mathbb{1}[\mathbf{X} \notin \mathcal{E}_n(s, 7)]\right] \leq \sqrt{\mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}}\left[\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_q^{2k}\right]}\sqrt{\Pr_{\mathbf{X}}[\mathbf{X} \notin \mathcal{E}_n(s, 7))]} \quad (17)$$

The hypotheses of Theorem 11 are satisfied, so $\sqrt{\Pr_{\mathbf{X}}[\mathbf{X} \notin \mathcal{E}_n(s, 7))]} \leq 2e^{-\frac{c_2}{2}n}$. Using Lemma 14 along with the identity $(a + b)^k \leq 2^{k-1}(a^k + b^k)$ we have that,

$$\mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}}[\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_q^{2k}] \leq \mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}}[\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_1^{2k}] \leq 2^{2k-1} \cdot \mathbb{E}_{\boldsymbol{\epsilon}}\left[\frac{(\sum_{i=1}^n \epsilon_i^2/n)^{2k}}{2^{2k}\lambda^{2k}} + 2^{2k}\|\boldsymbol{\beta}_0\|_1^{2k}\right]$$

Since the $\epsilon_i \sim \text{sG}(0, \sigma^2)$, $\epsilon_i^2 \sim \text{sE}(8\sigma^2, 8\sigma^2)$ by Lemma 8, so $Z = \sum_{i=1}^n \epsilon_i^2/n \sim \text{sE}(8\sigma^2, 8\sigma^2)$ satisfies the tail bound $\Pr[Z - \mathbb{E}[Z] \geq t] \leq \exp(-n/2 \min(t^2/(8\sigma^2)^2, t/(8\sigma^2)))$ since the ϵ_i are independent. Defining $c = 8\sigma^2$, we find by integrating the tail bound,

$$\mathbb{E}[Z^k] = \int_0^{\mathbb{E}[Z]} kt^{k-1} + \int_{\mathbb{E}[Z]}^c \exp(-n/2 \cdot t^2/c^2) + \int_c^\infty \exp(-n/2 \cdot t/c) \leq$$

$$(\sigma^2)^k + \frac{k2^{k/2-1}c^k\Gamma(k/2)}{n^{k/2}} + \frac{k2^k c^k \Gamma(k)}{n^k} \leq 2(\sigma^2)^k.$$

since $\mathbb{E}[Z] \leq \sigma^2$, and we choose $n^{2k/2} \geq 2(2k)2^{2k/2}(8^{2k})(2k)!$. Assembling, we have the bound

$$\mathbb{E}_{\mathbf{X},\epsilon}[\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_q^{2k}] \leq \frac{\sigma^{4k}}{\lambda^{2k}} + 2^{4k}\|\boldsymbol{\beta}_0\|_1^{2k} \tag{18}$$

Inserting the coarse bound in (18) into (17) and combining with (16) gives the result using subadditivity of $\sqrt{\cdot}$,

$$\mathbb{E}_{\mathbf{X},\epsilon}\left[\left\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\right\|_q^k\right] \leq \mathbb{E}_{\mathbf{X},\epsilon}\left[\|\hat{\boldsymbol{\beta}}_L(\lambda) - \boldsymbol{\beta}_0\|_q^k \mathbb{1}[\mathbf{X} \in \mathcal{E}_n(s,7)]\right] + \left(\frac{\sigma^{2k}}{\lambda^k} + 2^{2k}\|\boldsymbol{\beta}_0\|_1^k\right)\left(2e^{-\frac{c_2}{2}n}\right) \tag{19}$$

As previously noted the first term in Equation (19) is computed exactly as the final result of either Lemmas 10 or 13. $\quad\square$

F.2. Random Design Matrices and Ridge Regression Consistency

Here we collect several useful results we use to show consistency of the ridge regression estimator in the random design setting. There are several results showing risk bounds for ridge regression in the random design setting, see for example Hsu et al. (2012). Such results make assumptions which do not match our setting and also do not immediately imply control over the higher moments of the ℓ_2-error which are also needed in our setting. Accordingly, we use a similar approach to that used for the Lasso estimator to show appropriate non-asymptotic risk bounds (in expectation) for ridge regression.

To begin recall we define the ridge estimator $\hat{\boldsymbol{\beta}}_R(\lambda) = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2\right)$ which implies $\hat{\boldsymbol{\beta}}_R(\lambda) = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{y}$. Throughout we also use $\hat{\boldsymbol{\Sigma}}_n = \frac{\mathbf{X}^\top\mathbf{X}}{n}$, $\hat{\boldsymbol{\Sigma}}_\lambda = \frac{\mathbf{X}^\top\mathbf{X}}{n} + \frac{\lambda}{n}\mathbf{I}_p$ and $\boldsymbol{\Pi}_\lambda = \mathbf{I}_p - (\hat{\boldsymbol{\Sigma}}_\lambda)^{-1}\hat{\boldsymbol{\Sigma}}_n$. Note that under Assumption 1, $\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_0 = -\boldsymbol{\Pi}_\lambda\boldsymbol{\beta}_0 + \hat{\boldsymbol{\Sigma}}_\lambda^{-1}\mathbf{X}^\top\boldsymbol{\epsilon}/n$, which can be thought of as a standard bias-variance decomposition for the ridge estimator.

We first introduce a standard sub-Gaussian concentration result providing control on the fluctuations of the spectral norm of the design matrix which follows immediately from Wainwright, Theorem 6.5,

Lemma 16. *Let* $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *be i.i.d. random vectors satisfying Assumptions 2 and 3 with sample covariance* $\hat{\boldsymbol{\Sigma}}_n = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$, *then there exist universal constants* c_1, c_2, c_3 *such that for* $n \geq c_1\kappa^4 C_{cond}^2 p$,

$$\left\|\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}\right\|_2 \leq \frac{C_{\min}}{2}$$

with probability at least $1 - c_2 e^{-c_3 n/(\kappa^4 C_{cond}^2)}$.

With this result we first provide a conditional (on $\mathbf{X}$) risk bound for ridge regression. For convenience throughout this section we define the set of design matrices $\mathcal{E}_n = \{\mathbf{X} : \forall \mathbf{v} \text{ such that } \|\mathbf{v}\|_2 = 1, \mathbf{v}^\top\hat{\boldsymbol{\Sigma}}\mathbf{v} \geq \frac{C_{\min}}{2}\}$.

Lemma 17. *Let Assumptions 2 and 4 hold (with variance parameter renamed to* σ^2*) and assume a deterministic design matrix* $\mathbf{X} \in \mathcal{E}_n$ *and that* $n \geq p$*. Then if* $\hat{\boldsymbol{\beta}}_R(\lambda)$ *denotes the solution to the ridge regression program, with* $\lambda \leq \lambda_* = \arg\min_\lambda\left(\left(\frac{\lambda/n}{C_{\min}+\lambda/n}\right)^4\|\boldsymbol{\beta}_0\|_2^4 + \sigma^4 p^2/n^2\left(\frac{C_{\max}}{(C_{\min}+\lambda/n)^2}\right)^2\right)$,

$$\left(\mathbb{E}\left[\left\|\hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\right\|_2^4\right]\right)^{1/2} \leq O\left(\sigma^2\frac{C_{cond}}{C_{\min}}\frac{p}{n}\right).$$

Proof. Recall the standard bias variance decomposition $\hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0 = -\boldsymbol{\Pi}_\lambda\boldsymbol{\beta}_0 + \hat{\boldsymbol{\Sigma}}_\lambda^{-1}\mathbf{X}^\top\boldsymbol{\epsilon}/n$. So $\left\|\hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\right\|_2^4 \leq 64\left((\boldsymbol{\beta}_0\boldsymbol{\Pi}_\lambda^2\boldsymbol{\beta}_0)^2 + (\boldsymbol{\epsilon}^\top\mathbf{X}\hat{\boldsymbol{\Sigma}}_\lambda^{-1} \cdot \hat{\boldsymbol{\Sigma}}_\lambda^{-1}\mathbf{X}^\top\boldsymbol{\epsilon}/n^2)^2\right)$. Using the SVD of $\mathbf{X}/\sqrt{n} = \mathbf{U}^\top\boldsymbol{\Lambda}\mathbf{V}$ we see that $\hat{\boldsymbol{\Sigma}}_n = \mathbf{V}^\top\boldsymbol{\Lambda}^2\mathbf{V} = \mathbf{V}^\top\mathbf{D}\mathbf{V}$. Further, on the event $\mathcal{E}_n$ we have that $\frac{1}{2}C_{\min} \leq d_i \leq \frac{3}{2}C_{\max}$ for $i \in [p]$ where $d_i = \mathbf{D}_{ii}$ by the Weyl inequalities. So on $\mathcal{E}_n$, $\boldsymbol{\beta}_0^\top\boldsymbol{\Pi}_\lambda^2\boldsymbol{\beta}_0 = \boldsymbol{\beta}_0^\top\mathbf{V}^\top(\text{diag}(\frac{\lambda/n}{d_i+\lambda/n}))^2\mathbf{V}\boldsymbol{\beta}_0 \leq O((\frac{\lambda/n}{C_{\min}+\lambda/n})^2\|\boldsymbol{\beta}_0\|_2^2)$. Define $\mathbf{S} = \boldsymbol{\epsilon}^\top\mathbf{X}\hat{\boldsymbol{\Sigma}}_\lambda^{-1} \cdot \hat{\boldsymbol{\Sigma}}_\lambda^{-1}\mathbf{X}^\top\boldsymbol{\epsilon}/n$, we have that $\mathbf{S} = \mathbf{U}^\top\text{diag}(\frac{z_i}{(z_i+\lambda/n)^2})\mathbf{U} \preceq O(\mathbf{U}^\top\text{diag}(\frac{C_{\max}}{(C_{\min}+\lambda/n)^2})\mathbf{U})$ on $\mathcal{E}_n$, which also has at most rank p since $\boldsymbol{\Lambda}$ has at most p non-zero singular values. Hence applying Lemma 20 we find that $\mathbb{E}[(\boldsymbol{\epsilon}^\top\mathbf{S}\boldsymbol{\epsilon})^2] \leq O(\sigma^4 p^2(\frac{C_{\max}}{(C_{\min}+\lambda/n)^2})^2)$. Combining, gives that

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\right\|_2^4\right] \leq c_1\left((\frac{\lambda/n}{C_{\min}+\lambda/n})^4\|\boldsymbol{\beta}_0\|_2^4 + \sigma^4 p^2/n^2(\frac{C_{\max}}{(C_{\min}+\lambda/n)^2})^2\right).$$

for some universal constant c_1. Since by definition λ_* minimizes the upper bound in the above expression it is upper bounded by setting $\lambda = 0$ in the same expression so,

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\right\|_2^4\right] \leq O\left(\sigma^4 p^2 / n^2 (\frac{C_{\max}}{C_{\min}^2})^2\right).$$

We can further check that the upper bound is decreasing over the interval $[0, \lambda_*]$ and hence the conclusion follows. As an aside a short computation shows the optimal choice of $\lambda_*/p = (C_{cond} C_{\max} \frac{n}{p} \frac{\sigma^4}{\|\boldsymbol{\beta}_0\|_2^4})^{1/3}$. □

We now prove a simple result which provides a crude bound on the error of the ridge regression estimate we deploy when $\mathbf{X} \notin \mathcal{E}_n$.

Lemma 18. *Let $\hat{\boldsymbol{\beta}}_R(\lambda)$ be the solution of the ridge regression program $\hat{\boldsymbol{\beta}}_R(\lambda) = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$, then*

$$\left\|\hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\right\|_2^2 \leq 4\left(\|\boldsymbol{\epsilon}\|_2^2/\lambda + \|\boldsymbol{\beta}_0\|_2^2\right).$$

Proof. By definition we have that,

$$\left\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_R(\lambda)\right\|_2^2 + \lambda\left\|\hat{\boldsymbol{\beta}}_R(\lambda)\right\|_2^2 \leq \|\boldsymbol{\epsilon}\|_2^2 + \lambda\|\boldsymbol{\beta}_0\|_2^2 \implies \left\|\hat{\boldsymbol{\beta}}_R(\lambda)\right\|^2 \leq \|\boldsymbol{\epsilon}\|_2^2/\lambda + \|\boldsymbol{\beta}_0\|^2$$

So we obtain that,

$$\left\|\hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\right\|_2^2 \leq 2(\left\|\hat{\boldsymbol{\beta}}_R(\lambda)\right\|_2^2 + \|\boldsymbol{\beta}_0\|_2)^2 \leq 4(\|\boldsymbol{\epsilon}\|_2^2/\lambda + \|\boldsymbol{\beta}_0\|_2^2).$$

□

Finally, we prove the final result which will provide an unconditional risk bound in expectation for the ridge regression estimator,

Lemma 19. *Let Assumptions 1, 2, 3, 4 hold (with variance parameter renamed to σ^2). Then there exist universal constants $c_1, c_2, c_3 > 0$ such that if $n \geq c_1 \kappa^4 C_{cond}^2 p$, and $\hat{\boldsymbol{\beta}}_R(\lambda)$ a solution of the ridge regression program with $c_2 \frac{n^2 C_{\min}}{p C_{cond}} e^{-nc_3/\kappa^4 C_{cond}^2} \leq \lambda \leq \lambda_* = \arg\min_{\lambda}\left(\left(\frac{\lambda/n}{C_{\min}+\lambda/n}\right)^4\|\boldsymbol{\beta}_0\|_2^4 + \sigma^4 p^2 \left(\frac{C_{\max}}{(C_{\min}+\lambda/n)^2}\right)^2\right) = p(C_{cond}C_{\max}\frac{n}{p}\frac{\sigma^4}{\|\boldsymbol{\beta}_0\|_2^4})^{1/3}$*

$$\mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}}\left[\|\hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\|_2^4\right] \leq \mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}}\left[\|\hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\|_q^4 \mathbb{1}[\mathbf{X} \in \mathcal{E}_n]\right] + O\left((\frac{n^2\sigma^4}{\lambda^2} + \|\boldsymbol{\beta}_0\|_2^4)e^{-\frac{c_3}{\kappa^4 C_{cond}^2}n}\right).$$

Moreover if $\|\boldsymbol{\beta}_0\|_\infty = O(1)$ then,

$$\sqrt{\mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}}\left[\left\|\hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\right\|_2^4\right]} \leq O(\frac{\sigma^2 C_{cond}}{C_{\min}}\frac{p}{n}).$$

where the O hides universal constants in $C_{\max}, C_{\min}, C_{cond}, \kappa$ in the final statement.

Proof. Decomposing as

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\right\|_2^4\right] = \mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}}\left[\|\hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\|_2^4 \mathbb{1}[\mathbf{X} \in \mathcal{E}_n]\right] + \mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}}\left[\|\hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\|_q^4 \mathbb{1}[\mathbf{X} \notin \mathcal{E}_n]\right]$$

We can bound the second term explicitly using the Cauchy-Schwarz inequality as,

$$\mathbb{E}[\left\|\hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\right\|_2^4 \mathbb{1}[\mathbf{X} \notin \mathcal{E}_n]] \leq \sqrt{\mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}}\left[\|\hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\|_2^8\right]}\sqrt{\Pr_{\mathbf{X}}[\mathbf{X} \notin \mathcal{E}_n]} \leq O(\frac{n^2\sigma^4}{\lambda^2} + \|\boldsymbol{\beta}_0\|_2^4)e^{-\frac{c_3}{\kappa^4 C_{cond}^2}n}$$

using the crude upper bound from Lemma 18 to upper bound the first term and Lemma 16 to bound the probability in the second term.

For the second statement note that we can bound the first term using the using the independence of $\mathbf{X}, \boldsymbol{\epsilon}$ and Lemma 17, to conclude,

$$\sqrt{\mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}}\left[\|\hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\|_2^4 \mathbb{1}[\mathbf{X} \in \mathcal{E}_n]\right]} \leq O(\frac{\sigma^2 C_{\text{cond}}}{C_{\min}} \frac{p}{n}).$$

With the specific lower bound on λ in the theorem statement, when $\|\boldsymbol{\beta}_0\|_\infty/\sigma_\epsilon = O(1)$ and $n \gtrsim \kappa^4 C_{\text{cond}}^2 p$ we have,

$$\sqrt{O(\frac{n^2\sigma^4}{\lambda^2} + \|\boldsymbol{\beta}_0\|_2^4)e^{-\frac{c_3}{\kappa^4 C_{\text{cond}}^2}n}} \leq O(\frac{\sigma^2 C_{\text{cond}}}{C_{\min}} \frac{p}{n})$$

$\square$

Finally, we prove a simple matrix expectation upper bound,

Lemma 20. *Let* $\mathbf{S} \in \mathbb{R}^{n \times n}$ *be a (deterministic) p.s.d. matrix with rank at most* p *satisfying* $\|\mathbf{S}\|_2 \leq z$, *and let* $\boldsymbol{\epsilon} \in \mathbb{R}^n$ *satisfy Assumption 4. Then*

$$\mathbb{E}\left[(\boldsymbol{\epsilon}^\top \mathbf{S} \boldsymbol{\epsilon})^2\right] \leq O(\sigma^4 z^2 p^2).$$

Proof. This follows by a straightforward computation using the sub-Gaussianity of each ϵ_i:

$$\mathbb{E}\left[(\boldsymbol{\epsilon}^\top \mathbf{S} \boldsymbol{\epsilon})^2\right] \leq O(\sum_i S_{ii}^2 \mathbb{E}[\epsilon_i^4] + \sum_{i \neq j} S_{ij}^2 \mathbb{E}[\epsilon_i^2 \epsilon_j^2] + \sum_{i \neq j} S_{ii} S_{jj} \mathbb{E}[\epsilon_i^2 \epsilon_j^2]) \leq O(\sigma^4 \|S\|_F^2 + \sigma^4 \text{Tr}[S]^2) \leq O(\sigma^4 p^2 z^2).$$

$\square$

G. Experimental Details

G.1. Implementation Details

All algorithms were implemented in Python (with source code to be released to be upon publication). The open-source library scikit-learn was used to fit the Lasso estimator, the cross-validated Lasso estimators, and the random forest regression models used in the synthetic/real data experiments. The convex program for the JM-style estimator was solved using the open-source library CVXPY equipped with the MOSEK solver (Diamond & Boyd, 2016).

Note the debiased estimators presented require either refitting the auxiliary regression for $\mathbf{g}(\cdot)$ (i.e. the Lasso estimator or a random forest) in the case of the OM estimators, or resolving the convex program in Eq. (4) for each new test point $\mathbf{x}_\star$. Although this presents a computational overhead in both our synthetic and real-data experiments, such computations are trivially parallelizable across the test points $\mathbf{x}_\star$. As such, we used the open-source library Ray to parallelize training of the aforementioned models (Moritz et al., 2018). All experiments were run on 48-core instances with 256 GB of RAM.

G.2. Data Preprocessing and Cross-Validation Details

In all of the experiments (both synthetic and real data) the training covariates (in the design $\mathbf{X}$) was first centered and scaled to have features with mean zero and unit variance. Subsequently the vector of $\mathbf{y}$ values was also centered by subtracting its mean; that is $\mathbf{y} \to \mathbf{y} - \bar{\mathbf{y}}$. After any given model was fit the mean $\bar{\mathbf{y}}$ was added back to the (y-centered) prediction θ of the model. On account of this centering, the Lasso estimators were not explicitly fit with an intercept term (we found the performance was unchanged by not performing the demeaning and instead explicitly fitting the intercept for the Lasso baseline). In each case the cross-validated Lasso estimator was fit, the regularization parameter was selected by cross validation over a logarithmically spaced grid containing a 100 values spaced between 10^{-6} and 10^1. The cross-validated ridge estimator was fit by using leave-one-out cross-validation to select the regularization parameter over a logarithmically spaced grid containing a 100 values spaced between 10^{-2} and 10^6 for the synthetic experiments, while a range of 10^{-6} and 10^1 was used for the real data. The ℓ_1 and ℓ_1/ℓ_2 ratio parameter for the elastic net were also set using cross-validation by letting the ℓ_1 regularization parameter over a logarithmically spaced grid containing a 100 values spaced between 10^{-6} and 10^1, while the ℓ_1/ℓ_2 ratio parameter was allowed to range over $[.1, .5, .7, .9, .95, .99, 1]$. In the case of the real data experiments the random forest regressors (RF) used in the $\mathbf{g}(\cdot)$ models were fit using a default value of 50 estimators in each RF.

G.3. JM-style Estimator Details

Note that $\lambda_{\mathbf{w}}$ was chosen for the JM-style estimator using the heuristic to search for the smallest $\lambda_{\mathbf{w}}$ in a set for which the convex program in Eq. (4) is feasible. If no such value existed (i.e. all the programs were infeasible) we defaulted to simply predicting using the base Lasso regression in all cases (which is equivalent to using $\mathbf{w} = 0$).

G.4. OM Estimators Details

As described in the main text, the OM estimators use 2-fold data-splitting. Such a procedure can be sample-inefficient since only a fraction of the data is used in each stage of the procedure. For the OM methods used in the experiments we instead used a more general K-fold cross-fitting as described in (Chernozhukov et al., 2017), with $K = 5$ and $K = 10$.

The OM methods can be fit exactly as described in the paper with the following modifications. First the original dataset is split into K equally-sized folds we denote as $(\mathbf{X}_{\mathcal{I}_1}, \mathbf{y}_{\mathcal{I}_1}), \ldots, (\mathbf{X}_{\mathcal{I}_K}, \mathbf{y}_{\mathcal{I}_K})$; here the index sets range over the datapoints as $\mathcal{I}_1 = \{1, \ldots, \frac{n}{K}\}, \mathcal{I}_2 = \{\frac{n}{K} + 1, \ldots, \frac{2n}{K}\}$ etc... We also use $(\mathbf{X}_{\mathcal{I}_{-i}}, \mathbf{y}_{\mathcal{I}_{-i}})$ to describe K-leave-one-out subsets of the original folds which contain the union of datapoints in all *but* the $\mathcal{I}_i$th fold of data.

Then, K sets of first-stage regressions are trained on the K-leave-one-out subsets to produce $(\mathbf{f}^{-1}, \mathbf{g}^{-1}), \ldots, (\mathbf{f}^{-K}, \mathbf{g}^{-K})$; explicitly the pair $(\mathbf{f}^{-i}, \mathbf{g}^{-i})$ is fit on $(\mathbf{X}_{\mathcal{I}_{-i}}, \mathbf{y}_{\mathcal{I}_{-i}})$. Finally the empirical moment equations can be solved for $\hat{y}_{\text{OM}}$ by summing over the entire dataset, but evaluating the $(\mathbf{f}^{-i}, \mathbf{g}^{-i})$ model on only the ith fold:

$$\sum_{i \in K} \sum_{j \in \mathcal{I}_i} m(t_j, y_j, \hat{y}_{\text{OM}}, \mathbf{z}_j^\top \mathbf{f}^{-i}, \mathbf{g}^{-i}(\mathbf{z}_j)) = 0.$$

The estimator for the variance μ_2 can also be computed in an analogous fashion, $\sum_{i \in K} \sum_{j \in \mathcal{I}_i} t_j(t_j - \mathbf{g}^{-i}(\mathbf{z}_j))$. More details on this procedure can be found in Chernozhukov et al. (2017) and Mackey et al. (2017). Note that since K is chosen to be constant, our theoretical guarantees also apply to this estimator up to constant factors.

Also though the thresholding step (with the parameter τ) is used in our theoretical analysis to control against the denominator μ_2 being too small, we found in practice the estimate of μ_2 concentrated quickly and was quite stable. Hence we found explicitly implementing the thresholding step was unnecessary and we did not include this in our implementation.

G.4.1. OM q MOMENTS

In Section 3.2 we focus our analysis on the OM f moments but also introduce the first-order orthogonal q moments, whose practical efficacy we explore in our real data experiments. For completeness we include the details of the algorithm to predict with q-moments here. The primary difference with respect to the f-moments is with respect to how the $\mathbf{q}$ or $\mathbf{f}$ regression is fit, the $\mathbf{g}$ regression is handled identically. For simplicity, we present the algorithm in parallel to how the $\mathbf{f}$ moments are introduced in the main text (without the K-fold cross-fitting), although K-fold cross-fitting is used in practice exactly as described above.

After the data reparametrization we have $\mathbf{x}_i' = [t_i, \mathbf{z}_i] = (\mathbf{U}^{-1})^\top \mathbf{x}_i$. In the reparametrized basis, the linear model becomes,

$$y_i = \theta t_i + \mathbf{z}_i^\top \mathbf{f}_0 + \epsilon_i \qquad t_i = \mathbf{g}_0(\mathbf{z}_i) + \eta_i$$

where $\mathbf{q}_0(\mathbf{z}_i) = \theta \mathbf{g}_0(\mathbf{z}_i) + \mathbf{z}_i^\top \mathbf{f}_0$.

- The first fold $(\mathbf{X}^{(1)}, \mathbf{y}^{(1)})$ is used to run two *first-stage* regressions. We estimate $\mathbf{q}_0$ using a linear estimator (such as the Lasso) by directly regressing $\mathbf{y}^{(1)}$ onto $\mathbf{z}^{(1)}$ to produce the vector $\hat{\mathbf{q}}$. Second we estimate $\mathbf{g}_0(\cdot)$ by regressing $\mathbf{t}^{(1)}$ onto $\mathbf{z}^{(1)}$ to produce a regression model $\hat{\mathbf{g}}(\cdot) : \mathbb{R}^{p-1} \to \mathbb{R}$.

- Then, we estimate $\mathbb{E}[\eta_1^2]$ as $\mu_2 = \frac{1}{n/2} \sum_{i=n/2+1}^{n} (t_i - \hat{\mathbf{g}}(\mathbf{z}_i))^2$ where the sum is taken over the second fold of data; crucially $(t_i, \mathbf{z})$ are (statistically) independent of $\hat{\mathbf{g}}(\cdot)$ in this expression.

- If $\mu_2 \leq \tau$ for a threshold T we simply output $\hat{y}_{\text{OM}} = \mathbf{x}_\star^\top \hat{\boldsymbol{\beta}}$. If $\mu_2 \geq \tau$ we estimate θ by solving the empirical moment equation:

$$\sum_{i=n/2+1}^{n} m(t_i, y_i, \hat{y}_{\text{OM}}, \mathbf{z}_i^\top \hat{\mathbf{q}}, \hat{\mathbf{g}}(\mathbf{z}_i)) = 0 \implies \hat{y}_{\text{OM}} = \frac{\frac{1}{n/2} \sum_{i=n/2+1}^{n} (y_i - \mathbf{z}_i^\top \hat{\mathbf{q}})(t_i - \hat{\mathbf{g}}(\mathbf{z}_i))}{\mu_2}$$

where the sum is taken over the second fold of data and m is defined in (7).

G.4.2. SYNTHETIC DATA EXPERIMENT DETAILS

The experiments on synthetic data were conducted as described in the main text in Section 4. In each case for the JM-style estimator the base regression was fit using the cross-validated Lasso, while the auxiliary parameter for the regression was chosen to be the smaller of $\sqrt{\log p / n}$ and $0.01\sqrt{\log p / n}$ for which the convex program in Eq. (4) was feasible. The OM f moments were fit as described above using 5-fold cross-fitting with the Lasso estimator (with either theoretically-calibrated values for the hyperparameters or hyperparameters chosen by cross-validation) used for both the first-stage regressions.

In Section 4.1 all hyperparameters wer set to their theoretically-motivated values: $\lambda_{\boldsymbol{\beta}} = \lambda_{\mathbf{g}} = 4\sqrt{\log p / n}$ for the Lasso regressions, and, inspired by the feasibility heuristic of (Javanmard & Montanari, 2014), we set $\lambda_{\mathbf{w}}$ to the smallest value between $\sqrt{\log p / n}$ and $.01\sqrt{\log p / n}$ for which the JM-style program (4) was feasible. The RMSRE in each experiment was computed over 500 test datapoints (i.e., 500 independent $\mathbf{x}_\star$'s) generated from the training distribution; each experiment was repeated 20 times, and the average RMSRE is reported.

G.4.3. REAL DATA EXPERIMENT DETAILS

For the base regression procedures five-fold CV was used to select hyperparameters for the Lasso and elastic net estimators, while leave-one-out CV was used for ridge regression.

OM methods The OM f and q moments were implemented as above with 10-fold cross-fitting. However to exploit the generality of the OM framework in addition to allowing $\hat{\mathbf{g}}(\cdot)$ to be estimated via the cross-validated Lasso estimator, we also allowed $\hat{\mathbf{g}}(\cdot)$ to be estimated via random forest regression, and a $\mathbf{g} = 0$ baseline. However, note that $\hat{\mathbf{f}}$ and $\hat{\mathbf{q}}$ were *always* fit with the cross-validated Lasso (a linear estimator) since our primary purpose is to investigate the impacts of debiasing linear prediction with the $\hat{\mathbf{f}}$ and $\hat{\mathbf{q}}$ moments.

For each $\mathbf{x}_\star$ we fit a cross-validated Lasso estimator, a random forest regressor, and a $\hat{\mathbf{g}} = 0$ baseline on each of the K-leave-one-subsets of data. We adaptively chose between these models in a *data-dependent* fashion by selecting the method that produced the minimal (estimated) variance for $\hat{y}_{\mathrm{OM}}$. We used a plug-in estimate of the asymptotic variance which can be computed as,

$$\text{q-var(method)} = \frac{\sum_{i \in K} \sum_{j \in \mathcal{I}_i} (t_j - \mathbf{g}_{\mathrm{method}}^{-i}(\mathbf{z}_j))^2}{V}$$

and

$$\text{f-var(method)} = \frac{\sum_{i \in K} \sum_{j \in \mathcal{I}_i} t_i (t_j - \mathbf{g}_{\mathrm{method}}^{-i}(\mathbf{z}_j))}{V}$$

where $V_{\mathrm{method}} = \sum_{i \in K} \sum_{j \in \mathcal{I}_i} (t_j - \mathbf{g}_{\mathrm{method}}^{-i}(\mathbf{z}_j))^2 - (\sum_{i \in K} \sum_{j \in \mathcal{I}_i} (t_j - \mathbf{g}_{\mathrm{method}}^{-i}(\mathbf{z}_j))^2$ for each method. These asymptotic variance expressions can be computed from a general formula for the asymptotic variance from Mackey et al. (2017, Theorem 1). Upon selecting the appropriate $\hat{\mathbf{g}}(\cdot)$ method for either the f or q moments the algorithm proceeds as previously described with the given choice of $\hat{\mathbf{g}}(\cdot)$.

JM-style method For the real data experiments the $\lambda_{\mathbf{w}}$ for the JM-style estimator was selected by constructing a logarithmically-spaced grid of 100 values of $\lambda_{\mathbf{w}}$ between 10^{-7} and 10^2 and selecting the smallest value of $\lambda_{\mathbf{w}}$ for which the convex program in Eq. (4) was feasible.

Datasets All regression datasets, in this paper were downloaded from the publicly available UCI dataset repository (Dua & Graff, 2017). The triazines dataset was randomly split in an 80/20 train-test split and selected since $n_{train} \approx p$ for it. The other 4 datasets were selected due to the fact they can be naturally induced to have distributional shift. The Parkinsons and Wine datasets were selected exactly as in Chen et al. (2016). The Parkinsons dataset, where the task is to predict a jitter index, was split into train and test as in Chen et al. (2016), by splitting on the "age" feature of patients: $\leq 60 \to$ train and $> 60 \to$ test. The task for prediction in the Wine dataset, as in Chen et al. (2016), is to predict the acidity levels of wine but given training data comprised only of red wines with a test set comprised only of white wines. In the fertility dataset, where the task is to predict the fertility of a sample, we split into train and test by splitting upon the binary feature of whether patients were in the $18 - 36$ age group ($\to$ train) or not ($\to$ test). Finally, for the Forest Fires dataset, where the task it to predict the burned area of forest fires that occurred in Portugal during a roughly year-long period, we split into train/test based on the "month" feature of the fire: those occurring before the month of September ($\to$ train) and those after the month of September ($\to$ test).

Note in all the cases the feature that was split upon was not used as a covariate in the prediction task. In Table 2 we include further information these datasets,

Table 2. Information on Real Datasets.

Dataset	n_{train}	n_{test}	p	Distrib. Shift?
Fertility	69	31	8	Yes
Forest Fires	320	197	10	Yes
Parkinson	1877	3998	17	Yes
Wine	4898	1599	11	Yes
Triazines	139	47	60	No