Correlations and Prediction Performance of the LASSO Estimator

Aya Amenssag & Emily Ehrhardt *{aya.amenssag22,emily.ehrhardt24}@imperial.ac.uk*
Department of Mathematics
Imperial College London

Firas Darwish & Harleen Gulati *{firas.darwish@worc,harleen.gulati@univ}.ox.ac.uk*
Department of Statistics
University of Oxford

1 Introduction

We focus on the Gaussian linear regression setting

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \sigma^* \mathcal{N}_n(0, \mathbf{I}_n). \tag{1}$$

Here, $\boldsymbol{y} := (y_1, ..., y_n)^\top \in \mathbb{R}^n$ is the response vector and $\mathbf{X} := \left(\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(p)}\right) \in \mathbb{R}^{n \times p}$ is the design matrix, whose columns $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(p)}$ correspond to the p covariates respectively. The regression vector $\boldsymbol{\beta}^* = (\beta_1^*, ..., \beta_p^*)^\top \in \mathbb{R}^p$ is unknown, and one we wish to estimate. The $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^\top \in \mathbb{R}^n$ is random Gaussian noise with $\boldsymbol{\epsilon} \sim \sigma^* \mathcal{N}(\mathbf{0}, I_n)$ and a noise level $\sigma^* > 0$. We now present the LASSO estimator.

The LASSO estimator is any solution to the convex optimisation problem

$$\hat{\boldsymbol{\beta}}_\lambda^{\text{Lasso}} \in \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ ||\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1 \right\}. \tag{2}$$

The LASSO objective consists of two components: the term $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$ measures the data fit and the ℓ_1-penalty $\lambda\|\beta\|_1$ encourages sparsity and controls model complexity using the tuning parameter $\lambda > 0$.

We next introduce the mean squared prediction loss

$$\ell_n(\boldsymbol{\beta}, \boldsymbol{\beta}') = \frac{1}{n} ||\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}')||_2^2. \tag{3}$$

The prediction loss measures the discrepancy between the predictions generated by $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$. For instance, $\ell_n(\hat{\boldsymbol{\beta}}_\lambda^{\text{Lasso}}, \boldsymbol{\beta}^*)$ measures how close the LASSO predictions are to the true predictions. As n increases, we hope our predictions converge to the true predictions. The rate at which this convergence happens can be bounded, and much work has been done in this area (e.g. standard slow rate and fast rate bounds).

In this report, we review these standard bounds for the LASSO prediction loss and then discuss how the correlation of the covariates in the design matrix $\mathbf{X}$ can be exploited to improve these bounds. More specifically, the next section summarises these standard bounds, after which we discuss the different paradigms of how correlation between covariates can be measured and subsequently incorporated into the choice of tuning parameter λ. For each paradigm, we then observe how accounting for correlations affects the standard bounds and under which correlation regimes improvements occur.

2 Standard Bounds for LASSO Predictions

In this section, we first discuss general bounds for the predictive performance of the LASSO estimator, as presented for example in Lewis (2025). Broadly speaking, there are two types of bounds: slow rate bounds, which depend linearly on the tuning parameter λ and hold for arbitrary designs, and fast rate bounds, which depend quadratically on λ but only apply to weakly correlated designs.

A classical slow rate bound is stated in the following theorem:

Theorem 1 (Theorem 9 in Lewis (2025)). *Assume that the columns of the design matrix* $\mathbf{X}$ *are normalised such that* $\max_{j\in[p]} \|\boldsymbol{x}^{(j)}\|_2^2 \leq n$. *Then, for* $0 < \kappa < 1$ *and*

$$\lambda = 2\sqrt{2}\sigma^* \sqrt{n \log(2p/\kappa)}$$

it holds that

$$\ell_n(\hat{\boldsymbol{\beta}}_\lambda^{Lasso}, \boldsymbol{\beta}^*) \leq 4\sqrt{2}\sigma^* \left(\frac{\log(2p/\kappa)}{n}\right)^{1/2} \|\boldsymbol{\beta}^*\|_1. \tag{4}$$

This theorem does not impose any additional assumptions on the design matrix $\mathbf{X}$ and therefore also applies in settings where the covariates are highly correlated. However, the resulting rate scales only as $n^{-1/2}$. Faster bounds, scaling as n^{-1}, can be obtained when additional structural assumptions are imposed on the design matrix. A classical assumption enabling such fast rates is the restricted eigenvalue condition.

Definition 2 (Restricted Eigenvalue Condition). *Let* $\alpha \geq 1$ *and* $\gamma > 0$. *The matrix* $\mathbf{X}$ *satisfies the* (γ, α)*-restricted eigenvalue condition over* $S \subseteq [p], S \neq \emptyset$ *iff for all*

$$\boldsymbol{v} \in \mathbb{C}_\alpha(S) := \{\boldsymbol{v} \in \mathbb{R}^p \ : \ \|\boldsymbol{v}_{S^c}\|_1 \leq \alpha \|\boldsymbol{v}_S\|_1\}$$

it holds that

$$\gamma \|\boldsymbol{v}\|_2^2 \leq \frac{1}{n} \|\mathbf{X}\boldsymbol{v}\|_2^2.$$

Remark 3. *Note, that the restricted eigenvalue condition implies*

$$\lambda_{\min}(\mathbf{X}^\top \mathbf{X}) = \inf_{\boldsymbol{v} \neq 0} \frac{\|\mathbf{X}\boldsymbol{v}\|_2^2}{\|\boldsymbol{v}\|_2^2} \geq \gamma n > 0,$$

i.e. it enforces a strictly positive lower bound on the minimum eigenvalue of $\mathbf{X}^\top \mathbf{X}$. *Hence,* $\mathbf{X}$ *must have full rank and the columns of* $\mathbf{X}$ *cannot be highly correlated.*

Under this condition, the following fast rate bound can be obtained:

Theorem 4 (Theorem 12 in Lewis (2025)). *Assume that the columns of the design matrix* $\mathbf{X}$ *are normalised such that* $\max_{j\in[p]} \|\boldsymbol{x}^{(j)}\|_2^2 \leq n$ *and that* $\mathbf{X}$ *satisfies the* $(\gamma, 3)$*-restricted eigenvalue condition with respect to* $S^* = \{j \in [p] \ : \ \boldsymbol{\beta}_j^* \neq 0\}$ *for some* $\gamma > 0$. *Then, for* $0 < \kappa < 1$ *and*

$$\lambda = 4\sqrt{2}\sigma^* \sqrt{n \log(2p/\kappa)}$$

it holds that

$$\ell_n(\hat{\boldsymbol{\beta}}_\lambda^{Lasso}, \boldsymbol{\beta}^*) \leq \frac{72(\sigma^*)^2 \log(2p/\kappa) \|\boldsymbol{\beta}^*\|_0}{\gamma n} \tag{5}$$

with probability at least $1 - \kappa$.

However, because the restricted eigenvalue condition implicitly rules out strong correlations, this bound is only valid for weakly correlated designs.

In the next section, we review alternative results that specifically target highly correlated designs and yield bounds that are more favourable than the classical slow rate bounds.

3 Bounds Accounting for Correlations in the Design Matrix

In this section, we now discuss paradigms which consider the effect correlations of the covariates of $\mathbf{X}$ have on the bounds of the LASSO estimator. More precisely, we discuss how these paradigms measure the correlations of the covariates in $\mathbf{X}$. We then discuss how they incorporate this measurement into the tuning parameter λ and the implications this has on the bounds of the prediction loss. We observe how correlations can actually provide benefit to the bounds of the LASSO prediction.

3.1 Measuring Correlation Based on Entropy Numbers

In this section, we discuss a bound that applies exclusively to highly correlated designs, originally derived in van de Geer & Lederer (2013) and reviewed in Hebiri & Lederer (2013). Correlations are measured using covering numbers $N(\delta, \mathcal{F}, d)$ or rather the closely related entropy numbers $H(\delta, \mathcal{F}, d) := \log(N(\delta, \mathcal{F}, d))$.

Definition 5 (Covering Numbers)**.** *The covering number $N(\delta, \mathcal{F}, d)$ of a set $\mathcal{F}$ with respect to a metrics d and a radius δ is the minimal number of centres $\boldsymbol{v}_j \in \mathcal{F}$ such that the closed d-balls with radius $\delta > 0$ around $\boldsymbol{v}_j$ cover the set $\mathcal{F}$.*

The intuition is that for highly correlated designs, the symmetric convex hull of the columns of the design matrix $\text{sconv}\{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(p)}\}$ behaves like a low-dimensional set. Consequently, the covering (or entropy) numbers increase only slowly as the radius δ of the balls is decreased. Note that in this section we assume that the design matrix is normalised such that $\|\boldsymbol{x}^{(j)}\|_2 = \sqrt{n}$ for $j \in [p]$.

Assumption 6. *For a fixed $0 < \alpha < 1$ assume*

$$\log(1 + N(\sqrt{n}\delta, \mathit{sconv}\{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(p)}\}, \|\cdot\|_2)) \leq \left(\frac{A}{\delta}\right)^{2\alpha} \tag{6}$$

for all $0 < \delta \leq 1$.

Note that 6 is only satisfied by highly correlated designs. Here, the higher the correlations, the slower the covering numbers grow, and hence the smaller α can be chosen. Based on this assumption van de Geer & Lederer (2013) derive the following bound on the mean squared prediction error:

Theorem 7 (Theorem 3.1 in Hebiri & Lederer (2013) based on results from van de Geer & Lederer (2013))**.** *Let $0 < \alpha < 1$ be fixed. Under Assumption 6, there exists a value $C(\alpha, A)$ depending on α and A only such that for all $\kappa > 0$ and for*

$$\lambda = \left(2\sigma^* C(\alpha, A)\sqrt{n^\alpha \log(2/\kappa)}\right)^{\frac{2}{1+\alpha}} \|\boldsymbol{\beta}^*\|_1^{\frac{\alpha-1}{1+\alpha}},$$

it holds that

$$\ell_n(\hat{\boldsymbol{\beta}}_\lambda^{Lasso}, \boldsymbol{\beta}^*) \leq \frac{21}{2}\left(2\sigma^* C(\alpha, A)\sqrt{\log(2/\kappa)/n}\right)^{\frac{2}{1+\alpha}} \|\boldsymbol{\beta}^*\|_1^{\frac{2\alpha}{1+\alpha}} \tag{7}$$

with probability at least $1 - \kappa$.

The bound in equation 7 decays as $n^{-\frac{1}{1+\alpha}}$. Hence, for $0 < \alpha < 1$, it lies between the slow rate bounds and the fast rate bounds with respect to n. The smaller α (i.e. the higher the correlations), the closer this bound becomes to the fast rate bounds with respect to n. Moreover, unlike the classical bounds, equation 7 does not explicitly depend on p, which can improve the rate in high-dimensional settings.

This result indicates that highly correlated designs are not necessarily disadvantageous for the predictive performance of the LASSO estimator and that they allow for a smaller choice of the tuning parameter λ.

One main drawback of this result is that Assumption 6 is only satisfied for highly correlated designs and hence this approach does not provide any results for moderately correlated designs.

3.2 Measuring Correlation Based on Minimal Convex Combination Representations

In this section, Hebiri & Lederer (2013) introduce a unified framework for analysing LASSO prediction under arbitrary correlation structures, moving beyond the metric entropy framework from **Section 3.1** which is only well-suited for highly correlated designs.

Definition 8 (Correlation Function)**.** *The correlation function K is defined as*

$$\begin{aligned} K :\ & \mathbb{R}_0^+ \to \mathbb{N} \\ & x \mapsto K(x) := \min\{l \in \mathbb{N} :\ \exists\, \boldsymbol{v}^{(1)}, ..., \boldsymbol{v}^{(l)} \in \sqrt{n}S^{n-1}, \\ & \qquad \boldsymbol{x}^{(m)} \in (1+x)\mathit{sconv}\{\boldsymbol{v}^{(1)}, ..., \boldsymbol{v}^{(l)}\}\ \forall\, 1 \leq m \leq p\}, \end{aligned}$$

where S^{n-1} denotes the unit sphere in $\mathbb{R}^n$ and $\mathit{sconv}\{\cdot\}$ is the symmetrised convex hull, i.e.,

$$\mathit{sconv}\{\boldsymbol{v}^{(1)}, ..., \boldsymbol{v}^{(l)}\} = \left\{\sum_{i=1}^{l} \alpha_i \boldsymbol{v}^{(i)} :\ \sum_{i=1}^{l} |\alpha_i| \leq 1\right\}.$$

Thus, $K(x)$ is the smallest number of $\sqrt{n}$-norm template vectors whose symmetrised convex hull, scaled by $(1+x)$, contains every column of $\mathbf{X}$. The parameter $(1+x)$ dictates the approximation precision: a smaller x forces tighter approximation and hence increases $K(x)$. This means $K(x)$ measures the geometric complexity of the column set $\{\boldsymbol{x}^{(m)}\}_{m=1}^{p}$.

For a moderate choice of x, uncorrelated designs yield columns that are nearly orthogonal, so each $\boldsymbol{x}^{(m)}$ requires its own distinct direction for accurate approximation, in which case we get $K(x) \approx p$. For highly correlated designs, the columns lie in a low-dimensional subspace, yielding $K(x) \ll p$.

Definition 9 (Correlation Factor)**.** *The correlation factor K_κ distils the correlation information from $K(x)$ and is defined for $\kappa \in (0,1]$ as*

$$K_\kappa := \inf_{x \in \mathbb{R}_0^+} (1+x)\sqrt{\frac{\log\left(2K(x)/\kappa\right)}{\log(2p/\kappa)}}.$$

Since $K(0) \leq p$, we can immediately see that $K_\kappa \in (0,1]$. Moreover, we get $K_\kappa \approx 1$ for uncorrelated designs (since $K(x) \approx p$), and $K_\kappa \ll 1$ for highly correlated designs.

Leveraging this measure, Hebiri & Lederer (2013) obtain the bound in the following corollary:

Corollary 10 (Corollary 3.1 in Hebiri & Lederer (2013))**.** *For all $\kappa \in (0,1]$, and for*

$$\lambda = K_\kappa 2\sigma^* \sqrt{2n\log(2p/\kappa)},$$

it holds that

$$\ell_n(\hat{\boldsymbol{\beta}}_\lambda^{Lasso}, \boldsymbol{\beta}^*) \leq 4K_\kappa \sigma^* \sqrt{\frac{2\log(2p/\kappa)}{n}} ||\boldsymbol{\beta}^*||_1, \tag{8}$$

with probability at least $1-\kappa$.

This result holds for arbitrary correlation structures, in contrast to **Theorem 7** which is tailored to highly correlated designs, or the fast rate bound from **Theorem 4** which requires weakly correlated designs. More importantly, strong correlations allow for smaller values of K_κ, and hence a smaller tuning parameter λ and a more favourable prediction error bound. For weakly correlated designs with $K_\kappa \approx 1$, we recover the classical slow rate bound from **Theorem 1**.

Finally, the result requires no sparsity assumption on $\boldsymbol{\beta}^*$, and it remains valid even when $\boldsymbol{\beta}^*$ is only approximately sparse. However, the bound becomes suboptimal for very large $||\boldsymbol{\beta}^*||_1$.

3.3 Measuring Correlation Based on the Distance to the Linear Subspace of the Relevant Covariates

In this section, we discuss a paradigm which considers the geometry of the covariates as a way to measure the correlation of the covariates in the design matrix $\mathbf{X}$. As in the prior sections, we discuss how this measurement is incorporated into the tuning parameter λ to account for the effects the correlations of the covariates has on the LASSO prediction loss bounds.

We begin by introducing notation necessary to this section from Dalalyan et al. (2017). Firstly, we introduce $T \subset [p]$, where T can be considered to be the set of the indices of the relevant covariates. We then denote by $\mathbf{X}_T$ the matrix with only the columns of $\mathbf{X}$ indexed by the elements of T (or i.e. indexed by the relevant covariates). We denote by V_T the linear subspace of $\mathbb{R}^n$ spanned by the columns of $\mathbf{X}_T$. The orthogonal projector onto V_T is denoted as Π_T. Finally, we write $a_n \lesssim b_n$ for two positive sequences (a_n) and (b_n) when for some $c \in (0,\infty)$ it holds that $\overline{\lim}_{n\to\infty}(a_n/b_n) \leq c$.

We can now introduce the measure of correlation proposed in Dalalyan et al. (2017), ρ_T, which is the Euclidean distance between the (normalised) columns of $\boldsymbol{X}$ and V_T

$$\rho_T := n^{-1/2} \max_{j \in [p]} ||(\mathbf{I}_n - \Pi_T)\boldsymbol{x}^{(j)}||_2. \tag{9}$$

In other words, we find maximal distance between a covariate and its orthogonal projection onto the set V_T. For the relevant covariates, *i.e.* for $j \in T$, $(\mathbf{I}_n - \Pi_T)\boldsymbol{x}^{(j)}$ is trivially 0. If ρ_T is small, this then implies the irrelevant covariates are close to their orthogonal projections onto the linear subspace spanned by the relevant covariates. The covariates are therefore strongly correlated. On the contrary, if ρ_T is large, then

there exists an irrelevant covariate which is far away from its projection onto V_T. This implies the irrelevant covariate cannot be well explained by the relevant covariates, and therefore there is less correlation between the irrelevant covariate and the relevant covariates.

The measure ρ_T is computable and allows for an exhaustive characterisation of the LASSO prediction as a function of the correlations (Dalalyan et al., 2017). Intuitively we can interpret this as incorporating ρ_T into the LASSO prediction loss allows for predicting the full effects of correlations (*e.g.* high, moderate, low correlations) on the LASSO prediction loss. We now display how ρ_T is incorporated into the prediction loss, and the effects on the bounds of the prediction loss this implies.

We now discuss two main results from Dalalyan et al. (2017):

Theorem 11 (Theorem 4 in Dalalyan et al. (2017))**.** *Let $T \subset [p]$ be a set of indices and let $0 < \kappa$, $\gamma \geq 1$ be constants. Then, if*

$$\lambda \geq 2\gamma\sigma^* \rho_T \sqrt{2n \log(p/\kappa)}, \tag{10}$$

the LASSO fulfils

$$\ell_n(\hat{\boldsymbol{\beta}}_\lambda^{Lasso}, \boldsymbol{\beta}^*) + \frac{2(\gamma-1)\lambda}{\gamma} ||\hat{\boldsymbol{\beta}}_\lambda^{Lasso}||_1 \leq \inf_{\bar{\boldsymbol{\beta}} \in \mathbb{R}^p} \{\ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) + \frac{2(\gamma+1)\lambda}{\gamma} ||\bar{\boldsymbol{\beta}}||_1\} + \frac{2(\sigma^*)^2 \left(|T| + 2\log(1/\kappa)\right)}{n} \tag{11}$$

with probability at least $1 - 2\kappa$.

Note that in equation 10 the term ρ_T incorporates the level of correlations into the lower bound of the tuning parameter, allowing for considerably smaller tuning parameters if the covariates are highly correlated.

The obtained bound lies between the standard slow rate and fast rate bounds and especially for highly correlated designs fast bounds can be obtained as $n \to \infty$:

Corollary 12 (Corollary 1 in Dalalyan et al. (2017))**.** *Assume that $T_n \subset [p]$ is a set of indices (that may depend on the sample size n) such that all covariates $\{\boldsymbol{x}^{(j)} : j \in [p]\}$ are very close to the linear span of the set of vectors $\{\boldsymbol{x}^{(j)} : j \in T_n\}$ in the sense that $\rho_{T_n} \lesssim n^{-r}$ for a positive constant $r > 0$. Then, if the tuning parameter satisfies $\lambda \geq 2c\sigma^* \sqrt{\log(p)/n^{2r-1}}$ for a sufficiently large constant $c > 0$, the LASSO fulfils*

$$\ell_n(\hat{\boldsymbol{\beta}}_\lambda^{Lasso}, \boldsymbol{\beta}^*) \lesssim \left(\sqrt{\frac{\log(p)}{n^{2r+1}}} ||\boldsymbol{\beta}^*||_1 \right) \vee \frac{|T_n|}{n}$$

with high probability.

For instance, if the irrelevant covariates are within a Euclidean distance of 1 of the linear space spanned by the relevant covariates, then it holds that $r = 1/2$ and hence the LASSO achieves a fast rate $|T_n|/n$ up to logarithmic factors provided λ is chosen of order $\sqrt{n \log(p)}$ (with sufficiently large constants) (Dalalyan et al., 2017).

4 Discussion and Practical Implications

4.1 Algorithms

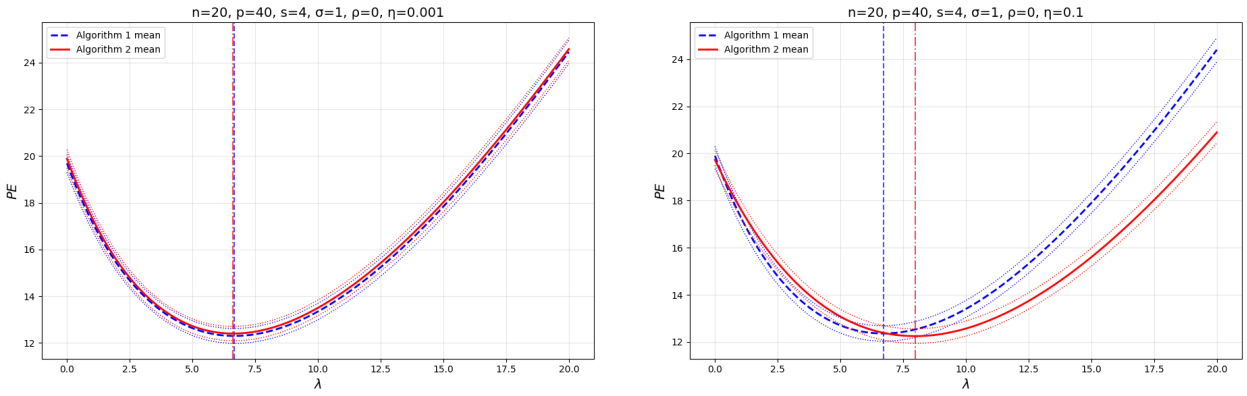
In order to experimentally test the ways in which the degree of correlation in the design matrix affects the predictive performance of the LASSO estimator as well as the optimal λ value, we use two algorithms drawn from Hebiri & Lederer (2013) (these can be found in the appendix).

Both of these algorithms allow us to simulate different settings where we control the degree of correlation in our design matrix. In **Algorithm** 1, this comes down to the ρ variable which defines the off-diagonal elements in the covariance matrix we use to generate the design matrix 1. In **Algorithm** 2, in addition to the ρ, we control the degree of correlation in our design matrix using η which controls how correlated the additional columns are to the original columns they are generated from 2. As in Hebiri & Lederer (2013), whenever we use either of these algorithms, we use the LARS algorithm to compute the LASSO solution due to the good behaviour of the algorithm when the variables are correlated.

4.2 Experiments

We now present the experiments we designed based on **Algorithm 1** and **Algorithm 2** in order to empirically explore the effect of correlation on predictive performance and optimal λ, as well as showcase our results.

Building on the design of Hebiri & Lederer (2013), we first run Algorithm 1 and Algorithm 2 for 1000 iterations and plot the mean values of the prediction errors against the tuning parameter λ. We set $n = 20, p = 40, s = 4, \sigma = 1, \rho = 0$. We set $\eta = 0.001$ and $\eta = 0.1$ to evaluate the impact of the degree of correlation in the added columns. We display the results in Figure 1.



(a) $\eta = 0.001$ (b) $\eta = 0.1$

Figure 1: The mean prediction error against tuning parameter λ over 1000 iterations for both Algorithm 1 and Algorithm 2 where $\eta = \{0.001, 0.1\}$. In both subfigures, Algorithm 1 is represented by the blue, dashed line and Algorithm 2 is represented by the red, solid line. The dotted lines for both algorithms give the confidence bounds.

Figure 1 highlight an interesting finding. The minimal prediction error of **Algorithm 1** and **Algorithm 2** are not very different across both figures, despite the fact that Algorithm 2 introduces $p^2 - p$ new impertinent variables that are correlated with the other columns in the initial design matrix. Even when we vary η from 0.001 (highly correlated variables) to 0.1 (less correlated variables), we find that this behaviour does not change.

Another valuable observation from Figure 1a is that the optimal λ value is the same across **Algorithm 1** and **Algorithm 2**. We also find that this is not the case in Figure 1b, where there is a slight difference between the two optimal λ values–despite attaining a similar mean prediction error across the 1000 iterations.

Additionally, we design another experiment based on the work of Hebiri & Lederer (2013) using Algorithm 1 to test the influence of correlations ρ. We iterate 1000 times and present the mean optimal tuning parameters $\overline{\lambda}^*$ and the mean minimal prediction error $\overline{PE}_{min}$. Holding $n = 20, p = 40, s = 4, \sigma = 1$, we vary ρ across the set $0.99, 0.9, 0$. We present the results of this experiment in 1 in the appendix.

From the results, we find that larger ρ values (where the correlation between variables is high) leads to smaller λ^* values. Furthermore, we see that the mean prediction errors for design matrices with higher correlation (ρ) is not higher than than for design matrices with lower correlation. In fact, $\overline{\text{PE}}_{\text{min}}$ is lower across the board when ρ is higher.

4.3 Theoretical Comparison

The theory presented above offers an explanation as to why correlation can improve prediction, although it may appear counter-intuitive, especially when the design has a favourable geometric structure. Three analytical frameworks characterise this geometry.

The **entropy-based approach** in Hebiri & Lederer (2013) provides the most theoretically refined framework, capturing global geometric properties of the design. However, it is difficult to compute in practice and is mainly useful for extreme correlation structures.

Hebiri & Lederer (2013) also introduces a more unified framework using the K_κ **correlation measure** which covers arbitrary correlation structures and interpolates between slow and fast rates, but it requires optimisation over correlation envelopes $K(x)$, which can be computationally demanding.

The most practical approach is the ρ_T **measure** presented by Dalalyan et al. (2017) which quantifies how well all covariates can be approximated by the linear span of a small subset T. When ρ_T is small, the design is effectively low-dimensional, even if p is much larger than n. It is easy to compute, and therefore the most useful in applications.

4.4 Correlation-Adaptive Tuning and Practical Consequences

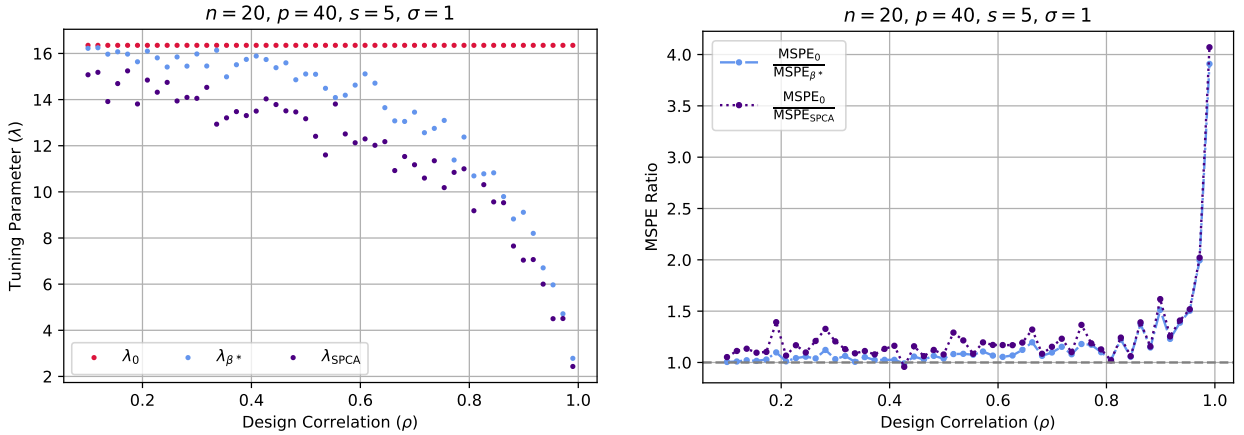
An important implication of this is how we choose the LASSO tuning parameter. The classical lower for a desired probability level κ given by $\lambda_0 = 2\sigma^*\sqrt{2n\log(p/\kappa)}$ can be overly conservative when covariates are strongly correlated, as it treats all directions in the design as equally noisy. In contrast, the correlation-adaptive lower bound on λ derived from Dalalyan et al. (2017) is given by $\lambda_T = 2\sigma^*\rho_T\sqrt{2n\log(p/\kappa)}$. This is automatically smaller when the design is highly correlated, since the effective dimensionality of the prediction problem decreases and hence requires less penalisation.

To study this effect, we consider two choices of the set $T \subset [p]$ used in the computation of ρ_T:

- $T_{\beta^*} = \{j \in [p] : \beta_j^* \neq 0\}$, *i.e.* the true support of $\boldsymbol{\beta}^*$. The associated tuning parameter is denoted by λ_{β^*}.

- Following Dalalyan et al. (2017), we use sparse PCA to construct a small set T_{SPCA} whose span approximates the full design well. The associated tuning parameter is denoted λ_{SPCA}.

Using Algorithm 1, we generate design matrices with identical dimensions but varying correlation levels (controlled by ρ). We compare prediction performance under the three lower bounds for λ. Figures 2a and 2b summarise the results.



(a) Tuning parameter comparison

(b) Prediction error comparison

Figure 2: Comparison of correlation-adaptive and universal tuning parameters and their predictive performance

Figure 2a shows that both λ_{β^*} and λ_{SPCA} decrease as correlation ρ increases, whereas the universal parameter λ_0 remains fixed. This reflects the fact that correlated designs lie closer to a low-dimensional subspace, which ρ_T successfully captures.

Figure 2b demonstrates that correlation-adaptive tuning tends to outperform universal tuning, particularly in moderately or highly correlated designs. The SPCA-based version often performs best, reflecting the advantage of modelling the geometry of X rather than relying solely on knowledge of the true support.

In practice, the above λ values are only lower bounds. Cross-validation is still needed to explore a wider grid, including sufficiently small values informed by correlation diagnostics such as $\rho_{T_{\text{SPCA}}}$.

4.5 Limitations and When Correlation-Adaptive Tuning Helps Most

The guarantees in Hebiri & Lederer (2013); Dalalyan et al. (2017) concern prediction error only. Highly correlated designs still hinder variable selection and support recovery. Moreover, computing correlation-adaptive tuning parameters can be demanding in very high dimensions, *e.g.* if repeated sparse PCA is required. Additionally, some correlation structures may prevent the LASSO from achieving fast rates altogether, regardless of tuning (see Example 2 to in Dalalyan et al. (2017)).

Correlation-adaptive tuning is most beneficial when classical fast rate assumptions (*e.g.* RE) fail. In weakly correlated designs, where these assumptions hold, standard tuning already achieves the desired $s\log(p)/n$ rates, and the gains from refined tuning are moderate. The strongest improvements occur in settings with structured weakly correlated designs (see Remark 3.1 in Hebiri & Lederer (2013)) and moderate or strong correlation, where universal tuning becomes too conservative but where the design still exhibits enough structure (e.g., small ρ_T) to allow favourable rates.

5 Conclusion

In this report, we reviewed classical prediction bounds for the LASSO estimator and highlighted how their performance depends on the correlation structure of the design matrix. While slow and fast rate bounds provide general guarantees, they are either too weak for practical use or rely on restrictive assumptions such as the restricted eigenvalue condition, which rules out strong correlations. We then examined three alternative approaches that explicitly account for correlation: one based on entropy numbers, one based on minimal convex combination representations and one based on the distance to the linear subspace of the relevant covariates. All three frameworks show that correlation can be exploited to improve prediction guarantees, often allowing for smaller choices of the tuning parameter and yielding bounds that lie in between slow and fast rates. Additionally, we run several experiments where we observe how optimizing our choice of the LASSO tuning parameter can lead to similar prediction errors in high and low correlation settings. These results demonstrate that strong correlations are not purely detrimental. When properly incorporated into the tuning parameter, they can lead to sharper bounds and more favourable prediction behaviour for the LASSO.

References

Arnak S. Dalalyan, Mohamed Hebiri, and Johannes Lederer. On the prediction performance of the Lasso. *Bernoulli*, 23(1):552 – 581, 2017. doi: https://doi.org/10.3150/15-BEJ756.

Mohamed Hebiri and Johannes Lederer. How Correlations Influence Lasso Prediction. *IEEE Transactions on Information Theory*, 59(3):1846–1854, 2013. doi: https://doi.org/10.1109/TIT.2012.2227680.

Rebecca Lewis. Modern statistical theory: High-dimensional regression and lasso. Lecture notes, 2025. Based on previous notes by George Deligiannidis.

Sara van de Geer and Johannes Lederer. The lasso, correlated design, and improved oracle inequalities. *IMS Collections*, 9:303–316, 2013. doi: https://doi.org/10.1214/12-IMSCOLL922.

A Algorithms

Algorithm 1: Algorithm 1 (Hebiri & Lederer, 2013)

Input: n (samples), p (variables), σ^* (noise level), s (# nonzero coefficients), $\rho \in [0,1]$ (correlation parameter)

Define the true coefficient vector $\boldsymbol{\beta}^*$ of length p with $\beta_i^* = \mathbb{1}_{[i \leq s]}$

Generate the design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ with rows sampled from $\mathcal{N}(0, \Sigma)$ with $\Sigma_{jk} = \mathbb{1}_{[j=k]} + \rho \cdot \mathbb{1}_{[j \neq k]}$;

Normalize X so that $(\mathbf{X}^\top \mathbf{X})_{jj} = n$ for all j;

Sample noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}_n)$;

Compute the response vector

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}^* + \sigma^* \boldsymbol{\epsilon}$$

Compute the Lasso estimator $\hat{\boldsymbol{\beta}}_\lambda^{\text{Lasso}}$ for a grid of tuning parameters λ;

For each λ, compute the prediction error

$$\|\mathbf{X}(\hat{\boldsymbol{\beta}}_\lambda^{\text{Lasso}} - \boldsymbol{\beta}^*)\|_2^2$$

Select the optimal tuning parameter

$$\lambda^* = \operatorname*{argmin}_\lambda \|\mathbf{X}(\hat{\boldsymbol{\beta}}_\lambda^{\text{Lasso}} - \boldsymbol{\beta}^*)\|_2^2$$

Output: Optimal tuning parameter λ^* and corresponding Lasso estimate $\hat{\boldsymbol{\beta}}_{\lambda^*}^{\text{Lasso}}$

Algorithm 2: Algorithm 2 (Hebiri & Lederer, 2013)

Input: n (samples), p (original variables), σ^* (noise level), s (number of nonzero coefficients), $\rho \in [0,1]$ (correlation of original design), η (correlation parameter for added columns)

Define the true coefficient vector $\boldsymbol{\beta}^*$ of length p with $\beta_i^* = \mathbb{1}_{[i \leq s]}$

Generate the initial design matrix $\mathbf{X}^{(0)} \in \mathbb{R}^{n \times p}$ as in Algorithm 1;

Normalize $\mathbf{X}^{(0)}$ so that $\left((\mathbf{X}^{(0)})^\top \mathbf{X}^{(0)}\right)_{jj} = n$ for all j;

for *each original column* $\boldsymbol{x}^{(0,j)}$, $j \in [p]$ **do**

Generate $p-1$ additional columns as

$$\boldsymbol{x}^{(j,k)} = \boldsymbol{x}^{(0,j)} + \eta N_{j,k}, \quad k \in [p-1],$$

where each $N_{j,k}$ is a standard normally distributed random vector in $\mathbb{R}^n$;

Form the augmented design matrix $\mathbf{X} \in \mathbb{R}^{n \times p^2}$ by stacking all original and additional columns;

Extend $\boldsymbol{\beta}^*$ by assigning coefficient 0 to all added columns (so there are $p^2 - p$ impertinent variables);

Normalize $\mathbf{X}$ so that $(\mathbf{X}^\top \mathbf{X})_{jj} = n$ for all j;

Sample noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}_n)$;

Compute the response vector

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}^* + \sigma^* \boldsymbol{\epsilon}$$

Compute the Lasso estimator $\hat{\boldsymbol{\beta}}_\lambda^{\text{Lasso}}$ on a grid of tuning parameters λ;

For each λ, compute the prediction error

$$\|\mathbf{X}(\hat{\boldsymbol{\beta}}_\lambda^{\text{Lasso}} - \boldsymbol{\beta}^*)\|_2^2$$

Choose the optimal tuning parameter

$$\lambda^* = \operatorname*{argmin}_\lambda \|\mathbf{X}(\hat{\boldsymbol{\beta}}_\lambda^{\text{Lasso}} - \boldsymbol{\beta}^*)\|_2^2$$

Output: Optimal tuning parameter λ^* and corresponding Lasso estimate $\hat{\boldsymbol{\beta}}_{\lambda^*}^{\text{Lasso}}$

B Algorithm 1 Grid Search Experiment

Table 1: Summary of Algorithm 1 results for various settings of $n, p, s, \sigma, and \rho$.

n	p	s	σ	ρ	$\bar{\lambda}_{\min}$	$\bar{\mathrm{PE}}_{\min}$
20	40	4	1	0.99	5.12 ± 0.17	2.27 ± 0.05
20	40	4	1	0.9	6.67 ± 0.16	6.43 ± 0.09
20	40	4	1	0.0	7.02 ± 0.08	11.63 ± 0.16
50	40	4	1	0.99	7.58 ± 0.24	3.59 ± 0.05
50	40	4	1	0.9	10.35 ± 0.22	9.24 ± 0.12
50	40	4	1	0.0	14.37 ± 0.10	12.19 ± 0.18
20	400	4	1	0.99	5.81 ± 0.17	2.67 ± 0.04
20	400	4	1	0.9	8.31 ± 0.18	8.63 ± 0.10
20	400	4	1	0.0	8.71 ± 0.13	14.89 ± 0.16
20	40	10	1	0.99	4.75 ± 0.17	4.92 ± 0.07
20	40	10	1	0.9	5.53 ± 0.16	11.34 ± 0.12
20	40	10	1	0.0	3.45 ± 0.07	16.67 ± 0.17
20	40	4	3	0.99	13.58 ± 0.33	8.52 ± 0.36
20	40	4	3	0.9	19.69 ± 0.25	22.81 ± 0.44
20	40	4	3	0.0	28.54 ± 0.11	58.33 ± 0.70