The Relative Instability of Model Comparison with Cross-validation

Alexandre Bayle
Harvard University
alexandre_bayle@g.harvard.edu

Lucas Janson
Harvard University
ljanson@fas.harvard.edu

Lester Mackey
Microsoft Research New England
lmackey@microsoft.com

Abstract

Existing work has shown that cross-validation (CV) can be used to provide an asymptotic confidence interval for the test error of a stable machine learning algorithm, and existing stability results for many popular algorithms can be applied to derive positive instances where such confidence intervals will be valid. However, in the common setting where CV is used to compare two algorithms, it becomes necessary to consider a notion of *relative* stability which cannot easily be derived from existing stability results, even for simple algorithms. To better understand relative stability and when CV provides valid confidence intervals for the test error difference of two algorithms, we study the soft-thresholded least squares algorithm, a close cousin of the Lasso. We prove that while stability holds when assessing the individual test error of this algorithm, relative stability *fails* to hold when comparing the test error of two such algorithms, even in a sparse low-dimensional linear model setting. Additionally, we empirically confirm the invalidity of CV confidence intervals for the test error difference when either soft-thresholding or the Lasso is used. In short, caution is needed when quantifying the uncertainty of CV estimates of the performance difference of two machine learning algorithms, even when both algorithms are individually stable.

1 Introduction

In the machine learning and statistics literature, cross-validation (CV) [27, 15] is routinely used to compare the performance of learning algorithms. In practice, it is not uncommon to pair CV's point estimates with uncertainty quantification in the form of estimated standard errors or putative confidence intervals. Yet the validity of such uncertainty quantification has been poorly understood until recently, and it is now understood to be closely related to notions of algorithmic stability [2, 4]. Stability of algorithms has long been studied in the learning theory literature, allowing existing stability results to be applied to CV uncertainty quantification for assessing the performance of a single algorithm. However, when comparing two algorithms' performances, their individual stabilities do not directly translate to the type of stability needed for valid CV uncertainty quantification, raising the question of when such uncertainty quantification is valid.

Our contributions This work lies at the interface of algorithmic stability and cross-validation. We demonstrate the importance of considering *relative* stability by studying the soft-thresholded least squares algorithm [13], or soft-thresholding (ST) for short, a close cousin of the Lasso [28]. In the canonical fixed-dimensional linear regression setting of Section 3, we tightly characterize the

Preprint. Under review.

components of relative stability and show that while assessment of a single ST fit satisfies relative stability (Theorem 1), comparison of two ST fits with similar (but different) tuning parameters does not (Theorem 2), calling into question the validity of CV confidence intervals for such a comparison. Simulations in Section 5 support these conclusions, showing that CV confidence intervals provide accurate coverage of the test error of a single ST fit even for moderate sample sizes, while they fail to cover the difference in test errors between two ST fits even for very large sample sizes. We empirically find the same dichotomy for the Lasso but not for ridge regression.

Related work The importance of the stability of an algorithm with respect to its generalization error [6] has prompted numerous studies of the stability of popular classes of algorithms [6, 14, 16, 8, 1]. Across the years, different notions of stability have been introduced [10, 11, 18, 21, 17, 20] and building upon the domain of algorithmic stability, multiple papers [17, 20, 8, 2, 4] have established interesting relationships between the theoretical properties of cross-validation and the stability properties of the algorithms involved. Austern and Zhou [2] and Bayle et al. [4] derive central limit theorems and consistent variance estimators for the CV estimator under sufficient conditions on the *loss stability* [20] or *mean-square stability* [17], which are known to decay to zero for a variety of algorithms. However, to our knowledge, no prior works have assessed the sufficient conditions for asymptotic normality in the case when the asymptotic variance in these central limit theorems goes to zero, as would be expected in the common scenario of comparing the performance of two algorithms that converge to the same prediction rule (e.g., if they are both consistent for the optimal prediction rule). This is the focus of this paper, leading to novel negative results about stability and validity of CV confidence intervals even in very regular settings. We note that some recent works have studied various other aspects of asymptotic distributional properties of CV [22, 23, 3], but none present negative results comparable to ours.

Notation For each $n \in \mathbb{N}$, we define the set $[n] \triangleq \{1, \ldots, n\}$. For deterministic sequences $(f_n)_n$ and $(g_n)_n$, we write $f_n = \omega(g_n)$ to mean that $g_n = o(f_n)$ as $n \to \infty$, we write $f_n = \Omega(g_n)$ to mean that $g_n = O(f_n)$ as $n \to \infty$, and we write $f_n = \Theta(g_n)$ to mean that $f_n = O(g_n)$ and $f_n = \Omega(g_n)$ as $n \to \infty$. Finally, we write $f_n \sim g_n$ to mean that $\frac{f_n}{g_n} \to 1$ as $n \to \infty$.

2 Preliminaries

Before presenting our results, we establish some necessary definitions, largely following the notation and language of [4].

We will consider a sequence $(Z_i)_{i\geq 0}$ of random data points taking values in a set $\mathcal{Z}$ and a scalar loss function $h_n(Z_0, \mathbf{Z})$ where $\mathbf{Z}$ is a training set of size n. A typical choice for h_n in the regression setting is squared error loss,

$$h_n(Z_0, \mathbf{Z}) = (Y_0 - \hat{f}(X_0; \mathbf{Z}))^2,$$

applied to the predicted response value of a test point $Z_0 = (X_0, Y_0)$, obtained from an algorithm fitting a prediction rule $\hat{f}(\cdot; \mathbf{Z})$ to training data $\mathbf{Z}$. When comparing the performance of two algorithms, we will choose h_n to be the difference between the losses of two prediction rules. In order to ensure a smooth read when we switch between the settings of single algorithm assessment and comparison of algorithms, we will make the distinction clear by adding a superscript to h_n: h_n^{sing} and h_n^{diff}, respectively. In addition, our asymptotic statements should all be interpreted as taking $n \to \infty$.

For the purpose of illustrating the importance of considering stability in a relative sense rather than an absolute sense, we will now define a notion of relative stability based on loss stability. We introduce the definition of loss stability in the case of algorithms that yield a learned predictor independent from the order of the training points, which will be our focus here, and we can then evaluate the impact of replacing only the first point in the training set.

Definition 1 (Relative loss stability)**.** *For $n > 0$, let Z_0 and $Z_1', Z_1, \ldots, Z_n$ be i.i.d. data points with $\mathbf{Z} = (Z_1, \ldots, Z_n)$ and $\mathbf{Z}' = (Z_1', Z_2, \ldots, Z_n)$. For any function $h_n : \mathcal{Z} \times \mathcal{Z}^n \to \mathbb{R}$, the* loss stability *[20] is defined as*

$$\gamma(h_n) \triangleq \mathbb{E}[(h_n(Z_0, \mathbf{Z}) - \mathbb{E}[h_n(Z_0, \mathbf{Z}) \mid \mathbf{Z}] - (h_n(Z_0, \mathbf{Z}') - \mathbb{E}[h_n(Z_0, \mathbf{Z}') \mid \mathbf{Z}']))^2].$$

We also define $\sigma^2(h_n) \triangleq \mathrm{Var}(\mathbb{E}[h_n(Z_0, \mathbf{Z}) \mid Z_0])$. *And finally we can define the* relative loss stability *as*

$$r(h_n) \triangleq \tfrac{n \cdot \gamma(h_n)}{\sigma^2(h_n)}.$$

We introduced these quantities for any function h_n, but we will generically refer to the loss stability and the relative loss stability of an algorithm or a comparison of algorithms when h_n is clear from context. Note that we include the factor of n in the definition of the relative loss stability because it facilitates reasoning about this quantity in a relative manner, allowing it to always be compared to 1. We will say that an algorithm or a comparison of algorithms satisfies the *relative loss stability condition* if $r(h_n) = o(1)$, which is equivalent to a key sufficient condition for the central limit theorem and consistent variance estimation for cross-validation proved in [4]. We will illustrate the importance of relative stability for CV by studying the soft-thresholded least squares regression algorithm in the linear regression setting.

Throughout, we will consider i.i.d. data points $Z_i = (X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}, i = 1, \ldots, n$ from the linear model

$$\mathbf{Y} = \mathbf{X}\beta^\star + \varepsilon, \quad \varepsilon \perp\!\!\!\perp X \tag{2.1}$$

parametrized by the unknown vector $\beta^\star \in \mathbb{R}^p$, where $\mathbf{Y} = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$ is the vector of response variables or targets, $\mathbf{X} = (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n \times p}$ is the matrix of regressors or features with $X_i \sim \mathcal{N}(0, \mathbf{I})$, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \in \mathbb{R}^n$ is the noise vector with normally distributed elements $\varepsilon_i \sim \mathcal{N}(0, \tau^2)$ for some $\tau > 0$. One of our proofs relies on the assumption $\|\beta^\star\|_0 < p$, and our simulations in Section 5 confirm its importance.

The loss function considered for a linear prediction rule will be the squared error loss

$$h_n^{\text{sing}}(Z_0, \mathbf{Z}) \triangleq (Y_0 - X_0^\top \hat{\beta})^2,$$

where the estimated parameter vector $\hat{\beta}$ is learned from the training set $\mathbf{Z} = (Z_1, \ldots, Z_n)$. When the focus is on the comparison of two prediction rules, the loss function will be defined as the difference of two such individual losses in the form

$$h_n^{\text{diff}}(Z_0, \mathbf{Z}) \triangleq (Y_0 - X_0^\top \hat{\beta}^{(1)})^2 - (Y_0 - X_0^\top \hat{\beta}^{(2)})^2$$

for $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$ both learned on the training set $\mathbf{Z}$.

A very classic way to estimate $\beta^\star$ is the ordinary least squares (OLS) estimator defined as

$$\hat{\beta}_{\text{OLS}} \triangleq (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

Note that the OLS estimator has a dependence on the sample size n. For the sake of simplicity, we will leave this dependence implicit, but it will underlie asymptotic results which involve $\hat{\beta}_{\text{OLS}}$. When we expect the parameter vector $\beta^\star$ to exhibit some level of sparsity, that is to say it has some number of zero coefficients, a popular estimator used is the Lasso estimator [28] for some choice of penalization parameter λ which determines the level of sparsity in the learned parameter vector. A simpler cousin of the Lasso that we adopt to ease our analysis is soft-thresholded least squares.

Definition 2 (Soft-thresholding (ST))**.** *We define the soft-thresholding estimator ST*(λ_n) *element-wise as*

$$\hat{\beta}_{\lambda_n, i} \triangleq \mathrm{sign}(\hat{\beta}_{\text{OLS},i})(|\hat{\beta}_{\text{OLS},i}| - \tfrac{\lambda_n}{n})_+, \quad i = 1, \ldots, p.$$

Remark ST is known to exactly match the Lasso when the features are orthogonal [28], which is approximately the case in our setting since our features are independent. Indeed, we will see that the theoretical lessons learned on ST will hold empirically equally well for the Lasso as for ST.

3 Main Results

We state the two main results of this paper below. The first one relates to ST in the setting of single algorithm assessment and details why it satisfies the relative loss stability condition, while the second one focuses on ST in the comparison setting and reveals the fact that under realistic conditions on the penalization parameters, it does not satisfy the relative loss stability condition.

For simulations with features and targets sampled in the same conditions as the theorems, we observed that the values selected for λ_n via CV are concentrated around a constant times $\sqrt{n}$. It therefore makes sense to compare two versions of ST with penalization of order $\sqrt{n}$, and we do so by setting the base level of penalization to λ_n of order $\sqrt{n}$ and parameterizing the difference in penalization of the ST algorithms by δ_n of order 1. Note that both λ_n and δ_n are assumed deterministic in the theorems, but we will present simulations with stochastic λ_n selected via inner cross-validation in Section 5. Under some regularity conditions on the features, Knight and Fu [19, Thm. 1] proved that choosing $\lambda_n = o(n)$ ensures weak consistency of the Lasso estimator for $\beta^\star$, i.e. it converges in probability to $\beta^\star$, and it is therefore natural that the regimes we study are always within this weak consistency regime. As for the $\sqrt{n}$ order of the penalization specific to our second result, it has been shown to be a regime of interest for variable selection consistency [29, 30].

Theorem 1 (Relative stability of individual soft-thresholding)**.** *Assume the linear model* (2.1)*, with feature and noise distributions as given immediately following its equation. For the single algorithm assessment of ST*(λ_n)*, the loss* h_n^{sing} *is defined as* $h_n^{\text{sing}}(Z_0, \mathbf{Z}) = (Y_0 - X_0^\top \hat{\beta}_{\lambda_n})^2$*. If* $\lambda_n = o(n)$*, then*

$$\sigma^2(h_n^{\text{sing}}) \to 2\tau^4\,, \quad \gamma(h_n^{\text{sing}}) \sim \tfrac{C}{n^2}$$

for a constant $C > 0$ *whose explicit expression is given in* (D.1)*, and thus ST satisfies the relative loss stability condition since*

$$r(h_n^{\text{sing}}) \sim \tfrac{C}{2\tau^4} \cdot \tfrac{1}{n} = o(1).$$

The proof of Theorem 1 can be found in Appendix B. Hardt et al. [16] proved stochastic gradient descent on convex objectives (of which ST is a special case) to have $O(1/n)$ *uniform stability* [6], which implies a loss stability of $O(1/n^2)$ by [17, Lem. 1] and [20, Lem. 2]. Thus, Theorem 1 proves a stronger result on ST's loss stability by establishing $1/n^2$ to be its exact rate.

Theorem 2 (Relative instability of soft-thresholding comparison)**.** *Assume the linear model* (2.1)*, with feature and noise distributions as given immediately following its equation, and* $\|\beta^\star\|_0 < p$*. For the algorithm comparison of ST*(λ_n) *with ST*($\lambda_n + \delta_n$)*, the loss* h_n^{diff} *is defined as* $h_n^{\text{diff}}(Z_0, \mathbf{Z}) = (Y_0 - X_0^\top \hat{\beta}_{\lambda_n})^2 - (Y_0 - X_0^\top \hat{\beta}_{\lambda_n + \delta_n})^2$*. If* $\lambda_n = O(\sqrt{n})$*,* $\lambda_n = \omega(1)$*, and* $\delta_n = \Theta(1)$*, then*

$$\tfrac{n^2}{\delta_n^2}\, \sigma^2(h_n^{\text{diff}}) \to 4\tau^2 \|\beta^\star\|_0\,, \quad \gamma(h_n^{\text{diff}}) = \Omega(\tfrac{1}{n^2\sqrt{n}})$$

and thus the ST comparison does not satisfy the relative loss stability condition since

$$r(h_n^{\text{diff}}) = \Omega(\sqrt{n}) \neq o(1).$$

The proof of Theorem 2 can be found in Appendix E. We can think of Theorem 2 as a stylized version of a setting where one wants to compare two similar machine learning algorithms, such as when the two only differ by a tuning parameter. Then, even if both algorithms are individually well-behaved, their comparison may not be.

4 Importance of Relative Stability for Cross-validation

To connect our results on relative stability back to CV and prepare for our numerical experiments, we need to introduce some further notation. We have been using n for the size of the training sets used in the iterations of cross-validation, while Bayle et al. [4] use it for the sample size of the larger set on which CV is run. For the sake of simplicity, we will write k instead of k_n to denote the number of folds even though it can depend on n (leave-one-out cross-validation corresponds to $k = n + 1$), and we will assume that $k - 1$ evenly divides n. The sample size of the larger set is then simply equal to $\frac{nk}{k-1}$.

For B a vector of indices in $[\frac{nk}{k-1}]$, we denote by Z_B the subvector of $(Z_1, \ldots, Z_{\frac{nk}{k-1}})$ which follows the ordering of B. When assigning points to the training set and validation set, we can refer to *train-validation splits* (B, B') based on the corresponding vectors of indices in $[\frac{nk}{k-1}]$. It is typically assumed that every data point is either in the training or validation set, that is B and B' form a partition of $[\frac{nk}{k-1}]$, leading to sizes n and $\frac{n}{k-1}$ for B and B', respectively.

Consider $\{(B_j, B'_j)\}_{j=1}^k$ a set of k train-validation splits such that $[\frac{nk}{k-1}]$ is partitioned into k folds by the validation indices $\{B'_j\}_{j=1}^k$, and a scalar loss function $h_n(Z_i, Z_B)$, we define the *k-fold cross-validation error*

$$\hat{R}_n \triangleq \frac{k-1}{nk} \sum_{j=1}^k \sum_{i \in B'_j} h_n(Z_i, Z_{B_j})$$

and the inferential target, the *k-fold test error*

$$R_n \triangleq \frac{k-1}{nk} \sum_{j=1}^k \sum_{i \in B'_j} \mathbb{E}[h_n(Z_i, Z_{B_j}) \mid Z_{B_j}]. \tag{4.1}$$

In our notation, Bayle et al. [4] use the stability condition $\gamma(h_n) = o(\frac{\sigma^2(h_n)}{n})$, equivalent to $r(h_n) = o(1)$, to prove the central limit theorem

$$\frac{\sqrt{\frac{nk}{k-1}}}{\sigma(h_n)}(\hat{R}_n - R_n) \xrightarrow{d} \mathcal{N}(0,1). \tag{4.2}$$

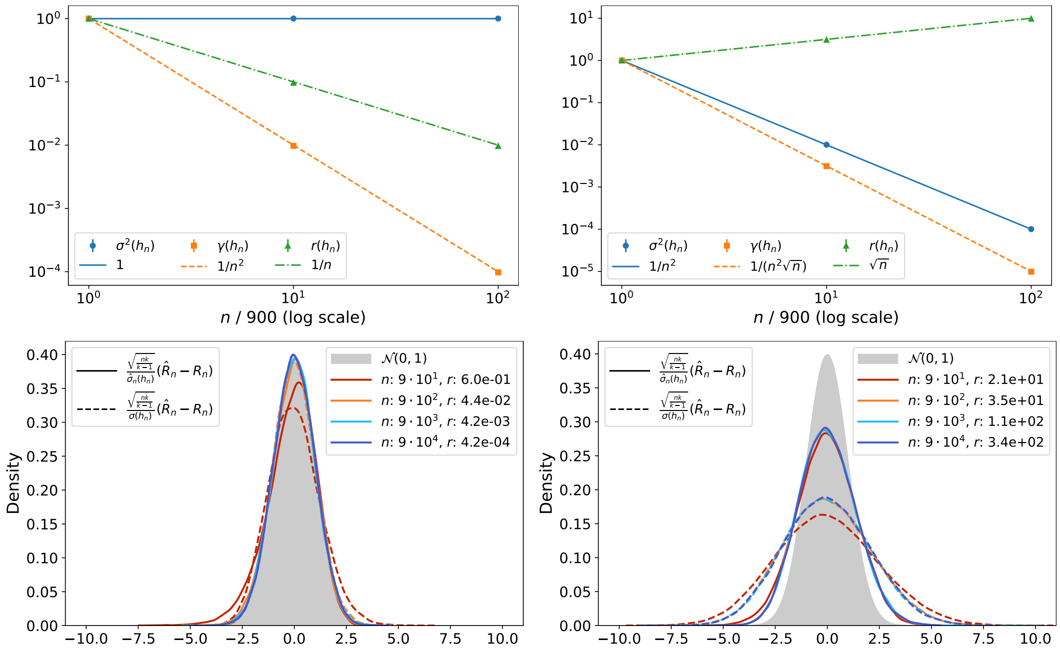
Along with an estimator $\hat{\sigma}_n^2(h_n)$ provided in Bayle et al. [4] and proved to be consistent for $\sigma^2(h_n)$ therein, this central limit theorem enables the construction of asymptotically valid confidence intervals for R_n. Note that it is indeed possible to use the training sample size in the denominator of the stability condition rather than the full sample size of the dataset on which CV is run, as $\frac{nk}{k-1} = \Theta(n)$ for any choice of $k > 1$.

When assessing a single algorithm, unless we are in a fully noiseless setting, we might expect $\sigma^2(h_n^{\text{sing}})$ to be of constant order in general. This means the loss stability condition simplifies to a condition on stability in the absolute sense $\gamma(h_n^{\text{sing}}) = o(1/n)$. For instance, we show in Lemma C.2 that in the linear model with noise, for any linear predictor satisfying some consistency condition, $\sigma^2(h_n^{\text{sing}})$ converges to a positive constant. However, when comparing two consistent algorithms, we then expect $\sigma^2(h_n^{\text{diff}})$ to go to 0 for algorithms whose performances become increasingly similar when the sample size grows, and this is when reasoning about stability in an absolute sense, as has been the focus in past literature, becomes insufficient. In fact, in Theorem 2 it turns out that $\gamma(h_n^{\text{diff}}) = O(1/n^2)$ (see Appendix G), so the ST comparison is loss stable in the absolute sense. But the *relative* loss stability condition fails because it properly accounts for the fact that $\sigma^2(h_n^{\text{diff}})$ goes to zero at a $1/n^2$ rate.

5 Numerical Experiments

We performed numerical experiments to empirically confirm the theoretical results of Section 3 for ST. We sampled the features from $\mathcal{N}(0, \mathbf{I})$, the target variables from the linear model (2.1) with parameter vector $\beta^\star = (3, 1, -5, 3, 0, 0, 0, 0, 0, 0)$ of dimension 10, and the independent noise terms from $\mathcal{N}(0, \tau^2)$ with $\tau = 10$. We fix $k = 10$. To satisfy the assumptions of Theorems 1 and 2, we choose $\lambda_n = \sqrt{n}$ for the base level of penalization, and when comparing algorithms, we set $\delta_n = 1$ for the difference in the penalization parameters. We used Monte Carlo estimation to compute both $\sigma^2(h_n)$ and $\gamma(h_n)$, leveraging Lemmas H.1 and H.2 proved in Appendix H. We also provide additional details about the experiments in Appendix H. Open-source Python code replicating all experiments can be found in the supplemental material.

We present two types of plot. The first type displays the rates for $\sigma^2(h_n)$, $\gamma(h_n)$ and $r(h_n)$ on the log–log scale by plotting their values with dots. To facilitate the visual identification of their rates, we plot lines for the corresponding rates. We display the values with a ± 2 standard error confidence band, with details on how to obtain it for $r(h_n)$ in Appendix H. Note that, thanks to the large number of Monte Carlo replications used, the error bars are very small and thus are not visible. For the second type, using kernel density estimation (KDE), we plot the probability density function across sample sizes of both $\frac{\sqrt{\frac{nk}{k-1}}}{\sigma(h_n)}(\hat{R}_n - R_n)$ and $\frac{\sqrt{\frac{nk}{k-1}}}{\hat{\sigma}_n(h_n)}(\hat{R}_n - R_n)$, where $\hat{\sigma}_n^2(h_n)$ is the within-fold variance estimator introduced in [2, Prop. 1] and proved to be consistent for $\sigma^2(h_n)$ under the relative loss stability condition in [4, Thm. 4]. We expect convergence in distribution to $\mathcal{N}(0,1)$ under the relative loss stability condition thanks to the combination of results of Bayle et al. [4, Thms. 1, 2, and 4], we thus shade the area below the curve of the probability density function of $\mathcal{N}(0,1)$ to make it clearer when the probability density function curves match or not. From its definition (4.1), note that R_n is straightforward to compute in the simulations thanks to Lemma H.2.



(a) Single algorithm

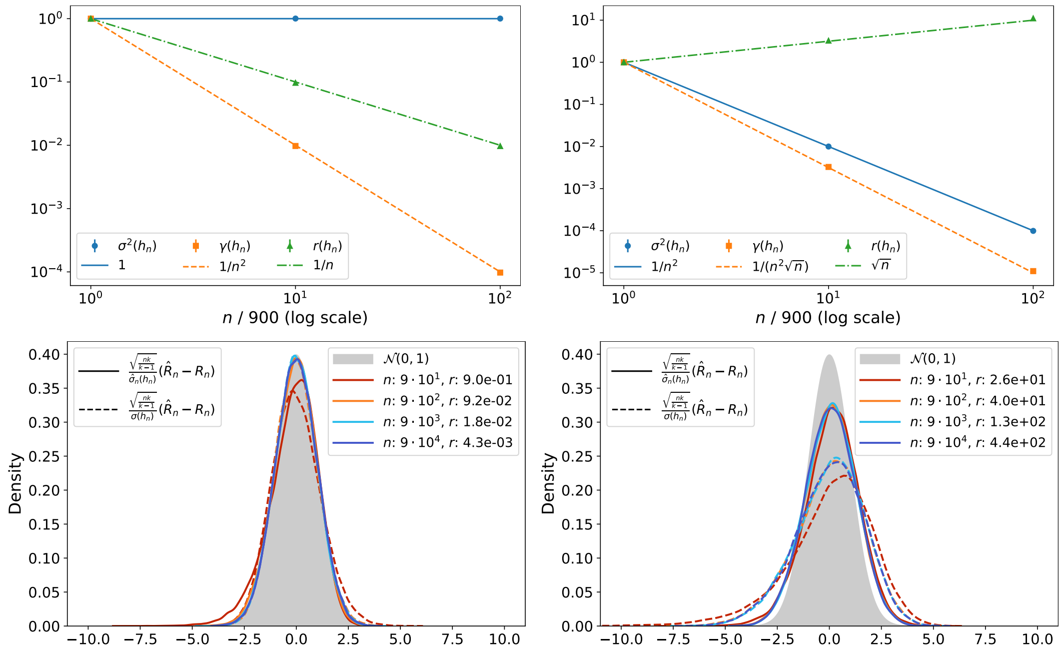
(b) Comparison

Figure 1: ST with $\lambda_n = \sqrt{n}$ when $\beta^\star = (3, 1, -5, 3, 0, 0, 0, 0, 0, 0)$. **Top:** $\sigma^2(h_n)$, $\gamma(h_n)$ and $r(h_n)$ all normalized by their values at $n = 900$. **Bottom:** (best viewed in color) KDE plots for $\frac{\sqrt{\frac{nk}{k-1}}}{\hat\sigma_n(h_n)}(\hat R_n - R_n)$ (solid curves) and $\frac{\sqrt{\frac{nk}{k-1}}}{\sigma(h_n)}(\hat R_n - R_n)$ (dashed curves).

For the ST estimator with $\lambda_n = \sqrt{n}$, the simulation results are presented in Figure 1. For the single algorithm assessment of ST, the rates of $\sigma^2(h_n^{\text{sing}})$, $\gamma(h_n^{\text{sing}})$ and $r(h_n^{\text{sing}})$ are constant order, $1/n^2$ order and $1/n$ order, respectively, as stated in Theorem 1, and for the algorithm comparison of ST, when $\delta_n = 1$, we have the expected $1/n^2$ rate for $\sigma^2(h_n^{\text{diff}})$ and we actually observe that $\gamma(h_n^{\text{diff}})$ and $r(h_n^{\text{diff}})$ seem to be scaling as $1/(n^2\sqrt{n})$ and $\sqrt{n}$, respectively, even though Theorem 2 only established them being Ω of these rates. As we can see for both choices of the dividing standard deviation in the KDE plots of Figure 1, the asymptotic distribution seems to be Gaussian, but the asymptotic variance does not go to 1 when the relative loss stability condition does not hold, that is to say in the comparison setting. An interesting observation is that despite $\hat\sigma_n^2(h_n^{\text{diff}})$ not being consistent for $\sigma^2(h_n^{\text{diff}})$ when the relative stability condition does not hold, it is actually overestimating and thus reduces the mismatch with the true variance of $\sqrt{\frac{nk}{k-1}}(\hat R_n - R_n)$, even if it is still significantly below.

To see if the theoretical results carried over to the Lasso, a close cousin of ST, we ran simulations for the Lasso with λ_n selected via an inner CV (see Appendix H) for each of the k iterations of the CV run, still with constant order $\delta_n = 1$ for the comparison. As mentioned in Section 3, we actually observed in simulations that the values selected for λ_n are concentrated around a constant times $\sqrt{n}$. The results for this new setting are displayed in Figure 2 and confirm that the same conclusions hold empirically as for ST.

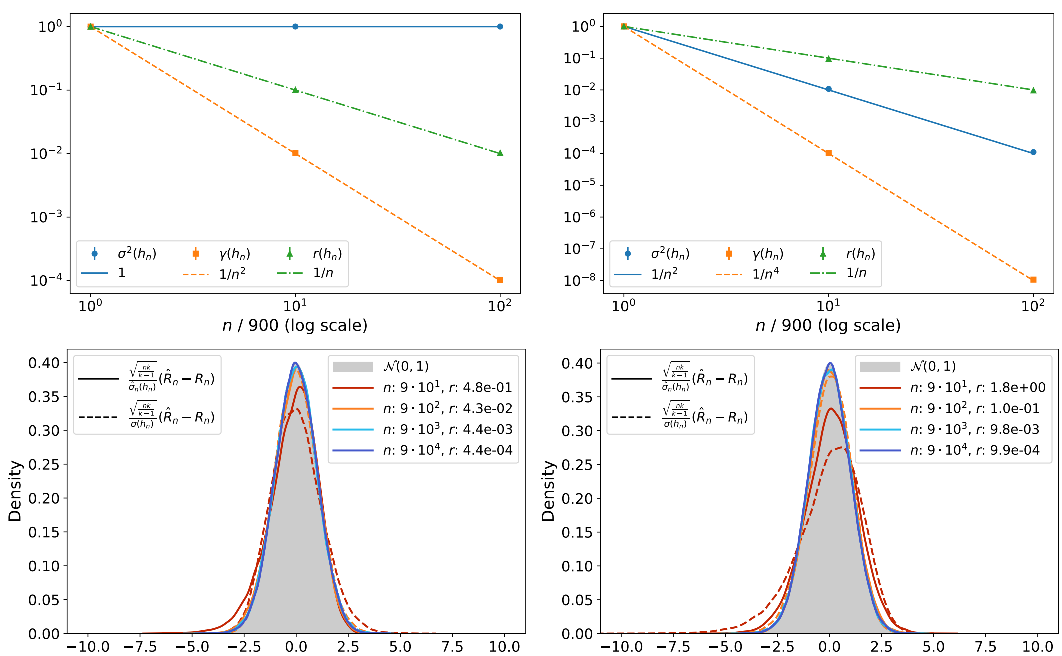
There are definitely instances when an algorithm satisfies the relative loss stability condition both in its individual form and in the comparison setting. One example of this is the ridge estimator and we present the corresponding simulations in Figure 3. Bousquet and Elisseeff [6] proved that ridge regression, with bounded targets, has $O(\frac{1}{n})$ uniform stability. This means it has $O(\frac{1}{n^2})$ loss stability by [17, Lem. 1] and [20, Lem. 2]. In the simulations, we see that for individual ridge, with no boundedness assumption, with isotropic features, loss stability scales as $1/n^2$ and the relative loss stability condition then holds since $\sigma^2(h_n^{\text{sing}})$ is of constant order. And loss stability scales as $1/n^4$ in the comparison setting, which, when compared to the observed $1/n^2$ rate of $\sigma^2(h_n^{\text{diff}})$, means the relative loss stability condition also holds for comparison.



(a) Single algorithm

(b) Comparison

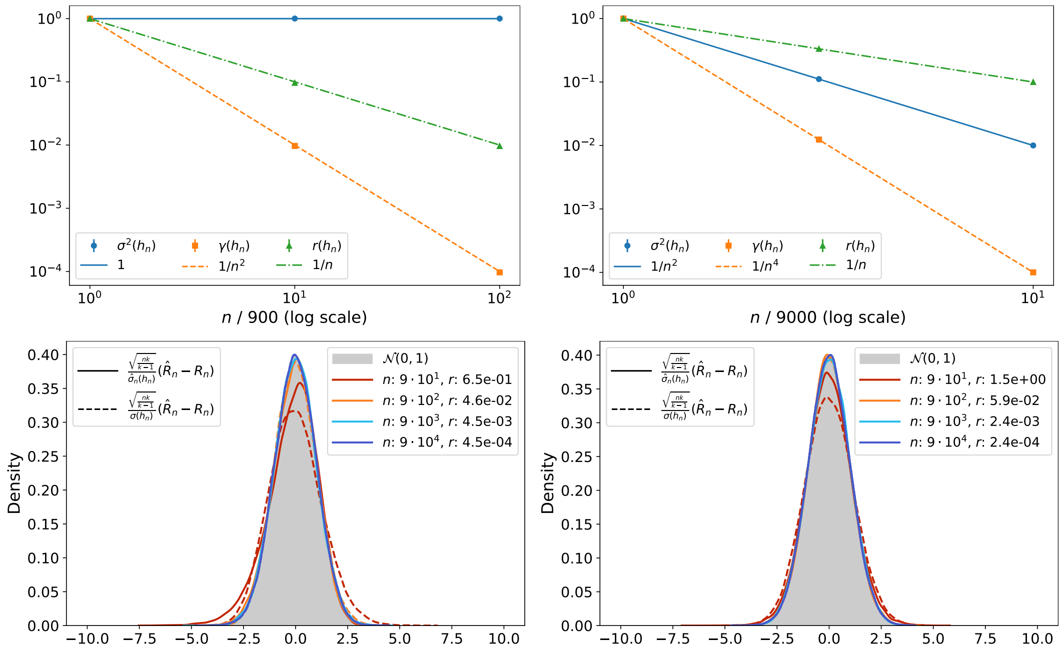
Figure 2: Lasso with cross-validated λ_n when $\beta^\star = (3,1,-5,3,0,0,0,0,0,0)$. **Top:** $\sigma^2(h_n)$, $\gamma(h_n)$ and $r(h_n)$ all normalized by their values at $n = 900$. **Bottom:** (best viewed in color) KDE plots for $\frac{\sqrt{\frac{nk}{k-1}}}{\hat{\sigma}_n(h_n)}(\hat{R}_n - R_n)$ (solid curves) and $\frac{\sqrt{\frac{nk}{k-1}}}{\sigma(h_n)}(\hat{R}_n - R_n)$ (dashed curves).



(a) Single algorithm

(b) Comparison

Figure 3: Ridge regression with $\lambda_n = \sqrt{n}$ when $\beta^\star = (3,1,-5,3,0,0,0,0,0,0)$. **Top:** $\sigma^2(h_n)$, $\gamma(h_n)$ and $r(h_n)$ all normalized by their values at $n = 900$. **Bottom:** (best viewed in color) KDE plots for $\frac{\sqrt{\frac{nk}{k-1}}}{\hat{\sigma}_n(h_n)}(\hat{R}_n - R_n)$ (solid curves) and $\frac{\sqrt{\frac{nk}{k-1}}}{\sigma(h_n)}(\hat{R}_n - R_n)$ (dashed curves).



(a) Single algorithm (b) Comparison

Figure 4: ST with $\lambda_n = \sqrt{n}$ when $\beta^\star = (3, 1, -5, 3, 4, -3, 10, 8, 5, 2)$. **Top:** $\sigma^2(h_n)$, $\gamma(h_n)$ and $r(h_n)$ all normalized by their values at $n = 900$ for single algorithm and at $n = 9000$ for comparison. **Bottom:** (best viewed in color) KDE plots for $\frac{\sqrt{\frac{nk}{k-1}}}{\hat{\sigma}_n(h_n)}(\hat{R}_n - R_n)$ (solid curves) and $\frac{\sqrt{\frac{nk}{k-1}}}{\sigma(h_n)}(\hat{R}_n - R_n)$ (dashed curves).

As a matter of fact, when $\beta^\star$ has no zero coefficients, the ST estimator can also be an example of an algorithm which satisfies the relative loss stability condition in both its individual form and in the comparison setting. The theory sheds light on the importance of the zero coefficients in the true parameter vector. When $\beta^\star$ has no zero coefficients, i.e. $\|\beta^\star\|_0 = p$, ST actually becomes stable for the algorithm comparison setting. The results of the simulations for this setting, with the choice $\beta^\star = (3, 1, -5, 3, 4, -3, 10, 8, 5, 2)$, are presented in Figure 4 and show how the convergence rate of $\gamma(h_n^{\text{diff}})$ changes compared to the $\|\beta^\star\|_0 < p$ setting. It now scales as $1/n^4$, which means that ST satisfies the relative loss stability condition $r(h_n^{\text{diff}}) = o(1)$ in the comparison setting, since $\frac{n^2}{\delta_n^2}\sigma^2(h_n^{\text{diff}})$ still goes to $4\tau^2\|\beta^\star\|_0$ when $\|\beta^\star\|_0 = p$. Nonetheless, we reiterate that even a single zero coefficient in $\beta^\star$ leads to instability for ST, and more generally Lasso, in the comparison setting.

6 Conclusion and Future Work

Cross-validation is a powerful tool, but given its widespread use for comparing and selecting models, scrutiny of its statistical properties is critical for safe model deployment. This work highlights the importance of relative stability for CV and the challenges posed by relative instability for model comparison. In particular, we proved that even simple, absolutely-stable learning algorithms can generate relatively unstable comparisons. In practice, this led to invalid and highly misleading confidence intervals for the test error difference with $\sigma^2(h_n^{\text{diff}})$ being well below the targeted variance of $\sqrt{\frac{nk}{k-1}}(\hat{R}_n - R_n)$. Since CV is often used to conduct formal hypothesis tests for an improvement in test error between two learning algorithms [12, 24, 25, 5, 9, 4], our work shows that such tests can be misleading even for simple, absolutely stable algorithms and that method developers and consumers should first verify the relative stability of a comparison before applying them.

However, this work is not without its limitations. For example, our analysis does not show that all model comparisons are relatively unstable. Indeed the experiments of Section 5 suggest that a second popular model, ridge regression, does generate relatively stable comparisons. Establishing broad, easily verified conditions under which an algorithm comparison is or is not relatively stable is an important direction for future work. Second, while we prove the relative instability of ST comparisons and demonstrate the invalidity of their CV confidence intervals, we leave open the question of whether relative instability always implies CV invalidity.

Finally, our presentation thus far has focused on identifying and proving instability and leaves us without a general solution for confidently comparing models. While we have shown that the CV central limit theorem (4.2) and hence the CV confidence interval construction of Bayle et al. [4] can break down in the presence of relative unstable comparisons, it is possible to produce an asymptotically conservative (and hence valid) confidence intervals for CV whenever the algorithms are individually stable in the following way:

Proposition 1 (Comparison coverage from single algorithm coverage)**.** *Let $\hat{R}_n^{(1)}, R_n^{(1)}$ be the cross-validation error and test error of algorithm $\mathcal{A}_1$, and $\hat{R}_n^{(2)}, R_n^{(2)}$ those of algorithm $\mathcal{A}_2$. To compare $\mathcal{A}_1$ and $\mathcal{A}_2$, if $[L_n^{(1)}, U_n^{(1)}]$ and $[L_n^{(2)}, U_n^{(2)}]$ are asymptotic $(1-\alpha/2)$-coverage confidence intervals for $R_n^{(1)}$ and $R_n^{(2)}$, respectively, then*

$$[L_n^{(1)} - U_n^{(2)}, U_n^{(1)} - L_n^{(2)}]$$

will asymptotically cover $R_n^{(1)} - R_n^{(2)}$ with probability at least $1-\alpha$.

Proof

$$\begin{aligned}
&\liminf_{n\to\infty} \mathbb{P}(R_n^{(1)} - R_n^{(2)} \in [L_n^{(1)} - U_n^{(2)}, U_n^{(1)} - L_n^{(2)}]) \\
&\quad \geq 1 - \limsup_{n\to\infty} \mathbb{P}(R_n^{(1)} \notin [L_n^{(1)}, U_n^{(1)}] \text{ or } R_n^{(2)} \notin [L_n^{(2)}, U_n^{(2)}]) \\
&\quad \geq 1 - \limsup_{n\to\infty} \mathbb{P}(R_n^{(1)} \notin [L_n^{(1)}, U_n^{(1)}]) + \limsup_{n\to\infty} \mathbb{P}(R_n^{(2)} \notin [L_n^{(2)}, U_n^{(2)}]) \\
&\quad \geq 1 - \alpha/2 - \alpha/2 = 1 - \alpha.
\end{aligned}$$

□

This approach would ensure valid asymptotic coverage under individual algorithm stability without requiring any additional stability assumption on the comparison. However, the interval could also be significantly wider than the interval derived from Bayle et al. [4], due to strong positive correlations between $\hat{R}_n^{(1)}$ and $\hat{R}_n^{(2)}$ ignored in the construction of Proposition 1. An open question for the reader is whether one can derive tighter confidence intervals for algorithm comparisons when it is only known that each algorithm is individually stable.

References

[1] Arsov, N., Pavlovski, M., and Kocarev, L. (2019). Stability of decision trees and logistic regression. *arXiv preprint arXiv:1903.00816v1*.

[2] Austern, M. and Zhou, W. (2020). Asymptotics of Cross-Validation. *arXiv preprint arXiv:2001.11111v2*.

[3] Bates, S., Hastie, T., and Tibshirani, R. (2024). Cross-validation: what does it estimate and how well does it do it? *Journal of the American Statistical Association*, 119(546):1434–1445.

[4] Bayle, P., Bayle, A., Janson, L., and Mackey, L. (2020). Cross-validation confidence intervals for test error. *Advances in Neural Information Processing Systems*, 33:16339–16350.

[5] Bouckaert, R. R. and Frank, E. (2004). Evaluating the replicability of significance tests for comparing learning algorithms. In *PAKDD*, pages 3–12. Springer.

[6] Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, 2:499–526.

[7] Burkardt, J. (2023). The truncated normal distribution. https://people.sc.fsu.edu/~jburkardt/presentations/truncated_normal.pdf.

[8] Celisse, A. and Guedj, B. (2016). Stability revisited: new generalisation bounds for the Leave-one-Out. *arXiv preprint arXiv:1608.06412v1*.

[9] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

[10] Devroye, L. and Wagner, T. (1979a). Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Transactions on Information Theory*, 25(2):202–207.

[11] Devroye, L. and Wagner, T. (1979b). Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604.

[12] Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.

[13] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

[14] Elisseeff, A., Evgeniou, T., and Pontil, M. (2005). Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6:55–79.

[15] Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350):320–328.

[16] Hardt, M., Recht, B., and Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33rd International Conference on Machine Learning - Volume 48*, ICML'16, pages 1225–1234. JMLR.org.

[17] Kale, S., Kumar, R., and Vassilvitskii, S. (2011). Cross-validation and mean-square stability. In *Proceedings of the Second Symposium on Innovations in Computer Science (ICS2011)*. Citeseer.

[18] Kearns, M. and Ron, D. (1999). Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11(6):1427–1453.

[19] Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378.

[20] Kumar, R., Lokshtanov, D., Vassilvitskii, S., and Vattani, A. (2013). Near-optimal bounds for cross-validation via loss stability. In *International Conference on Machine Learning*, pages 27–35.

[21] Kutin, S. and Niyogi, P. (2002). Almost-everywhere algorithmic stability and generalization error. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02, pages 275–282, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[22] Lei, J. (2020). Cross-validation with confidence. *Journal of the American Statistical Association*, 115(532):1978–1997.

[23] Li, J. (2023). Asymptotics of k-fold cross validation. *Journal of Artificial Intelligence Research*, 78:491–526.

[24] Lim, T.-S., Loh, W.-Y., and Shih, Y.-S. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Mach. Learn.*, 40(3):203–228.

[25] Nadeau, C. and Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, 52(3):239–281.

[26] Orjebin, E., Liquet, B., and Nazarathy, Y. (2014). A recursive formula for the moments of a truncated univariate normal distribution. https://people.smp.uq.edu.au/YoniNazarathy/teaching_projects/studentWork/EricOrjebin_TruncatedNormalMoments.pdf.

[27] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147.

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

[29] Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202.

[30] Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.

Appendix Contents

A Additional Notation

Let $\xrightarrow{\text{a.s.}}$ denote almost sure convergence. Let $\mathbb{1}[A]$ denote the indicator function of a subset A. We will denote by Φ the cumulative distribution function of the standard Normal and by φ its probability density function. We define the sign function as $\mathrm{sign}(x) = \frac{x}{|x|}\mathbb{1}[x \neq 0]$ and the positive part as $x_+ = \max(x, 0)$. We write $\mathbf{M} \sim W_p^{-1}(\Sigma, n)$ to indicate $\mathbf{M}$ follows the inverse-Wishart distribution with n degrees of freedom and scale matrix $\Sigma \in \mathbb{R}^{p \times p}$.

B Proof of Theorem 1: Relative stability of individual soft-thresholding

Theorem 1 follows immediately from the following two propositions, proved in Appendices C and D, respectively.

Proposition B.1 (Convergence of $\sigma^2(h_n^{\mathrm{sing}})$ for ST(λ_n))**.** *Assume the linear model (2.1), with feature and noise distributions as given immediately following its equation. If $\lambda_n = o(n)$, then $\sigma^2(h_n^{\mathrm{sing}}) \to 2\tau^4$.*

Proposition B.2 (Convergence rate of $\gamma(h_n^{\mathrm{sing}})$ for ST(λ_n))**.** *Assume the linear model (2.1), with feature and noise distributions as given immediately following its equation. If $\lambda_n = o(n)$, then $\gamma(h_n^{\mathrm{sing}}) \sim \frac{C}{n^2}$ for a constant $C > 0$ whose explicit expression is given in (D.1).*

C Proof of Proposition B.1: Convergence of $\sigma^2(h_n^{\mathrm{sing}})$ for ST(λ_n)

We start by introducing a lemma which provides equations that will prove useful in the single algorithm setting.

Lemma C.1 (Useful equations for single linear predictor)**.** *When defining $h_n(Z_0, \mathbf{Z}) = (Y_0 - X_0^\top \hat{\beta})^2$, we have:*

$$\begin{aligned}
h_n(Z_0, \mathbf{Z}) &= Y_0^2 - 2Y_0 X_0^\top \hat{\beta} + \mathrm{tr}(X_0 X_0^\top \hat{\beta}\hat{\beta}^\top) \\
\mathbb{E}[h_n(Z_0, \mathbf{Z}) \mid Z_0] &= Y_0^2 - 2Y_0 X_0^\top \mathbb{E}[\hat{\beta}] + \mathrm{tr}(X_0 X_0^\top \mathbb{E}[\hat{\beta}\hat{\beta}^\top]) \\
\mathbb{E}[h_n(Z_0, \mathbf{Z}) \mid \mathbf{Z}] &= \mathbb{E}[Y_0^2] - 2\beta^{\star\top}\mathbb{E}[X_0 X_0^\top]\hat{\beta} + \mathrm{tr}(\mathbb{E}[X_0 X_0^\top]\hat{\beta}\hat{\beta}^\top) \\
\mathbb{E}[h_n(Z_0, \mathbf{Z})] &= \mathbb{E}[Y_0^2] - 2\beta^{\star\top}\mathbb{E}[X_0 X_0^\top]\mathbb{E}[\hat{\beta}] + \mathrm{tr}(\mathbb{E}[X_0 X_0^\top]\mathbb{E}[\hat{\beta}\hat{\beta}^\top]) \\
\sigma^2(h_n) &= \mathbb{E}[(Y_0^2 - \mathbb{E}[Y_0^2] - 2(Y_0 X_0^\top - \beta^{\star\top}\mathbb{E}[X_0 X_0^\top])\mathbb{E}[\hat{\beta}] \\
&\quad + \mathrm{tr}((X_0 X_0^\top - \mathbb{E}[X_0 X_0^\top])\mathbb{E}[\hat{\beta}\hat{\beta}^\top]))^2] \\
\gamma(h_n) &= \mathbb{E}[(2(Y_0 X_0^\top - \beta^{\star\top}\mathbb{E}[X_0 X_0^\top])(\hat{\beta}' - \hat{\beta}) \\
&\quad + \mathrm{tr}((X_0 X_0^\top - \mathbb{E}[X_0 X_0^\top])(\hat{\beta}\hat{\beta}^\top - \hat{\beta}'\hat{\beta}'^\top)))^2]
\end{aligned}$$

where $\hat{\beta}'$ is the linear predictor learned on a training set $\mathbf{Z}'$ that is the same as $\mathbf{Z}$ except for the first point Z_1 being replaced by an i.i.d copy Z_1'.

Proof

$$\begin{aligned}
h_n(Z_0, \mathbf{Z}) &= (Y_0 - X_0^\top \hat{\beta})^2 \\
&= Y_0^2 - 2Y_0 X_0^\top \hat{\beta} + (X_0^\top \hat{\beta})^2 \\
&= Y_0^2 - 2Y_0 X_0^\top \hat{\beta} + X_0^\top \hat{\beta}\hat{\beta}^\top X_0 \\
&= Y_0^2 - 2Y_0 X_0^\top \hat{\beta} + \mathrm{tr}(X_0 X_0^\top \hat{\beta}\hat{\beta}^\top)
\end{aligned}$$

Note that $\mathbb{E}[Y_0 X_0^\top] = \mathbb{E}[\mathbb{E}[Y_0 \mid X_0] X_0^\top] = \mathbb{E}[X_0^\top \beta^\star X_0^\top] = \beta^{\star\top}\mathbb{E}[X_0 X_0^\top]$.

Since $\hat{\beta}$ is only a function of $\mathbf{Z}$, the independence of Z_0 and $\mathbf{Z}$ yields the next three equations.

The fifth equation comes from noticing

$$\sigma^2(h_n) = \mathrm{Var}(\mathbb{E}[h_n(Z_0, \mathbf{Z}) \mid Z_0]) = \mathbb{E}[(\mathbb{E}[h_n(Z_0, \mathbf{Z}) \mid Z_0] - \mathbb{E}[h_n(Z_0, \mathbf{Z})])^2].$$

And the last one comes from the definition of $\gamma(h_n)$ as

$$\gamma(h_n) = \mathbb{E}[(h(Z_0, \mathbf{Z}) - h(Z_0, \mathbf{Z}') - (\mathbb{E}[h(Z_0, \mathbf{Z}) \mid \mathbf{Z}] - \mathbb{E}[h(Z_0, \mathbf{Z}') \mid \mathbf{Z}']))^2].$$

□

In addition to giving a first glimpse into the differences between the single algorithm and comparison settings, the following lemma plays an important role in our proof via its result for a single linear predictor.

Lemma C.2 (Convergence of $\sigma^2(h_n^{\mathrm{sing}})$ and $\sigma^2(h_n^{\mathrm{diff}})$)**.** *Assume the features are drawn i.i.d. from a distribution with mean 0 and identity covariance matrix. For a single linear predictor, if we have consistency in the form of $\mathbb{E}[\hat{\beta}_n] \to \beta^\star$ and $\mathbb{E}[\hat{\beta}_n\hat{\beta}_n^\top] \to \beta^\star\beta^{\star\top}$, then $\sigma^2(h_n^{\mathrm{sing}}) \to 2\tau^4$, where τ^2 is the variance of the noise term in the linear model (2.1). For two linear predictors, if we have $\mathbb{E}[\hat{\beta}_n^{(1)} - \hat{\beta}_n^{(2)}] \to 0$ and $\mathbb{E}[\hat{\beta}_n^{(1)}\hat{\beta}_n^{(1)\top} - \hat{\beta}_n^{(2)}\hat{\beta}_n^{(2)\top}] \to 0$, then $\sigma^2(h_n^{\mathrm{diff}}) \to 0$.*

Proof Let $Y_0 = X_0^\top \beta^\star + \varepsilon_0$ be the response variable with $\mathrm{Var}(\varepsilon_0) = \tau^2$. Using the information on the distribution of X_0 and the independence of X_0 and ε_0, note that

$$\mathbb{E}[Y_0^2] = \mathrm{Var}(Y_0) + \mathbb{E}[Y_0]^2 = \mathrm{Var}(X_0^\top \beta^\star + \varepsilon_0) + 0 = \beta^{\star\top}\mathrm{Var}(X_0)\beta^\star + \mathrm{Var}(\varepsilon_0) = \beta^{\star\top}\beta^\star + \tau^2.$$

For the single linear predictor, starting from the expression of $\sigma^2(h_n)$ in Lemma C.1, since $\mathbb{E}[\hat{\beta}_n]$ and $\mathbb{E}[\hat{\beta}_n\hat{\beta}_n^\top]$ are non-random, we can expand the square, use linearity of expectation, take the limits and factorize back to obtain the convergence

$$\begin{aligned}
\sigma^2(h_n^{\mathrm{sing}}) &= \mathbb{E}[(Y_0^2 - \mathbb{E}[Y_0^2] - 2(Y_0 X_0^\top - \beta^{\star\top}\mathbb{E}[X_0 X_0^\top])\mathbb{E}[\hat{\beta}_n] + \mathrm{tr}((X_0 X_0^\top - \mathbb{E}[X_0 X_0^\top])\mathbb{E}[\hat{\beta}_n\hat{\beta}_n^\top]))^2] \\
&\to \mathbb{E}[(Y_0^2 - \mathbb{E}[Y_0^2] - 2(Y_0 X_0^\top - \beta^{\star\top}\mathbb{E}[X_0 X_0^\top])\beta^\star + \mathrm{tr}((X_0 X_0^\top - \mathbb{E}[X_0 X_0^\top])\beta^\star\beta^{\star\top}))^2] \\
&= \mathbb{E}[(Y_0^2 - \beta^{\star\top}\beta^\star - \tau^2 - 2Y_0 X_0^\top \beta^\star + 2\beta^{\star\top}\beta^\star + \mathrm{tr}((X_0 X_0^\top \beta^\star\beta^{\star\top} - \beta^\star\beta^{\star\top}))^2] \\
&= \mathbb{E}[((X_0^\top \beta^\star + \varepsilon_0)^2 - \beta^{\star\top}\beta^\star - \tau^2 - 2(X_0^\top \beta^\star + \varepsilon_0)X_0^\top \beta^\star + 2\beta^{\star\top}\beta^\star + (X_0^\top \beta^\star)^2 - \beta^{\star\top}\beta^\star)^2] \\
&= \mathbb{E}[(\varepsilon_0^2 - \tau^2)^2] = \mathrm{Var}(\varepsilon_0^2) = \mathbb{E}[\varepsilon_0^4] - \mathbb{E}[\varepsilon_0^2]^2 = 3\tau^4 - \tau^4 = 2\tau^4.
\end{aligned}$$

Similarly, we derive the second result with two linear predictors by starting from the expression of $\sigma^2(h_n)$ in Lemma F.1. $\square$

We will show that $\mathbb{E}[\hat{\beta}_{\lambda_n}] \to \beta^\star$ and $\mathbb{E}[\hat{\beta}_{\lambda_n}\hat{\beta}_{\lambda_n}^\top] \to \beta^\star\beta^{\star\top}$ in order to obtain the convergence of $\sigma^2(h_n^{\text{sing}})$ as an application of Lemma C.2.

We have for $i = 1, \dots, p$,

$$\begin{aligned}\hat{\beta}_{\lambda_n,i} &= \text{sign}(\hat{\beta}_{\text{OLS},i})(|\hat{\beta}_{\text{OLS},i}| - \tfrac{\lambda_n}{n})_+ \\ &= \text{sign}(\hat{\beta}_{\text{OLS},i})\begin{cases}|\hat{\beta}_{\text{OLS},i}| - \frac{\lambda_n}{n} & \text{if } |\hat{\beta}_{\text{OLS},i}| \geq \frac{\lambda_n}{n} \\ 0 & \text{if } |\hat{\beta}_{\text{OLS},i}| < \frac{\lambda_n}{n}\end{cases}.\end{aligned}$$

A classic result for the OLS estimator is $\hat{\beta}_{\text{OLS}} \mid \mathbf{X} \sim \mathcal{N}(\beta^\star, \tau^2(\mathbf{X}^\top\mathbf{X})^{-1})$. We can write $\hat{\beta}_{\text{OLS},i} = \beta_i^\star + \tilde{\tau}_n Z$ where $\tilde{\tau}_n = \frac{\tau}{\sqrt{n}}\sqrt{(\frac{\mathbf{X}^\top\mathbf{X}}{n})^{-1}_{i,i}}$ and $Z \mid \mathbf{X} \sim \mathcal{N}(0,1)$. Note that we could have i as a subscript of $\tilde{\tau}_n$ and Z, but we will only consider one i at a time in our computations and we can thus omit this subscript for both of them for the sake of notational simplicity, and we will also omit it for some additional notation we define in the rest of the proof.

We now show that $\mathbb{E}[\hat{\beta}_{\lambda_n,i}] \to \beta_i^\star$.

Using the law of total expectation,

$$\begin{aligned}\mathbb{E}[\hat{\beta}_{\lambda_n,i} \mid \mathbf{X}] = \mathbb{E}[\hat{\beta}_{\text{OLS},i} - \tfrac{\lambda_n}{n} \mid \hat{\beta}_{\text{OLS},i} \geq \tfrac{\lambda_n}{n}, \mathbf{X}]\, \mathbb{P}(\hat{\beta}_{\text{OLS},i} \geq \tfrac{\lambda_n}{n} \mid \mathbf{X}) \\ + \mathbb{E}[\hat{\beta}_{\text{OLS},i} + \tfrac{\lambda_n}{n} \mid \hat{\beta}_{\text{OLS},i} \leq -\tfrac{\lambda_n}{n}, \mathbf{X}]\, \mathbb{P}(\hat{\beta}_{\text{OLS},i} \leq -\tfrac{\lambda_n}{n} \mid \mathbf{X}).\end{aligned}$$

Define $\alpha_n^{(1)} = \frac{1}{\tilde{\tau}_n}(\frac{\lambda_n}{n} - \beta_i^\star)$ and $\alpha_n^{(2)} = \frac{1}{\tilde{\tau}_n}(\frac{\lambda_n}{n} + \beta_i^\star)$.

The first probability is equal to

$$\mathbb{P}(Z \geq \alpha_n^{(1)} \mid \mathbf{X}) = 1 - \Phi(\alpha_n^{(1)})$$

and the second probability to

$$\mathbb{P}(Z \leq -\alpha_n^{(2)} \mid \mathbf{X}) = \Phi(-\alpha_n^{(2)}) = 1 - \Phi(\alpha_n^{(2)}).$$

Using the first moment of the truncated normal [7], we have

$$\begin{aligned}\mathbb{E}[\hat{\beta}_{\text{OLS},i} - \tfrac{\lambda_n}{n} \mid \hat{\beta}_{\text{OLS},i} \geq \tfrac{\lambda_n}{n}, \mathbf{X}] &= \beta_i^\star - \tfrac{\lambda_n}{n} + \tilde{\tau}_n\, \mathbb{E}[Z \mid Z \geq \alpha_n^{(1)}, \mathbf{X}] \\ &= \beta_i^\star - \tfrac{\lambda_n}{n} + \tilde{\tau}_n \tfrac{\varphi(\alpha_n^{(1)})}{1-\Phi(\alpha_n^{(1)})}\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[\hat{\beta}_{\text{OLS},i} + \tfrac{\lambda_n}{n} \mid \hat{\beta}_{\text{OLS},i} \leq -\tfrac{\lambda_n}{n}, \mathbf{X}] &= \beta_i^\star + \tfrac{\lambda_n}{n} + \tilde{\tau}_n \mathbb{E}[Z \mid Z \leq -\alpha_n^{(2)}] \\ &= \beta_i^\star + \tfrac{\lambda_n}{n} - \tilde{\tau}_n \tfrac{\varphi(-\alpha_n^{(2)})}{\Phi(-\alpha_n^{(2)})}.\end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{E}[\hat{\beta}_{\lambda_n,i} \mid \mathbf{X}] &= \mathbb{E}[\hat{\beta}_{\text{OLS},i} - \tfrac{\lambda_n}{n} \mid \hat{\beta}_{\text{OLS},i} \geq \tfrac{\lambda_n}{n}, \mathbf{X}]\, \mathbb{P}(\hat{\beta}_{\text{OLS},i} \geq \tfrac{\lambda_n}{n} \mid \mathbf{X}) \\ &\quad + \mathbb{E}[\hat{\beta}_{\text{OLS},i} + \tfrac{\lambda_n}{n} \mid \hat{\beta}_{\text{OLS},i} \leq -\tfrac{\lambda_n}{n}, \mathbf{X}]\, \mathbb{P}(\hat{\beta}_{\text{OLS},i} \leq -\tfrac{\lambda_n}{n} \mid \mathbf{X}) \\ &= (\beta_i^\star - \tfrac{\lambda_n}{n})(1 - \Phi(\alpha_n^{(1)})) + \tilde{\tau}_n\varphi(\alpha_n^{(1)}) + (\beta_i^\star + \tfrac{\lambda_n}{n})\Phi(-\alpha_n^{(2)}) - \tilde{\tau}_n\varphi(-\alpha_n^{(2)}) \\ &= (\beta_i^\star - \tfrac{\lambda_n}{n})(1 - \Phi(\alpha_n^{(1)})) + (\beta_i^\star + \tfrac{\lambda_n}{n})\Phi(-\alpha_n^{(2)}) + \tilde{\tau}_n(\varphi(\alpha_n^{(1)}) - \varphi(-\alpha_n^{(2)})).\end{aligned}$$

Note that $\varphi'(x) = -x\varphi(x)$. A straightforward study of the behavior of the function $x \mapsto x\varphi(x)$ shows it is bounded. We denote the maximum of its absolute value by M.

Using the mean value inequality for φ, we have

$$\begin{aligned}|\tilde{\tau}_n(\varphi(\alpha_n^{(1)}) - \varphi(-\alpha_n^{(2)}))| &\leq \tilde{\tau}_n|\alpha_n^{(1)} - (-\alpha_n^{(2)})| \cdot \max_{[-\alpha_n^{(2)},\alpha_n^{(1)}]} |\varphi'| \\ &\leq M\tilde{\tau}_n|\alpha_n^{(1)} - (-\alpha_n^{(2)})| \\ &= M\tilde{\tau}_n \tfrac{1}{\tilde{\tau}_n}(\tfrac{\lambda_n}{n} - \beta_i^\star + \tfrac{\lambda_n}{n} + \beta_i^\star) \\ &= 2M\tfrac{\lambda_n}{n}.\end{aligned}$$

Therefore, since $\lambda_n = o(n)$, $\tilde{\tau}_n(\varphi(\alpha_n^{(1)}) - \varphi(-\alpha_n^{(2)}))$ goes to 0 in L^1.

We first consider $\beta_i^\star > 0$.

Since $\frac{\mathbf{X}^\top\mathbf{X}}{n} \xrightarrow{\text{a.s.}} \mathbb{E}[X_0X_0^\top]$ (strong law of large numbers), and $\lambda_n = o(n)$, we have $\tilde{\tau}_n \xrightarrow{\text{a.s.}} 0^+$, and using the continuous mapping theorem, $\alpha_n^{(1)} \xrightarrow{\text{a.s.}} -\infty$ and $\alpha_n^{(2)} \xrightarrow{\text{a.s.}} +\infty$. Φ is continuous bounded so we get L^1 convergence of $\Phi(\alpha_n^{(1)})$ and $\Phi(-\alpha_n^{(2)})$ to 0. By putting everything together, we obtain

$$\mathbb{E}[\hat{\beta}_{\lambda_n,i}] = \mathbb{E}[\mathbb{E}[\hat{\beta}_{\lambda_n,i} \mid \mathbf{X}]] \to \beta_i^\star.$$

When $\beta_i^\star < 0$, we show in a similar manner that $\mathbb{E}[\hat{\beta}_{\lambda_n,i}] \to \beta_i^\star$.

If $\beta_i^\star = 0$, $\alpha_n^{(1)} = \alpha_n^{(2)}$ so $1 - \Phi(\alpha_n^{(1)}) = \Phi(-\alpha_n^{(2)})$ and $\varphi(\alpha_n^{(1)}) = \varphi(-\alpha_n^{(2)})$ which leads to $\mathbb{E}[\hat{\beta}_{\lambda_n,i} \mid \mathbf{X}] = 0$ and thus $\mathbb{E}[\hat{\beta}_{\lambda_n,i}] = 0$.

Thus, we have convergence component-wise and can conclude $\mathbb{E}[\hat{\beta}_{\lambda_n}] \to \beta^\star$.

We now show that $\mathbb{E}[\hat{\beta}_{\lambda_n,i}\hat{\beta}_{\lambda_n,j}] \to \beta_i^\star\beta_j^\star$.

Note that

$$\mathbb{E}[\hat{\beta}_{\lambda_n,i}\hat{\beta}_{\lambda_n,j} - \beta_i^\star\beta_j^\star] = \mathbb{E}[(\hat{\beta}_{\lambda_n,i} - \beta_i^\star)\hat{\beta}_{\lambda_n,j}] + \beta_i^\star\mathbb{E}[\hat{\beta}_{\lambda_n,j} - \beta_j^\star]$$

where, using Cauchy–Schwarz and the fact that $(a+b)^2 \leq 2(a^2+b^2)$,

$$|\mathbb{E}[(\hat{\beta}_{\lambda_n,i} - \beta_i^\star)\hat{\beta}_{\lambda_n,j}]| \leq \sqrt{\mathbb{E}[(\hat{\beta}_{\lambda_n,i} - \beta_i^\star)^2]\mathbb{E}[\hat{\beta}_{\lambda_n,j}^2]} \leq \sqrt{\mathbb{E}[(\hat{\beta}_{\lambda_n,i} - \beta_i^\star)^2]2(\mathbb{E}[(\hat{\beta}_{\lambda_n,j} - \beta_j^\star)^2] + \beta_j^{\star 2})}.$$

Therefore, proving $\mathbb{E}[\hat{\beta}_{\lambda_n,i}\hat{\beta}_{\lambda_n,j}] \to \beta_i^\star\beta_j^\star$ for all i, j comes down to proving $\mathbb{E}[(\hat{\beta}_{\lambda_n,i} - \beta_i^\star)^2] \to 0$ for all i given that we have already shown $\mathbb{E}[\hat{\beta}_{\lambda_n,i}] \to \beta_i^\star$ for all i.

As a reminder, we have

$$\hat{\beta}_{\lambda_n,i} = \text{sign}(\hat{\beta}_{\text{OLS},i})\begin{cases}|\hat{\beta}_{\text{OLS},i}| - \frac{\lambda_n}{n} & \text{if } |\hat{\beta}_{\text{OLS},i}| \geq \frac{\lambda_n}{n} \\ 0 & \text{if } |\hat{\beta}_{\text{OLS},i}| < \frac{\lambda_n}{n}\end{cases}$$

thus

$$\begin{aligned}&\mathbb{E}[(\hat{\beta}_{\lambda_n,i} - \beta_i^\star)^2 \mid \mathbf{X}] \\ &= \mathbb{E}[(\hat{\beta}_{\text{OLS},i} - \beta_i^\star - \tfrac{\lambda_n}{n})^2 \mid \hat{\beta}_{\text{OLS},i} \geq \tfrac{\lambda_n}{n}, \mathbf{X}]\,\mathbb{P}(\hat{\beta}_{\text{OLS},i} \geq \tfrac{\lambda_n}{n} \mid \mathbf{X}) \\ &\quad + \mathbb{E}[(\hat{\beta}_{\text{OLS},i} - \beta_i^\star + \tfrac{\lambda_n}{n})^2 \mid \hat{\beta}_{\text{OLS},i} \leq -\tfrac{\lambda_n}{n}, \mathbf{X}]\,\mathbb{P}(\hat{\beta}_{\text{OLS},i} \leq -\tfrac{\lambda_n}{n} \mid \mathbf{X}).\end{aligned}$$

Using the second moment of the truncated normal [26], we have

$$\begin{aligned}&\mathbb{E}[(\hat{\beta}_{\text{OLS},i} - \beta_i^\star - \tfrac{\lambda_n}{n})^2 \mid \hat{\beta}_{\text{OLS},i} \geq \tfrac{\lambda_n}{n}, \mathbf{X}] \\ &= \mathbb{E}[(\tilde{\tau}_n Z - \tfrac{\lambda_n}{n})^2 \mid Z \geq \alpha_n^{(1)}, \mathbf{X}] \\ &= \tilde{\tau}_n^2\,\mathbb{E}[Z^2 \mid Z \geq \alpha_n^{(1)}, \mathbf{X}] - 2\tilde{\tau}_n\tfrac{\lambda_n}{n}\mathbb{E}[Z \mid Z \geq \alpha_n^{(1)}, \mathbf{X}] + \tfrac{\lambda_n^2}{n^2} \\ &= \tilde{\tau}_n^2(1 + \tfrac{\alpha_n^{(1)}\varphi(\alpha_n^{(1)})}{1-\Phi(\alpha_n^{(1)})}) - 2\tilde{\tau}_n\tfrac{\lambda_n}{n}\tfrac{\varphi(\alpha_n^{(1)})}{1-\Phi(\alpha_n^{(1)})} + \tfrac{\lambda_n^2}{n^2}\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}[(\hat\beta_{\mathrm{OLS},i} - \beta_i^\star + \tfrac{\lambda_n}{n})^2 \mid \hat\beta_{\mathrm{OLS},i} \le -\tfrac{\lambda_n}{n}, \mathbf{X}] \\
&= \tilde\tau_n^2\, \mathbb{E}[Z^2 \mid Z \le -\alpha_n^{(2)}, \mathbf{X}] + 2\tilde\tau_n \tfrac{\lambda_n}{n} \mathbb{E}[Z \mid Z \le -\alpha_n^{(2)}, \mathbf{X}] + \tfrac{\lambda_n^2}{n^2} \\
&= \tilde\tau_n^2 (1 + \tfrac{\alpha_n^{(2)} \varphi(-\alpha_n^{(2)})}{\Phi(-\alpha_n^{(2)})}) - 2\tilde\tau_n \tfrac{\lambda_n}{n} \tfrac{\varphi(-\alpha_n^{(2)})}{\Phi(-\alpha_n^{(2)})} + \tfrac{\lambda_n^2}{n^2}.
\end{aligned}$$

Thus

$$\begin{aligned}
&\mathbb{E}[(\hat\beta_{\lambda_n,i} - \beta_i^\star)^2 \mid \mathbf{X}] \\
&= \mathbb{E}[(\hat\beta_{\mathrm{OLS},i} - \beta_i^\star - \tfrac{\lambda_n}{n})^2 \mid \hat\beta_{\mathrm{OLS},i} \ge \tfrac{\lambda_n}{n}, \mathbf{X}]\, \mathbb{P}(\hat\beta_{\mathrm{OLS},i} \ge \tfrac{\lambda_n}{n} \mid \mathbf{X}) \\
&\quad + \mathbb{E}[(\hat\beta_{\mathrm{OLS},i} - \beta_i^\star + \tfrac{\lambda_n}{n})^2 \mid \hat\beta_{\mathrm{OLS},i} \le -\tfrac{\lambda_n}{n}, \mathbf{X}]\, \mathbb{P}(\hat\beta_{\mathrm{OLS},i} \le -\tfrac{\lambda_n}{n} \mid \mathbf{X}) \\
&= \tilde\tau_n^2 (1 - \Phi(\alpha_n^{(1)}) + \alpha_n^{(1)} \varphi(\alpha_n^{(1)})) - 2\tilde\tau_n \tfrac{\lambda_n}{n} \varphi(\alpha_n^{(1)}) + \tfrac{\lambda_n^2}{n^2}(1 - \Phi(\alpha_n^{(1)})) \\
&\quad + \tilde\tau_n^2 (\Phi(-\alpha_n^{(2)}) + \alpha_n^{(2)} \varphi(-\alpha_n^{(2)})) - 2\tilde\tau_n \tfrac{\lambda_n}{n} \varphi(-\alpha_n^{(2)}) + \tfrac{\lambda_n^2}{n^2} \Phi(-\alpha_n^{(2)}).
\end{aligned}$$

For $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I})$, we know $(\mathbf{X}^\top \mathbf{X})^{-1} \sim W_p^{-1}(\mathbf{I}, n)$, therefore $\mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1}] = \frac{\mathbf{I}}{n-p-1}$ and $\mathbb{E}[(\frac{\mathbf{X}^\top \mathbf{X}}{n})^{-1}_{i,i}] = \frac{n}{n-p-1} = o(n)$.

Thus, using Jensen's inequality, $\mathbb{E}[\sqrt{(\frac{\mathbf{X}^\top \mathbf{X}}{n})^{-1}_{i,i}}] \le \sqrt{\mathbb{E}[(\frac{\mathbf{X}^\top \mathbf{X}}{n})^{-1}_{i,i}]} = \sqrt{\frac{n}{n-p-1}} = o(\sqrt{n})$.

As a reminder, $\tilde\tau_n = \frac{\tau}{\sqrt{n}} \sqrt{(\frac{\mathbf{X}^\top \mathbf{X}}{n})^{-1}_{i,i}}$. We then have L^1 convergence of both $\tilde\tau_n$ and $\tilde\tau_n^2$ to 0. As previously mentioned, the function $x \mapsto x\varphi(x)$ is bounded. Since Φ and φ are also bounded, and $\lambda_n = o(n)$, then

$$\mathbb{E}[(\hat\beta_{\lambda_n,i} - \beta_i^\star)^2] = \mathbb{E}[\mathbb{E}[(\hat\beta_{\lambda_n,i} - \beta_i^\star)^2 \mid \mathbf{X}]] \to 0.$$

Therefore, we get

$$\mathbb{E}[\hat\beta_{\lambda_n} \hat\beta_{\lambda_n}^\top] \to \beta^\star \beta^{\star\top}.$$

We can then conclude that $\sigma^2(h_n^{\text{sing}}) \to 2\tau^4$ by Lemma C.2.

D Proof of Proposition B.2: Convergence rate of $\gamma(h_n^{\text{sing}})$ for ST(λ_n)

As a reminder, to study the loss stability, we consider $Z_1' = (X_1', Y_1')$ an i.i.d. copy of $Z_1 = (X_1, Y_1)$ used as replacement for the first point of the training set.

Define the vector $V \triangleq (Y_1' - X_1'^\top \beta^\star) X_1' - (Y_1 - X_1^\top \beta^\star) X_1$ and the symmetric matrix $M \triangleq -(V\beta^{\star\top} + \beta^\star V^\top)$.

Starting from the expression of $\gamma(h_n)$ in Lemma C.1 and using the fact that $X_0 \sim \mathcal{N}(0, \mathbf{I})$, we have

$$\gamma(h_n^{\text{sing}}) = \mathbb{E}[(2(Y_0 X_0^\top - \beta^{\star\top})(\hat\beta'_{\lambda_n} - \hat\beta_{\lambda_n}) + \operatorname{tr}((X_0 X_0^\top - \mathbf{I})(\hat\beta_{\lambda_n} \hat\beta_{\lambda_n}^\top - \hat\beta'_{\lambda_n} \hat\beta'^\top_{\lambda_n})))^2].$$

We will show that

$$\gamma(h_n^{\text{sing}}) \sim \tfrac{1}{n^2} \mathbb{E}[(2(Y_0 X_0^\top - \beta^{\star\top})V + \operatorname{tr}((X_0 X_0^\top - \mathbf{I})M))^2].$$

by proving that the difference

$$\begin{aligned}
W_n \triangleq\ & (2(Y_0 X_0^\top - \beta^{\star\top})(\hat\beta'_{\lambda_n} - \hat\beta_{\lambda_n}) + \operatorname{tr}((X_0 X_0^\top - \mathbf{I})(\hat\beta_{\lambda_n} \hat\beta_{\lambda_n}^\top - \hat\beta'_{\lambda_n} \hat\beta'^\top_{\lambda_n})))^2 \\
& - (2(Y_0 X_0^\top - \beta^{\star\top})\tfrac{V}{n} + \operatorname{tr}((X_0 X_0^\top - \mathbf{I})\tfrac{M}{n}))^2
\end{aligned}$$

goes to 0 in L^1.

Since $a^2 - b^2 = (a-b)(a+b)$, we have

$$W_n = (D_{n,1} + D_{n,2})(S_{n,1} + S_{n,2}).$$

where

$$
\begin{aligned}
D_{n,1} &\triangleq 2(Y_0 X_0^\top - \beta^{\star\top})(\hat\beta'_{\lambda_n} - \hat\beta_{\lambda_n} - \tfrac{V}{n}),\\
D_{n,2} &\triangleq \operatorname{tr}((X_0 X_0^\top - \mathbf{I})(\hat\beta_{\lambda_n}\hat\beta_{\lambda_n}^\top - \hat\beta'_{\lambda_n}\hat\beta'^\top_{\lambda_n} - \tfrac{M}{n})),\\
S_{n,1} &\triangleq 2(Y_0 X_0^\top - \beta^{\star\top})(\hat\beta'_{\lambda_n} - \hat\beta_{\lambda_n} + \tfrac{V}{n}),\\
S_{n,2} &\triangleq \operatorname{tr}((X_0 X_0^\top - \mathbf{I})(\hat\beta_{\lambda_n}\hat\beta_{\lambda_n}^\top - \hat\beta'_{\lambda_n}\hat\beta'^\top_{\lambda_n} + \tfrac{M}{n})).
\end{aligned}
$$

Using Cauchy–Schwarz and the fact that $(a+b)^2 \le 2(a^2+b^2)$,

$$
\begin{aligned}
\mathbb{E}[|W_n|] &\le \sqrt{\mathbb{E}[(D_{n,1}+D_{n,2})^2]\,\mathbb{E}[(S_{n,1}+S_{n,2})^2]}\\
&\le 2\sqrt{\mathbb{E}[D_{n,1}^2 + D_{n,2}^2]\,\mathbb{E}[S_{n,1}^2+S_{n,2}^2]}.
\end{aligned}
$$

To obtain convergence of W_n to 0 in L^1, we will thus prove that $\mathbb{E}[D_{n,1}^2] \to 0$, $\mathbb{E}[S_{n,1}^2] = O(1)$, $\mathbb{E}[D_{n,2}^2] \to 0$ and $\mathbb{E}[S_{n,2}^2] = O(1)$.

We have

$$
\begin{aligned}
\mathbb{E}[D_{n,1}^2] &= \mathbb{E}[4(Y_0 X_0^\top - \beta^{\star\top})(\hat\beta'_{\lambda_n} - \hat\beta_{\lambda_n} - \tfrac{V}{n})(\hat\beta'_{\lambda_n} - \hat\beta_{\lambda_n} - \tfrac{V}{n})^\top (Y_0 X_0 - \beta^\star)]\\
&= \mathbb{E}[4\operatorname{tr}((Y_0 X_0^\top - \beta^{\star\top})(\hat\beta'_{\lambda_n} - \hat\beta_{\lambda_n} - \tfrac{V}{n})(\hat\beta'_{\lambda_n} - \hat\beta_{\lambda_n} - \tfrac{V}{n})^\top (Y_0 X_0 - \beta^\star))]\\
&= \mathbb{E}[4\operatorname{tr}((Y_0 X_0 - \beta^\star)(Y_0 X_0^\top - \beta^{\star\top})(\hat\beta'_{\lambda_n} - \hat\beta_{\lambda_n} - \tfrac{V}{n})(\hat\beta'_{\lambda_n} - \hat\beta_{\lambda_n} - \tfrac{V}{n})^\top)]\\
&= 4\operatorname{tr}(\mathbb{E}[(Y_0 X_0 - \beta^\star)(Y_0 X_0^\top - \beta^{\star\top})(\hat\beta'_{\lambda_n} - \hat\beta_{\lambda_n} - \tfrac{V}{n})(\hat\beta'_{\lambda_n} - \hat\beta_{\lambda_n} - \tfrac{V}{n})^\top])\\
&= 4\operatorname{tr}(\mathbb{E}[(Y_0 X_0 - \beta^\star)(Y_0 X_0^\top - \beta^{\star\top})]\mathbb{E}[(\hat\beta'_{\lambda_n} - \hat\beta_{\lambda_n} - \tfrac{V}{n})(\hat\beta'_{\lambda_n} - \hat\beta_{\lambda_n} - \tfrac{V}{n})^\top])
\end{aligned}
$$

as $\hat\beta'_{\lambda_n} - \hat\beta_{\lambda_n} - \frac{V}{n}$ is a function of the training points and using independence of Z_0 from the training points.

By Cauchy–Schwarz, for all i, j,

$$
\mathbb{E}[|(\hat\beta'_{\lambda_n,i} - \hat\beta_{\lambda_n,i} - \tfrac{V_i}{n})(\hat\beta'_{\lambda_n,j} - \hat\beta_{\lambda_n,j} - \tfrac{V_j}{n})|] \le \sqrt{\mathbb{E}[(\hat\beta'_{\lambda_n,i} - \hat\beta_{\lambda_n,i} - \tfrac{V_i}{n})^2]\mathbb{E}[(\hat\beta'_{\lambda_n,j} - \hat\beta_{\lambda_n,j} - \tfrac{V_j}{n})^2]}
$$

thus, if we show $\mathbb{E}[(\hat\beta'_{\lambda_n,i} - \hat\beta_{\lambda_n,i} - \frac{V_i}{n})^2] \to 0$ for all i, then we obtain

$$
\mathbb{E}[(\hat\beta'_{\lambda_n} - \hat\beta_{\lambda_n} - \tfrac{V}{n})(\hat\beta'_{\lambda_n} - \hat\beta_{\lambda_n} - \tfrac{V}{n})^\top] \to 0
$$

and therefore $\mathbb{E}[D_{n,1}^2] \to 0$. We are going to hold off on proving $\mathbb{E}[(\hat\beta'_{\lambda_n,i} - \hat\beta_{\lambda_n,i} - \frac{V_i}{n})^2] \to 0$ as we will actually show the stronger convergence $\mathbb{E}[(\hat\beta'_{\lambda_n,i} - \hat\beta_{\lambda_n,i} - \frac{V_i}{n})^4] \to 0$ in the context of proving $\mathbb{E}[D_{n,2}^2] \to 0$.

With similar computations and upper-bounding, we can show that $\mathbb{E}[S_{n,1}^2] = O(1)$ if we prove that for all i, $\mathbb{E}[(\hat\beta'_{\lambda_n,i} - \hat\beta_{\lambda_n,i} + \frac{V_i}{n})^2] = O(1)$.

As we have shown in Appendix C that the soft-thresholding Lasso estimator is consistent for $\beta^\star$ in L^2 when $\lambda_n = o(n)$, both $\mathbb{E}[\hat\beta_{\lambda_n,i}^2]$ and $\mathbb{E}[\hat\beta'^2_{\lambda_n,i}]$ are bounded and thus $\mathbb{E}[(\hat\beta'_{\lambda_n,i} - \hat\beta_{\lambda_n,i} + \frac{V_i}{n})^2] = O(1)$ since $(\hat\beta'_{\lambda_n,i} - \hat\beta_{\lambda_n,i} + \frac{V_i}{n})^2 \le 3(\hat\beta'^2_{\lambda_n,i} + \hat\beta^2_{\lambda_n,i} + \frac{V_i^2}{n^2})$ by Cauchy–Schwarz.

We now focus on proving $\mathbb{E}[D_{n,2}^2] \to 0$.

We have

$$
\begin{aligned}
D_{n,2} &= \operatorname{tr}((X_0 X_0^\top - \mathbf{I})(\hat\beta_{\lambda_n}\hat\beta_{\lambda_n}^\top - \hat\beta'_{\lambda_n}\hat\beta'^\top_{\lambda_n} - \tfrac{M}{n}))\\
&= X_0^\top(\hat\beta_{\lambda_n}\hat\beta_{\lambda_n}^\top - \hat\beta'_{\lambda_n}\hat\beta'^\top_{\lambda_n} - \tfrac{M}{n})X_0 - \operatorname{tr}(\hat\beta_{\lambda_n}\hat\beta_{\lambda_n}^\top - \hat\beta'_{\lambda_n}\hat\beta'^\top_{\lambda_n} - \tfrac{M}{n})\\
&= \textstyle\sum_{i,j}(X_{0,i}X_{0,j} - \mathbb{1}[i=j])(\hat\beta_{\lambda_n,i}\hat\beta_{\lambda_n,j} - \hat\beta'_{\lambda_n,i}\hat\beta'_{\lambda_n,j} - \frac{M_{i,j}}{n})\\
&= \textstyle\sum_{i,j} U_{i,j}(\hat\beta_{\lambda_n,i}\hat\beta_{\lambda_n,j} - \hat\beta'_{\lambda_n,i}\hat\beta'_{\lambda_n,j} - \frac{M_{i,j}}{n})
\end{aligned}
$$

where $U_{i,j} \triangleq X_{0,i}X_{0,j} - \mathbb{1}[i=j]$, and thus

$$D_{n,2}^2 = \sum_{i,j,k,l} U_{i,j}U_{k,l}(\hat{\beta}_{\lambda_n,i}\hat{\beta}_{\lambda_n,j} - \hat{\beta}'_{\lambda_n,i}\hat{\beta}'_{\lambda_n,j} - \tfrac{M_{i,j}}{n})(\hat{\beta}_{\lambda_n,k}\hat{\beta}_{\lambda_n,l} - \hat{\beta}'_{\lambda_n,k}\hat{\beta}'_{\lambda_n,l} - \tfrac{M_{k,l}}{n}).$$

Using independence of Z_0 and the training points, we have

$$\mathbb{E}[D_{n,2}^2] = \sum_{i,j,k,l} \mathbb{E}[U_{i,j}U_{k,l}]\mathbb{E}[(\hat{\beta}_{\lambda_n,i}\hat{\beta}_{\lambda_n,j} - \hat{\beta}'_{\lambda_n,i}\hat{\beta}'_{\lambda_n,j} - \tfrac{M_{i,j}}{n})(\hat{\beta}_{\lambda_n,k}\hat{\beta}_{\lambda_n,l} - \hat{\beta}'_{\lambda_n,k}\hat{\beta}'_{\lambda_n,l} - \tfrac{M_{k,l}}{n})]$$

where, using Cauchy–Schwarz,

$$\begin{aligned}&\mathbb{E}[|(\hat{\beta}_{\lambda_n,i}\hat{\beta}_{\lambda_n,j} - \hat{\beta}'_{\lambda_n,i}\hat{\beta}'_{\lambda_n,j} - \tfrac{M_{i,j}}{n})(\hat{\beta}_{\lambda_n,k}\hat{\beta}_{\lambda_n,l} - \hat{\beta}'_{\lambda_n,k}\hat{\beta}'_{\lambda_n,l} - \tfrac{M_{k,l}}{n})|]\\ &\leq \sqrt{\mathbb{E}[(\hat{\beta}_{\lambda_n,i}\hat{\beta}_{\lambda_n,j} - \hat{\beta}'_{\lambda_n,i}\hat{\beta}'_{\lambda_n,j} - \tfrac{M_{i,j}}{n})^2]\mathbb{E}[(\hat{\beta}_{\lambda_n,k}\hat{\beta}_{\lambda_n,l} - \hat{\beta}'_{\lambda_n,k}\hat{\beta}'_{\lambda_n,l} - \tfrac{M_{k,l}}{n})^2]}.\end{aligned}$$

We thus want to show $\mathbb{E}[(\hat{\beta}_{\lambda_n,i}\hat{\beta}_{\lambda_n,j} - \hat{\beta}'_{\lambda_n,i}\hat{\beta}'_{\lambda_n,j} - \tfrac{M_{i,j}}{n})^2] \to 0$ for all i,j.

Since $M = -(V\beta^{\star\top} + \beta^\star V^\top)$, we have $M_{i,j} = -V_i\beta_j^\star - \beta_i^\star V_j$ and then

$$\begin{aligned}&\hat{\beta}_{\lambda_n,i}\hat{\beta}_{\lambda_n,j} - \hat{\beta}'_{\lambda_n,i}\hat{\beta}'_{\lambda_n,j} - \tfrac{M_{i,j}}{n}\\ &= \hat{\beta}_{\lambda_n,i}\hat{\beta}_{\lambda_n,j} - \hat{\beta}'_{\lambda_n,i}\hat{\beta}'_{\lambda_n,j} + \tfrac{V_i}{n}\beta_j^\star + \beta_i^\star\tfrac{V_j}{n}\\ &= -(\hat{\beta}'_{\lambda_n,i} - \hat{\beta}_{\lambda_n,i} - \tfrac{V_i}{n})\hat{\beta}_{\lambda_n,j} - \hat{\beta}'_{\lambda_n,i}(\hat{\beta}'_{\lambda_n,j} - \hat{\beta}_{\lambda_n,j} - \tfrac{V_j}{n}) - \tfrac{V_i}{n}(\hat{\beta}_{\lambda_n,j} - \beta_j^\star) - (\hat{\beta}'_{\lambda_n,i} - \beta_i^\star)\tfrac{V_j}{n}.\end{aligned}$$

By Cauchy–Schwarz,

$$\begin{aligned}&(\hat{\beta}_{\lambda_n,i}\hat{\beta}_{\lambda_n,j} - \hat{\beta}'_{\lambda_n,i}\hat{\beta}'_{\lambda_n,j} - \tfrac{M_{i,j}}{n})^2\\ &= ((\hat{\beta}'_{\lambda_n,i} - \hat{\beta}_{\lambda_n,i} - \tfrac{V_i}{n})\hat{\beta}_{\lambda_n,j} + \hat{\beta}'_{\lambda_n,i}(\hat{\beta}'_{\lambda_n,j} - \hat{\beta}_{\lambda_n,j} - \tfrac{V_j}{n}) + \tfrac{V_i}{n}(\hat{\beta}_{\lambda_n,j} - \beta_j^\star) + (\hat{\beta}'_{\lambda_n,i} - \beta_i^\star)\tfrac{V_j}{n})^2\\ &\leq 4((\hat{\beta}'_{\lambda_n,i} - \hat{\beta}_{\lambda_n,i} - \tfrac{V_i}{n})^2\hat{\beta}_{\lambda_n,j}^2 + \hat{\beta}'^2_{\lambda_n,i}(\hat{\beta}'_{\lambda_n,j} - \hat{\beta}_{\lambda_n,j} - \tfrac{V_j}{n})^2 + \tfrac{V_i^2}{n^2}(\hat{\beta}_{\lambda_n,j} - \beta_j^\star)^2 + (\hat{\beta}'_{\lambda_n,i} - \beta_i^\star)^2\tfrac{V_j^2}{n^2})\end{aligned}$$

and the probability version of Cauchy–Schwarz yields

$$\begin{aligned}&\mathbb{E}[(\hat{\beta}_{\lambda_n,i}\hat{\beta}_{\lambda_n,j} - \hat{\beta}'_{\lambda_n,i}\hat{\beta}'_{\lambda_n,j} - \tfrac{M_{i,j}}{n})^2]\\ &\leq 4\Big(\sqrt{\mathbb{E}[(\hat{\beta}'_{\lambda_n,i} - \hat{\beta}_{\lambda_n,i} - \tfrac{V_i}{n})^4]\mathbb{E}[\hat{\beta}^4_{\lambda_n,j}]} + \sqrt{\mathbb{E}[\hat{\beta}'^4_{\lambda_n,i}]\mathbb{E}[(\hat{\beta}'_{\lambda_n,j} - \hat{\beta}_{\lambda_n,j} - \tfrac{V_j}{n})^4]}\\ &\quad + \sqrt{\tfrac{\mathbb{E}[V_i^4]}{n^4}\mathbb{E}[(\hat{\beta}_{\lambda_n,j} - \beta_j^\star)^4]} + \sqrt{\mathbb{E}[(\hat{\beta}'_{\lambda_n,i} - \beta_i^\star)^4]\tfrac{\mathbb{E}[V_j^4]}{n^4}}\Big).\end{aligned}$$

Hence, we will get $\mathbb{E}[D_{n,2}^2] \to 0$ if we prove that for all i

- $\mathbb{E}[(\hat{\beta}_{\lambda_n,i} - \beta_i^\star)^4] \to 0$, the proof will be the same for $\mathbb{E}[(\hat{\beta}'_{\lambda_n,i} - \beta_i^\star)^4] \to 0$,
- $\mathbb{E}[(\hat{\beta}'_{\lambda_n,i} - \hat{\beta}_{\lambda_n,i} - \tfrac{V_i}{n})^4] \to 0$.

Note that we will automatically get L^2 convergence of $\hat{\beta}'_{\lambda_n,i} - \hat{\beta}_{\lambda_n,i} - \frac{V_i}{n}$ to 0 for all i, which implies $\mathbb{E}[D_{n,1}^2] \to 0$ as mentioned earlier.

We now introduce a lemma that will allow us to upper-bound quantities of interest.

Lemma D.1 (Hölder corollary). *For integers $k, \ell \geq 2$, for $(a_1, \ldots, a_k) \in \mathbb{R}^k$, we have the following inequality*

$$(\textstyle\sum_{i=1}^k |a_i|)^\ell \leq k^{\ell-1}\sum_{i=1}^k |a_i|^\ell.$$

Proof For $(x_1, \ldots, x_k), (y_1, \ldots, y_k) \in \mathbb{R}^k$ and $p, q \in (1, +\infty)$ such that $\frac{1}{p} + \frac{1}{q} = 1$, Hölder's inequality gives us

$$\textstyle\sum_{i=1}^k |x_i y_i| \leq (\sum_{i=1}^k |x_i|^p)^{\frac{1}{p}}(\sum_{i=1}^k |y_i|^q)^{\frac{1}{q}}$$

and therefore the lemma is an application of it with $x_i = a_i, y_i = 1, p = \ell$. $\square$

Combining Lemma D.1 for $\ell = 4$ with similar computations and upper-bounding as above, we can show that $\mathbb{E}[S_{n,2}^2]$ is bounded if for all i, $\mathbb{E}[\hat{\beta}_{\lambda_n,i}^4]$ and $\mathbb{E}[\hat{\beta}'^4_{\lambda_n,i}]$ are bounded, which automatically comes from the L^4 convergence of the soft-thresholding Lasso estimator to $\beta^\star$ needed for $\mathbb{E}[D_{n,2}^2] \to 0$.

We start by showing $\mathbb{E}[(\hat{\beta}_{\lambda_n,i} - \beta_j^\star)^4] \to 0$.

As a reminder, we have

$$\hat{\beta}_{\lambda_n,i} = \text{sign}(\hat{\beta}_{\text{OLS},i}) \begin{cases} |\hat{\beta}_{\text{OLS},i}| - \frac{\lambda_n}{n} & \text{if } |\hat{\beta}_{\text{OLS},i}| \geq \frac{\lambda_n}{n} \\ 0 & \text{if } |\hat{\beta}_{\text{OLS},i}| < \frac{\lambda_n}{n} \end{cases}$$

thus, using $(a+b)^4 \leq 8(a^4 + b^4)$, which is an application of Lemma D.1 for $\ell = 4$,

$$\begin{aligned}
&\mathbb{E}[(\hat{\beta}_{\lambda_n,i} - \beta_i^\star)^4 \mid \mathbf{X}] \\
&= \mathbb{E}[(\hat{\beta}_{\text{OLS},i} - \beta_i^\star - \tfrac{\lambda_n}{n})^4 \mid \hat{\beta}_{\text{OLS},i} \geq \tfrac{\lambda_n}{n}, \mathbf{X}]\mathbb{P}(\hat{\beta}_{\text{OLS},i} \geq \tfrac{\lambda_n}{n} \mid \mathbf{X}) \\
&\quad + \mathbb{E}[(\hat{\beta}_{\text{OLS},i} - \beta_i^\star + \tfrac{\lambda_n}{n})^4 \mid \hat{\beta}_{\text{OLS},i} \leq -\tfrac{\lambda_n}{n}, \mathbf{X}]\mathbb{P}(\hat{\beta}_{\text{OLS},i} \leq -\tfrac{\lambda_n}{n} \mid \mathbf{X}) \\
&\leq 8(\mathbb{E}[(\hat{\beta}_{\text{OLS},i} - \beta_i^\star)^4 \mid \hat{\beta}_{\text{OLS},i} \geq \tfrac{\lambda_n}{n}, \mathbf{X}] + \tfrac{\lambda_n^4}{n^4})\mathbb{P}(\hat{\beta}_{\text{OLS},i} \geq \tfrac{\lambda_n}{n} \mid \mathbf{X}) \\
&\quad + 8(\mathbb{E}[(\hat{\beta}_{\text{OLS},i} - \beta_i^\star)^4 \mid \hat{\beta}_{\text{OLS},i} \leq -\tfrac{\lambda_n}{n}, \mathbf{X}] + \tfrac{\lambda_n^4}{n^4})\mathbb{P}(\hat{\beta}_{\text{OLS},i} \leq -\tfrac{\lambda_n}{n} \mid \mathbf{X}).
\end{aligned}$$

Since $\hat{\beta}_{\text{OLS}} \mid \mathbf{X} \sim \mathcal{N}(\beta^\star, \tau^2(\mathbf{X}^\top\mathbf{X})^{-1})$, we can write $\hat{\beta}_{\text{OLS},i} = \beta_i^\star + \tilde{\tau}_n Z$ where $\tilde{\tau}_n = \frac{\tau}{\sqrt{n}}\sqrt{(\frac{\mathbf{X}^\top\mathbf{X}}{n})^{-1}_{i,i}}$ and $Z \mid \mathbf{X} \sim \mathcal{N}(0,1)$. Note that we could have i as a subscript of $\tilde{\tau}_n$ and Z, but we will only consider one i at a time in our computations and we can thus omit this subscript for both of them for the sake of notational simplicity, and we will also omit it for some additional notation we define in the rest of the proof.

Define $\alpha_n^{(1)} = \frac{1}{\tilde{\tau}_n}(\frac{\lambda_n}{n} - \beta_i^\star)$ and $\alpha_n^{(2)} = \frac{1}{\tilde{\tau}_n}(\frac{\lambda_n}{n} + \beta_i^\star)$.

Using the fourth moment of the truncated normal [26], we have

$$\begin{aligned}
&\mathbb{E}[(\hat{\beta}_{\text{OLS},i} - \beta_i^\star)^4 \mid \hat{\beta}_{\text{OLS},i} \geq \tfrac{\lambda_n}{n}, \mathbf{X}] \\
&= \mathbb{E}[(\tilde{\tau}_n Z)^4 \mid Z \geq \alpha_n^{(1)}, \mathbf{X}] \\
&= \tilde{\tau}_n^4\, \mathbb{E}[Z^4 \mid Z \geq \alpha_n^{(1)}, \mathbf{X}] \\
&= \tilde{\tau}_n^4 (3 + \tfrac{((\alpha_n^{(1)})^3 + 3\alpha_n^{(1)})\varphi(\alpha_n^{(1)})}{1 - \Phi(\alpha_n^{(1)})})
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}[(\hat{\beta}_{\text{OLS},i} - \beta_i^\star)^4 \mid \hat{\beta}_{\text{OLS},i} \leq -\tfrac{\lambda_n}{n}, \mathbf{X}] \\
&= \tilde{\tau}_n^4\, \mathbb{E}[Z^4 \mid Z \leq -\alpha_n^{(2)}, \mathbf{X}] \\
&= \tilde{\tau}_n^4 (3 + \tfrac{((\alpha_n^{(2)})^3 + \alpha_n^{(2)})\varphi(-\alpha_n^{(2)})}{\Phi(-\alpha_n^{(2)})}).
\end{aligned}$$

Since $\mathbb{P}(\hat{\beta}_{\text{OLS},i} \geq \frac{\lambda_n}{n} \mid \mathbf{X}) = 1 - \Phi(\alpha_n^{(1)})$ and $\mathbb{P}(\hat{\beta}_{\text{OLS},i} \leq -\frac{\lambda_n}{n} \mid \mathbf{X}) = \Phi(-\alpha_n^{(2)})$,

$$\begin{aligned}
&\mathbb{E}[(\hat{\beta}_{\lambda_n,i} - \beta_i^\star)^4 \mid \mathbf{X}] \\
&\leq 8(\mathbb{E}[(\hat{\beta}_{\text{OLS},i} - \beta_i^\star)^4 \mid \hat{\beta}_{\text{OLS},i} \geq \tfrac{\lambda_n}{n}, \mathbf{X}] + \tfrac{\lambda_n^4}{n^4})\mathbb{P}(\hat{\beta}_{\text{OLS},i} \geq \tfrac{\lambda_n}{n} \mid \mathbf{X}) \\
&\quad + 8(\mathbb{E}[(\hat{\beta}_{\text{OLS},i} - \beta_i^\star)^4 \mid \hat{\beta}_{\text{OLS},i} \leq -\tfrac{\lambda_n}{n}, \mathbf{X}] + \tfrac{\lambda_n^4}{n^4})\mathbb{P}(\hat{\beta}_{\text{OLS},i} \leq -\tfrac{\lambda_n}{n} \mid \mathbf{X}) \\
&= 8(3\tilde{\tau}_n^4(1 - \Phi(\alpha_n^{(1)})) + \tilde{\tau}_n^4((\alpha_n^{(1)})^3 + 3\alpha_n^{(1)})\varphi(\alpha_n^{(1)}) + \tfrac{\lambda_n^4}{n^4}(1 - \Phi(\alpha_n^{(1)}))) \\
&\quad + 8(3\tilde{\tau}_n^4\Phi(-\alpha_n^{(2)}) + \tilde{\tau}_n^4((\alpha_n^{(2)})^3 + \alpha_n^{(2)})\varphi(-\alpha_n^{(2)}) + \tfrac{\lambda_n^4}{n^4}\Phi(-\alpha_n^{(2)})).
\end{aligned}$$

For $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I})$, we know $(\mathbf{X}^\top\mathbf{X})^{-1} \sim W_p^{-1}(\mathbf{I}, n)$ and then the diagonal element $(\mathbf{X}^\top\mathbf{X})^{-1}_{i,i}$ follows an inverse gamma distribution with shape parameter $\frac{n-p+1}{2}$ and scale parameter $\frac{1}{2}$. Therefore, $\mathbb{E}[((\mathbf{X}^\top\mathbf{X})^{-1}_{i,i})^2] = \frac{1}{(n-p-1)(n-p-3)}$ and $\mathbb{E}[((\frac{\mathbf{X}^\top\mathbf{X}}{n})^{-1}_{i,i})^2] = \frac{n^2}{(n-p-1)(n-p-3)} = o(n^2)$.

As a reminder, $\tilde{\tau}_n = \frac{\tau}{\sqrt{n}}\sqrt{(\frac{\mathbf{X}^\top\mathbf{X}}{n})^{-1}_{i,i}}$. We then have L^1 convergence of $\tilde{\tau}_n^4$ to 0. As previously mentioned, the function $x \mapsto x\varphi(x)$ is bounded. Similarly, a straightforward study of the behavior of the function $x \mapsto x^3\varphi(x)$ shows it is bounded. Since Φ is also bounded, and $\lambda_n = o(n)$, then

$$\mathbb{E}[(\hat{\beta}_{\lambda_n,i} - \beta_i^\star)^4] = \mathbb{E}[\mathbb{E}[(\hat{\beta}_{\lambda_n,i} - \beta_i^\star)^4 \mid \mathbf{X}]] \to 0.$$

We now show that $\mathbb{E}[(\hat{\beta}'_{\lambda_n,i} - \hat{\beta}_{\lambda_n,i} - \frac{V_i}{n})^4] \to 0$.

We have

$$\begin{aligned}
&\hat{\beta}'_{\lambda_n,i} - \hat{\beta}_{\lambda_n,i} \\
&= \operatorname{sign}(\hat{\beta}'_{\text{OLS},i})(|\hat{\beta}'_{\text{OLS},i}| - \tfrac{\lambda_n}{n})_+ - \operatorname{sign}(\hat{\beta}_{\text{OLS},i})(|\hat{\beta}_{\text{OLS},i}| - \tfrac{\lambda_n}{n})_+ \\
&= \operatorname{sign}(\hat{\beta}'_{\text{OLS},i})\begin{cases} |\hat{\beta}'_{\text{OLS},i}| - \frac{\lambda_n}{n} & \text{if } |\hat{\beta}'_{\text{OLS},i}| \geq \frac{\lambda_n}{n} \\ 0 & \text{if } |\hat{\beta}'_{\text{OLS},i}| < \frac{\lambda_n}{n} \end{cases} - \operatorname{sign}(\hat{\beta}_{\text{OLS},i})\begin{cases} |\hat{\beta}_{\text{OLS},i}| - \frac{\lambda_n}{n} & \text{if } |\hat{\beta}_{\text{OLS},i}| \geq \frac{\lambda_n}{n} \\ 0 & \text{if } |\hat{\beta}_{\text{OLS},i}| < \frac{\lambda_n}{n} \end{cases} \\
&= \begin{cases} \hat{\beta}'_{\text{OLS},i} - \frac{\lambda_n}{n} & \text{if } \hat{\beta}'_{\text{OLS},i} \geq \frac{\lambda_n}{n} \\ \hat{\beta}'_{\text{OLS},i} + \frac{\lambda_n}{n} & \text{if } \hat{\beta}'_{\text{OLS},i} \leq -\frac{\lambda_n}{n} \\ 0 & \text{if } |\hat{\beta}'_{\text{OLS},i}| < \frac{\lambda_n}{n} \end{cases} - \begin{cases} \hat{\beta}_{\text{OLS},i} - \frac{\lambda_n}{n} & \text{if } \hat{\beta}_{\text{OLS},i} \geq \frac{\lambda_n}{n} \\ \hat{\beta}_{\text{OLS},i} + \frac{\lambda_n}{n} & \text{if } \hat{\beta}_{\text{OLS},i} \leq -\frac{\lambda_n}{n} \\ 0 & \text{if } |\hat{\beta}_{\text{OLS},i}| < \frac{\lambda_n}{n} \end{cases}.
\end{aligned}$$

As an intermediate step, we need to show $\hat{\beta}'_{\text{OLS}} - \hat{\beta}_{\text{OLS}} - \frac{V}{n} \xrightarrow{\text{a.s.}} 0$.

Let $\tilde{\mathbf{X}} \triangleq (X_2, \ldots, X_n)^\top$ be the matrix of regressors for the training points except for the first one that is being changed.

We have

$$\begin{aligned}
&\hat{\beta}'_{\text{OLS}} - \hat{\beta}_{\text{OLS}} \\
&= (\mathbf{X}'^\top\mathbf{X}')^{-1}\mathbf{X}'^\top\mathbf{Y}' - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} \\
&= (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + X_1'X_1'^\top)^{-1}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{Y}} + Y_1'X_1') - (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + X_1X_1^\top)^{-1}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{Y}} + Y_1X_1) \\
&= [(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + X_1'X_1'^\top)^{-1} - (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + X_1X_1^\top)^{-1}]\tilde{\mathbf{X}}^\top\tilde{\mathbf{Y}} \\
&\quad + (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + X_1'X_1'^\top)^{-1}Y_1'X_1' - (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + X_1X_1^\top)^{-1}Y_1X_1.
\end{aligned}$$

Using the Sherman–Morrison–Woodbury formula,

$$\begin{aligned}
(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + X_1X_1^\top)^{-1} &= (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1} - (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}X_1(\mathbf{I} + X_1^\top(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}X_1)^{-1}X_1^\top(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1} \\
&= \tfrac{1}{n}(\tfrac{\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}}{n})^{-1} - \tfrac{1}{n^2}(\tfrac{\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}}{n})^{-1}X_1(\mathbf{I} + \tfrac{1}{n}X_1^\top(\tfrac{\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}}{n})^{-1}X_1)^{-1}X_1^\top(\tfrac{\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}}{n})^{-1} \\
&= \tfrac{1}{n}A_n - \tfrac{1}{n^2}B_n
\end{aligned}$$

where, by the strong law of large numbers,

- $A_n \triangleq (\frac{\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}}{n})^{-1} \xrightarrow{\text{a.s.}} \mathbb{E}[X_0X_0^\top]^{-1} = \mathbf{I}$,
- $B_n \triangleq (\frac{\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}}{n})^{-1}X_1(\mathbf{I} + \frac{1}{n}X_1^\top(\frac{\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}}{n})^{-1}X_1)^{-1}X_1^\top(\frac{\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}}{n})^{-1} \xrightarrow{\text{a.s.}} X_1X_1^\top$.

Similarly,

$$(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} + X_1'X_1'^\top)^{-1} = \tfrac{1}{n}A_n - \tfrac{1}{n^2}B_n'$$

with

$$B_n' \triangleq (\tfrac{\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}}{n})^{-1}X_1'(\mathbf{I} + \tfrac{1}{n}X_1'^\top(\tfrac{\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}}{n})^{-1}X_1')^{-1}X_1'^\top(\tfrac{\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}}{n})^{-1} \xrightarrow{\text{a.s.}} X_1'X_1'^\top.$$

Then

$$\begin{aligned}
\hat{\beta}'_{\text{OLS}} - \hat{\beta}_{\text{OLS}} &= \tfrac{1}{n^2}(B_n - B_n')\tilde{\mathbf{X}}^\top\tilde{\mathbf{Y}} + (\tfrac{1}{n}A_n - \tfrac{1}{n^2}B_n')Y_1'X_1' - (\tfrac{1}{n}A_n - \tfrac{1}{n^2}B_n)Y_1X_1 \\
&= \tfrac{1}{n}(B_n - B_n')\tfrac{\tilde{\mathbf{X}}^\top\tilde{\mathbf{Y}}}{n} + (\tfrac{1}{n}A_n - \tfrac{1}{n^2}B_n')Y_1'X_1' - (\tfrac{1}{n}A_n - \tfrac{1}{n^2}B_n)Y_1X_1
\end{aligned}$$

where $\frac{\tilde{\mathbf{X}}^\top \tilde{\mathbf{Y}}}{n} \xrightarrow{\text{a.s.}} \mathbb{E}[Y_0 X_0] = \beta^\star$, by the strong law of large numbers.

Therefore,

$$\begin{aligned} n(\hat{\beta}'_{\text{OLS}} - \hat{\beta}_{\text{OLS}}) &\xrightarrow{\text{a.s.}} (X_1 X_1^\top - X'_1 X_1'^\top)\beta^\star + Y'_1 X'_1 - Y_1 X_1 \\ &= (Y'_1 - X_1'^\top \beta^\star) X'_1 - (Y_1 - X_1^\top \beta^\star) X_1 \\ &= V. \end{aligned}$$

We can write

$$\begin{aligned} (\hat{\beta}'_{\lambda_n,i} - \hat{\beta}_{\lambda_n,i} - \tfrac{V_i}{n})^4 = &(\hat{\beta}'_{\text{OLS},i} - \hat{\beta}_{\text{OLS},i} - \tfrac{V_i}{n})^4 \, \mathbb{1}\left[\hat{\beta}_{\text{OLS},i} \geq \tfrac{\lambda_n}{n}, \hat{\beta}'_{\text{OLS},i} \geq \tfrac{\lambda_n}{n}\right] \\ &+ (\hat{\beta}'_{\text{OLS},i} - \hat{\beta}_{\text{OLS},i} - \tfrac{V_i}{n})^4 \, \mathbb{1}\left[\hat{\beta}_{\text{OLS},i} \leq -\tfrac{\lambda_n}{n}, \hat{\beta}'_{\text{OLS},i} \leq -\tfrac{\lambda_n}{n}\right] \\ &+ (\hat{\beta}'_{\text{OLS},i} - \hat{\beta}_{\text{OLS},i} - 2\tfrac{\lambda_n}{n} - \tfrac{V_i}{n})^4 \, \mathbb{1}\left[\hat{\beta}_{\text{OLS},i} \leq -\tfrac{\lambda_n}{n}, \hat{\beta}'_{\text{OLS},i} \geq \tfrac{\lambda_n}{n}\right] \\ &+ (\hat{\beta}'_{\text{OLS},i} - \hat{\beta}_{\text{OLS},i} + 2\tfrac{\lambda_n}{n} - \tfrac{V_i}{n})^4 \, \mathbb{1}\left[\hat{\beta}_{\text{OLS},i} \geq \tfrac{\lambda_n}{n}, \hat{\beta}'_{\text{OLS},i} \leq -\tfrac{\lambda_n}{n}\right] \\ &+ (\hat{\beta}'_{\text{OLS},i} - \tfrac{\lambda_n}{n} - \tfrac{V_i}{n})^4 \, \mathbb{1}\left[|\hat{\beta}_{\text{OLS},i}| < \tfrac{\lambda_n}{n}, \hat{\beta}'_{\text{OLS},i} \geq \tfrac{\lambda_n}{n}\right] \\ &+ (\hat{\beta}'_{\text{OLS},i} + \tfrac{\lambda_n}{n} - \tfrac{V_i}{n})^4 \, \mathbb{1}\left[|\hat{\beta}_{\text{OLS},i}| < \tfrac{\lambda_n}{n}, \hat{\beta}'_{\text{OLS},i} \leq -\tfrac{\lambda_n}{n}\right] \\ &+ (\hat{\beta}_{\text{OLS},i} - \tfrac{\lambda_n}{n} + \tfrac{V_i}{n})^4 \, \mathbb{1}\left[\hat{\beta}_{\text{OLS},i} \geq \tfrac{\lambda_n}{n}, |\hat{\beta}'_{\text{OLS},i}| < \tfrac{\lambda_n}{n}\right] \\ &+ (\hat{\beta}_{\text{OLS},i} + \tfrac{\lambda_n}{n} + \tfrac{V_i}{n})^4 \, \mathbb{1}\left[\hat{\beta}_{\text{OLS},i} \leq -\tfrac{\lambda_n}{n}, |\hat{\beta}'_{\text{OLS},i}| < \tfrac{\lambda_n}{n}\right] \\ &+ (\tfrac{V_i}{n})^4 \, \mathbb{1}\left[|\hat{\beta}_{\text{OLS},i}| < \tfrac{\lambda_n}{n}, |\hat{\beta}'_{\text{OLS},i}| < \tfrac{\lambda_n}{n}\right] \end{aligned}$$

and we have a similar expression for $(\hat{\beta}'_{\lambda_n,i} - \hat{\beta}_{\lambda_n,i} - \frac{V_i}{n})^6$ with terms taken to the sixth power.

Since $\hat{\beta}_{\text{OLS}} \mid \mathbf{X} \sim \mathcal{N}(\beta^\star, \tau^2(\mathbf{X}^\top\mathbf{X})^{-1})$ and we can bound the central moments of a Normal with the powers of its variance, there exists $C > 0$ such that $\mathbb{E}[(\hat{\beta}_{\text{OLS},i} - \beta^\star_i)^6 \mid \mathbf{X}] \leq C(\tau^2(\mathbf{X}^\top\mathbf{X})^{-1}_{i,i})^3 = C\tau^6((\mathbf{X}^\top\mathbf{X})^{-1}_{i,i})^3$.

For $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I})$, we know $(\mathbf{X}^\top\mathbf{X})^{-1} \sim W_p^{-1}(\mathbf{I}, n)$ and then the diagonal element $(\mathbf{X}^\top\mathbf{X})^{-1}_{i,i}$ follows an inverse gamma distribution with shape parameter $\frac{n-p+1}{2}$ and scale parameter $\frac{1}{2}$. Therefore, $\mathbb{E}[((\mathbf{X}^\top\mathbf{X})^{-1}_{i,i})^3] = \frac{1}{(n-p-1)(n-p-3)(n-p-5)}$, which means $\mathbb{E}[(\hat{\beta}_{\text{OLS},i} - \beta^\star_i)^6]$ and thus $\mathbb{E}[\hat{\beta}^6_{\text{OLS},i}]$, by an application of Lemma D.1 for $\ell = 6$, are bounded. Similarly, $\mathbb{E}[\hat{\beta}'^6_{\text{OLS},i}]$ is bounded.

Consequently, since $\lambda_n = o(n)$ and $\mathbb{E}[\hat{\beta}^6_{\text{OLS},i}]$ and $\mathbb{E}[\hat{\beta}'^6_{\text{OLS},i}]$ are bounded, the almost sure convergence of the fourth moment turns into L^1 convergence to 0.

Therefore,

$$\begin{aligned} \gamma(h_n^{\text{sing}}) &\sim \tfrac{1}{n^2}\mathbb{E}[(2(Y_0 X_0^\top - \beta^{\star\top})V + \text{tr}((X_0 X_0^\top - \mathbf{I})M))^2] \\ &= \tfrac{1}{n^2}\mathbb{E}[(2Y_0 X_0^\top V - 2\beta^{\star\top} V - \text{tr}((X_0 X_0^\top - \mathbf{I})(V\beta^{\star\top} + \beta^\star V^\top)))^2] \\ &= \tfrac{1}{n^2}\mathbb{E}[(2Y_0 X_0^\top V - 2X_0^\top \beta^\star X_0^\top V)^2] \\ &= \tfrac{1}{n^2}\mathbb{E}[(2(Y_0 - X_0^\top \beta^\star) X_0^\top V)^2] \end{aligned} \tag{D.1}$$

where $V = (Y'_1 - X_1'^\top \beta^\star) X'_1 - (Y_1 - X_1^\top \beta^\star) X_1$.

E Proof of Theorem 2: Relative instability of soft-thresholding comparison

Theorem 2 follows immediately from the following two propositions, proved in Appendices F and G, respectively. Note that the first proposition holds for $\lambda_n = o(n)$ and $\delta_n = o(n)$, and does not require the assumption $\|\beta^\star\|_0 < p$, which makes this proposition a stronger result than what is needed for the proof of Theorem 2 assuming $\lambda_n = O(\sqrt{n})$, $\lambda_n = \omega(1)$, $\delta_n = \Theta(1)$ and $\|\beta^\star\|_0 < p$.

Proposition E.1 (Convergence rate of $\sigma^2(h_n^{\text{diff}})$ for comparison of $\text{ST}(\lambda_n)$ with $\text{ST}(\lambda_n+\delta_n)$)**.** *Assume the linear model* (2.1)*, with feature and noise distributions as given immediately following its equation. If $\lambda_n = o(n)$ and $\delta_n = o(n)$, then $\frac{n^2}{\delta_n^2}\,\sigma^2(h_n^{\text{diff}}) \to 4\tau^2\|\beta\|_0$.*

Proposition E.2 (Lower-bounding rate of $\gamma(h_n^{\text{diff}})$ for comparison of $\text{ST}(\lambda_n)$ with $\text{ST}(\lambda_n+\delta_n)$)**.** *Assume the linear model* (2.1)*, with feature and noise distributions as given immediately following its equation, and $\|\beta^\star\|_0 < p$. If $\lambda_n = O(\sqrt{n})$, $\lambda_n = \omega(1)$, and $\delta_n = \Theta(1)$, then $\gamma(h_n^{\text{diff}}) = \Omega(\frac{\delta_n^2}{n^2\sqrt{n}})$.*

F Proof of Proposition E.1: Convergence rate of $\sigma^2(h_n^{\text{diff}})$ for comparison of ST(λ_n) with ST$(\lambda_n+\delta_n)$

We will show that

$$\frac{n}{\delta_n}\mathbb{E}[\hat\beta_{\lambda_n+\delta_n} - \hat\beta_{\lambda_n}] \to -\text{sign}(\beta^\star)$$

and

$$\frac{n}{\delta_n}\mathbb{E}[\hat\beta_{\lambda_n+\delta_n}\hat\beta_{\lambda_n+\delta_n}^\top - \hat\beta_{\lambda_n}\hat\beta_{\lambda_n}^\top] \to -(\text{sign}(\beta^\star)\beta^{\star\top} + \beta^\star\text{sign}(\beta^\star)^\top)$$

where $\text{sign}(\beta^\star) = (\text{sign}(\beta_i^\star))_{i\in[p]}$.

We have for $i = 1,\dots,p$,

$$\begin{aligned}\hat\beta_{\lambda_n+\delta_n,i} - \hat\beta_{\lambda_n,i} &= \text{sign}(\hat\beta_{\text{OLS},i})(|\hat\beta_{\text{OLS},i}| - \tfrac{\lambda_n+\delta_n}{n})_+ - \text{sign}(\hat\beta_{\text{OLS},i})(|\hat\beta_{\text{OLS},i}| - \tfrac{\lambda_n}{n})_+ \\ &= -\text{sign}(\hat\beta_{\text{OLS},i})\begin{cases}\frac{\delta_n}{n} & \text{if } |\hat\beta_{\text{OLS},i}| > \frac{\lambda_n+\delta_n}{n} \\ |\hat\beta_{\text{OLS},i}| - \frac{\lambda_n}{n} & \text{if } |\hat\beta_{\text{OLS},i}| \in [\frac{\lambda_n}{n}, \frac{\lambda_n+\delta_n}{n}] \\ 0 & \text{if } |\hat\beta_{\text{OLS},i}| < \frac{\lambda_n}{n}\end{cases}.\end{aligned}$$

Since $\hat\beta_{\text{OLS}} \mid \mathbf{X} \sim \mathcal{N}(\beta^\star, \tau^2(\mathbf{X}^\top\mathbf{X})^{-1})$, we can write $\hat\beta_{\text{OLS},i} = \beta_i^\star + \tilde\tau_n Z$ where $\tilde\tau_n = \frac{\tau}{\sqrt{n}}\sqrt{(\frac{\mathbf{X}^\top\mathbf{X}}{n})^{-1}_{i,i}}$ and $Z \mid \mathbf{X} \sim \mathcal{N}(0,1)$. Note that we could have i as a subscript of $\tilde\tau_n$ and Z, but we will only consider one i at a time in our computations and we can thus omit this subscript for both of them for the sake of notational simplicity, and we will also omit it for some additional notation we define in the rest of the proof.

We now show that $\frac{n}{\delta_n}\mathbb{E}[\hat\beta_{\lambda_n+\delta_n,i} - \hat\beta_{\lambda_n,i}] \to -\text{sign}(\beta_i^\star)$.

Using the law of total expectation,

$$\begin{aligned}&\mathbb{E}[\hat\beta_{\lambda_n+\delta_n,i} - \hat\beta_{\lambda_n,i} \mid \mathbf{X}] \\ &= -\tfrac{\delta_n}{n}\mathbb{P}(\hat\beta_{\text{OLS},i} > \tfrac{\lambda_n+\delta_n}{n} \mid \mathbf{X}) + \tfrac{\delta_n}{n}\mathbb{P}(\hat\beta_{\text{OLS},i} < -\tfrac{\lambda_n+\delta_n}{n} \mid \mathbf{X}) \\ &\quad - \mathbb{E}[\hat\beta_{\text{OLS},i} - \tfrac{\lambda_n}{n} \mid \hat\beta_{\text{OLS},i} \in [\tfrac{\lambda_n}{n}, \tfrac{\lambda_n+\delta_n}{n}], \mathbf{X}]\,\mathbb{P}(\hat\beta_{\text{OLS},i} \in [\tfrac{\lambda_n}{n}, \tfrac{\lambda_n+\delta_n}{n}] \mid \mathbf{X}) \\ &\quad - \mathbb{E}[\hat\beta_{\text{OLS},i} + \tfrac{\lambda_n}{n} \mid \hat\beta_{\text{OLS},i} \in [-\tfrac{\lambda_n+\delta_n}{n}, -\tfrac{\lambda_n}{n}], \mathbf{X}]\,\mathbb{P}(\hat\beta_{\text{OLS},i} \in [-\tfrac{\lambda_n+\delta_n}{n}, -\tfrac{\lambda_n}{n}] \mid \mathbf{X})\end{aligned}$$

Define $\alpha_n^{(1)} = \frac{1}{\tilde\tau_n}(\frac{\lambda_n}{n} - \beta_i^\star)$, $\alpha_n^{(2)} = \frac{1}{\tilde\tau_n}(\frac{\lambda_n}{n} + \beta_i^\star)$, $\theta_n^{(1)} = \frac{1}{\tilde\tau_n}(\frac{\lambda_n+\delta_n}{n} - \beta_i^\star)$ and $\theta_n^{(2)} = \frac{1}{\tilde\tau_n}(\frac{\lambda_n+\delta_n}{n} + \beta_i^\star)$.

In the order they appear, the four probabilities above are equal to

$$\mathbb{P}(Z > \theta_n^{(1)} \mid \mathbf{X}) = 1 - \Phi(\theta_n^{(1)}),$$

$$\mathbb{P}(Z < -\theta_n^{(2)} \mid \mathbf{X}) = \Phi(-\theta_n^{(2)}),$$

$$\mathbb{P}(Z \in [\alpha_n^{(1)}, \theta_n^{(1)}] \mid \mathbf{X}) = \Phi(\theta_n^{(1)}) - \Phi(\alpha_n^{(1)}),$$

$$\mathbb{P}(Z \in [-\theta_n^{(2)}, -\alpha_n^{(2)}] \mid \mathbf{X}) = \Phi(-\alpha_n^{(2)}) - \Phi(-\theta_n^{(2)}).$$

Using the first moment of the truncated normal [7], we have

$$\begin{aligned}\mathbb{E}[\hat\beta_{\mathrm{OLS},i} - \tfrac{\lambda_n}{n} \mid \hat\beta_{\mathrm{OLS},i} \in [\tfrac{\lambda_n}{n}, \tfrac{\lambda_n+\delta_n}{n}], \mathbf{X}] &= \beta_i^\star - \tfrac{\lambda_n}{n} + \tilde\tau_n\, \mathbb{E}[Z \mid Z \in [\alpha_n^{(1)}, \theta_n^{(1)}], \mathbf{X}] \\ &= \beta_i^\star - \tfrac{\lambda_n}{n} - \tilde\tau_n \tfrac{\varphi(\theta_n^{(1)}) - \varphi(\alpha_n^{(1)})}{\Phi(\theta_n^{(1)}) - \Phi(\alpha_n^{(1)})}\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[\hat\beta_{\mathrm{OLS},i} + \tfrac{\lambda_n}{n} \mid \hat\beta_{\mathrm{OLS},i} \in [-\tfrac{\lambda_n+\delta_n}{n}, -\tfrac{\lambda_n}{n}], \mathbf{X}] &= \beta_i^\star + \tfrac{\lambda_n}{n} + \tilde\tau_n\, \mathbb{E}[Z \mid Z \in [-\theta_n^{(2)}, -\alpha_n^{(2)}], \mathbf{X}] \\ &= \beta_i^\star + \tfrac{\lambda_n}{n} - \tilde\tau_n \tfrac{\varphi(-\alpha_n^{(2)}) - \varphi(-\theta_n^{(2)})}{\Phi(-\alpha_n^{(2)}) - \Phi(-\theta_n^{(2)})}.\end{aligned}$$

Therefore

$$\begin{aligned}&\mathbb{E}[\hat\beta_{\lambda_n+\delta_n,i} - \hat\beta_{\lambda_n,i} \mid \mathbf{X}] \\ &= -\tfrac{\delta_n}{n}\mathbb{P}(\hat\beta_{\mathrm{OLS},i} > \tfrac{\lambda_n+\delta_n}{n} \mid \mathbf{X}) + \tfrac{\delta_n}{n}\mathbb{P}(\hat\beta_{\mathrm{OLS},i} < -\tfrac{\lambda_n+\delta_n}{n} \mid \mathbf{X}) \\ &\quad - \mathbb{E}[\hat\beta_{\mathrm{OLS},i} - \tfrac{\lambda_n}{n} \mid \hat\beta_{\mathrm{OLS},i} \in [\tfrac{\lambda_n}{n}, \tfrac{\lambda_n+\delta_n}{n}], \mathbf{X}]\, \mathbb{P}(\hat\beta_{\mathrm{OLS},i} \in [\tfrac{\lambda_n}{n}, \tfrac{\lambda_n+\delta_n}{n}] \mid \mathbf{X}) \\ &\quad - \mathbb{E}[\hat\beta_{\mathrm{OLS},i} + \tfrac{\lambda_n}{n} \mid \hat\beta_{\mathrm{OLS},i} \in [-\tfrac{\lambda_n+\delta_n}{n}, -\tfrac{\lambda_n}{n}], \mathbf{X}]\, \mathbb{P}(\hat\beta_{\mathrm{OLS},i} \in [-\tfrac{\lambda_n+\delta_n}{n}, -\tfrac{\lambda_n}{n}] \mid \mathbf{X}) \\ &= -\tfrac{\delta_n}{n}(1 - \Phi(\theta_n^{(1)})) + \tfrac{\delta_n}{n}\Phi(-\theta_n^{(2)}) \\ &\quad - (\beta_i^\star - \tfrac{\lambda_n}{n})(\Phi(\theta_n^{(1)}) - \Phi(\alpha_n^{(1)})) + \tilde\tau_n(\varphi(\theta_n^{(1)}) - \varphi(\alpha_n^{(1)})) \\ &\quad - (\beta_i^\star + \tfrac{\lambda_n}{n})(\Phi(-\alpha_n^{(2)}) - \Phi(-\theta_n^{(2)})) + \tilde\tau_n(\varphi(-\alpha_n^{(2)}) - \varphi(-\theta_n^{(2)})) \\ &= -\tfrac{\delta_n}{n}(1 - \Phi(\theta_n^{(1)})) + \tfrac{\delta_n}{n}\Phi(-\theta_n^{(2)}) \\ &\quad - (\beta_i^\star - \tfrac{\lambda_n}{n})(\theta_n^{(1)} - \alpha_n^{(1)})\Phi'(c_n^{(1)}) + \tilde\tau_n(\theta_n^{(1)} - \alpha_n^{(1)})\varphi'(d_n^{(1)}) \\ &\quad - (\beta_i^\star + \tfrac{\lambda_n}{n})(\theta_n^{(2)} - \alpha_n^{(2)})\Phi'(-c_n^{(2)}) + \tilde\tau_n(\theta_n^{(2)} - \alpha_n^{(2)})\varphi'(-d_n^{(2)})\end{aligned}$$

where $c_n^{(1)}, d_n^{(1)} \in [\alpha_n^{(1)}, \theta_n^{(1)}]$ and $c_n^{(2)}, d_n^{(2)} \in [\alpha_n^{(2)}, \theta_n^{(2)}]$ using first-order Taylor expansions.

We have $\theta_n^{(1)} - \alpha_n^{(1)} = \theta_n^{(2)} - \alpha_n^{(2)} = \frac{1}{\tilde\tau_n}\frac{\delta_n}{n}$, $\Phi' = \varphi$ and $\varphi'(x) = -x\varphi(x)$, thus

$$\begin{aligned}\mathbb{E}[\hat\beta_{\lambda_n+\delta_n,i} - \hat\beta_{\lambda_n,i} \mid \mathbf{X}] &= -\tfrac{\delta_n}{n}(1 - \Phi(\theta_n^{(1)})) + \tfrac{\delta_n}{n}\Phi(-\theta_n^{(2)}) \\ &\quad - (\beta_i^\star - \tfrac{\lambda_n}{n})\tfrac{1}{\tilde\tau_n}\tfrac{\delta_n}{n}\varphi(c_n^{(1)}) - \tilde\tau_n\tfrac{1}{\tilde\tau_n}\tfrac{\delta_n}{n} d_n^{(1)}\varphi(d_n^{(1)}) \\ &\quad - (\beta_i^\star + \tfrac{\lambda_n}{n})\tfrac{1}{\tilde\tau_n}\tfrac{\delta_n}{n}\varphi(-c_n^{(2)}) - \tilde\tau_n\tfrac{1}{\tilde\tau_n}\tfrac{\delta_n}{n}(-d_n^{(2)}\varphi(-d_n^{(2)})) \\ &= -\tfrac{\delta_n}{n}(1 - \Phi(\theta_n^{(1)})) + \tfrac{\delta_n}{n}\Phi(-\theta_n^{(2)}) \\ &\quad - (\beta_i^\star - \tfrac{\lambda_n}{n})\tfrac{1}{\tilde\tau_n}\tfrac{\delta_n}{n}\varphi(c_n^{(1)}) - \tfrac{\delta_n}{n} d_n^{(1)}\varphi(d_n^{(1)}) \\ &\quad - (\beta_i^\star + \tfrac{\lambda_n}{n})\tfrac{1}{\tilde\tau_n}\tfrac{\delta_n}{n}\varphi(-c_n^{(2)}) - \tfrac{\delta_n}{n}(-d_n^{(2)}\varphi(-d_n^{(2)})) \\ &= -\tfrac{\delta_n}{n}(1 - \Phi(\theta_n^{(1)})) + \tfrac{\delta_n}{n}\Phi(-\theta_n^{(2)}) \\ &\quad + \tfrac{\delta_n}{n}\alpha_n^{(1)}\varphi(c_n^{(1)}) - \tfrac{\delta_n}{n} d_n^{(1)}\varphi(d_n^{(1)}) \\ &\quad - \tfrac{\delta_n}{n}\alpha_n^{(2)}\varphi(-c_n^{(2)}) - \tfrac{\delta_n}{n}(-d_n^{(2)}\varphi(-d_n^{(2)})).\end{aligned}$$

We first consider $\beta_i^\star > 0$.

Since $\lambda_n = o(n)$ and $\delta_n = o(n)$, for n large enough, $\frac{\lambda_n+\delta_n}{n} < \beta_i^\star$, so $\alpha_n^{(1)} \leq \theta_n^{(1)} < 0$, thus for $c_n^{(1)} \in [\alpha_n^{(1)}, \theta_n^{(1)}]$, we have $|\alpha_n^{(1)}\varphi(c_n^{(1)})| \leq |\alpha_n^{(1)}|\varphi(\theta_n^{(1)}) = |\frac{\alpha_n^{(1)}}{\theta_n^{(1)}}||\theta_n^{(1)}|\varphi(\theta_n^{(1)})$, where the ratio $\frac{\alpha_n^{(1)}}{\theta_n^{(1)}} = \frac{\frac{\lambda_n}{n} - \beta_i^\star}{\frac{\lambda_n+\delta_n}{n} - \beta_i^\star}$ is deterministic and goes to 1.

As $-\theta_n^{(2)} \leq -\alpha_n^{(2)} < 0$, for $c_n^{(1)} \in [-\theta_n^{(2)}, -\alpha_n^{(2)}]$, we have $|-\alpha_n^{(2)}\varphi(-c_n^{(2)})| \leq |-\alpha_n^{(2)}\varphi(-\alpha_n^{(2)})|$.

Since $\frac{\mathbf{X}^\top\mathbf{X}}{n} \xrightarrow{\text{a.s.}} \mathbb{E}[X_0 X_0^\top]$ (strong law of large numbers), $\lambda_n = o(n)$ and $\delta_n = o(n)$, we have $\tilde\tau_n \xrightarrow{\text{a.s.}} 0^+$, and using the continuous mapping theorem, $\alpha_n^{(1)} \xrightarrow{\text{a.s.}} -\infty$, $\theta_n^{(1)} \xrightarrow{\text{a.s.}} -\infty$, $\alpha_n^{(2)} \xrightarrow{\text{a.s.}} +\infty$ and $\theta_n^{(2)} \xrightarrow{\text{a.s.}} +\infty$. We then also have $d_n^{(1)} \xrightarrow{\text{a.s.}} -\infty$ and $d_n^{(2)} \xrightarrow{\text{a.s.}} +\infty$.

Φ and $x \mapsto x\varphi(x)$ are continuous bounded functions so we get L^1 convergence of $\Phi(\theta_n^{(1)})$, $\Phi(-\theta_n^{(2)})$, $\theta_n^{(1)}\varphi(\theta_n^{(1)})$, $-\alpha_n^{(2)}\varphi(-\alpha_n^{(2)})$, $d_n^{(1)}\varphi(d_n^{(1)})$ and $-d_n^{(2)}\varphi(-d_n^{(2)})$ to 0. By putting everything together, we obtain

$$\frac{n}{\delta_n}\mathbb{E}[\hat{\beta}_{\lambda_n+\delta_n,i} - \hat{\beta}_{\lambda_n,i}] = \frac{n}{\delta_n}\mathbb{E}[\mathbb{E}[\hat{\beta}_{\lambda_n+\delta_n,i} - \hat{\beta}_{\lambda_n,i} \mid \mathbf{X}]] \to -1 = -\text{sign}(\beta_i^\star).$$

When $\beta_i^\star < 0$, we show in a similar manner that

$$\frac{n}{\delta_n}\mathbb{E}[\hat{\beta}_{\lambda_n+\delta_n,i} - \hat{\beta}_{\lambda_n,i}] \to 1 = -\text{sign}(\beta_i^\star).$$

If $\beta_i^\star = 0$, $\alpha_n^{(1)} = \alpha_n^{(2)}$ and $\theta_n^{(1)} = \theta_n^{(2)}$ so $1 - \Phi(\alpha_n^{(1)}) = \Phi(-\alpha_n^{(2)})$, $\varphi(\alpha_n^{(1)}) = \varphi(-\alpha_n^{(2)})$, $1 - \Phi(\theta_n^{(1)}) = \Phi(-\theta_n^{(2)})$and $\varphi(\theta_n^{(1)}) = \varphi(-\theta_n^{(2)})$ which leads to

$$\begin{aligned}
&\mathbb{E}[\hat{\beta}_{\lambda_n+\delta_n,i} - \hat{\beta}_{\lambda_n,i} \mid \mathbf{X}] \\
&= -\tfrac{\delta_n}{n}(1 - \Phi(\theta_n^{(1)})) + \tfrac{\delta_n}{n}\Phi(-\theta_n^{(2)}) \\
&\quad - (\beta_i^\star - \tfrac{\lambda_n}{n})(\Phi(\theta_n^{(1)}) - \Phi(\alpha_n^{(1)})) + \tilde{\tau}_n(\varphi(\theta_n^{(1)}) - \varphi(\alpha_n^{(1)})) \\
&\quad - (\beta_i^\star + \tfrac{\lambda_n}{n})(\Phi(-\alpha_n^{(2)}) - \Phi(-\theta_n^{(2)})) + \tilde{\tau}_n(\varphi(-\alpha_n^{(2)}) - \varphi(-\theta_n^{(2)})) \\
&= 0
\end{aligned}$$

and thus $\mathbb{E}[\hat{\beta}_{\lambda_n+\delta_n,i} - \hat{\beta}_{\lambda_n,i}] = 0 = \text{sign}(\beta_i^\star)$.

Thus, we have convergence component-wise and can conclude $\frac{n}{\delta_n}\mathbb{E}[\hat{\beta}_{\lambda_n+\delta_n} - \hat{\beta}_{\lambda_n}] \to -\text{sign}(\beta^\star)$.

We now show that $\frac{n}{\delta_n}\mathbb{E}[\hat{\beta}_{\lambda_n+\delta_n,i}\hat{\beta}_{\lambda_n+\delta_n,j} - \hat{\beta}_{\lambda_n,i}\hat{\beta}_{\lambda_n,j}] \to -(\text{sign}(\beta_i^\star)\beta_j^\star + \beta_i^\star\text{sign}(\beta_j^\star))$.

Note that

$$\begin{aligned}
&\mathbb{E}[\tfrac{n}{\delta_n}(\hat{\beta}_{\lambda_n+\delta_n,i}\hat{\beta}_{\lambda_n+\delta_n,j} - \hat{\beta}_{\lambda_n,i}\hat{\beta}_{\lambda_n,j}) + \text{sign}(\beta_i^\star)\beta_j^\star + \beta_i^\star\text{sign}(\beta_j^\star)] \\
&= \mathbb{E}[\tfrac{n}{\delta_n}(\hat{\beta}_{\lambda_n+\delta_n,i} - \hat{\beta}_{\lambda_n,i})\hat{\beta}_{\lambda_n+\delta_n,j} + \text{sign}(\beta_i^\star)\beta_j^\star] + \mathbb{E}[\hat{\beta}_{\lambda_n,i}\,\tfrac{n}{\delta_n}(\hat{\beta}_{\lambda_n+\delta_n,j} - \hat{\beta}_{\lambda_n,j}) + \beta_i^\star\text{sign}(\beta_j^\star)]
\end{aligned}$$

with

$$\begin{aligned}
&\mathbb{E}[\tfrac{n}{\delta_n}(\hat{\beta}_{\lambda_n+\delta_n,i} - \hat{\beta}_{\lambda_n,i})\hat{\beta}_{\lambda_n+\delta_n,j} + \text{sign}(\beta_i^\star)\beta_j^\star] \\
&= \mathbb{E}[(\tfrac{n}{\delta_n}(\hat{\beta}_{\lambda_n+\delta_n,i} - \hat{\beta}_{\lambda_n,i}) + \text{sign}(\beta_i^\star))(\hat{\beta}_{\lambda_n+\delta_n,j} - \beta_j^\star)] \\
&\quad + \beta_j^\star\,\mathbb{E}[\tfrac{n}{\delta_n}(\hat{\beta}_{\lambda_n+\delta_n,i} - \hat{\beta}_{\lambda_n,i}) + \text{sign}(\beta_i^\star)] - \text{sign}(\beta_i^\star)\mathbb{E}[\hat{\beta}_{\lambda_n+\delta_n,j} - \beta_j^\star]
\end{aligned}$$

where, using Cauchy–Schwarz,

$$\begin{aligned}
&\mathbb{E}[(\tfrac{n}{\delta_n}(\hat{\beta}_{\lambda_n+\delta_n,i} - \hat{\beta}_{\lambda_n,i}) + \text{sign}(\beta_i^\star))(\hat{\beta}_{\lambda_n+\delta_n,j} - \beta_j^\star)] \\
&\le \sqrt{\mathbb{E}[(\tfrac{n}{\delta_n}(\hat{\beta}_{\lambda_n+\delta_n,i} - \hat{\beta}_{\lambda_n,i}) + \text{sign}(\beta_i^\star))^2]\mathbb{E}[(\hat{\beta}_{\lambda_n+\delta_n,j} - \beta_j^\star)^2]}.
\end{aligned}$$

We can do the same with $\mathbb{E}[\hat{\beta}_{\lambda_n,i}\,\frac{n}{\delta_n}(\hat{\beta}_{\lambda_n+\delta_n,j} - \hat{\beta}_{\lambda_n,j}) + \beta_i^\star\text{sign}(\beta_j^\star)]$.

Therefore, proving $\mathbb{E}[\frac{n}{\delta_n}(\hat{\beta}_{\lambda_n+\delta_n,i}\hat{\beta}_{\lambda_n+\delta_n,j} - \hat{\beta}_{\lambda_n,i}\hat{\beta}_{\lambda_n,j}) + \text{sign}(\beta_i^\star)\beta_j^\star + \beta_i^\star\text{sign}(\beta_j^\star)] \to 0$ for all i, j comes down to proving $\mathbb{E}[(\frac{n}{\delta_n}(\hat{\beta}_{\lambda_n+\delta_n,i} - \hat{\beta}_{\lambda_n,i}) + \text{sign}(\beta_i^\star))^2] = O(1)$ for all i given that we have already shown for all i, accounting for the fact that both λ_n and δ_n are $o(n)$,

- $\mathbb{E}[\hat{\beta}_{\lambda_n,i}] \to \beta_i^\star$ and $\mathbb{E}[\hat{\beta}_{\lambda_n+\delta_n,i}] \to \beta_i^\star$,
- $\mathbb{E}[(\hat{\beta}_{\lambda_n,i} - \beta_i^\star)^2] \to 0$ and $\mathbb{E}[(\hat{\beta}_{\lambda_n+\delta_n,i} - \beta_i^\star)^2] \to 0$,
- $\frac{n}{\delta_n}\mathbb{E}[\hat{\beta}_{\lambda_n+\delta_n,i} - \hat{\beta}_{\lambda_n,i}] \to -\text{sign}(\beta_i^\star)$.

The first two bullet points were proved in Appendix C and the third one earlier in this proof.

As a reminder, we have

$$\hat\beta_{\lambda_n+\delta_n,i} - \hat\beta_{\lambda_n,i} = -\text{sign}(\hat\beta_{\text{OLS},i})\begin{cases}\frac{\delta_n}{n} & \text{if } |\hat\beta_{\text{OLS},i}| > \frac{\lambda_n+\delta_n}{n}\\ |\hat\beta_{\text{OLS},i}| - \frac{\lambda_n}{n} & \text{if } |\hat\beta_{\text{OLS},i}| \in [\frac{\lambda_n}{n}, \frac{\lambda_n+\delta_n}{n}]\\ 0 & \text{if } |\hat\beta_{\text{OLS},i}| < \frac{\lambda_n}{n}\end{cases}$$

thus

$$(\hat\beta_{\lambda_n+\delta_n,i} - \hat\beta_{\lambda_n,i})^2 \le \tfrac{\delta_n^2}{n^2}$$

and

$$(\tfrac{n}{\delta_n}(\hat\beta_{\lambda_n+\delta_n,i} - \hat\beta_{\lambda_n,i}) + \text{sign}(\beta^\star_i))^2 \le 2(\tfrac{n^2}{\delta_n^2}(\hat\beta_{\lambda_n+\delta_n,i} - \hat\beta_{\lambda_n,i})^2 + \text{sign}(\beta^\star_i)^2) \le 4.$$

Hence, $\mathbb{E}[(\frac{n}{\delta_n}(\hat\beta_{\lambda_n+\delta_n,i} - \hat\beta_{\lambda_n,i}) + \text{sign}(\beta^\star_i))^2] = O(1)$.

Therefore, we get

$$\frac{n}{\delta_n}\mathbb{E}[\hat\beta_{\lambda_n+\delta_n}\hat\beta_{\lambda_n+\delta_n}^\top - \hat\beta_{\lambda_n}\hat\beta_{\lambda_n}^\top] \to -(\text{sign}(\beta^\star)\beta^{\star\top} + \beta^\star\text{sign}(\beta^\star)^\top).$$

The following lemma provides key equations in the comparison setting.

Lemma F.1 (Useful equations for comparison of two linear predictors)**.** *When defining $h_n(Z_0, \mathbf{Z}) = (Y_0 - X_0^\top\hat\beta^{(1)})^2 - (Y_0 - X_0^\top\hat\beta^{(2)})^2$, we have:*

$$\begin{aligned}
h_n(Z_0,\mathbf{Z}) &= 2Y_0X_0^\top(\hat\beta^{(2)} - \hat\beta^{(1)}) + \text{tr}(X_0X_0^\top(\hat\beta^{(1)}\hat\beta^{(1)\top} - \hat\beta^{(2)}\hat\beta^{(2)\top}))\\
\mathbb{E}[h_n(Z_0,\mathbf{Z}) \mid Z_0] &= 2Y_0X_0^\top\mathbb{E}[\hat\beta^{(2)} - \hat\beta^{(1)}] + \text{tr}(X_0X_0^\top\mathbb{E}[\hat\beta^{(1)}\hat\beta^{(1)\top} - \hat\beta^{(2)}\hat\beta^{(2)\top}])\\
\mathbb{E}[h_n(Z_0,\mathbf{Z}) \mid \mathbf{Z}] &= 2\beta^{\star\top}\mathbb{E}[X_0X_0^\top](\hat\beta^{(2)} - \hat\beta^{(1)}) + \text{tr}(\mathbb{E}[X_0X_0^\top](\hat\beta^{(1)}\hat\beta^{(1)\top} - \hat\beta^{(2)}\hat\beta^{(2)\top}))\\
\mathbb{E}[h_n(Z_0,\mathbf{Z})] &= 2\beta^{\star\top}\mathbb{E}[X_0X_0^\top]\mathbb{E}[\hat\beta^{(2)} - \hat\beta^{(1)}] + \text{tr}(\mathbb{E}[X_0X_0^\top]\mathbb{E}[\hat\beta^{(1)}\hat\beta^{(1)\top} - \hat\beta^{(2)}\hat\beta^{(2)\top}])\\
\sigma^2(h_n) &= \mathbb{E}[(2(Y_0X_0^\top - \beta^{\star\top}\mathbb{E}[X_0X_0^\top])\mathbb{E}[\hat\beta^{(2)} - \hat\beta^{(1)}]\\
&\qquad + \text{tr}((X_0X_0^\top - \mathbb{E}[X_0X_0^\top])\mathbb{E}[\hat\beta^{(1)}\hat\beta^{(1)\top} - \hat\beta^{(2)}\hat\beta^{(2)\top}]))^2]\\
\gamma(h_n) &= \mathbb{E}[(2(Y_0X_0^\top - \beta^{\star\top}\mathbb{E}[X_0X_0^\top])(\hat\beta^{(2)} - \hat\beta^{(1)} - (\hat\beta'^{(2)} - \hat\beta'^{(1)}))\\
&\qquad + \text{tr}((X_0X_0^\top - \mathbb{E}[X_0X_0^\top])(\hat\beta^{(1)}\hat\beta^{(1)\top} - \hat\beta^{(2)}\hat\beta^{(2)\top} - (\hat\beta'^{(1)}\hat\beta'^{(1)\top} - \hat\beta'^{(2)}\hat\beta'^{(2)\top}))))^2]
\end{aligned}$$

where $\hat\beta'^{(1)}$ and $\hat\beta'^{(2)}$ are the linear predictor counterparts of $\hat\beta^{(1)}$ and $\hat\beta^{(2)}$, but learned on a training set $\mathbf{Z}'$ that is the same as $\mathbf{Z}$ except for the first point Z_1 being replaced by an i.i.d copy Z_1'.

Proof The first equation follows from the first equation of Lemma C.1. The remaining equations are then derived from there using the same arguments as those mentioned in Lemma C.1. ☐

Starting from the expression of $\sigma^2(h_n)$ in Lemma F.1, since $\frac{n}{\delta_n}\mathbb{E}[\hat\beta_{\lambda_n+\delta_n} - \hat\beta_{\lambda_n}]$ and $\frac{n}{\delta_n}\mathbb{E}[\hat\beta_{\lambda_n+\delta_n}\hat\beta_{\lambda_n+\delta_n}^\top - \hat\beta_{\lambda_n}\hat\beta_{\lambda_n}^\top]$ are non-random, we can expand the square, use linearity of expectation, take the limits and factorize back to obtain the following convergence

$$\frac{n^2}{\delta_n^2}\sigma^2(h_n^{\text{diff}}) \to \mathbb{E}[(2(Y_0X_0^\top - \beta^{\star\top}\mathbb{E}[X_0X_0^\top])(-\text{sign}(\beta^\star)) + \text{tr}((X_0X_0^\top - \mathbb{E}[X_0X_0^\top])(\text{sign}(\beta^\star)\beta^{\star\top} + \beta^\star\text{sign}(\beta^\star)^\top)))^2]$$

where, for $Y_0 = X_0^\top\beta^\star + \varepsilon_0$ with $\mathbb{E}[X_0] = 0$ and $\text{Var}(X_0) = \mathbf{I}$,

$$\begin{aligned}
&\mathbb{E}[(2(Y_0X_0^\top - \beta^{\star\top}\mathbb{E}[X_0X_0^\top])(-\text{sign}(\beta^\star)) + \text{tr}((X_0X_0^\top - \mathbb{E}[X_0X_0^\top])(\text{sign}(\beta^\star)\beta^{\star\top} + \beta^\star\text{sign}(\beta^\star)^\top)))^2]\\
&= \mathbb{E}[(-2Y_0X_0^\top\text{sign}(\beta^\star) + 2\beta^{\star\top}\text{sign}(\beta^\star) + 2X_0^\top\beta^\star X_0^\top\text{sign}(\beta^\star) - 2\beta^{\star\top}\text{sign}(\beta^\star)))^2]\\
&= \mathbb{E}[(-2\varepsilon_0X_0^\top\text{sign}(\beta^\star))^2]\\
&= 4\mathbb{E}[\varepsilon_0^2]\mathbb{E}[(X_0^\top\text{sign}(\beta^\star))^2] \quad \text{by independence of } \varepsilon_0, X_0\\
&= 4\tau^2\|\beta^\star\|_0
\end{aligned}$$

since

$$\mathbb{E}[(X_0^\top\text{sign}(\beta^\star))^2] = \text{Var}(\text{sign}(\beta^\star)^\top X_0) = \text{sign}(\beta^\star)^\top\text{Var}(X_0)\text{sign}(\beta^\star) = \text{sign}(\beta^\star)^\top\text{sign}(\beta^\star) = \|\beta^\star\|_0.$$

We can then conclude that $\frac{n^2}{\delta_n^2}\,\sigma^2(h_n^{\text{diff}}) \to 4\tau^2\|\beta^\star\|_0$.

G Proof of Proposition E.2: Lower-bounding rate of $\gamma(h_n^{\text{diff}})$ for comparison of ST(λ_n) with ST$(\lambda_n + \delta_n)$

Starting from the expression for $\gamma(h_n)$ stated in Lemma F.1, we have

$$\gamma(h_n^{\text{diff}}) = \mathbb{E}[(2(Y_0 X_0^\top - \beta^{\star\top}\mathbb{E}[X_0 X_0^\top])\nu_n + \text{tr}((X_0 X_0^\top - \mathbb{E}[X_0 X_0^\top])\Psi_n))^2].$$

where

- $\nu_n \triangleq \hat{\beta}_{\lambda_n+\delta_n} - \hat{\beta}_{\lambda_n} - (\hat{\beta}'_{\lambda_n+\delta_n} - \hat{\beta}'_{\lambda_n})$,
- $\Psi_n \triangleq \hat{\beta}_{\lambda_n}\hat{\beta}_{\lambda_n}^\top - \hat{\beta}_{\lambda_n+\delta_n}\hat{\beta}_{\lambda_n+\delta_n}^\top - (\hat{\beta}'_{\lambda_n}\hat{\beta}'^\top_{\lambda_n} - \hat{\beta}'_{\lambda_n+\delta_n}\hat{\beta}'^\top_{\lambda_n+\delta_n})$.

$\mathbb{E}[X_0 X_0^\top] = \mathbf{I}$ since the features are drawn from $\mathcal{N}(0, \mathbf{I})$, and using independence of Z_0 from the training points, we have

$$\begin{aligned}
\gamma(h_n^{\text{diff}}) &= \mathbb{E}[(2\textstyle\sum_i (Y_0 X_{0,i} - \beta_i^\star)\nu_{n,i} + \sum_{i,j}(X_{0,i}X_{0,j} - \mathbb{1}[i=j])\Psi_{n,i,j}))^2] \\
&= 4\textstyle\sum_i \mathbb{E}[(Y_0 X_{0,i} - \beta_i^\star)^2]\mathbb{E}[\nu_{n,i}^2] \\
&\quad + 4\textstyle\sum_{i\neq j}\mathbb{E}[(Y_0 X_{0,i} - \beta_i^\star)(Y_0 X_{0,j} - \beta_j^\star)]\mathbb{E}[\nu_{n,i}\nu_{n,j}] \\
&\quad + 4\textstyle\sum_{i,j,k}\mathbb{E}[(Y_0 X_{0,i} - \beta_i^\star)(X_{0,j}X_{0,k} - \mathbb{1}[j=k])]\mathbb{E}[\nu_{n,i}\Psi_{n,j,k}] \\
&\quad + \textstyle\sum_{i,j,k,l}\mathbb{E}[(X_{0,i}X_{0,j} - \mathbb{1}[i=j])(X_{0,k}X_{0,l} - \mathbb{1}[k=l])]\mathbb{E}[\Psi_{n,i,j}\Psi_{n,k,l}].
\end{aligned}$$

Since $Y_0 = X_0^\top\beta^\star + \varepsilon_0 = \sum_k X_{0,k}\beta_k^\star + \varepsilon_0$ with $X_0 \sim \mathcal{N}(0, \mathbf{I})$ and $\varepsilon_0 \perp\!\!\!\perp X_0$, we have

$$\mathbb{E}[Y_0 X_{0,i}] = \beta_i^\star \mathbb{E}[X_{0,i}^2] + \textstyle\sum_{k\neq i}\beta_k^\star\mathbb{E}[X_{0,i}X_{0,k}] + \mathbb{E}[\varepsilon_0 X_{0,i}] = \beta_i^\star$$

and $Y_0^2 = \sum_{k,l} X_{0,k}X_{0,l}\beta_k^\star\beta_l^\star + 2\varepsilon_0\sum_k X_{0,k}\beta_k^\star + \varepsilon_0^2$, so for $i \neq j$,

$$\mathbb{E}[Y_0^2 X_{0,i}X_{0,j}] = \textstyle\sum_{k,l}\mathbb{E}[X_{0,i}X_{0,j}X_{0,k}X_{0,l}]\beta_k^\star\beta_l^\star + 2\sum_k \mathbb{E}[\varepsilon_0 X_{0,i}X_{0,j}X_{0,k}]\beta_k^\star + \mathbb{E}[\varepsilon_0^2 X_{0,i}X_{0,j}] = 2\beta_i^\star\beta_j^\star$$

since the expectation in the first sum is equal to 1 when $k = i, l = j$ or $k = j, l = i$, and equal to 0 otherwise, and thus, for $i \neq j$,

$$\mathbb{E}[(Y_0 X_{0,i} - \beta_i^\star)(Y_0 X_{0,j} - \beta_j^\star)] = \mathbb{E}[Y_0^2 X_{0,i}X_{0,j}] - \beta_i^\star\mathbb{E}[Y_0 X_{0,j}] - \beta_j^\star\mathbb{E}[Y_0 X_{0,i}] + \beta_i^\star\beta_j^\star = \beta_i^\star\beta_j^\star.$$

For the case $i = j$,

$$\begin{aligned}
\mathbb{E}[Y_0^2 X_{0,i}^2] &= \textstyle\sum_{k,l}\mathbb{E}[X_{0,i}^2 X_{0,k}X_{0,l}]\beta_k^\star\beta_l^\star + 2\sum_k \mathbb{E}[\varepsilon_0 X_{0,i}^2 X_{0,k}]\beta_k^\star + \mathbb{E}[\varepsilon_0^2 X_{0,i}^2] \\
&= \mathbb{E}[X_{0,i}^4]\beta_i^{\star 2} + \textstyle\sum_{k\neq i}\mathbb{E}[X_{0,i}^2 X_{0,k}^2]\beta_k^{\star 2} + \tau^2 \\
&= \mathbb{E}[X_{0,i}^4]\beta_i^{\star 2} + \textstyle\sum_{k\neq i}\beta_k^{\star 2} + \tau^2
\end{aligned}$$

and then, for $\beta_i^\star = 0$,

$$\mathbb{E}[(Y_0 X_{0,i} - \beta_i^\star)^2] = \mathbb{E}[Y_0^2 X_{0,i}^2] = \textstyle\sum_{k\neq i}\beta_k^{\star 2} + \tau^2 \geq \tau^2 > 0.$$

Therefore

$$\begin{aligned}
\gamma(h_n^{\text{diff}}) &= 4\textstyle\sum_{i,\beta_i^\star=0}\mathbb{E}[Y_0^2 X_{0,i}^2]\mathbb{E}[\nu_{n,i}^2] \\
&\quad + 4\textstyle\sum_{i,\beta_i^\star\neq 0}\mathbb{E}[(Y_0 X_{0,i} - \beta_i^\star)^2]\mathbb{E}[\nu_{n,i}^2] \\
&\quad + 4\textstyle\sum_{i\neq j,\beta_i^\star\neq 0,\beta_j^\star\neq 0}\beta_i^\star\beta_j^\star\mathbb{E}[\nu_{n,i}\nu_{n,j}] \\
&\quad + 4\textstyle\sum_{i,j,k}\mathbb{E}[(Y_0 X_{0,i} - \beta_i^\star)(X_{0,j}X_{0,k} - \mathbb{1}[j=k])]\mathbb{E}[\nu_{n,i}\Psi_{n,j,k}] \\
&\quad + \textstyle\sum_{i,j,k,l}\mathbb{E}[(X_{0,i}X_{0,j} - \mathbb{1}[i=j])(X_{0,k}X_{0,l} - \mathbb{1}[k=l])]\mathbb{E}[\Psi_{n,i,j}\Psi_{n,k,l}].
\end{aligned}$$

where importantly we were able to remove the i, j terms in the third sum when $\beta_i^\star = 0$ or $\beta_j^\star = 0$.

We will now prove the following results:

- $\mathbb{E}[\nu_{n,i}^2] = O(\frac{\delta_n^2}{n^2})$ for all i,

- $\mathbb{E}[\nu_{n,i}^2] = \Omega(\frac{\delta_n^2}{n^2\sqrt{n}})$ for all i such that $\beta_i^\star = 0$,
- $\mathbb{E}[\nu_{n,i}^2] = o(\frac{\delta_n^2}{n^2\sqrt{n}})$ for all i such that $\beta_i^\star \neq 0$,
- $\mathbb{E}[\Psi_{n,i,j}^2] = O(\frac{\delta_n^2}{n^4})$ for all i, j.

Once we prove these, Cauchy–Schwarz will yield the following upper-bounding rates for terms appearing in the expression of $\gamma(h_n^{\text{diff}})$:

- for i, j such that $\beta_i^\star \neq 0$ and $\beta_j^\star \neq 0$, $|\mathbb{E}[\nu_{n,i}\nu_{n,j}]| \leq \sqrt{\mathbb{E}[\nu_{n,i}^2]\mathbb{E}[\nu_{n,j}^2]} = o(\frac{\delta_n^2}{n^2\sqrt{n}})$,
- $|\mathbb{E}[\nu_{n,i}\Psi_{n,j,k}]| \leq \sqrt{\mathbb{E}[\nu_{n,i}^2]\mathbb{E}[\Psi_{n,j,k}^2]} = O(\sqrt{\frac{\delta_n^2}{n^2}\frac{\delta_n^2}{n^4}}) = O(\frac{\delta_n^2}{n^3}) = o(\frac{\delta_n^2}{n^2\sqrt{n}})$,
- $|\mathbb{E}[\Psi_{n,i,j}\Psi_{n,k,l}]| \leq \sqrt{\mathbb{E}[\Psi_{n,i,j}^2]\mathbb{E}[\Psi_{n,k,l}^2]} = O(\sqrt{\frac{\delta_n^2}{n^4}\frac{\delta_n^2}{n^4}}) = O(\frac{\delta_n^2}{n^4}) = o(\frac{\delta_n^2}{n^2\sqrt{n}})$,

and it will therefore be clear that $\gamma(h_n^{\text{diff}}) = \Omega(\frac{\delta_n^2}{n^2\sqrt{n}})$ as the terms of leading order in $\gamma(h_n^{\text{diff}})$ will be the $\mathbb{E}[\nu_{n,i}^2]$ terms for i such that $\beta_i^\star = 0$.

We will now prove the first result $\mathbb{E}[\nu_{n,i}^2] = O(\frac{\delta_n^2}{n^2})$ for all i.

We have

$$\begin{aligned}
\nu_{n,i} &= \hat{\beta}_{\lambda_n+\delta_n,i} - \hat{\beta}_{\lambda_n,i} - (\hat{\beta}'_{\lambda_n+\delta_n,i} - \hat{\beta}'_{\lambda_n,i}) \\
&= \operatorname{sign}(\hat{\beta}_{\text{OLS},i})(|\hat{\beta}_{\text{OLS},i}| - \tfrac{\lambda_n+\delta_n}{n})_+ - \operatorname{sign}(\hat{\beta}_{\text{OLS},i})(|\hat{\beta}_{\text{OLS},i}| - \tfrac{\lambda_n}{n})_+ \\
&\quad - (\operatorname{sign}(\hat{\beta}'_{\text{OLS},i})(|\hat{\beta}'_{\text{OLS},i}| - \tfrac{\lambda_n+\delta_n}{n})_+ - \operatorname{sign}(\hat{\beta}'_{\text{OLS},i})(|\hat{\beta}'_{\text{OLS},i}| - \tfrac{\lambda_n}{n})_+) \\
&= \operatorname{sign}(\hat{\beta}_{\text{OLS},i})\begin{cases} -\frac{\delta_n}{n} & \text{if } |\hat{\beta}_{\text{OLS},i}| > \frac{\lambda_n+\delta_n}{n} \\ \frac{\lambda_n}{n} - |\hat{\beta}_{\text{OLS},i}| & \text{if } |\hat{\beta}_{\text{OLS},i}| \in [\frac{\lambda_n}{n}, \frac{\lambda_n+\delta_n}{n}] \\ 0 & \text{if } |\hat{\beta}_{\text{OLS},i}| < \frac{\lambda_n}{n} \end{cases} \\
&\qquad - \operatorname{sign}(\hat{\beta}'_{\text{OLS},i})\begin{cases} -\frac{\delta_n}{n} & \text{if } |\hat{\beta}'_{\text{OLS},i}| > \frac{\lambda_n+\delta_n}{n} \\ \frac{\lambda_n}{n} - |\hat{\beta}'_{\text{OLS},i}| & \text{if } |\hat{\beta}'_{\text{OLS},i}| \in [\frac{\lambda_n}{n}, \frac{\lambda_n+\delta_n}{n}] \\ 0 & \text{if } |\hat{\beta}'_{\text{OLS},i}| < \frac{\lambda_n}{n} \end{cases}.
\end{aligned}$$

We can observe that both $|\hat{\beta}_{\lambda_n+\delta_n,i} - \hat{\beta}_{\lambda_n,i}|$ and $|\hat{\beta}'_{\lambda_n+\delta_n,i} - \hat{\beta}'_{\lambda_n,i}|$ are upper-bounded by $\frac{\delta_n}{n}$ and thus $\nu_{n,i}^2 \leq 4\frac{\delta_n^2}{n^2}$, which implies $\mathbb{E}[\nu_{n,i}^2] = O(\frac{\delta_n^2}{n^2})$ for all i.

We will now prove the second result $\mathbb{E}[\nu_{n,i}^2] = \Omega(\frac{\delta_n^2}{n^2\sqrt{n}})$ for all i such that $\beta_i^\star = 0$.

Based on the previous expression, we can further detail $\nu_{n,i}$ as follows

$$\nu_{n,i} = \begin{cases} -\frac{\delta_n}{n} & \text{if } \hat{\beta}_{\text{OLS},i} > \frac{\lambda_n+\delta_n}{n} \\ \frac{\delta_n}{n} & \text{if } \hat{\beta}_{\text{OLS},i} < -\frac{\lambda_n+\delta_n}{n} \\ \frac{\lambda_n}{n} - \hat{\beta}_{\text{OLS},i} & \text{if } \hat{\beta}_{\text{OLS},i} \in [\frac{\lambda_n}{n}, \frac{\lambda_n+\delta_n}{n}] \\ -\frac{\lambda_n}{n} - \hat{\beta}_{\text{OLS},i} & \text{if } \hat{\beta}_{\text{OLS},i} \in [-\frac{\lambda_n+\delta_n}{n}, -\frac{\lambda_n}{n}] \\ 0 & \text{if } |\hat{\beta}_{\text{OLS},i}| < \frac{\lambda_n}{n} \end{cases} - \begin{cases} -\frac{\delta_n}{n} & \text{if } \hat{\beta}'_{\text{OLS},i} > \frac{\lambda_n+\delta_n}{n} \\ \frac{\delta_n}{n} & \text{if } \hat{\beta}'_{\text{OLS},i} < -\frac{\lambda_n+\delta_n}{n} \\ \frac{\lambda_n}{n} - \hat{\beta}'_{\text{OLS},i} & \text{if } \hat{\beta}'_{\text{OLS},i} \in [\frac{\lambda_n}{n}, \frac{\lambda_n+\delta_n}{n}] \\ -\frac{\lambda_n}{n} - \hat{\beta}'_{\text{OLS},i} & \text{if } \hat{\beta}'_{\text{OLS},i} \in [-\frac{\lambda_n+\delta_n}{n}, -\frac{\lambda_n}{n}] \\ 0 & \text{if } |\hat{\beta}'_{\text{OLS},i}| < \frac{\lambda_n}{n} \end{cases}$$

which means there are 25 possible cases that form a partition and we can write $\nu_{n,i}$ as the sum of 25 terms that are of the form: an indicator of one of the 25 events multiplied by the value of $\nu_{n,i}$ for this event. We can then similarly write $\nu_{n,i}^2$ as the sum of 25 terms that are of the form: an indicator of one of the 25 events multiplied by the value of $\nu_{n,i}^2$ for this event.

We can then lower-bound $\mathbb{E}[\nu_{n,i}^2]$ by the expectation of any one of the 25 terms since they are all non-negative. In particular, we can do it using the term coming from the combination of the first case on the left side and the last case on the right side

$$\begin{aligned}
\mathbb{E}[\nu_{n,i}^2] &\geq \mathbb{E}[\tfrac{\delta_n^2}{n^2}\mathbb{1}\left[\hat{\beta}_{\text{OLS},i} > \tfrac{\lambda_n+\delta_n}{n}, |\hat{\beta}'_{\text{OLS},i}| < \tfrac{\lambda_n}{n}\right]] \\
&= \tfrac{\delta_n^2}{n^2}\mathbb{P}(\hat{\beta}_{\text{OLS},i} > \tfrac{\lambda_n+\delta_n}{n}, |\hat{\beta}'_{\text{OLS},i}| < \tfrac{\lambda_n}{n}).
\end{aligned}$$

Since $\lambda_n = \omega(1)$ and $\delta_n = \Theta(1)$, $\frac{\lambda_n}{n} - \frac{\delta_n}{n} > 0$ for n large enough, and we then have $\{\hat\beta_{\text{OLS},i} > \hat\beta'_{\text{OLS},i} + 2\frac{\delta_n}{n}, \hat\beta'_{\text{OLS},i} \in [\frac{\lambda_n}{n} - \frac{\delta_n}{n}, \frac{\lambda_n}{n}]\} \subseteq \{\hat\beta_{\text{OLS},i} > \frac{\lambda_n+\delta_n}{n}, |\hat\beta'_{\text{OLS},i}| < \frac{\lambda_n}{n}\}$, therefore

$$\begin{aligned}
&\mathbb{P}(\hat\beta_{\text{OLS},i} > \tfrac{\lambda_n+\delta_n}{n}, |\hat\beta'_{\text{OLS},i}| < \tfrac{\lambda_n}{n}) \\
&\geq \mathbb{P}(\hat\beta_{\text{OLS},i} > \hat\beta'_{\text{OLS},i} + 2\tfrac{\delta_n}{n}, \hat\beta'_{\text{OLS},i} \in [\tfrac{\lambda_n}{n} - \tfrac{\delta_n}{n}, \tfrac{\lambda_n}{n}]) \\
&= \mathbb{P}(n(\hat\beta'_{\text{OLS},i} - \hat\beta_{\text{OLS},i}) < -2\delta_n, \hat\beta'_{\text{OLS},i} \in [\tfrac{\lambda_n}{n} - \tfrac{\delta_n}{n}, \tfrac{\lambda_n}{n}]).
\end{aligned}$$

We have

$$\begin{aligned}
\mathrm{Cov}(\hat\beta_{\text{OLS},i}, \hat\beta'_{\text{OLS},i} \mid \mathbf{X}, \mathbf{X}') &= \mathrm{Cov}(\beta^\star + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\varepsilon, \beta^\star + (\mathbf{X}'^\top\mathbf{X}')^{-1}\mathbf{X}'^\top\varepsilon' \mid \mathbf{X}, \mathbf{X}') \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \mathrm{Cov}(\varepsilon, \varepsilon')\mathbf{X}'(\mathbf{X}'^\top\mathbf{X}')^{-1} \\
&= \tau^2(\mathbf{X}^\top\mathbf{X})^{-1}\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}(\mathbf{X}'^\top\mathbf{X}')^{-1}
\end{aligned}$$

where $\tilde{\mathbf{X}} \triangleq (X_2, \ldots, X_n)^\top$ is the matrix of regressors for the training points except for the first one that is being changed, since $\mathrm{Cov}(\varepsilon_i, \varepsilon'_j)$ is equal to τ^2 if $i = j \geq 2$ and 0 otherwise. Then

$$\begin{aligned}
\mathrm{Cov}(\hat\beta'_{\text{OLS},i} - \hat\beta_{\text{OLS},i}, \hat\beta'_{\text{OLS},i} \mid \mathbf{X}, \mathbf{X}') &= \tau^2(\mathbf{X}'^\top\mathbf{X}')^{-1} - \tau^2(\mathbf{X}^\top\mathbf{X})^{-1}\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}(\mathbf{X}'^\top\mathbf{X}')^{-1} \\
&= \tau^2(\mathbf{I} - (\mathbf{X}^\top\mathbf{X})^{-1}\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})(\mathbf{X}'^\top\mathbf{X}')^{-1}.
\end{aligned}$$

Hence, the bivariate normal vector $(\hat\beta'_{\text{OLS},i} - \hat\beta_{\text{OLS},i}, \hat\beta'_{\text{OLS},i})$ has uncorrelated components in the limit, with zero correlation being equivalent to independence for multivariate normal vectors. Since $n(\hat\beta'_{\text{OLS}} - \hat\beta_{\text{OLS}}) \xrightarrow{\text{a.s.}} V \triangleq (Y'_1 - X_1'^\top\beta^\star)X'_1 - (Y_1 - X_1^\top\beta^\star)X_1$, proved in Appendix D, and $\delta_n = \Theta(1)$, we have

$$\mathbb{P}(n(\hat\beta'_{\text{OLS},i} - \hat\beta_{\text{OLS},i}) < -2\delta_n, \hat\beta'_{\text{OLS},i} \in [\tfrac{\lambda_n}{n} - \tfrac{\delta_n}{n}, \tfrac{\lambda_n}{n}]) = \Theta(\mathbb{P}(\hat\beta'_{\text{OLS},i} \in [\tfrac{\lambda_n}{n} - \tfrac{\delta_n}{n}, \tfrac{\lambda_n}{n}])).$$

We can then focus on the rate of $\mathbb{P}(\hat\beta'_{\text{OLS},i} \in [\frac{\lambda_n}{n} - \frac{\delta_n}{n}, \frac{\lambda_n}{n}])$.

$$\mathbb{P}(\hat\beta'_{\text{OLS},i} \in [\tfrac{\lambda_n}{n} - \tfrac{\delta_n}{n}, \tfrac{\lambda_n}{n}]) = \mathbb{E}[\mathbb{P}(\hat\beta'_{\text{OLS},i} \in [\tfrac{\lambda_n}{n} - \tfrac{\delta_n}{n}, \tfrac{\lambda_n}{n}] \mid \mathbf{X}')]$$

where, using $\beta^\star_i = 0$ and $\hat\beta'_{\text{OLS}} \mid \mathbf{X} \sim \mathcal{N}(\beta^\star, \tau^2(\mathbf{X}'^\top\mathbf{X}')^{-1})$ and defining $\tilde\tau'_n = \frac{\tau}{\sqrt{n}}\sqrt{(\frac{\mathbf{X}'^\top\mathbf{X}'}{n})^{-1}_{i,i}}$,

$$\begin{aligned}
\mathbb{P}(\hat\beta'_{\text{OLS},i} \in [\tfrac{\lambda_n}{n} - \tfrac{\delta_n}{n}, \tfrac{\lambda_n}{n}] \mid \mathbf{X}') &= \Phi(\tfrac{1}{\tilde\tau'_n}\tfrac{\lambda_n}{n}) - \Phi(\tfrac{1}{\tilde\tau'_n}(\tfrac{\lambda_n}{n} - \tfrac{\delta_n}{n})) \\
&= \tfrac{1}{\tilde\tau'_n}\tfrac{\delta_n}{n}\varphi(\tfrac{1}{\tilde\tau'_n}\tfrac{\lambda_n}{n}) - \tfrac{1}{2}\tfrac{1}{\tilde\tau'^2_n}\tfrac{\delta_n^2}{n^2}\varphi'(c_n)
\end{aligned}$$

for $c_n \in [\frac{1}{\tilde\tau'_n}(\frac{\lambda_n}{n} - \frac{\delta_n}{n}), \frac{1}{\tilde\tau'_n}\frac{\lambda_n}{n}]$ by a second-order Taylor expansion. We have

$$\tfrac{1}{\tilde\tau'_n}\tfrac{\delta_n}{n}\varphi(\tfrac{1}{\tilde\tau'_n}\tfrac{\lambda_n}{n}) = \tfrac{\delta_n}{\sqrt{n}}\tfrac{1}{\tilde\tau'_n}\tfrac{\sqrt{n}}{n}\varphi(\tfrac{1}{\tilde\tau'_n}\tfrac{\lambda_n}{n})$$

whose expectation is $\Theta(\frac{1}{\sqrt{n}})$ since $\lambda_n = O(\sqrt{n})$ and $\delta_n = \Theta(1)$ yield $\frac{\delta_n}{\sqrt{n}} = \Theta(\frac{1}{\sqrt{n}})$ and $\mathbb{E}[\frac{1}{\tilde\tau'_n}\frac{\sqrt{n}}{n}\varphi(\frac{1}{\tilde\tau'_n}\frac{\lambda_n}{n})] = \Theta(1)$.

As for the second part of the Taylor expansion, its expectation is a $o(\frac{1}{\sqrt{n}})$ since φ' is bounded, $\delta_n = \Theta(1)$ and we have

$$\mathbb{E}[\tfrac{1}{\tilde\tau'^2_n}] = \tfrac{1}{\tau^2}\mathbb{E}[((\mathbf{X}'^\top\mathbf{X}')^{-1}_{i,i})^{-1}] = n - p + 1$$

using the fact that for $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I})$, we know $(\mathbf{X}^\top\mathbf{X})^{-1} \sim W_p^{-1}(\mathbf{I}, n)$ and then the diagonal element $(\mathbf{X}^\top\mathbf{X})^{-1}_{i,i}$ follows an inverse gamma distribution with shape parameter $\frac{n-p+1}{2}$ and scale parameter $\frac{1}{2}$, and the expectation of the reciprocal of an inverse gamma distributed variable is the ratio of the shape and the scale.

We can then conclude that

$$\mathbb{P}(\hat\beta'_{\text{OLS},i} \in [\tfrac{\lambda_n}{n} - \tfrac{\delta_n}{n}, \tfrac{\lambda_n}{n}]) = \mathbb{E}[\mathbb{P}(\hat\beta'_{\text{OLS},i} \in [\tfrac{\lambda_n}{n} - \tfrac{\delta_n}{n}, \tfrac{\lambda_n}{n}] \mid \mathbf{X}')] = \Theta(\tfrac{1}{\sqrt{n}})$$

and thus $\mathbb{E}[\nu_{n,i}^2] = \Omega(\frac{\delta_n^2}{n^2\sqrt{n}})$.

We will now prove the third result $\mathbb{E}[\nu_{n,i}^2] = o(\frac{\delta_n^2}{n^2\sqrt{n}})$ for all i such that $\beta_i^\star \neq 0$.

Consider i such that $\beta_i^\star > 0$, since the combination of the first case on the left side and the first case on the right side in the expression of $\nu_{n,i}$ corresponds to a value of 0 for $\nu_{n,i}$, we can write $\nu_{n,i}^2$ as the sum of 24 terms that are of the form: an indicator of one of the 24 other events multiplied by the value of $\nu_{n,i}^2$ for this event. Since $\nu_{n,i}^2 \leq 4\frac{\delta_n^2}{n^2}$, we can upper-bound $\nu_{n,i}^2$ by $4\frac{\delta_n^2}{n^2}$ multiplied by the sum of the 24 indicators and we then need to show that all 24 indicators have an expectation which is $o(\frac{1}{\sqrt{n}})$. Using $\mathbb{E}[\mathbb{1}[A]] = \mathbb{P}(A)$, $\mathbb{P}(A \cap B) \leq \min(\mathbb{P}(A), \mathbb{P}(B))$ and the fact that $\hat{\beta}_{\text{OLS}}$ and $\hat{\beta}'_{\text{OLS}}$ have the same unconditional distribution, we can upper-bound all 24 indicator expectations by one of the following four probabilities

- $\mathbb{P}(\hat{\beta}_{\text{OLS},i} < -\frac{\lambda_n+\delta_n}{n}) = \mathbb{E}[\mathbb{P}(\hat{\beta}_{\text{OLS},i} < -\frac{\lambda_n+\delta_n}{n} \mid \mathbf{X})]$,
- $\mathbb{P}(\hat{\beta}_{\text{OLS},i} \in [\frac{\lambda_n}{n}, \frac{\lambda_n+\delta_n}{n}]) = \mathbb{E}[\mathbb{P}(\hat{\beta}_{\text{OLS},i} \in [\frac{\lambda_n}{n}, \frac{\lambda_n+\delta_n}{n}] \mid \mathbf{X})]$,
- $\mathbb{P}(\hat{\beta}_{\text{OLS},i} \in [-\frac{\lambda_n+\delta_n}{n}, -\frac{\lambda_n}{n}]) = \mathbb{E}[\mathbb{P}(\hat{\beta}_{\text{OLS},i} \in [-\frac{\lambda_n+\delta_n}{n}, -\frac{\lambda_n}{n}] \mid \mathbf{X})]$,
- $\mathbb{P}(|\hat{\beta}_{\text{OLS},i}| < \frac{\lambda_n}{n}) = \mathbb{E}[\mathbb{P}(|\hat{\beta}_{\text{OLS},i}| < \frac{\lambda_n}{n} \mid \mathbf{X})]$.

Since $\hat{\beta}_{\text{OLS}} \mid \mathbf{X} \sim \mathcal{N}(\beta^\star, \tau^2(\mathbf{X}^\top\mathbf{X})^{-1})$, we can write $\hat{\beta}_{\text{OLS},i} = \beta_i^\star + \tilde{\tau}_n Z$ where $\tilde{\tau}_n = \frac{\tau}{\sqrt{n}}\sqrt{(\frac{\mathbf{X}^\top\mathbf{X}}{n})^{-1}_{i,i}}$ and $Z \mid \mathbf{X} \sim \mathcal{N}(0,1)$. Note that we could have i as a subscript of $\tilde{\tau}_n$ and Z, but we will only consider one i at a time in our computations and we can thus omit this subscript for both of them for the sake of notational simplicity, and we will also omit it for some additional notation we define in the rest of the proof.

Define $\alpha_n^{(1)} = \frac{1}{\tilde{\tau}_n}(\frac{\lambda_n}{n} - \beta_i^\star)$, $\alpha_n^{(2)} = \frac{1}{\tilde{\tau}_n}(\frac{\lambda_n}{n} + \beta_i^\star)$, $\theta_n^{(1)} = \frac{1}{\tilde{\tau}_n}(\frac{\lambda_n+\delta_n}{n} - \beta_i^\star)$ and $\theta_n^{(2)} = \frac{1}{\tilde{\tau}_n}(\frac{\lambda_n+\delta_n}{n} + \beta_i^\star)$.

In the order they appear, the four conditional probabilities above are equal to

$$\begin{gathered}
\mathbb{P}(Z < -\theta_n^{(2)} \mid \mathbf{X}) = \Phi(-\theta_n^{(2)}), \\
\mathbb{P}(Z \in [\alpha_n^{(1)}, \theta_n^{(1)}] \mid \mathbf{X}) = \Phi(\theta_n^{(1)}) - \Phi(\alpha_n^{(1)}), \\
\mathbb{P}(Z \in [-\theta_n^{(2)}, -\alpha_n^{(2)}] \mid \mathbf{X}) = \Phi(-\alpha_n^{(2)}) - \Phi(-\theta_n^{(2)}), \\
\mathbb{P}(Z \in [-\alpha_n^{(2)}, \alpha_n^{(1)}] \mid \mathbf{X}) = \Phi(\alpha_n^{(1)}) - \Phi(-\alpha_n^{(2)}).
\end{gathered}$$

Since $\frac{\mathbf{X}^\top\mathbf{X}}{n} \xrightarrow{\text{a.s.}} \mathbb{E}[X_0 X_0^\top]$ (strong law of large numbers), $\lambda_n = o(n)$ and $\delta_n = o(n)$, we have $\tilde{\tau}_n \xrightarrow{\text{a.s.}} 0^+$, and using the continuous mapping theorem, $\alpha_n^{(1)} \xrightarrow{\text{a.s.}} -\infty$, $\theta_n^{(1)} \xrightarrow{\text{a.s.}} -\infty$, $\alpha_n^{(2)} \xrightarrow{\text{a.s.}} +\infty$ and $\theta_n^{(2)} \xrightarrow{\text{a.s.}} +\infty$ as $\beta_i^\star > 0$.

If we show that $\sqrt{n}\,\Phi(\alpha_n^{(1)})$ goes to 0 in L^1, then all other similar convergences will follow and we will get that all four unconditional probabilities listed above are $o(\frac{1}{\sqrt{n}})$ and thus $\mathbb{E}[\nu_{n,i}^2] = o(\frac{\delta_n^2}{n^2\sqrt{n}})$ for all i such that $\beta_i^\star > 0$.

We have

$$\sqrt{n}\,\Phi(\alpha_n^{(1)}) = \frac{\sqrt{n}}{\alpha_n^{(1)}} \cdot \alpha_n^{(1)}\Phi(\alpha_n^{(1)})$$

thus, by Cauchy–Schwarz,

$$\mathbb{E}[|\sqrt{n}\,\Phi(\alpha_n^{(1)})|] \leq \sqrt{\mathbb{E}\left[\frac{n}{(\alpha_n^{(1)})^2}\right]\mathbb{E}\Big[(\alpha_n^{(1)}\Phi(\alpha_n^{(1)}))^2\Big]}.$$

$\alpha_n^{(1)} \xrightarrow{\text{a.s.}} -\infty$ so $\alpha_n^{(1)}\Phi(\alpha_n^{(1)}) \xrightarrow{\text{a.s.}} 0$. This comes from the fact that $x\,\Phi(x) \to 0$ for $x \to -\infty$, as we notice that for $x < 0$, we have $0 < -x\,\Phi(x) = -x\,(1 - \Phi(-x)) = -x\int_{-x}^{+\infty}\varphi(t)dt \leq \int_{-x}^{+\infty} t\varphi(t)dt$ where this last expression goes to 0 when $x \to -\infty$.

Since $\lambda_n = o(n)$ and $\delta_n = o(n)$, for n large enough, $\frac{\lambda_n+\delta_n}{n} < \beta_i^\star$, so $\alpha_n^{(1)} \leq \theta_n^{(1)} < 0$. Since the function $x \mapsto x\,\Phi(x)$ is continuous bounded for $x < 0$, we get L^1 convergence of $(\alpha_n^{(1)}\Phi(\alpha_n^{(1)}))^2$ to 0.

Moreover, $\frac{n}{(\alpha_n^{(1)})^2} = \frac{n\tilde{\tau}_n^2}{(\frac{\lambda_n}{n}-\beta_i^\star)^2} = \frac{\tau^2}{(\frac{\lambda_n}{n}-\beta_i^\star)^2}(\frac{\mathbf{X}^\top\mathbf{X}}{n})^{-1}_{i,i}$ and it is thus sufficient to have $\mathbb{E}[(\frac{\mathbf{X}^\top\mathbf{X}}{n})^{-1}_{i,i}] = O(1)$, which is the case for features drawn i.i.d. from $\mathcal{N}(0,\mathbf{I})$ as $\mathbb{E}[(\frac{\mathbf{X}^\top\mathbf{X}}{n})^{-1}_{i,i}] = \frac{n}{n-p-1}$.

Hence, $\sqrt{n}\,\Phi(\alpha_n^{(1)})$ goes to 0 in L^1.

The proof is similar for i such that $\beta_i^\star < 0$.

Finally, we show the fourth result $\mathbb{E}[\Psi_{n,i,j}^2] = O(\frac{\delta_n^2}{n^4})$ for all i,j or equivalently $\mathbb{E}[\Psi_{n,i,j}^2] = O(\frac{\delta_n^4}{n^4})$ since $\delta_n = \Theta(1)$.

Similarly to previous computations and upper-bounding with Cauchy–Schwarz, we can upper-bound $\mathbb{E}[\Psi_{n,i,j}^2]$ using products of $\mathbb{E}[\nu_{n,i}^4]$ and the fourth moment of $\hat{\beta}_{\lambda_n,i}$, $\hat{\beta}_{\lambda_n+\delta_n,i}$, $\hat{\beta}'_{\lambda_n,i}$ or $\hat{\beta}'_{\lambda_n+\delta_n,i}$ and their counterparts for j. Since $\nu_{n,i}^2 \leq 4\frac{\delta_n^2}{n^2}$, we have $\mathbb{E}[\nu_{n,i}^4] = O(\frac{\delta_n^4}{n^4})$. Additionally, the fourth moments are bounded as we showed the L^4 consistency of soft-thresholding for $\beta^\star$. This gives us $\mathbb{E}[\Psi_{n,i,j}^2] = O(\frac{\delta_n^4}{n^4})$.

With the four results proved, we can conclude that $\gamma(h_n^{\text{diff}}) = \Omega(\frac{\delta_n^2}{n^2\sqrt{n}})$.

H Experimental Setup Details

We provide additional details about the numerical experiments presented in Section 5.

In our simulations, we work with the following sample sizes for the full set size $\frac{nk}{k-1}$: 100, 1,000, 10,000, 100,000, which means n takes the following values: 90, 900, 9,000, 90,000.

For simulations with the Lasso estimator, we used the implementation from `scikit-learn`. For the KDE plots, we called `kdeplot` from the `seaborn` library.

We perform 50,000 replications to sample from $\frac{\sqrt{\frac{nk}{k-1}}}{\sigma(h_n)}(\hat{R}_n - R_n)$ and $\frac{\sqrt{\frac{nk}{k-1}}}{\hat{\sigma}_n(h_n)}(\hat{R}_n - R_n)$. We ensured reproducibility by setting random seeds at the start of all replications.

Regarding the inner cross-validation used to determine λ_n in each iteration of the outer cross-validation, we performed an adaptive grid search via $(k-1)$-fold cross-validation on the training set of size n, based on the initial split of the cross-validation on the full set of size $\frac{nk}{k-1}$. For the adaptive grid search scheme, we started with powers of 10, identified the best choice of penalization, subdivided around this choice with 10 values with an exponential scaling, and did so 3 additional times to identify the optimal penalization with precision.

We now introduce two lemmas that allow us to properly estimate $\sigma^2(h_n)$, $\gamma(h_n)$ and R_n.

Lemma H.1 ($\sigma^2(h_n)$ rewriting for Monte Carlo estimation)**.**

$$\sigma^2(h_n) = \mathbb{E}[h_n(Z_0,\mathbf{Z})(h_n(Z_0,\tilde{\mathbf{Z}}) - h_n(\tilde{Z}_0,\tilde{\mathbf{Z}}))]$$

where $\tilde{Z}_0$ and $\tilde{\mathbf{Z}}$ are independent draws from the same distribution as Z_0 and $\mathbf{Z}$, respectively.

Proof

$$\begin{aligned}
\sigma^2(h_n) &= \mathrm{Var}(\mathbb{E}[h_n(Z_0,\mathbf{Z}) \mid Z_0]) \\
&= \mathbb{E}[\mathbb{E}[h_n(Z_0,\mathbf{Z}) \mid Z_0]^2] - \mathbb{E}[h_n(Z_0,\mathbf{Z})]^2 \\
&= \mathbb{E}[\mathbb{E}[h_n(Z_0,\mathbf{Z})h_n(Z_0,\tilde{\mathbf{Z}}) \mid Z_0]] - \mathbb{E}[h_n(Z_0,\mathbf{Z})h_n(\tilde{Z}_0,\tilde{\mathbf{Z}})] \\
&= \mathbb{E}[h_n(Z_0,\mathbf{Z})h_n(Z_0,\tilde{\mathbf{Z}})] - \mathbb{E}[h_n(Z_0,\mathbf{Z})h_n(\tilde{Z}_0,\tilde{\mathbf{Z}})] \\
&= \mathbb{E}[h_n(Z_0,\mathbf{Z})(h_n(Z_0,\tilde{\mathbf{Z}}) - h_n(\tilde{Z}_0,\tilde{\mathbf{Z}}))]
\end{aligned}$$

□

Lemma H.2 (Conditional expectation and $\gamma(h_n)$ rewriting for Monte Carlo estimation)**.** *If the features are drawn from a distribution with mean 0 and identity covariance matrix, we have*

$$\mathbb{E}[h_n^{\text{sing}}(Z_0, \mathbf{Z}) \mid \mathbf{Z}] = \tau^2 + \|\beta^\star - \hat{\beta}\|_2^2,$$

and thus

$$\mathbb{E}[h_n^{\text{diff}}(Z_0, \mathbf{Z}) \mid \mathbf{Z}] = \|\beta^\star - \hat{\beta}_1\|_2^2 - \|\beta^\star - \hat{\beta}^{(2)}\|_2^2,$$

$$\gamma(h_n^{\text{sing}}) = \mathbb{E}[(h_n^{\text{sing}}(Z_0, \mathbf{Z}) - \|\beta^\star - \hat{\beta}\|_2^2 - (h_n^{\text{sing}}(Z_0, \mathbf{Z}') - \|\beta^\star - \hat{\beta}'\|_2^2))^2],$$

and

$$\gamma(h_n^{\text{diff}}) = \mathbb{E}[(h_n^{\text{diff}}(Z_0, \mathbf{Z}) - \|\beta^\star - \hat{\beta}_1\|_2^2 + \|\beta^\star - \hat{\beta}^{(2)}\|_2^2 - (h_n^{\text{diff}}(Z_0, \mathbf{Z}') - \|\beta^\star - \hat{\beta}_1'\|_2^2 + \|\beta^\star - \hat{\beta}'^{(2)}\|_2^2))^2].$$

Proof Starting from the expression of $\mathbb{E}[h_n(Z_0, \mathbf{Z}) \mid \mathbf{Z}]$ in Lemma C.1, we have

$$\begin{aligned}
\mathbb{E}[h_n^{\text{sing}}(Z_0, \mathbf{Z}) \mid \mathbf{Z}] &= \mathbb{E}[Y_0^2] - 2\beta^{\star\top}\mathbb{E}[X_0 X_0^\top]\hat{\beta} + \operatorname{tr}(\mathbb{E}[X_0 X_0^\top]\hat{\beta}\hat{\beta}^\top) \\
&= \operatorname{Var}(Y_0) + \mathbb{E}[Y_0]^2 - 2\beta^{\star\top}\mathbb{E}[X_0 X_0^\top]\hat{\beta} + \operatorname{tr}(\mathbb{E}[X_0 X_0^\top]\hat{\beta}\hat{\beta}^\top) \\
&= \beta^{\star\top}\operatorname{Var}(X_0)\beta^\star + \tau^2 + (\mathbb{E}[X_0]^\top\beta^\star)^2 - 2\beta^{\star\top}\mathbb{E}[X_0 X_0^\top]\hat{\beta} + \operatorname{tr}(\mathbb{E}[X_0 X_0^\top]\hat{\beta}\hat{\beta}^\top) \\
&= \beta^{\star\top}\operatorname{Var}(X_0)\beta^\star + \tau^2 + \beta^{\star\top}\mathbb{E}[X_0]\mathbb{E}[X_0]^\top\beta^\star - 2\beta^{\star\top}\mathbb{E}[X_0 X_0^\top]\hat{\beta} + \hat{\beta}^\top\mathbb{E}[X_0 X_0^\top]\hat{\beta} \\
&= \tau^2 + \beta^{\star\top}\mathbb{E}[X_0 X_0^\top]\beta^\star + \hat{\beta}^\top\mathbb{E}[X_0 X_0^\top]\hat{\beta} - 2\beta^{\star\top}\mathbb{E}[X_0 X_0^\top]\hat{\beta} \\
&= \tau^2 + (\beta^\star - \hat{\beta})^\top\mathbb{E}[X_0 X_0^\top](\beta^\star - \hat{\beta}) \\
&= \tau^2 + \|\beta^\star - \hat{\beta}\|_2^2
\end{aligned}$$

since the features are drawn from a distribution with mean 0 and identity covariance matrix. The other three expressions follow from the definition of the quantities. □

The Monte Carlo estimation of $\sigma^2(h_n)$ and $\gamma(h_n)$ is based on 5,000,000 replications when using deterministic λ_n, but on 1,000,000 when λ_n is selected via inner cross-validation due to computational complexity. Based on the Monte Carlo standard errors obtained for $\sigma^2(h_n)$ and $\gamma(h_n)$, we applied the Delta method as follows to obtain a standard error for $r(h_n) = \frac{n \cdot \gamma(h_n)}{\sigma^2(h_n)}$. We define $f(x, y) = \frac{nx}{y}$ and we denote by M the number of Monte Carlo replications used to estimate $\sigma^2(h_n)$ and $\gamma(h_n)$. Starting from the Monte Carlo standard errors $\frac{\sigma_x}{\sqrt{M}}$ of $\sigma^2(h_n)$ and $\frac{\sigma_y}{\sqrt{M}}$ of $\gamma(h_n)$, and using $\nabla f = (\frac{n}{y}, -\frac{nx}{y^2})$, we get to a standard error for $r(h_n)$ by computing

$$\nabla f(x, y)^\top \operatorname{diag}(\sigma_x^2, \sigma_y^2)\nabla f(x, y) = \frac{n^2\sigma_x^2}{y^2} + \frac{n^2 x^2 \sigma_y^2}{y^4}.$$

Denoting the Monte Carlo estimates of $\sigma^2(h_n)$ and $\gamma(h_n)$ by $\hat{x}$ and $\hat{y}$, respectively, the standard error we use for $r(h_n)$ is then

$$\frac{1}{\sqrt{M}}\sqrt{\frac{n^2\sigma_x^2}{\hat{y}^2} + \frac{n^2\hat{x}^2\sigma_y^2}{\hat{y}^4}}.$$