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 semi-parametric 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,

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

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

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: **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 asympotically-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

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

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

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 (CIs) for treatment effects.

These pioneering works all focus on improved CI 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.

While past work on transductive regression has demonstrated both empirical and theoretical benefits over induction when many unlabeled test points are simultaneously available (Belkin et al., 2006; Alquier & Hebiri, 2012; Bellec et al., 2018; Chapelle et al., 2000; Cortes & Mohri, 2007; Cortes et al., 2008), none of these works have demonstrated a significant benefit, either empirical or theoretical, from transduction given access to only a single test point. For example, the works (Belkin et al., 2006; Chapelle et al., 2000), while theoretically motivated, provide no formal guarantees on transductive predictive performance and only show empirical benefits for large unlabeled test sets. The transductive Lasso analyses of Alquier & Hebiri (2012); Bellec et al. (2018) provide prediction error bounds identical to those of the inductive Lasso, where only the restricted-eigenvalue constant is potentially improved by transduction. Neither analysis improves the dimension dependence of Lasso prediction in the SP setting to provide $O(1/n)$ rates. The formal analysis of Cortes & Mohri (2007); Cortes et al. (2008) only guarantees small error when the number of unlabeled test points is large. Our aim is to develop single point transductive prediction procedures that improve upon the standard inductive approaches both in theory and in practice.

Our approach also bears some resemblance to semi-supervised learning (SSL) – improving the predictive power of an inductive learner by observing additional unlabelled examples (see, e.g., Zhu, 2005; Bellec et al., 2018). Conventionally, SSL benefits from access to a large pool of unlabeled points drawn from the same distribution as the training data. In contrast, our procedures receive access to only a single arbitrary test point $\mathbf{x}_\star$ (we make no assumption about its distribution), and our aim is accurate prediction for that point. We are unaware of SSL results that benefit significantly from access to single unlabeled point $\mathbf{x}_\star$.

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,* $\mathbb{E}[\exp(\mathbf{v}^\top \mathbf{x}_i)] \leq \exp\left(\kappa^2 \|\boldsymbol{\Sigma}^{1/2}\mathbf{v}\|^2/2\right)$ *.*

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

Throughout, we 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 define $[p] = \{1, \ldots, p\}$ and $p \vee n = \max(p, n)$. 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 $\boldsymbol{\beta}_0$, $\text{supp}(\boldsymbol{\beta}_0) = \{j : (\boldsymbol{\beta}_0)_j \neq 0\}$ and $\mathbb{B}_0(s)$ refers to the set of s-sparse vectors in $\mathbb{R}^p$. We 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 $\hat{\boldsymbol{\beta}}_R(\lambda) \triangleq \text{argmin}_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$ 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* $n \geq p \geq 20$,

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

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* $\text{SNR} \triangleq \|\boldsymbol{\beta}_0\|_2^2/\sigma_\epsilon^2$,

$$\lambda_* \triangleq \operatorname{argmin}_\lambda \mathbb{E}[\langle \tilde{\mathbf{x}}, \hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0 \rangle^2] =$$
$$\operatorname{argmin}_\lambda \mathbb{E}[\|\hat{\boldsymbol{\beta}}_R(\lambda) - \boldsymbol{\beta}_0\|_2^2] = \tfrac{p}{\text{SNR}}, \textit{ and, for } n \geq \tfrac{1}{6}\tfrac{p}{\text{SNR}},$$
$$\mathbb{E}[\langle \mathbf{x}_\star, \hat{\boldsymbol{\beta}}_R(\lambda_*) - \boldsymbol{\beta}_0 \rangle^2] \geq \tfrac{p^2}{n\text{SNR}} \cdot \|\mathbf{x}_\star\|_2^2 \cdot \tfrac{\sigma_\epsilon^2}{n} \cdot \tfrac{\cos(\mathbf{x}_\star, \boldsymbol{\beta}_0)^2}{784}.$$

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 $\mathbb{E}[(\tilde{\mathbf{y}} - \tilde{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}_R(\lambda))^2] - \sigma_\epsilon^2$ for a new datapoint $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ 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 $\text{SNR} = \frac{1}{6}\frac{p}{n}$ yields

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

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

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

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* $s \geq 0$*, 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* $p \geq 20$,[2] *then there exist universal constants* $c_{1:3}$ *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* $q \in [1, \infty]$,

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

[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.

[2] The cutoff at 20 is arbitrary and can be decreased.

We make several comments regarding these results. First, 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 (2019, 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 for Lasso prediction error (Wainwright, 2019, Example 7.14). 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 $(\hat{y} - \langle \mathbf{x}_\star, \boldsymbol{\beta}_0 \rangle)^2$. Throughout this section, the $O(\cdot)$ masks constants depending only on $\kappa, C_{\min}, C_{\max}, C_{\text{cond}}$.

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[3]:

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

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

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 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 $n \geq c_1 a^2 \log(2e(p \vee n))$,*

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

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

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.3.

Corollary 3. *Under the conditions of Theorem 3, consider the JM-style estimator (3) with pilot estimate $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_L(\lambda)$ with $\lambda \geq 80\sigma_\epsilon \sqrt{\frac{\log\left(2ep/s_{\boldsymbol{\beta}_0}\right)}{n}}$. 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 $n \geq c_2 \max\{\frac{s_{\boldsymbol{\beta}_0}\kappa^4}{C_{\min}}, a^2\} \log(2e(p \vee n))$,*

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

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

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 estimator 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 (3) and (4) 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 $\hat{y}_{\text{JM}}$ (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 $\kappa, C_1, C_2, C_3, C_{\min}, C_{\max}, C_{cond}$.

The proof can be found in Appendix D.2.

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,

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

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):

$$\begin{aligned} \textbf{f moments: } & m(t_i, y_i, \theta, \mathbf{z}_i^\top \mathbf{f}, \mathbf{g}(\mathbf{z}_i)) = \\ & \quad (y_i - t_i\theta - \mathbf{z}_i^\top \mathbf{f})(t_i - \mathbf{g}(\mathbf{z}_i)) \qquad (7) \\ \textbf{q moments: } & m(t_i, y_i, \theta, \mathbf{q}(\mathbf{z}_i), \mathbf{g}(\mathbf{z}_i)) = \\ & \quad (y_i - \mathbf{q}(\mathbf{z}_i) - \theta(t_i - \mathbf{g}(\mathbf{z}_i)))(t_i - \mathbf{g}(\mathbf{z}_i)). \end{aligned}$$

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 $\mathbf{f}$ moments since the analysis is similar for the $\mathbf{q}$, 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[4] disjoint, equal-sized folds $(\mathbf{X}^{(1)}, \mathbf{y}^{(1)}) = \{(\mathbf{x}_i, y_i) : i \in \{1, \dots, \frac{n}{2}\}\}$ and $(\mathbf{X}^{(2)}, \mathbf{y}^{(2)}) = \{(\mathbf{x}_i, y_i) : i \in \{\frac{n}{2}+1, \dots, n\}\}$. Then,

- 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 $\hat{y}_{\text{OM}} = \mathbf{x}_\star^\top \hat{\boldsymbol{\beta}}$. If $\mu_2 \geq \tau$ we estimate θ_0 by solving the empirical moment equation:

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

  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 $\hat{y}_{\text{OM}}$. 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[5]:

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

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 $\mathbf{g}_0 = \operatorname{argmin}_{\mathbf{g}} \mathbb{E}[(t_1 - \mathbf{z}_1^\top \mathbf{g})^2]$. Then the thresholded orthogonal ML estimator $\hat{y}_{\text{OM}}$ of (8) with $\tau = \frac{1}{4}\sigma_\eta^2$ satisfies*

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

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.

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 $c_{1:5}$ 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 $n \geq c_5 \kappa^4 C_{cond}^2 p$,*

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

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

[4] 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.

[5] 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.

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 $n \geq \frac{c_1\kappa^4}{C_{\min}}\max\{s_{\boldsymbol{\beta}_0}, s_{\mathbf{g}}\}\log(2ep)$,*

$$\mathbb{E}[(\hat{y}_{\text{OM}} - \mathbf{x}_\star^\top \boldsymbol{\beta}_0)^2] \leq$$

$$\|\mathbf{x}_\star\|_2^2\left[O(\tfrac{\sigma_\epsilon^2}{\sigma_\eta^2 n}) + O(\tfrac{\lambda_{\boldsymbol{\beta}}^2\lambda_{\mathbf{g}}^2 s_{\boldsymbol{\beta}_0}s_{\mathbf{g}_0}}{(\sigma_\eta^2)^2}) + O(\tfrac{\lambda_{\boldsymbol{\beta}}^2 s_{\boldsymbol{\beta}_0}\sigma_\eta^2 + \lambda_{\mathbf{g}}^2 s_{\mathbf{g}_0}\sigma_\epsilon^2}{(\sigma_\eta^2)^2 n})\right].$$

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 $n = \Omega(p^2)$, $\text{SNR} = o(\frac{p^2}{n})$, $\cos(\mathbf{x}_\star, \boldsymbol{\beta}_0)^2 = \Theta(1)$ and $\text{SNR} \gtrsim \frac{p}{n}$. Then, the upper bound[6] 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 $\text{SNR} = \frac{1}{6}\frac{p}{n}$. 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 provide complete experimental set-up details in Appendix G and code replicating all experiments at https://github.com/nileshtrip/SPTransducPredCode.

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, 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 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 $\frac{p}{s_{\boldsymbol{\beta}_0}} = 2$).

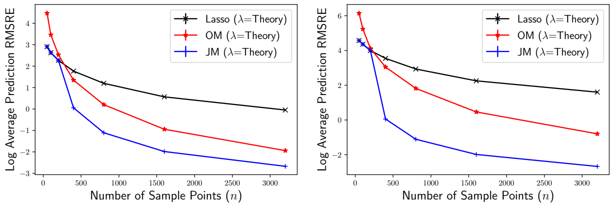

Figure 1. Lasso vs. OM and JM Lasso prediction without distribution shift. Hyperparameters are set according to theory (see Section 4.1). Left: $p = 200, s_{\boldsymbol{\beta}_0} = 20$. Right: $p = 200, s_{\boldsymbol{\beta}_0} = 100$.

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 plots in Figure 2 show the OM estimator with λ_*-ridge pilot provides significant gains over the baseline λ_*-ridge estimator.

In Figure 4 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

[6] 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.

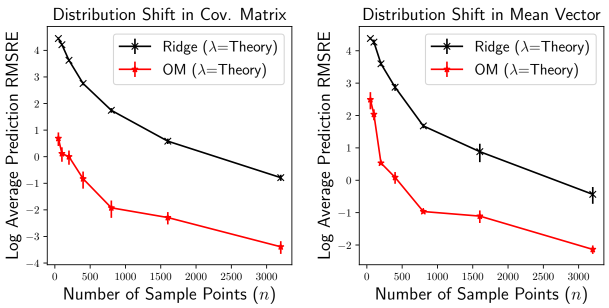

Figure 2. Ridge vs. OM ridge prediction ($p = 200$) under train-test distribution shift. Hyperparameters are set according to theory.

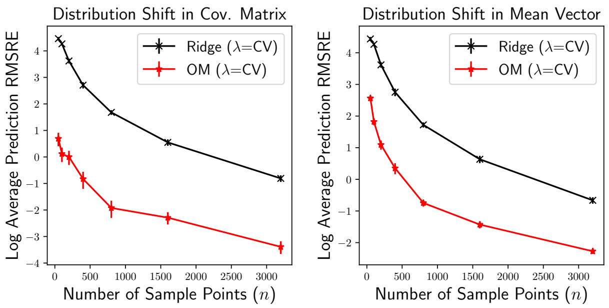

Figure 3. Ridge vs. OM ridge prediction ($p = 200$) under train-test distribution shift. Hyperparameters are set according to CV.

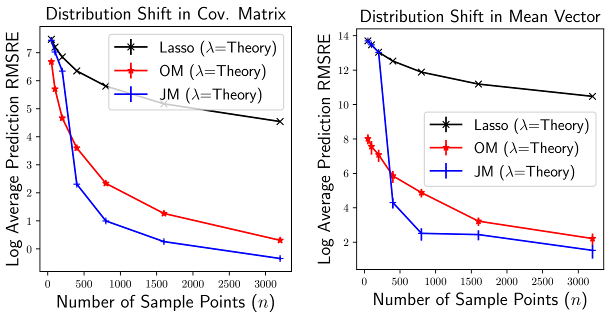

Figure 4. Lasso vs. OM and JM Lasso prediction ($p = 200$) under mean ($s_{\beta_0} = 100$) or covariance ($s_{\beta_0} = 20$) train-test distribution shifts. Hyperparameters are set according to theory.

mean shifting, we generated $\mathbf{x}_\star \sim \mathcal{N}(10\boldsymbol{\beta}_0, \mathbf{I}_p)$ for each problem instance. The first and second plots in Figure 4 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 Figure 3 and Figure 5 compares CV-tuned ridge or Lasso to OM and JM with CV-tuned base procedures—showing the benefit of transduction in this practical setting where regularization hyperparameters are chosen by CV. As the first and second plots in Figure 3 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 first and second plots in Figure 5 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 $\mathbf{x}_\star$.

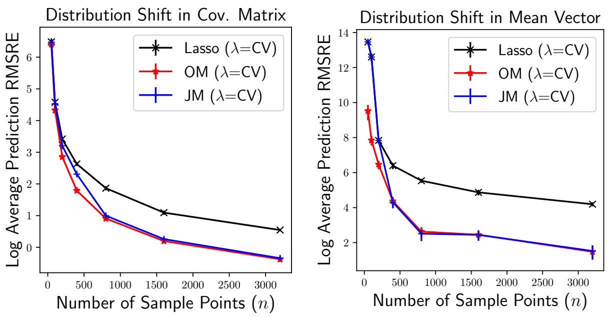

Figure 5. Lasso vs. OM and JM Lasso prediction ($p = 200$) under mean ($s_{\beta_0} = 100$) or covariance ($s_{\beta_0} = 20$) train-test distribution shifts. Hyperparameters are set according to CV.

4.3. Improving Cross-validated Prediction

Motivated by our findings on synthetic data, we next report the performance of our methods on 5 real datasets with and without distribution shift. We also include the popular elastic net estimator as a base regression procedure alongside ridge and the Lasso. 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 auxiliary $\mathbf{g}$ 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. The $\mathbf{f}$ and $\mathbf{q}$ regressions were always fit with Lasso, ridge, or elastic net estimation. See Appendix G for further details on the methodology and datasets from the UCI dataset repository (Dua & Graff, 2017).

In Table 1 we see that the OM estimators generically provide gains over the CV Lasso, CV ridge, 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 $\boldsymbol{\beta}_0$. OLS fares worse than most methods on each dataset due to an increase in variance. In contrast, our proposed transductive procedures limit the variance intro-

Table 1. Test set RMSE of OLS; CV-tuned ridge, Lasso, and elastic net; OM and JM transductive CV-tuned ridge, Lasso, and elastic net; and prior transductive approaches (TD Lasso, Ridge, and KNN) on real-world datasets. All hyperparameters are set via CV. Error bars represent a delta method interval based on ±1 standard error of the mean squared error 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	0.4092±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
TD Lasso (Alquier & Hebiri, 2012)	0.9813±0.0154	12.2535±0.1358	82.0657±36.0320	0.4092±0.0716	0.1483±0.0237
TD Ridge (Chapelle et al., 2000)	0.8411±0.0004	12.2534±0.0021	82.0664±2.567	0.4089±0.0128	0.1735±0.0004
TD KNN (Cortes & Mohri, 2007)	0.8345±0.0153	12.3326±0.1447	81.9467±35.8340	**0.3845±0.0760**	0.1510±0.0240

duced by targeting a single parameter of interest, $\langle \mathbf{x}_\star, \boldsymbol{\beta}_0 \rangle$.

Finally, we evaluated three existing transductive prediction methods—the transductive Lasso (TD Lasso) of (Alquier & Hebiri, 2012; Bellec et al., 2018), transductive ridge regression (TD Ridge) (Chapelle et al., 2000), and transductive ridge regression with local (kernel) neighbor labelling (TD KNN) (Cortes & Mohri, 2007)—on each dataset, tuning all hyperparameters via CV. TD Lasso does not significantly improve upon the Lasso baseline on any dataset. TD Ridge only improves upon the baselines on Wine but is outperformed by OM q. TD KNN also underperforms OM q on every dataset except Fertility.

5. Discussion and Future Work

We presented two single point transductive prediction procedures that, given advanced knowledge of a test point, can significantly improve the prediction error of an inductive learner. We provided theoretical guarantees for these procedures and demonstrated their practical utility, especially under distribution shift, on synthetic and real data. Promising directions for future work include improving 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 models such as generalized linear models and kernel machines.

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.

Belkin, M., Niyogi, P., and Sindhwani, V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(Nov):2399–2434, 2006.

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.

Chapelle, O., Vapnik, V., and Weston, J. Transductive inference for estimating values of functions. In *Advances in Neural Information Processing Systems*, pp. 421–427, 2000.

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.

Cortes, C. and Mohri, M. On transductive regression. In *Advances in Neural Information Processing Systems*, pp. 305–312, 2007.

Cortes, C., Mohri, M., Pechyony, D., and Rastogi, A. Stability of transductive regression algorithms. In *Proceedings of the 25th international conference on Machine learning*, pp. 176–183, 2008.

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. 2019.

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.