On Lasso refitting strategies

Evgenii Chzhen*1,2, Mohamed Hebiri†1, and Joseph Salmon‡2

1LAMA, Université Paris-Est Marne-la-Vallée, Université Paris-Est, 5 Boulevard Descartes, 77420, Champs-sur-Marne, France
2LTCI, Télécom ParisTech, Université Paris-Saclay, 46 rue Barrault, 75013, Paris, France,

Abstract

A well-know drawback of \(\ell_1\)-penalized estimators is the systematic shrinkage of the large coefficients towards zero. A simple remedy is to treat Lasso as a model-selection procedure and to perform a second refitting step on the selected support. In this work we formalize the notion of refitting and provide oracle bounds for arbitrary refitting procedures of the Lasso solution. One of the most widely used refitting techniques which is based on least-squares may bring a problem of interpretability, since the signs of the refitted estimator might be flipped with respect to the original estimator. This problem arises from the fact that the least-square refitting considers only the support of the Lasso solution, avoiding any information about signs or amplitudes. To this end we define a sign-consistent refitting as an arbitrary refitting procedure, preserving the signs of the first step Lasso solution and provide Oracle inequalities for such estimators. Finally, we consider special refitting strategies: Bregman Lasso and Boosted Lasso. Bregman Lasso has a fruitful property to converge to the sign-consistent least-squares refitting (least-squares with sign constraints), which provides with greater interpretability. We additionally study the Bregman Lasso refitting in the case of orthogonal design, providing with simple intuition behind the proposed method. Boosted Lasso, in contrast, considers information about magnitudes of the first Lasso step and allows to develop better oracle rates for prediction. Finally, we conduct an extensive numerical study to show advantages of one approach over others in different synthetic and semi-real scenarios.

1 Introduction

Least absolute shrinkage and selection operator (Lasso), introduced by Tibshirani [1996], became a popular method in high dimensional statistics due to competitive numerical solvers (as it is a convex program) and fruitful statistical guaranties [Bickel et al., 2009, Koltchinskii, 2011]. However, the shrinkage of large magnitudes towards zero, observed in practice, may affect the overall conclusion about the model. Different remedies were proposed to overcome this affect, all of them having their advantages and disadvantages. For instance, one may consider a non-convex penalty instead of the \(\ell_1\) regularization [Fan and Li, 2001, Zhang, 2010, Gasso et al., 2009]: this approach increases a computational complexity and might be not applicable in large-scale scenarios. Another way to avoid the underestimation of the coefficients is to perform the second least-squares refitting step based on the first step Lasso solution [Belloni and Chernozhukov, 2013, Lederer, 2013]: such an approach brings the problem of interpretability, since the coefficients may switch signs with respect to the original Lasso solution. A lot

*evgenii.chzhen@phystech.edu 
†mohamed.hebiri@u-pem.fr 
‡joseph.salmon@telecom-paristech.fr
of theoretical and applied works are devoted to the study of least-squares refitting of an arbitrary first step estimator [Belloni and Chernozhukov, 2013, Lederer, 2013, Deledalle et al., 2015, 2017].

Unlike their approach we are rather interested in the refitting of Lasso estimator, it allows us to use previous theoretical analysis provided for Lasso to derive guarantees for a wide class of refitting strategies. In Section 2 we introduce notation, used throughout the article, and the basic Lasso theory is partly covered in Section 3. Readers who are familiar with the Lasso theory may skip Section 3 and proceed to the following sections. Section 4, is concerned with our theoretical framework, where we define a refitting strategy as an estimator which reaches a lower mean square error (MSE), compared to the first step Lasso solution. For this family of refitting strategies we show that the rates for prediction are bounded by the Lasso rates plus an \( \ell_1 \)-norm of difference between Lasso estimator and the refitted vector. Inspired by this result we propose to use an additional information, provided by the Lasso solution, for refitting. It leads to a least-squares refitting with constraints, to avoid an explosion of the refitted coefficients. This estimator can be seen as Boosted Lasso strategy (see Section 4.2 and more particularly Lemma 3) allows us to develop better prediction bounds compared to the classical Lasso bounds.

Additionally, we propose another family of refitting strategies in Section 4.1, which restricts the possibility to switch signs with respect to the first step Lasso solution in addition to lower MSE. For every refitting strategy in this family we provide a unified oracle inequality stated in Theorem 3 showing minimax rates under the same assumptions as oracle inequalities for Lasso. We introduce Bregman Lasso, which can be seen as a generalization of Bregman Iterations [Osher et al., 2005, 2016], widely used method in compressed sensing settings and has a strong connection with the method proposed by Brinkmann et al. [2016]. Analyzing the Bregman Lasso in case of denoising model (Section 4.3.2) we provide useful insights on the proposed method. Additionally, we show that Bregman Lasso is a refitting strategy converging to Sign-Least-Squares Lasso, which can be tracked back in Brinkmann et al. [2016]. This estimator restricts the possibility to flip signs, minimizing MSE meanwhile. For Bregman Lasso, we conduct an intensive analysis in the orthogonal design case which is of independent interest and that exhibits some interesting interpretation of this method, and makes some analogies between Bregman Lasso and well-known existing thresholding methods (such as soft/hard/firm-thresholding). Moreover, an important part of the present contribution is the extension of the proposed Bregman Lasso refitting strategy to the exponential family. This part of the work is covered in Section 4.3.5.

Finally, we conduct an extensive numerical study of different post-Lasso refitting strategies to show advantages of different estimators in various scenarios.

## 2 Framework and notation

The standard Euclidean norm is written \[ \| \cdot \|_2 \], the \( \ell_1 \)-norm \[ \| \cdot \|_1 \], and the \( \ell_\infty \)-norm \[ \| \cdot \|_\infty \]. For any integer \( d \in \mathbb{N} \), we denote by \([d]\) the set \([1, \ldots, d]\) and by \( Q^\top \) the transpose of a matrix \( Q \), and \( I_d \in \mathbb{R}^{d \times d} \) is the identity matrix of size \( d \). For two real numbers \( a, b \in \mathbb{R} \) we defined by \( a \vee b \) the maximum between \( a \) and \( b \). For any vectors \( a, b \in \mathbb{R}^p \) we denote by \( \langle a, b \rangle = a^\top b \) the Euclidean inner product and by \( a \odot b \) the element wise (Hadamard) product of two vectors. Our approach is valid for a broad class of models, but to avoid digression, we study the prediction performance of the Lasso and refitting strategies only for Gaussian linear regression models with deterministic design. More specifically, we consider \( n \) random observations \( y_1, \ldots, y_n \in \mathbb{R} \) and fixed covariates \( x_1, \ldots, x_n \in \mathbb{R}^p \). We further assume that there is a regression vector \( \beta^* \in \mathbb{R}^p \) which satisfies the following relation:

\[
y = X \beta^* + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n),
\]

where \( y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n \) is the response vector and \( X = (x_1, \ldots, x_n)^\top \in \mathbb{R}^{n \times p} \) the design matrix. We additionally assume, that the columns of \( X \) are normalized in such a way that for all \( j \in [p] \) we have \( \|X_j\|_2^2 = n \), where \( X_j \) is \( j^{th} \) column of the matrix \( X \). For any set \( E \subset [p] \), we denote by \( E^c \) the
complement to $E$ (i.e., $E \cup E^c = [p]$) and by $X_E$ the matrix obtained from the matrix $X$ by erasing all the columns whose indexes are not in $E$. Similarly, for any $\beta \in \mathbb{R}^p$ we write $\beta_E$ to denote the vector obtained from $\beta$ by erasing all the components whose indexes are not in $E$. For all vectors $\beta \in \mathbb{R}^p$ we write $\text{supp}(\beta) \subset [p]$ for the support of the vector $\beta$, i.e., $\text{supp}(\beta) = \{j \in [p]: \beta_j \neq 0\}$. For every real-valued function $f : \mathbb{R}^p \mapsto \mathbb{R}$ we say that $g \in \mathbb{R}^p$ is a subgradient of $f$ at $x \in \mathbb{R}^p$ if $f(y) \geq f(x) + \langle g, y - x \rangle$ for all $y \in \mathbb{R}^p$. The set of all subgradients of $f$ at $x \in \mathbb{R}^p$ is called subdifferential of $f$ at $x \in \mathbb{R}^p$ and written as $\partial f(x)$. We also remind, that the subdifferential of the $\ell_1$-norm $\partial \| \cdot \|_1$ is a set valued vector function $\text{sign}(\cdot) = (\text{sign}(\cdot)_1, \ldots, \text{sign}(\cdot)_p)^T$, defined element-wise by

$$\forall \beta \in \mathbb{R}^p, \forall j \in [p], \quad \text{sign}(\beta)_j = \begin{cases} \{1\}, & \beta_j > 0 \\ \{-1\}, & \beta_j < 0 \\ \{-1, 1\}, & \beta_j = 0 \end{cases}.$$  \hfill (2)

Also, we assume that the unknown vector $\beta^*$ is sparse, i.e., $\text{supp}(\beta^*) = S$ has small cardinality $s$ compared to $n$ and $p$. To estimate $\beta^*$ we first minimize the negative log-likelihood with $\ell_1$ penalty [Tibshirani, 1996], which is equivalent for a fixed $\lambda > 0$ to the following optimization problem

$$\hat{\beta} \in \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \| y - X\beta \|_2^2 + \lambda \| \beta \|_1.$$ \hfill (3)

We additionally remind the KKT conditions (necessary and sufficient in this case) for the Equation (3):

**Lemma 1** (KKT conditions for Lasso). The Karush-Kuhn-Tucker conditions [Boyd and Vandenberghe, 2004] for the Lasso problem Equation (3) read as follows:

$$0 \in \lambda \text{sign}(\hat{\beta}) - \frac{1}{n} X^T (y - X\hat{\beta}) ,$$

or equivalently: there exists $\hat{\rho} \in \text{sign}(\hat{\beta})$, such that

$$\frac{1}{n} X^T (y - X\hat{\beta}) = \lambda \hat{\rho}.$$ \hfill (4)

### 3 Lasso theory

In this section we provide one of the classical Lasso oracle inequalities. To this end, we introduce the restricted eigenvalue condition [Bickel et al., 2009], a widely used assumption on the design matrix $X$.

**Definition 1** (Bickel et al. [2009]). We say that $X \in \mathbb{R}^{n \times p}$ satisfies the Restricted Eigenvalue condition RE$(c_0, s)$, where $c_0 > 0$ and $s \in [p]$, if $\exists \kappa(c_0, s) > 0$ such that for all $J \subset [p]$ with $|J| \leq s$ we have for all $\Delta \in \mathbb{R}^p$

$$\|\Delta_J\|_1 \leq c_0 \|\Delta\|_1 \quad \Rightarrow \quad \frac{\|X\Delta\|_2^2}{n\|\Delta\|_2^2} \geq \kappa^2(c_0, s) .$$

We also state below some classical concentration bound on tail of sup of Gaussian random variables.

**Lemma 2.** Let $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ and $X \in \mathbb{R}^{n \times p}$ be such that $\forall j \in [p]$ we have $\|X_j\|_2^2 = n$, hence with probability at least $1 - \delta$ we have

$$\|X^T \epsilon/\sqrt{n}\|_\infty \leq \lambda/2 ,$$

where $\lambda = 2\sigma \sqrt{2 \log (p/\delta)/n}$. 

3
The following theorem is a starting point of our analysis. Therefore, and for the sake of completeness we state its proof in the Appendix. We mention that similar techniques of the proof can be found in [Giraud, 2014, Dalalyan et al., 2017].

**Theorem 1.** If the design matrix \( X \in \mathbb{R}^{n \times p} \) satisfies the restricted eigenvalue condition \( \text{RE}(3, s) \) and \( \lambda = 2\sigma \sqrt{2 \log (p/\delta)} / n \) for every \( \delta \in (0, 1) \), then with probability \( 1 - \delta \) the following bound holds

\[
\frac{1}{n} \|X(\beta^* - \hat{\beta})\|_2^2 \leq \frac{9\lambda^2 s}{4\kappa^2(3,s)} ,
\]

where \( \hat{\beta} \) is a Lasso solution with tuning parameter \( \lambda \).

**Proof.** See supplementary materials for the proof.

---

### 4 Refitting strategies

Underestimation of large coefficients by Lasso and other \( \ell_1 \)-penalized estimators have long been well known by practitioners, and simple remedies have been proposed on case by case analysis. One of such approaches is Least-Squares refitting - widely used in high-dimensional regression to reduce the bias of the coefficients and consists in performing a least-squares re-estimation of the non-zero coefficients of the solution. Such a procedure is theoretically analyzed by Belloni and Chernozhukov [2013] applied to an arbitrary first-step estimator. Lederer [2013], showed that blind Least-Squares refitting of the Lasso solution is not advised in all possible scenarios and developed a refitting criteria. Deledalle et al. [2015, 2017] are mostly concerned with the practical aspects of refitting and provide efficient numerical procedures to perform refitting simultaneously along with the original estimator. In contrast, here we are interested in arbitrary refitting of the Lasso solution, which allows to exploit Lasso theory to provide new insights. In this section we study a general framework for refitting techniques. We define a refitting of a Lasso solution \( \bar{\beta} \) as:

**Definition 2.** Let \( \hat{\beta} = \hat{\beta}(\lambda) \) be a Lasso solution with regularization \( \lambda \). We call a vector \( \bar{\beta} = \bar{\beta}(\hat{\beta}, X, y) \) a refitting of the Lasso \( \hat{\beta} \) if it reduces the original loss function, namely if:

\[
\|y - X\bar{\beta}\|_2 \leq \|y - X\hat{\beta}\|_2.
\]

Notice, that one should not expect a superior performance of such a refitting, since the only information available for \( \bar{\beta} \) is the mean square error. Additionally notice, that the least-squares solution is obviously a refitting strategy for every Lasso solution. However, defining and analyzing such a family provides with interesting insights and serves as the step towards more thoughtful refitting strategies.

**Theorem 2.** If the design matrix \( X \in \mathbb{R}^{n \times p} \) satisfies the restricted eigenvalue condition \( \text{RE}(3, s) \) and \( \lambda = 2\sigma \sqrt{2 \log (p/\delta)} / n \) for some \( \delta \in (0, 1) \), then with probability \( 1 - \delta \) the following bound holds

\[
\frac{1}{n} \|X(\beta^* - \bar{\beta})\|_2^2 \leq \frac{1}{n} \|X(\beta^* - \hat{\beta})\|_2^2 + \lambda \|\bar{\beta} - \hat{\beta}\|_1 ,
\]

where \( \bar{\beta} \) is a refitting of Lasso solution \( \hat{\beta} \).

**Proof.** Substituting \( y = X\beta^* + \epsilon \) into Definition 2 to get

\[
\frac{1}{2n} \left\| X(\bar{\beta} - \beta^*) \right\|_2^2 - \frac{1}{2n} \left\| X(\beta^* - \hat{\beta}) \right\|_2^2 \leq \frac{1}{n} \epsilon^T X(\bar{\beta} - \hat{\beta}) .
\]

Hence, we can conclude using Hölder's inequality on the event \( \|\epsilon^T X/n\|_\infty \leq \lambda/2 \).

---

Previous theorem shows, that by controlling the \( \ell_1 \)-distance between refitting and Lasso solution, one might obtain satisfying performance, we discuss this idea in Section 4.2.
4.1 Sign consistent refitting strategies

Previous section is concerned with an arbitrary refitting strategy, which only uses the information about the mean square error of the Lasso solution. Here, we are interested in a more sophisticated family of refitting strategies, which additionally exploit the information provided by the sign of the Lasso solution. Such an approach has some similarities with the methods introduced by Brinkmann et al. [2016], even though the authors had a different motivation.

**Definition 3.** Let \( \hat{\beta} = \hat{\beta}(\lambda) \) be a Lasso solution with regularization \( \lambda \). We call a \( \hat{\beta} = \hat{\beta}(\beta, X, y) \) sign-consistent refitting of the Lasso solution \( \hat{\beta} \) if

\[
\|y - X\hat{\beta}\|_2 \leq \|y - X\beta^*\|_2 \quad \text{(refitting)},
\]

\[
\frac{1}{n} X^\top (y - X\hat{\beta}) \in \lambda \text{sign}(\hat{\beta}) \quad \text{(sign-consistency)}.
\]

**Remark 1.** The sign-consistency property in Definition 3 ensures that a sign-consistent refitting vector \( \hat{\beta} \) satisfies for all \( j \in [p] \) the following conditions

- if \( n^{-1}X_j^\top (y - X\hat{\beta}) = \lambda \), hence \( \hat{\beta}_j \geq 0 \);
- if \( n^{-1}X_j^\top (y - X\hat{\beta}) = -\lambda \), hence \( \hat{\beta}_j \leq 0 \);
- if \( n^{-1}|X_j^\top (y - X\hat{\beta})| < \lambda \), hence \( \hat{\beta}_j = 0 \).

Notice, that the equations in the previous remark are exactly the first-order optimality conditions for the Lasso problem written component-wise. Since two estimators \( \hat{\beta} \) (Lasso) and \( \hat{\beta} \) (sign-refitting) share the same subgradient we are able to re-use proof techniques similar to the one of Theorem 1 to provide an oracle inequality. The next theorem shows that the definition of the sign-refitting allows to develop oracle rates without any additional assumptions, except classical ones used for Lasso bounds. The proof of this theorem is instructive to understand our refitting strategy so that we incorporate it to the main body of this document.

**Theorem 3.** If the design matrix \( X \in \mathbb{R}^{n \times p} \) satisfies the restricted eigenvalue condition \( \text{RE}(3, s) \) and \( \lambda = 6\sigma \sqrt{2 \log (p/\delta)}/n \) for some \( \delta \in (0, 1) \), then with probability \( 1 - \delta \) the following bound holds

\[
\frac{1}{n} \|X(\beta^* - \hat{\beta})\|_2^2 + \frac{1}{n} \|X(\beta^* - \hat{\beta})\|_2^2 \leq \lambda^2 s \left( \frac{9}{\kappa^2(3, s)} + \frac{16}{\kappa^2(1, s)} \right),
\]

where \( \hat{\beta} \) is a sign-consistent refitting of Lasso solution \( \hat{\beta} \).

**Proof.** Let \( \hat{\beta} \) be a sign-consistent refitting of \( \hat{\beta} = \hat{\beta}(\lambda) \) - Lasso solution with the tuning parameter \( \lambda \), we first define

\[
\Delta = \hat{\beta} - \hat{\beta}, \quad \hat{\Delta} = \beta^* - \hat{\beta}, \quad \hat{\Delta} = \beta^* - \hat{\beta}.
\]

We start with the KKT conditions for Lasso (Lemma 1). Since the refitting is sign-consistent we have \( \langle \hat{\beta}, \hat{\rho} \rangle = \|\hat{\beta}\|_1 \), therefore one can write

\[
\frac{1}{n} (\beta^* - \hat{\beta})^\top X^\top (y - X\hat{\beta}) = \lambda \langle \beta^* - \hat{\beta}, \hat{\rho} \rangle \leq \lambda (\|\beta^*\|_1 - \|\hat{\beta}\|_1) \quad \text{(8)},
\]

\[
\frac{1}{n} (\beta^* - \hat{\beta})^\top X^\top (y - X\hat{\beta}) = \lambda \langle \beta^* - \hat{\beta}, \hat{\rho} \rangle \leq \lambda (\|\beta^*\|_1 - \|\hat{\beta}\|_1) \quad \text{(9)}.
\]
Substituting the model in Equation (1), we get
\[
\frac{1}{n} (\beta^* - \hat{\beta})^\top X^\top X (\beta^* - \hat{\beta}) \leq \lambda (\|\beta^*\|_1 - \|\hat{\beta}\|_1) + \frac{1}{n} \varepsilon^\top X (\hat{\beta} - \beta^*) ,
\]
\[
\frac{1}{n} (\beta^* - \hat{\beta})^\top X^\top X (\beta^* - \hat{\beta}) \leq \lambda (\|\beta^*\|_1 - \|\hat{\beta}\|_1) + \frac{1}{n} \varepsilon^\top X (\hat{\beta} - \beta^*) .
\]

Therefore, we have three ingredients to derive our final result: the first one, given in Eq. (8), relies on the subgradient property of the \(\ell_1\)-norm applied to the Lasso solution; the second one, given in Eq. (9), relies on the same property applied to the sign-consistent refitting; and the third one is coming from the definition of a refitting strategy Definition 2.

\[
\frac{1}{n} \|X\vec{\Delta}\|_2^2 \leq \lambda (\|\beta^*\|_1 - \|\hat{\beta}\|_1) + \frac{1}{n} \varepsilon^\top X (\hat{\beta} - \beta^*) ,
\]
\[
\frac{1}{2n} \|X\hat{\Delta}\|_2^2 + \frac{1}{2n} \|X\hat{\Delta}\|_2^2 - \frac{1}{2n} \|X\Delta\|_2^2 \leq \lambda (\|\beta^*\|_1 - \|\hat{\beta}\|_1) + \frac{1}{n} \varepsilon^\top X (\hat{\beta} - \beta^*) ,
\]
\[
\frac{1}{2n} \|X\hat{\Delta}\|_2^2 - \frac{1}{2n} \|X\hat{\Delta}\|_2^2 \leq \frac{1}{n} \varepsilon^\top X (\hat{\beta} - \beta^*) .
\]

Multiplying the second inequality by 1/2 and summing the three equations we get
\[
\frac{3}{4n} \|X\hat{\Delta}\|_2^2 + \frac{3}{4n} \|X\Delta\|_2^2 \leq \frac{1}{4n} \|X\Delta\|_2^2 + \lambda \left( \frac{3}{2} \|\beta^*\|_1 - \frac{1}{2} \|\hat{\beta}\|_1 - \|\hat{\beta}\|_1 \right)
\]
\[
+ \frac{3}{2n} \varepsilon^\top X (\hat{\beta} - \beta^*) . \quad (10)
\]

Now observe that
\[
\|X\Delta\|_2^2 - \|X\hat{\Delta}\|_2^2 = \|X\Delta\|_2^2 - 2\hat{\Delta}^\top X^\top X \hat{\Delta}
\]
\[
\leq \|X\Delta\|_2^2 + 2\|X\hat{\Delta}\|_2 \|X\hat{\Delta}\|_2
\]
\[
\leq \|X\hat{\Delta}\|_2^2 + \|X\Delta\|_2^2 + \|X\hat{\Delta}\|_2^2 ,
\]
where we have again used \(2ab \leq a^2 + b^2\) in the last inequality. Then, subtracting \(\frac{3}{4n}\|X\hat{\Delta}\|_2^2\) from the both sides in Equation (10) and using previous inequality, we get
\[
\frac{1}{n} \|X\hat{\Delta}\|_2^2 + \frac{1}{n} \|X\Delta\|_2^2 \leq \lambda \left( 6 \|\beta^*\|_1 - 2\|\hat{\beta}\|_1 - 4\|\hat{\beta}\|_1 \right) + \frac{6}{n} \varepsilon^\top X (\hat{\beta} - \beta^*)
\]
\[
\leq \lambda \left( 6 \|\beta^*_S\|_1 - 2\|\hat{\beta}_S\|_1 + 2\|\Delta_S\|_1 - 2\|\Delta_S\|_1 - 4\|\hat{\beta}_S\|_1 + 4\|\Delta_S\|_1 - 4\|\Delta_S\|_1 \right) + \frac{6}{n} \varepsilon^\top X (\hat{\beta} - \beta^*)
\]
\[
= \lambda \left( 2\|\Delta_S\|_1 - 2\|\Delta_S\|_1 \right) + \lambda \left( 4\|\Delta_S\|_1 - 4\|\Delta_S\|_1 \right) + \frac{6}{n} \varepsilon^\top X (\hat{\beta} - \beta^*) ,
\]
where in the last inequality we used the fact that \(\|\beta\|_1 \geq \|\beta^*_S\|_1 - \|\beta^*_S - \beta_S\|_1 + \|\beta_S\|_1\) for any vector \(\beta\) and \(\beta^*\) with \(\text{supp}(\beta^*) = S\). Now let us restrict our attention on the event where \(\|\varepsilon^\top X/n\|_\infty \leq \lambda/6\), hence we can write
\[
\frac{1}{n} \|X\hat{\Delta}\|_2^2 + \frac{1}{n} \|X\Delta\|_2^2 \leq \lambda (3\|\Delta_S\|_1 - \|\Delta_S\|_1) + \lambda (4\|\Delta_S\|_1 - 4\|\Delta_S\|_1) .
\]
notice that both $4\|\hat{\Delta}_S\|_1 - 4\|\hat{\Delta}_{S'}\|_1$ and $3\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S'}\|_1$ cannot be negative simultaneously, if one of them is negative we can simply erase it. W.l.o.g. we can assume that both terms are positive, hence we have
\[
3\|\hat{\Delta}_S\|_1 \geq \|\hat{\Delta}_{S'}\|_1
\]
\[
\|\hat{\Delta}_S\|_1 \geq \|\hat{\Delta}_{S'}\|_1
\]
which allows us to write
\[
\frac{1}{n}\|X\hat{\Delta}\|_2^2 + \frac{1}{n}\|X\hat{\Delta}\|_2^2 \leq \lambda\{3\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S'}\|_1\} + \lambda\{4\|\hat{\Delta}_S\|_1 - 4\|\hat{\Delta}_{S'}\|_1\}
\leq 3\lambda\|\hat{\Delta}_S\|_1 + 4\lambda\|\hat{\Delta}_{S'}\|_1 \leq 3\lambda\sqrt{5}\|\hat{\Delta}_S\|_2 + 4\lambda\sqrt{5}\|\hat{\Delta}_{S'}\|_2
\leq \frac{3\lambda\sqrt{5}}{\kappa(3,s)\sqrt{n}}\|X\hat{\Delta}\|_2 + \frac{4\lambda\sqrt{5}}{\kappa(1,s)\sqrt{n}}\|X\hat{\Delta}\|_2
\leq \frac{9\lambda^2s}{2\kappa^2(3,s)} + \frac{1}{2n}\|X\hat{\Delta}\|_2^2 + \frac{16\lambda^2s}{2\kappa^2(1,s)} + \frac{1}{2n}\|X\hat{\Delta}\|_2^2
\]
again in the last inequality we used $2ab \leq a^2 + b^2$. Hence, we can write
\[
\frac{1}{2n}\|X(\beta^* - \hat{\beta})\|_2^2 + \frac{1}{2n}\|X(\beta^* - \hat{\beta})\|_2^2 \leq \lambda^2s\left(\frac{9}{2\kappa^2(3,s)} + \frac{8}{\kappa^2(1,s)}\right)
\] (11)
\[
\square
\]

4.2 Boosted Lasso

Notice that to prove Theorem 2, we used only the refitting property Definition 2, one of the possible candidates is the following refitting step
\[
\hat{\beta} \in \arg\min_{\beta \in \Gamma} \frac{1}{2n}\|y - X\beta\|_2^2
\] (12)
where $\Gamma = \{\beta \in \mathbb{R}^p : \|\beta - \hat{\beta}\|_1 \leq \delta \lambda\}$ and $\delta = |\text{supp}(\hat{\beta})|$. Intuitively, the Lasso coefficients are shrunk towards zero by a value proportional to the tuning parameter $\lambda$. Since there are $\delta$ non-zero coefficient in the Lasso solution, the proposed refitting strategy tries to "unshrink" $\delta$ non-zero coefficients. As we measure the "shrinkage" factor globally with the $\ell_1$-norm, it is natural (inspired by the orthogonal design) to set this factor as $\delta \lambda$. This motivates our choice of the feasible set $\Gamma$. This estimator is legitimate following results given in [Bickel et al., 2009, Theorem 7.2]:

**Theorem 4 (Bickel et al. [2009]).** If the design matrix $X \in \mathbb{R}^{n \times p}$ satisfies the restricted eigenvalue condition $RE(3, s)$ and $\lambda = 2\sigma\sqrt{2\log(p/\delta)/n}$ for some $\delta \in (0, 1)$, then with probability $1 - \delta$ the following bound holds
\[
\delta \leq \frac{9s\phi_{\max}}{\kappa^2(3,s)}
\] (13)
where $\phi_{\max}$ is the maximal eigenvalue of $X^\top X/n$.

The proof of this theorem is a direct application of Theorem 7.2 in [Bickel et al., 2009] together with Theorem 1 to bound $\frac{1}{n}\|X(\beta^* - \hat{\beta})\|_2^2$, we then omit it here. For the estimator, described in Equation (12), we can state the following result, which relies on using both Theorem 4 and Theorem 2:
Corollary 1. If the design matrix $X \in \mathbb{R}^{n \times p}$ satisfies the restricted eigenvalue condition $RE(3, s)$ and $\lambda = 2\sigma \sqrt{2 \log (p/\delta)/n}$ for some $\delta \in (0, 1)$, then with probability $1 - \delta$ the following bound holds
\[
\frac{1}{n} \|X(\beta^* - \hat{\beta})\|_2^2 \leq \frac{\lambda^2 \delta}{\kappa^2(3, s)} \left( \frac{9}{4} + 9\phi_{\text{max}} \right).
\]

Note that the control over the magnitudes of the coefficients allows to develop desirable rates for the refitted estimator. The description of the set $\Gamma$ in the definition of the refitting (12) motivates us to consider the following Boosted Lasso estimator [Bühlmann and Yu, 2003]:

**Definition 4** (Boosted Lasso). For any $\lambda_1, \lambda_2 > 0$ we call $\hat{\beta}^{\lambda_1, \lambda_2}$ a Boosted Lasso refitting if it is a solution of
\[
\hat{\beta}^{\lambda_1, \lambda_2} \in \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda_2 \|\beta - \hat{\beta}^{\lambda_1}\|_1,
\]
where $\hat{\beta}^{\lambda_1}$ is the Lasso solution with tuning parameter $\lambda_1$.

**Remark 2.** Indeed, procedure (15) is equivalent to Lasso boosting. To see that, let $\hat{\Delta} = \beta - \hat{\beta}$. Then, with a change of variable we get:
\[
\hat{\Delta} \in \arg\min_{\Delta \in \mathbb{R}^p} \frac{1}{2n} \|(y - X\hat{\beta}^{\lambda_1}) - X\Delta\|_2^2 + \lambda_2 \|\Delta\|_1,
\]
and finally $\hat{\beta}^{\lambda_1, \lambda_2} = \hat{\Delta} + \hat{\beta}^{\lambda_1}$.

It is known result that in the Lasso case there exists a critical value $\lambda_{1,\text{max}} = \|X^T y/n\|_\infty$ such that $\hat{\beta}^{\lambda_1} = 0$ iff $\lambda_1 \geq \lambda_{1,\text{max}}$, due to the previous remark, the Boosted Lasso can be written as a Lasso problem and we can give the result of the same nature:

**Proposition 1.** If $\lambda_1 < \lambda_{1,\text{max}}$, then the solution of Boosted Lasso satisfies: $\hat{\beta}^{\lambda_1, \lambda_2} = \hat{\beta}^{\lambda_1}$ iff $\lambda_2 \geq \lambda_1$. Moreover, if $\lambda_1 \geq \lambda_{1,\text{max}}$, then the Boosted Lasso estimator is simply the Lasso estimator with tuning parameter $\lambda_2$.

**Proof.** For the second part we notice that $\hat{\beta}^{\lambda_1} = 0$ (since $\lambda_1 \geq \lambda_{1,\text{max}}$) and the statement follows from Equation (15) with $\hat{\beta}^{\lambda_1} = 0$. For the first part we can compute the critical value for Equation (15) (it is a Lasso problem), which is given by $\lambda_{2,\text{max}} = \|X^T \hat{y}/n\|_\infty$, where $\hat{y} = y - X\hat{\beta}^{\lambda_1}$. Notice that thanks to Lemma 1 we can write
\[
\lambda_{2,\text{max}} = \left\| \frac{1}{n} X^T \hat{y} \right\|_\infty = \left\| \frac{1}{n} X^T (y - X\hat{\beta}^{\lambda_1}) \right\|_\infty = \lambda_1 \|\hat{\beta}^{\lambda_1}\|_\infty,
\]
for some $\hat{\beta}^{\lambda_1} \in \text{sign}(\hat{\beta}^{\lambda_1})$, since $\lambda_1 < \lambda_{1,\text{max}}$, hence $\hat{\beta}^{\lambda_1} \neq 0$ and $\|\hat{\beta}^{\lambda_1}\|_\infty = 1$, which concludes the proof.

Control over the $\ell_1$-norm of the difference, allows to develop oracle inequality with minimax rates (Corollary 1), similar result but somehow stronger can be shown for the regularized version (Boosted Lasso):

**Lemma 3.** For any $\hat{\beta}^{\lambda_1}$ on the event $\mathcal{A} = \{\|e^T X/n\|_\infty \leq \lambda_1/2\}$ we have:
\[
\frac{1}{n} \|X(\beta^* - \hat{\beta}^{\lambda_1, \lambda_2})\|_2^2 + (2\lambda_2 - \lambda_1) \|\hat{\beta}^{\lambda_1, \lambda_2} - \hat{\beta}^{\lambda_1}\|_1 \leq \frac{1}{n} \|X(\beta^* - \hat{\beta}^{\lambda_1})\|_2^2.
\]
Proof. By optimality of $\hat{\beta}^{\lambda_1,\lambda_2}$ in Eq. (14)
\[
\frac{1}{2n} \| y - X\hat{\beta}^{\lambda_1,\lambda_2} \|_2^2 + \lambda_2 \| \hat{\beta}^{\lambda_1,\lambda_2} - \hat{\beta}^{\lambda_1} \|_1 \leq \frac{1}{2n} \| y - X\hat{\beta}^{\lambda_1} \|_2^2 ,
\]
hence on the event $\| e^T X/n \|_\infty \leq \lambda_1/2$ and using Hölder’s inequality, we can write
\[
\frac{1}{2n} \| X(\beta^* - \hat{\beta}^{\lambda_1,\lambda_2}) \|_2^2 + \lambda_2 \| \hat{\beta}^{\lambda_1,\lambda_2} - \hat{\beta}^{\lambda_1} \|_1 \leq \frac{1}{2n} \| X(\beta^* - \hat{\beta}^{\lambda_1}) \|_2^2 + \frac{1}{n} e^T X(\hat{\beta}^{\lambda_1,\lambda_2} - \hat{\beta}^{\lambda_1})
\]
\[
\leq \frac{1}{2n} \| X(\beta^* - \hat{\beta}^{\lambda_1}) \|_2^2 + \frac{1}{2} \| \hat{\beta}^{\lambda_1,\lambda_2} - \hat{\beta}^{\lambda_1} \|_1 .
\]

Combining Lemma 3 and Theorem 1 we can state the following corollary

Corollary 2. If the design matrix $X \in \mathbb{R}^{n \times p}$ satisfies the restricted eigenvalue condition $RE(3, s)$ and
\[\lambda_1 = 2\sigma \sqrt{2 \log (p/\delta)/n}\] for some $\delta \in (0, 1)$, then with probability $1 - \delta$ the following bound holds
\[
\frac{1}{n} \| X(\beta^* - \hat{\beta}) \|_2^2 + (2\lambda_2 - \lambda_1) \| \hat{\beta} - \hat{\beta}^{\lambda_1} \|_1 \leq \frac{9\lambda_1^2 s}{4k^2(3, s)} ,
\]
where $\hat{\beta} = \hat{\beta}^{\lambda_1,\lambda_2}$ is a boosting refitting of Lasso solution $\hat{\beta} = \hat{\beta}^{\lambda_1}$

The Boosted Lasso is obviously a refitting strategy, and the previous result shows that it is worth
refitting in terms of prediction error. Moreover, Proposition 1 and Corollary 2 suggest to select $\lambda_2 \in (\lambda_1/2, \lambda_1)$, such a choice allows to improve the Lasso prediction accuracy with high probability. Besides,
the result can be applied iteratively as it does not depend on the choice of $\hat{\beta}^{\lambda_1}$, hence it works for any
iteration step.

4.3 Bregman Lasso

We first remind the definition of the Bregman divergence associated with the $\ell_1$-norm

Definition 5 (Bregman divergence for the $\ell_1$-norm). For any $z, w \in \mathbb{R}^p$ and any $\rho \in \partial \| w \|_1$, the Bregman divergence for the $\ell_1$-norm is defined as
\[
D^\rho_{\ell_1}(z, w) = \| z \|_1 - \| w \|_1 - \langle \rho, z - w \rangle \geq 0, \quad \rho \in \partial \| w \|_1 .
\]  

In [Osher et al., 2005], the authors proposed the Bregman Iterations procedure, originally designed to
improve iso-TV results. In Lasso case, Bregman Iterations has the following expression: for a fixed $\lambda > 0$
\[
\hat{\rho}_0 = \hat{\beta}_0 = 0 \rho ,
\]
\[
\hat{\beta}_k = \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \| y - X\beta \|_2^2 + \lambda D^\hat{\beta}_{\ell_1} (\beta, \hat{\beta}_{k-1}) ,
\]
\[
\hat{\rho}_k = \hat{\rho}_{k-1} + \frac{1}{\lambda n} X^T (y - X \hat{\beta}_k) ,
\]
where $D^\rho_{\ell_1}(\cdot, \cdot)$ is the Bregman divergence defined in Eq. (17). The Bregman Iterations can be seen as a
discretization of Bregman Inverse Scale Space (ISS), which is analyzed by Osher et al. [2016], where
authors provided statistical guarantees for the ISS dynamic. One of the drawbacks of such an approach is
the iterative nature of the algorithm: one need to tune the number of iterations $k$ and the regularization
parameter $\lambda > 0$. In recent work of Brinkmann et al. [2016], authors proposed another closely related algorithm, which performs, for a given Lasso solution $\hat{\beta}^{\lambda_1}$, the following refitting

$$
\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \|y - X\beta\|_2^2
$$

$$
s.t. \ D_{\ell_1}^{\hat{\beta}^{\lambda_1}} (\beta, \hat{\beta}^{\lambda_1}) = 0 .
$$

Unlike the previous approaches, in this section we consider the Lasso solution $\hat{\beta}^{\lambda_1}$ and the following Bregman Lasso refitting strategy, defined as

**Definition 6 (Bregman Lasso).** For any $\lambda_1, \lambda_2 > 0$ we call $\hat{\beta}^{\lambda_1, \lambda_2}$ a Bregman Lasso refitting if it is a solution of

$$
\hat{\beta}^{\lambda_1, \lambda_2} \in \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda_2 D_{\ell_1}^{\hat{\beta}^{\lambda_1}} (\beta, \hat{\beta}^{\lambda_1}) ,
$$

where $\hat{\beta}^{\lambda_1}$ is the Lasso solution with tuning parameter $\lambda_1$ and $D_{\ell_1}^{\hat{\beta}^{\lambda_1}} (\beta, \hat{\beta}^{\lambda_1})$ is the Bregman divergence given in Equation (17).

Considering the particular case of orthogonal design, we show in Section 4.3.2 that Bregman Lasso refitting is a generalization of both approaches. In particular for this design, the Bregman Lasso refitting can recover Bregman Iterations for any $k$. Also, Bregman Lasso is computationally more appealing since it only requires evaluating two Lasso problems, while the Bregman Iterations would require $k$ evaluations.

We start by introducing some basic properties of the Bregman divergence associated with the $\ell_1$-norm.

**Lemma 4.** Let $z, w \in \mathbb{R}^p$, and denote $\rho \in \partial \|w\|_1$ a subgradient of the $\ell_1$-norm evaluated at $w \in \mathbb{R}^p$, therefore the following properties hold independent of the choice of $\rho$

1. $z \mapsto D_{\ell_1}^{\rho}(z, w)$ is convex for all $w$,
2. $D_{\ell_1}^{\rho}(z, w) = \|z\|_1 - \langle \rho, z \rangle$,
3. $0 \leq D_{\ell_1}^{\rho}(z, w) \leq 2\|z\|_1$,
4. $D_{\ell_1}^{\rho}(z, w) = \sum_{i=1}^p \|z_i\|_1 - z_i \rho_i = \sum_{i=1}^p D_{\ell_1}^{\rho}(z_i, w_i)$, where $z_i, w_i, \rho_i$ are the $i^{th}$ components of $z, w, \rho$,
5. If $\text{sign}(z) = \text{sign}(w)$, therefore $D_{\ell_1}^{\rho}(z, w) = 0$.

**Proof.** See supplementary material for details.

There is a simple geometrical interpretation of Bregman distance, associated with the $\ell_1$-norm, which is illustrated by Figure 1. According to Property 2 of Lemma 4, Bregman divergence in Equation (20) can be evaluated as

$$
D_{\ell_1}^{\hat{\beta}^{\lambda_1}} (\beta, \hat{\beta}^{\lambda_1}) = \|\beta\|_1 - \langle \hat{\beta}^{\lambda_1}, \beta \rangle ,
$$

where $\hat{\beta}^{\lambda_1}$ is a fixed subgradient of the $\ell_1$-norm evaluated at $\hat{\beta}^{\lambda_1}$. Since, the subdifferential of the $\ell_1$-norm is not uniquely defined when evaluated at zero, it is important to fix the way to pick a subgradient. Possible, and probably the most obvious way to evaluate the subgradient is to write the KKT conditions for Problem (20) and fix $\hat{\beta}^{\lambda_1}$ as follows

$$
\hat{\beta}^{\lambda_1} = \frac{1}{\lambda_1 n} X^\top (y - X\hat{\beta}^{\lambda_1}) .
$$

Starting from here we stick to this choice of the subgradient.
Figure 1: Geometrical interpretation of Bregman divergence. On the left plot the subgradient is uniquely defined. On the right plot the subgradient is a set, therefore the Bregman divergence can be any number from 0 to $2|z|$ depending on the choice of $\rho$.

**Proposition 2.** With the choice of the subgradient as in Equation (21), the Bregman Lasso refitting step can be evaluated as

$$\hat{\beta}^{\lambda_1, \lambda_2} \in \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \| \bar{y} - X\beta \|_2^2 + \lambda_2 \| \beta \|_1 ,$$

where $\bar{y} = y + \frac{\lambda_1}{\lambda_2} (y - X\hat{\beta}^{\lambda_1})$.

**Proof.** One can write the following sequence of equalities

$$\frac{1}{2n} \| y - X\beta \|_2^2 + \lambda_2 D^{\phi_{\lambda_1}}_{\ell_1} (\beta, \hat{\beta}^{\lambda_1}) = \frac{1}{2n} \| y \|_2^2 - \frac{1}{n} \langle X^T y, \beta \rangle + \frac{1}{2n} \| X\beta \|_2^2 + \lambda_2 D^{\phi_{\lambda_1}}_{\ell_1} (\beta, \hat{\beta}^{\lambda_1})$$

$$= \frac{1}{2n} \| y \|_2^2 - \frac{1}{n} \langle X^T \bar{y}, \beta \rangle + \frac{1}{2n} \| X\beta \|_2^2 + \lambda_2 \| \beta \|_1 ,$$

where to get the last equality we used Property 2 and the choice of subgradient given in Equation (21). Noticing the following relation

$$\frac{1}{2n} \| y \|_2^2 - \frac{1}{n} \langle X^T \bar{y}, \beta \rangle + \frac{1}{2n} \| X\beta \|_2^2 + \lambda_2 \| \beta \|_1$$

$$= \frac{1}{2n} \| \bar{y} \|_2^2 - \frac{1}{2n} \| \bar{y} \|_2^2 ,$$

and using the fact that $y, \bar{y}$ are independent of $\beta$, we conclude.

**Lemma 5.** Bregman Lasso refitting $\hat{\beta}^{\lambda_1, \lambda_2}$ is a refitting strategy in the sense of Definition 2.

**Proof.** The proof of this lemma relies on Equation (20), hence we can write

$$\frac{1}{2n} \| y - X\hat{\beta}^{\lambda_1, \lambda_2} \|_2^2 + \lambda_2 D^{\phi_{\lambda_1}}_{\ell_1} (\hat{\beta}^{\lambda_1, \lambda_2}, \hat{\beta}^{\lambda_1}) \leq \frac{1}{2n} \| y - X\hat{\beta}^{\lambda_1} \|_2^2 ,$$

so that we get the result since Bregman divergences are non negative.
4.3.1 Geometrical interpretation

It is known that (3) can be equivalently written in the following form
\[
\hat{\beta}_{T_1} \in \arg\min_{\hat{\beta} \in \mathbb{R}^p} \frac{1}{2n} \|y - X\hat{\beta}\|_2^2 \\
\text{s.t.} \quad \|\hat{\beta}\|_1 \leq T_1,
\]
for some \( T_1 = T_1(\lambda_1) \geq 0 \). Similarly, we can write the Bregman Lasso refitting step Equation (20) in the constrained form as
\[
\hat{\beta}_{T_1,T_2} \in \arg\min_{\hat{\beta} \in \mathbb{R}^p} \frac{1}{2n} \|y - X\hat{\beta}\|_2^2 \\
\text{s.t.} \quad D_{\ell_1}^{\hat{\beta}_{T_1}}(\beta, \hat{\beta}_{T_1}) \leq T_2,
\]
where \( \hat{\beta}_{T_1} \) is a solution of the constrained version given in Equation (24). The choice \( T_2 = 0 \) is considered in [Brinkmann et al., 2016], which corresponds to \( \lambda_2 \) being large enough in Equation (20). To provide a geometrical intuition we consider the simple case where \( n \geq p \).

• When \( \hat{\beta}_{T_2} > 0 > \hat{\beta}_{T_1} \), the feasible set of Problem (25) is given by
\[
\{((\beta_1, \beta_2)^T \in \mathbb{R}^2 : |\beta_1| - \beta_1 + |\beta_2| - \hat{\beta}_{T_2} \beta_2 \leq T_2\}.
\]
Notice that any positive value for \( \beta_1 \) is admissible in this case. We illustrate this phenomenon on Figure 2a, Figure 2b and Figure 2c for various values of \( T_2 \).

• When \( \hat{\beta}_{T_1} > 0 > \hat{\beta}_{T_2} > 0 \), the feasible set of Problem (25) is given by
\[
\{((\beta_1, \beta_2)^T \in \mathbb{R}^2 : |\beta_1| - \beta_1 + |\beta_2| - \beta_2 \leq T_2\}.
\]
Notice that every pair \( (\beta_1, \beta_2)^T \in \mathbb{R}^2 \) is inside of the feasible set. We illustrate this phenomenon on Figure 2d.

In Lasso case, if \( T_1 = 0 \) there is only one trivial solution \( \hat{\beta}_{T_1} = 0 \), this is not the case for the refitting step. Indeed, if \( T_2 = 0 \) the feasible set of the refitting step might be non-trivial, depending on the first step Lasso solution. More precisely, if there exists \( j_1 \in [p] \), such that \( \hat{\beta}_{j_1} \neq 0 \) the feasible set of the refitting step always contains \{\( \beta : \beta_{j_1} \leq 0, \beta_{j_2} = 0 \forall j_2 \neq j_1 \) or \{\( \beta : \beta_{j_1} \geq 0, \beta_{j_2} = 0 \forall j_2 \neq j_1 \), depending on the sign of \( \beta_{j_1} \).

4.3.2 Orthogonal design

In this section we investigate some important properties of the algorithm in Eq. (20) for the denoising model (i.e., we drop the statistical convention and instead assume that \( n = p \) and \( X = \text{Id}_p \))
\[
y = \beta^* + \epsilon,
\]
where \( y, \beta^* \in \mathbb{R}^p, \epsilon \sim \mathcal{N}(0, \sigma^2 \text{Id}_p) \). The following estimators correspond to Eq. (3) and Eq. (20) respectively
\[
\hat{\beta}^{\lambda_1} = \arg\min_{\hat{\beta} \in \mathbb{R}^p} \frac{1}{2} \|y - \hat{\beta}\|_2^2 + \lambda_1 \|\hat{\beta}\|_1,
\]
\[
\hat{\beta}^{\lambda_1,\lambda_2} = \arg\min_{\hat{\beta} \in \mathbb{R}^p} \frac{1}{2} \|y - \hat{\beta}\|_2^2 + \lambda_2 \left(\|\hat{\beta}\|_1 - \langle \hat{\beta}^{\lambda_1}, \beta \rangle\right),
\]

12
Feasible set of the refitting step compared to $\ell_1$-ball

(a) Sparse Lasso solution, $T_2$ - small.
(b) Sparse Lasso solution, $T_2$ - large.
(c) Sparse Lasso solution, $T_2 = 0$.
(d) Dense Lasso solution.

Figure 2: Feasible set of the first and the refitting steps. Here, the ellipses are levels of the objective function and $\hat{\beta}^{LS}$ is the least squares estimator. Geometry of the feasible set for the refitting step is described in terms of $T_2$ and the subgradient $\hat{\rho}^{T_1}$.

where the subgradient $\hat{\rho}^{\lambda_1}$ is given by Eq. (21) and simplifies to

$$\hat{\rho}^{\lambda_1} = \frac{y - \hat{\beta}^{\lambda_1}}{\lambda_1}. \quad (29)$$

We remind that the solution $\hat{\beta}^{\lambda_1}$ of Lasso with orthogonal design given in Eq. (27) relies on the soft-threshold operator $\text{ST}(\cdot, \cdot)$, defined component-wise for any $j \in [p]$ by

$$\text{ST}(y, \lambda_1)_j = \text{sign}(y_j)(|y_j| - \lambda_1)_+ \quad (30)$$

The next proposition shows that the subgradient follows the signs of the observed signal $y$.

**Proposition 3.** For all $\lambda_1 > 0$ we have $\text{sign}(y_j) = \text{sign}(\hat{\rho}^{\lambda_1}_j), \forall j \in [p]$. 

13
Figure 3: The solution of the refitting step is equivalent to the MCP penalty. Extreme cases include both hard and soft threshold operators.

Proof. Since \( \hat{\beta}^{\lambda} = \text{ST}(y, \lambda_1) \), we can equivalently rewrite Eq. (29) as

\[
\forall i \in [p] \quad \hat{\rho}_i^{\lambda} = \begin{cases} 
\text{sign}(y_i), & |y_i| \geq \lambda_1 \\
\frac{y_i}{\lambda_1}, & |y_i| < \lambda_1
\end{cases}.
\]

(31)

The following results states that the Bregman refitting also enjoys close properties to the Lasso but thresholds a translated response vector. This relation can be obviously proved using Equation (22) of Proposition 2, which establishes that the Bregman Lasso can be seen as Lasso solution applied to a modified signal.

Proposition 4. The solution of the Bregman Lasso refitting step in Equation (28) relies on the soft-threshold operator and reads:

\[
\hat{\beta}^{\lambda_1, \lambda_2} = \text{ST}(y + \lambda_2 \hat{\rho}^{\lambda_1}, \lambda_2).
\]

It is worth mentioning, that in the Lasso case, there exists a so called \( \lambda_{1,\text{max}} \), which is the smallest value of regularization parameter for which the solution \( \hat{\beta}^{\lambda_1} = 0 \) for all \( \lambda_1 > \lambda_{1,\text{max}} \). However, even though the refitting step can be formulated as Lasso problem, there is not such parameter \( \lambda_{2,\text{max}} \). It can be counter intuitive on the first sight, but since the parameter \( \lambda_2 \) is present inside the data-fitting term, it becomes clear that such extreme value does not exist.

In the sequel we provide another interpretation of the Bregman Lasso refitting in terms of firm-thresholding operator, introduced and analyzed by Gao and Bruce [1997]. We additionally mention, that the firm-thresholding operator is the solution of the least squares problem penalized with MCP regularization [Zhang, 2010] (for orthogonal design), which is a non-convex problem. Additionally, the firm-thresholding operator outperforms soft/hard-threshold in terms of bias-variance trade-off, see [Gao and Bruce, 1997] for the theoretical and numerical analysis.

Proposition 5. The solution of the Bregman Lasso refitting step in case of orthogonal design is given by

\[
\hat{\beta}^{\lambda_1, \lambda_2} = \text{FT}(y, \lambda_H, 1 + \frac{\lambda_1}{\lambda_2})
\]

where \( \lambda_H = (1/\lambda_1 + 1/\lambda_2)^{-1} \) and the firm-thresholding operator FT is defined for \( \mu > 0, \gamma > 1 \) component-wise for any \( j \in [p] \) by:

\[
\text{FT}(y, \mu, \gamma)_j = \begin{cases} 
\frac{\gamma}{\gamma - 1} \text{ST}(y_j, \mu), & |y_j| \leq \mu \gamma \\
y_j, & |y_j| > \mu \gamma
\end{cases}.
\]
In the Bregman Lasso (orthogonal) case \( \mu = \lambda_H \) and \( \gamma = 1 + \lambda_1/\lambda_2 = \lambda_1/\lambda_H \).

**Proof.** First we notice that the optimization problem in Equation (28) is separable i.e., can be solved component-wise and that \( \mu \gamma = \lambda_H (1 + \lambda_1/\lambda_2) = \lambda_1 \) and \( \gamma - 1 = \frac{1 + \lambda_1/\lambda_2 - 1}{(1 + \lambda_1/\lambda_2)} = 1 + \lambda_2/\lambda_1 \). For each \( j \in [p] \) the solution of the refitting step Eq. (20) is given by

\[
\hat{\beta}_j^{\lambda_1,\lambda_2} = \text{sign}(y_j) \left( \left| y_j + \lambda_2 \frac{\lambda_1}{\lambda_2} (y_j - \hat{\beta}_j^{\lambda_1}) \right| - \lambda_2 \right)_+, 
\]

where we used the fact that \( \text{sign}(y_j) = \text{sign}(y_j + \lambda_2 \hat{\beta}_j^{\lambda_1}) \), a result that follows from Proposition 3. We consider the following cases:

- \( y_j > 0, y_j \geq \lambda_1 \), hence \( \hat{\beta}_j^{\lambda_1} = y_j - \lambda_1 \) and

\[
\text{sign}(y_j) \left( \left| y_j + \lambda_2 \frac{\lambda_1}{\lambda_2} (y_j - \hat{\beta}_j^{\lambda_1}) \right| - \lambda_2 \right)_+ = \left( y_j + \lambda_2 \frac{\lambda_1}{\lambda_2} (y_j - y_j + \lambda_1) - \lambda_2 \right)_+ = y_j + \lambda_2 \frac{\lambda_1}{\lambda_2} (y_j - \lambda_H) + \lambda_2,
\]

which holds for all positive values of \( \lambda_2 \).

- \( y_j > 0, y_j < \lambda_1 \), hence \( \hat{\beta}_j^{\lambda_1} = 0 \) and

\[
\text{sign}(y_j) \left( \left| y_j + \lambda_2 \frac{\lambda_1}{\lambda_2} (y_j - \hat{\beta}_j^{\lambda_1}) \right| - \lambda_2 \right)_+ = \left( y_j + \lambda_2 \frac{\lambda_1}{\lambda_2} y_j - \lambda_2 \right)_+ = \left( y_j (1 + \frac{\lambda_2}{\lambda_1}) - \lambda_2 \right)_+ = \left( 1 + \frac{\lambda_2}{\lambda_1} \right) (y_j - \lambda_H)_+.
\]

The case of negative \( y_j \) is proved in the same manner. \( \Box \)

**Remark 3.** We notice the following behavior of the solution, described in Proposition 5:

- \( \hat{\beta}_j^{\lambda_1,\lambda_2} \to \text{HT}(y, \lambda_1) \), when \( \lambda_2 \to \infty \),

- \( \hat{\beta}_j^{\lambda_1,\lambda_2} \to \text{ST}(y, \lambda_2) \), when \( \lambda_1 \to \infty \).

These properties are illustrated on Figures 3a and 3b for several values of \( \lambda_1 \) and \( \lambda_2 \).

For the sake of completeness, we additionally provide the analysis for the Bregman Iterations (orthogonal case) in the form of Equation (18) to give insights on our motivation to consider 2-step iteration with two tuning parameters \( \lambda_1, \lambda_2 \). Similar analysis can be found in [Yin et al., 2008, Xu and Osher, 2007]. Following the proof of Proposition 4 one can prove by induction that the following result holds for the Bregman Iterations defined in Equation (18).

**Proposition 6.** For a fixed \( \lambda > 0 \) and for all \( k > 0 \), the solution generated by the Bregman Iterations Eq. (18) is given by

\[
\hat{\beta}_{k+1} = \text{ST}(y + \lambda \hat{\rho}_k, \lambda).
\]

We can additionally prove that for every \( k > 1 \), the solution of Bregman Iterations is given by the firm-threshold operator.
Proposition 7. For a fixed $\lambda > 0$ and for all $k > 1$, the solution generated by the Bregman Iterations Eq. (18) is given by

$$\hat{\beta}_{k+1} = \text{FT}(y, \frac{\lambda}{k+1}, \frac{k+1}{k}) = \begin{cases} (k+1)\text{ST}(y, \frac{1}{k+1}), & \text{if } |y| \leq \frac{1}{k} \\ y_j & \text{if } |y| > \frac{1}{k} \end{cases}.$$  

Proof. We prove this result component-wise (for arbitrary component $j \in [p]$) and additionally assume that $y_j > 0$, the case of negative $y_j$ being similar. The statement clearly holds for $k = 2$, since it is sufficient to use Proposition 5 with $\lambda_1 = \lambda_2$. We assume that the statement holds up to $k > 2$, hence using previous result we can write

$$\hat{\beta}_{k+1,j} = \text{ST}(y_j + \lambda \hat{\beta}_{k,j}, \lambda),$$

reminding that $\rho_0 = 0$, with the inductive assumption and the definition of the subgradient in Eq. (18) we have

$$\hat{\beta}_{k,j} = \frac{1}{\lambda}(y_j - \text{ST}(y_j, \lambda) + \sum_{m=2}^{k} (y_j - m\text{ST}(y_j, \frac{1}{m}))) = \frac{1}{\lambda}(y_j + (k-1)y_j) = \frac{k}{\lambda}y_j.$$

Therefore, when $\frac{1}{k} < y_j \leq \frac{1}{k+1}$, we have $\hat{\beta}_{k+1,j} = \text{ST}(y_j + \lambda \hat{\beta}_{k,j}, \lambda) = y_j = \text{FT}(y, \frac{\lambda}{k+1}, \frac{k+1}{k}).$

If $0 < y_j \leq \frac{1}{k}$ it means that for all $m \leq k$ iterations $\hat{\beta}_{m,j} = 0$, hence

$$\hat{\beta}_{k,j} = \frac{1}{\lambda}(y_j - \text{ST}(y_j, \lambda) + \sum_{m=2}^{k} (y_j - m\text{ST}(y_j, \frac{1}{m}))) = \frac{1}{\lambda}(y_j + (k-1)y_j) = \frac{k}{\lambda}y_j.$$

Therefore, when $0 < y_j \leq \frac{1}{k}$, we have $\hat{\beta}_{k+1,j} = \text{ST}(y_j + ky_j, \lambda) = (y_j + ky_j - \lambda)_+ = (k+1)(y_j - \frac{1}{k+1})_+ = \text{FT}(y, \frac{1}{k+1}, \frac{k+1}{k}).$

If $\frac{1}{m^*} < y_j \leq \frac{1}{m-1}$, for some $2 \leq m^* < k$ we know by the induction assumption that for all $m > m^*$ the estimation is given by $\hat{\beta}_{m,j} = y_j$, for all $m < m^*$ the estimation is given by $\hat{\beta}_{m,j} = 0$ and for $m = m^*$ we have $\hat{\beta}_{m^*,j} = m^*(y_j - \frac{1}{m^*})$, hence we can write

$$\hat{\beta}_{k,j} = \frac{1}{\lambda}(y_j - \text{ST}(y_j, \lambda) + \sum_{m=2}^{m^*} (y_j - m\text{ST}(y_j, \frac{1}{m}))) = \frac{1}{\lambda}(y_j + \sum_{m=2}^{m^*} y_j - m^*(y_j - \frac{1}{m^*})) = 1.$$

Comparing with the firm-threshold definition provide the expected result. \qed

Analyzing the previous result helps motivating the introduction of the Bregman Lasso refitting in the form of Eq. (20). Indeed, for orthogonal design, Bregman Lasso refitting generalizes Bregman Iterations Eq. (18) in the sense that for each pair $(k, \lambda)$ there exists a pair $(\lambda_1 = \frac{1}{k+1}, \lambda_2 = \lambda)$ such that $\hat{\beta}_k = \hat{\beta}_{\lambda_1, \lambda_2}$. In particular, Bregman Lasso refitting includes all possible solution of Bregman Iterations Eq. (18), while the reverse is not true, for instance when $k$ is not an integer.
### 4.3.3 Arbitrary design

In this section we provide some generalizations of the results obtained in Section 4.3.2. We first notice that in the case of orthogonal design, discussed in Section 4.3.2, for a given vector \( y \), there exists a value of the regularization parameter \( \lambda_2 \) such that, the refitting step Eq. (20) is equivalent to hard-thresholding. One might expect that for the case of arbitrary design there exists such a value of \( \lambda_2 \) that the refitting step Eq. (20) is equivalent to a least-squares refitting on the support obtained via the first lasso type step as it adds a sign constraints. Yet, the estimator obtained via the refitting step slightly differs from the simple least-squares refitting. To state our main result of this section, let us introduce some notation.

Consider \( \hat{\rho}^{\lambda_1} \), defined in Eq. (21), we define the equicorrelation set [Tibshirani, 2013] as

\[
E^{\lambda_1} = \{ j \in [p] : |\hat{\rho}_j^{\lambda_1}| = 1 \} .
\]

(32)

We will omit \( \lambda_1 \) in \( E^{\lambda_1} \) and write \( E \) instead for simplicity. We associate the following Sign-Least-Squares Lasso refitting step with the equicorrelation set given by Eq. (32)

**Definition 7 (Sign-Least-Squares Lasso).** For any \( \lambda_1 > 0 \) we call \( \hat{\beta}^{\text{SLS}} \) a Sign-Least-Squares Lasso refitting if it satisfies

\[
\begin{align*}
\hat{\beta}^{\text{SLS}}_E = 0 \\
\hat{\beta}^{\text{SLS}}_E \in \underset{\beta \in \mathbb{R}^{|E|} : \hat{\rho}^{\lambda_1} \circ \beta \geq 0}{\text{arg min}} \frac{1}{2n} \|y - X_E \beta_E\|_2^2,
\end{align*}
\]

where \( |E| \) is the cardinality of \( E \) and \( \hat{\rho}^{\lambda_1} \) is the Lasso subgradient defined in Equation (21).

Notice, that the refitting is performed on the equicorrelation set and not on the support of the Lasso solution. Possible motivation to consider the equicorrelation set instead of the support can be described by uniqueness issues of the Lasso [Tibshirani, 2013], while the equicorrelation set (and the signs) is always uniquely defined.

**Proposition 8.** The Sign-Least-Squares Lasso is a sign consistent refitting strategy of the Lasso solution \( \hat{\beta}^{\lambda_1} \) in the sense of Definition 3.

**Proof.** Notice that we have the following relation

\[
\{ \beta_E \in \mathbb{R}^{|E|} : \hat{\rho}^{\lambda_1} \circ \beta_E \geq 0 \} = \{ \beta_E \in \mathbb{R}^{|E|} : \hat{\rho}^{\lambda_1} \in \text{sign}(\beta) \},
\]

and using Remark 1 we conclude. \( \square \)

### 4.3.4 Bregman Lasso as Sign-Least-Squares Lasso

In this section we provide connections between the Bregman Lasso and the Sign-Least-Squares Lasso. Since Problem (33) is convex and Slater’s condition is satisfied therefore the KKT conditions state [Boyd and Vandenberghe, 2004] that there exist \( \hat{\beta}^{\text{SLS}}_E \in \mathbb{R}^{|E|} \) and \( \hat{\mu} \in \mathbb{R}^{|E|} \), such that

\[
\begin{align*}
\frac{1}{n} X_E^\top(y - X_E \hat{\beta}^{\text{SLS}}_E) + \hat{\rho}^{\lambda_1} \circ \hat{\mu} &= 0 \\
\hat{\mu} \circ \hat{\beta}^{\text{SLS}}_E \circ \hat{\rho}^{\lambda_1} &= 0 \\
\hat{\mu} &\geq 0
\end{align*}
\]

(34)

so that we can provide the following connection:

**Theorem 5.** For any signal vector \( y \in \mathbb{R}^n \) and any design matrix \( X \in \mathbb{R}^{n \times p} \) there exists \( \lambda_0 > 0 \) such that for any \( \lambda_2 \geq \lambda_0 \) the solution of the refitting step Eq. (20) is given by \( \hat{\beta}^{\lambda_1, \lambda_2} = \hat{\beta}^{\text{SLS}}. \)
The proof the this theorem is below. However we mention that it states that given $\lambda_1 > 0$, Bregman Lasso and Sign-Least-Squares Lasso often coincide. The value of the parameter $\lambda_0$ in Theorem 5 is explicit and can be found at the end of the following proof.

**Proof.** We start by writing the KKT conditions for Bregman Lasso, Eq. (20):

\[
\begin{cases}
\hat{\beta}_E^{\lambda_1} + \frac{1}{\lambda_2} X_E^T (y - X_E \hat{\beta}_E^{\lambda_1}) \in \text{sign}(\hat{\beta}_E^{\lambda_1}) \\
\hat{\beta}_E^{\lambda_1} + \frac{1}{\lambda_2} X_E^T (y - X_E \hat{\beta}_E^{\lambda_1}) \in \text{sign}(\hat{\beta}_E^{\lambda_1})
\end{cases}
\]  

These are necessary and sufficient conditions for a vector $\beta$ to be a solution of Bregman Lasso (20). Therefore, it is sufficient to check if the vector defined as $\hat{\beta}_{\lambda_1, \lambda_2} = \hat{\beta}_{\text{SLS}}$ satisfies Eq. (35) for some $\lambda_2$. Substituting $\hat{\beta}_E^{\lambda_1, \lambda_2} = \hat{\beta}_E^{\text{SLS}}$ and $\hat{\beta}_E^{\lambda_1, \lambda_2} = 0$ we arrive at

\[
\begin{cases}
\left(\hat{\beta}_E^{\lambda_1} + \frac{1}{\lambda_2} X_E^T (y - X_E \hat{\beta}_E^{\lambda_1}) \in \text{sign}(\hat{\beta}_E^{\lambda_1})
\end{cases}
\]  

Since $\hat{\beta}_E^{\text{SLS}}$ satisfies the KKT conditions, given in Eq. (34), we simplify the previous system as

\[
\begin{cases}
\hat{\beta}_E^{\lambda_1} \circ (1_{|E|} - \frac{1}{\lambda_2} \mu) \in \text{sign}(\hat{\beta}_E^{\text{SLS}}) \\
\hat{\beta}_E^{\lambda_1} + \frac{1}{\lambda_2} X_E^T (y - X_E \hat{\beta}_E^{\text{SLS}}) \in \text{sign}(0)
\end{cases}
\]  

where $1_{|E|} = (1, \ldots, 1)^T \in \mathbb{R}^{|E|}$. We notice that the second line in Eq. (37) can be satisfied for $\lambda_2$ large enough, since for each $j \in E^c$ we have $|\hat{\beta}_j^{\lambda_1}| < 1$ and $\hat{\beta}_E^{\text{SLS}}$ does not depend on the value of $\lambda_2$. Denote $\lambda$ the smallest $\lambda_2$ such that the second condition in Eq. (37) is satisfied. Now, notice that the first line can be studied element-wise, hence we study two distinct cases, where we take $j \in E$.

- $\mu_j = 0 \implies \hat{\beta}_j^{\lambda_1} \in \text{sign}(\hat{\beta}_E^{\text{SLS}})_j$, which holds due to the definition of $\hat{\beta}_E^{\text{SLS}}$.
- $\mu_j > 0 \implies (\hat{\beta}_E^{\text{SLS}})_j = 0 \implies$ if $\lambda_2 \geq \hat{\mu}_j$ then $|\hat{\beta}_j^{\lambda_1} (1 - 1/\lambda_2)| \leq 1$ and Eq. (37) is satisfied.

Setting $\lambda_0 = \lambda \vee \hat{\mu}_{\text{max}}$, where $\hat{\mu}_{\text{max}} = \max_{j \in [n]} |\hat{\mu}_j|$ concludes the proof.

### 4.3.5 Natural extensions

One can easily extend the Bregman Lasso refitting to any $\ell_1$-penalized methods such as logistic regression, Poisson regression, etc. Such a refitting strategy might be relevant when the estimation of the underlying probabilities is of the same importance as the classification itself. Similar ideas were addressed by Shi et al. [2012], where the authors proposed a new regularization path for the logistic regression inspired by the Bregman Iterations detailed in Equation (18). In this part, we propose a simple extension of the Bregman Lasso to models from the exponential family. To this end, we consider the setting of Generalized Linear Models (GLM) where for every $i \in [n]$ the observed samples are following a natural exponential family distribution with covariates $x_i \in \mathbb{R}^p$ and responses $y_i \in \mathbb{R}$, that is

\[
\forall i \in [n], \quad p_{y_i|x_i}(y_i; \theta) = h(y_i) \exp \left( y_i \langle x_i, \beta^* \rangle - b(\langle x_i, \beta^* \rangle) \right),
\]  

for some unknown $\beta^* \in \mathbb{R}^p$, where $b$ and $h$ are some continuously differentiable functions. In particular, one can show the following relation

\[
\mathbb{E}[y_i|x_i] = b'(\langle x_i, \beta^* \rangle),
\]  

18
where \( b'(\cdot) \) is a derivative of the function \( b(\cdot) \). For any vector \( a = (a_1, \ldots, a_n)^\top \in \mathbb{R}^n \) it is convenient to write \( b'(a) = (b'(a_1), \ldots, b'(a_n))^\top \). We form the observation vector \( y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n \) where \( n \) is the number of available samples and the design matrix \( X = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{n \times p} \) which has \( p \) covariates, and \( n \) observations. To estimate \( \beta^* \) we first minimize the negative log-likelihood with \( \ell_1 \)-penalty [Tibshirani, 1996], which is equivalent for a fixed \( \lambda_1 > 0 \) to the following optimization problem

\[
\hat{\beta}^{\lambda_1} = \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( -y_i \langle x_i, \beta \rangle + b(\langle x_i, \beta \rangle) \right) + \lambda_1 \| \beta \|_1 \right\} .
\]  

(40)

Similarly, Bregman type refitting (for a fixed \( \lambda_2 > 0 \)) can be performed as

\[
\hat{\beta}^{\lambda_1, \lambda_2} = \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( -y_i \langle x_i, \beta \rangle + b(\langle x_i, \beta \rangle) \right) + \lambda_2 D_{\ell_1}^{\lambda_1}(\beta, \hat{\beta}^{\lambda_1}) \right\} ,
\]  

(41)

where the subgradient is given by

\[
\hat{\beta}^{\lambda_1} = \frac{1}{\lambda_1 n} X^\top (y - b'(X\hat{\beta}^{\lambda_1})) .
\]  

(42)

Possible examples are:

- Logistic regression: \( b : t \mapsto \ln(1 + e^t) \) and \( b' : t \mapsto \frac{1}{1 + e^t} \);
- Poisson regression: \( b : t \mapsto e^t \) and \( b' : t \mapsto \frac{1}{1 + e^t} \);

5 Experiments

To evaluate each scenario, we consider the following 'oracle' performance measures (in the sense that in practice one can not evaluate them):

- Prediction: \( \| X(\beta^* - \hat{\beta}) \|_2^2 \);
For our synthetic experiments we generate the design matrix $X \in \mathbb{R}^{n \times p}$ in the following manner

$$X_j = \sqrt{n} \frac{\kappa \zeta + (1 - \kappa) \xi_j}{\| \kappa \zeta + (1 - \kappa) \xi_j \|_2},$$

(43)

5.1 Synthetic data

For our synthetic experiments we generate the design matrix $X \in \mathbb{R}^{n \times p}$ in the following manner

$$X_j = \sqrt{n} \frac{\kappa \zeta + (1 - \kappa) \xi_j}{\| \kappa \zeta + (1 - \kappa) \xi_j \|_2},$$

(43)
Table 1: Description of considered estimators in the experiments. The subgradient $\hat{\beta}^{\lambda_1}$, associated with the Lasso solution, is evaluated according to Equation (21) with $\hat{\beta}^{\lambda_1} = \hat{\beta}^{Lasso}$.

where $\zeta, \xi_1, \ldots, \xi_p \sim \mathcal{N}(0, I_p)$ are independent standard normal vectors. The level of correlations is determined by the parameter $\kappa \in [0, 1]$. We additionally set the underlying vector as

$$
\beta^* = \left(1, \ldots, 1, 0, \ldots, 0\right)^T \in \mathbb{R}^p,
$$

and the noise vector $\varepsilon \sim \mathcal{N}(0, I_n)$.

Hence, each scenario in the synthetic data section is described by the following parameters: number of observations: $n$, number of features: $p$, sparsity level of $\beta^*$: $s$, level of correlations: $\kappa$, noise level: $\sigma$. We fix $n = 40, p = 200, s = 4, \sigma = 0.5$, and use the following values of correlations $\kappa = 0.3; 0.5; 0.7$, representing low, average and high correlation settings respectively. The results are reported on Figure 4, Figure 5 and Figure 6. We first notice that the Boosted Lasso does not give any significant improvement over the Lasso. Other three refitting strategies can improve the first step Lasso solution. In case of modest correlations inside the design matrix, the improvement can be considered as significant.
Moreover, the Sign-Least-Squares Lasso outperforms the Least-Squares Lasso and the Bregman Lasso in average, in addition to the greater interpretability due to the sign preserving properties.

5.2 Semi-real data

For our experimental study, we generate semi-real datasets, following an approach presented in [Bühlmann and Mandozzi, 2014]. The data are generated from the model Equation (1), where the design matrix $X$ is obtained from the leukemia dataset with $n = 72$. We consider the following 4 parameters to describe the settings of our experiments: $p$ - number of covariates, $s$ - sparsity of the vector $\beta^*$, SNR - signal to noise ratio and the correlation settings - way to generate support of the vector $\beta^*$. All the plots are averaged over hundred runs of the simulation process with fixed values of parameters given in Table 2. During each simulation round we choose the first $p$ columns from the leukemia dataset. Additionally we set the signal to noise ratio, defined as

$$
\text{SNR} = \frac{\|X\beta^*\|_2}{\sqrt{n}\sigma^*},
$$

\hspace{1cm} (44)

to control the noise level $\sigma^2$ in the model Equation (1). Additionally the true cardinality of $\beta^*$ is set to $s$ and each non-zero component of $\beta^*$ is set to one or minus one with equal probability. Finally, the support of the vector $\beta^*$ is formed following two scenarios: normal correlations and high correlations. For normal correlations we choose randomly $s$ components out of $p$ and for high correlations scenario, the first component is chosen randomly and the remaining $s - 1$ having the highest Pearson correlation
with the first one. Additional scenarios can be found in Appendix, in the main text we provide only three cases due to the space limitation.

On Figure 7 we notice that in the low noise (SNR = 8) and very sparse (s = 5) scenario the overall performance is similar to the synthetic dataset, discussed above.

Meanwhile, the conclusion is different for Figure 8, where the noise level (SNR = 2) and the sparsity level (s = 20) are high. In this scenario we observe that simple Lasso outperforms in average all the refitting strategies in terms of estimation error and the TP rate, Boosted Lasso provides with better estimation and the Sign-Least-Squares Lasso achieves better results in terms of all the other measures. We additionally emphasize, that the Least-Squares Lasso shows large variance and might fail to improve the estimation in some cases.

Similar conclusions as for previous scenario can be made for Figure 9, where the sparsity level (s = 20) is still high, but the noise level is reduced (SNR = 8).

We conclude by pointing out that performing or not performing the refitting depends on the measure of interest and on the underlying (unknown in practice) scenario. Boosted Lasso agrees with our theoretical results, as it only improves the Lasso estimator in terms of prediction, but it fails to improve any other measure except True Positive rate (due to higher output sparsity). Least-Squares Lasso may show an undesirable performance in some scenarios and bring the problem of interpretability. Sign-Least-Squares Lasso and Bregman Lasso showed consistent and satisfying performance, however Sign-Least-Squares Lasso outperforms the Bregman Lasso in average. We additionally emphasize that our results are valid for cross-validation on MSE. Other choices are possible (BIC, AIC, AV_p) and may possibly provide with different overall conclusion.

**Conclusion**

In this article we introduced a simple framework to provide additional statistical guarantees on the refitting and sign-consistent refitting of Lasso solutions. We demonstrated that every sign-consistent refitting strategy satisfies an oracle inequality under the same assumptions as the Lasso bounds. We theoretically analyzed two refitting strategies: Boosted Lasso and Bregman Lasso, which are easy to implement as they
require only Lasso solver. It appeared that the Bregman Lasso converges to the Sign-Least-Squares Lasso—a particular refitting strategy with sign preserving properties. Experimental results show the advantages of sign-consistent strategies over the simple Least-Squares Lasso. Possible extension of this work is to consider other families of the refitting strategies by either adding an additional information provided by the Lasso or replacing the sign-preserving property. Another interesting road is to use our framework to provide oracle bounds for estimation and feature selection.

Acknowledgment

This work was partially supported by “Laboratoire d’excellence Bézout of Université Paris Est” and by “Chair Machine Learning for Big Data at Télécom ParisTech”.

References

A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013. 1, 2, 4, 21

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009. 1, 3, 7

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004. 3, 17

E-M. Brinkmann, M. Burger, J. Rasch, and C. Sutour. Bias-Reduction in Variational Regularization. Technical report, 2016. 2, 5, 10, 12, 21

P. Bühlmann and J. Mandozzi. High-dimensional variable screening and bias in subsequent inference, with an empirical comparison. *Computational Statistics*, 29(3):407–430, 2014. 22
P. Bühlmann and B. Yu. Boosting with the L2 loss: regression and classification. *J. Am. Statist. Assoc.*, 98(462):324–339, 2003.

A. Dalalyan, M. Hebiri, and J. Lederer. On the prediction performance of the lasso. *Bernoulli*, 23(1):552–581, 2017.

C.-A. Deledalle, N. Papadakis, and J. Salmon. On debiasing restoration algorithms: applications to total-variation and nonlocal-means. In *SSVM*, pages 129–141, 2015.

C.-A. Deledalle, N. Papadakis, J. Salmon, and S. Vaiter. CLEAR: Covariant LEAst-square Re-fitting with applications to image restoration. *SIAM J. Imaging Sci.*, 10(1):243–284, 2017.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360, 2001.

H-Y. Gao and A.G. Bruce. Waveshrink with firm shrinkage. *Statistica Sinica*, pages 855–874, 1997.

G. Gasso, A. Rakotomamonjy, and S. Canu. Recovering sparse signals with non-convex penalties and DC programming. *IEEE Trans. Sig. Process.*, 57(12):4686–4698, 2009.

C. Giraud. *Introduction to High-Dimensional Statistics*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability. Chapman and Hall/CRC, 2014.

V. Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2011.

J. Lederer. Trust, but verify: benefits and pitfalls of least-squares refitting in high dimensions. *arXiv preprint arXiv:1306.0113*, 2013.

S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin. An iterative regularization method for total variation-based image restoration. *Multiscale Modeling & Simulation*, 4(2):460–489, 2005.

S. Osher, F. Ruan, J. Xiong, Y. Yao, and W. Yin. Sparse recovery via differential inclusions. *Appl. Comput. Harmon. Anal.*, 2016.

J. Shi, W. Yin, and S. Osher. A new regularization path for logistic regression via linearized bregman. Technical report, 2012.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

R. J. Tibshirani. The lasso problem and uniqueness. *Electron. J. Stat.*, 7:1456–1490, 2013.

J. Xu and S. Osher. Iterative regularization and nonlinear inverse scale space applied to wavelet-based denoising. *IEEE Transactions on Image Processing*, 16(2):534–544, 2007.

W. Yin, S. Osher, D. Goldfarb, and J. Darbon. Bregman iterative algorithms for l1-minimization with applications to compressed sensing. *SIAM J. Imaging Sci.*, 1(1):143–168, 2008.

C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942, 2010.
A Proofs

Theorem 1. We start from the KKT conditions for Lasso Lemma 1, noticing that \( \langle \hat{\rho}, \hat{\beta} \rangle = \| \hat{\beta} \|_1 \) and for every \( \beta \in \mathbb{R}^p \) we have \( \langle \hat{\rho}, \beta \rangle \leq \| \beta \|_1 \) since \( \| \beta \|_\infty \leq 1 \), we can write

\[
\frac{1}{n} (\beta - \hat{\beta})^\top X^\top (y - X \hat{\beta}) = \langle \beta - \hat{\beta}, \hat{\rho} \rangle \leq \lambda (\| \beta \|_1 - \| \hat{\beta} \|_1) ,
\]

hence we have

\[
\frac{1}{n} (\beta - \hat{\beta})^\top X^\top X (\beta^* - \hat{\beta}) \leq \lambda (\| \beta \|_1 - \| \hat{\beta} \|_1) + \frac{1}{n} \epsilon^\top X (\hat{\beta} - \beta) .
\]

Using the fact that \( 2 \langle x, y \rangle = \| x \|_2^2 + \| y \|_2^2 - \| x - y \|_2^2 \) (for all \( x, y \in \mathbb{R}^p \)), we can derive the following inequality

\[
\frac{1}{2n} \| X (\beta - \hat{\beta}) \|_2^2 + \frac{1}{2n} \| X (\beta^* - \hat{\beta}) \|_2^2 \leq \frac{1}{2n} \| X (\beta - \beta^*) \|_2^2 \leq \lambda (\| \beta \|_1 - \| \hat{\beta} \|_1) + \frac{1}{n} \epsilon^\top X (\hat{\beta} - \beta) .
\]

From Lemma 2, with probability at least \( 1 - \delta \) we have \( \| X^\top \epsilon/n \|_\infty \leq \lambda/2 \), therefore we derive

\[
\frac{1}{n} \| X (\beta - \hat{\beta}) \|_2^2 + \frac{1}{n} \| X (\beta^* - \hat{\beta}) \|_2^2 \leq \frac{1}{2n} \| X (\beta - \beta^*) \|_2^2 \leq \lambda (\| \beta \|_1 - \| \hat{\beta} \|_1) + \frac{1}{n} \| \beta - \hat{\beta} \|_2 ,
\]

now let us set \( \hat{\Delta} = \hat{\beta} - \beta \) and let \( J \in [p] \) be any set such that \( |J| \leq s \). Using the triangle inequality and the decomposability of the \( \ell_1 \)-norm, we can write

\[
\| \hat{\beta} - \beta \|_1 + 2\| \beta \|_1 - 2\| \hat{\beta} \|_1 \leq 3\| \hat{\Delta} \|_1 - 2\| \hat{\Delta} \|_1 + 4\| \beta \|_1 .
\]

We consider two cases

- if \( 3\| \hat{\Delta} \|_1 - 2\| \hat{\Delta} \|_1 \leq 0 \), then Eq. (46) implies

\[
\frac{1}{n} \| X (\hat{\beta} - \beta^*) \|_2^2 \leq \frac{1}{n} \| X (\beta - \beta^*) \|_2^2 \leq \frac{1}{n} \| X (\beta - \beta^*) \|_2^2 + 4\lambda \| \beta \|_1 ,
\]

- if \( 3\| \hat{\Delta} \|_1 - 2\| \hat{\Delta} \|_1 > 0 \), then RE(3, s) implies

\[
\| \hat{\Delta} \|_2 \leq \frac{\| X \hat{\Delta} \|_2^2}{4\epsilon(3, s)} ,
\]

and hence in view of Eq. (46) we get

\[
\frac{1}{n} \| X (\hat{\beta} - \beta^*) \|_2^2 \leq \frac{1}{n} \| X (\beta - \beta^*) \|_2^2 + \frac{3\lambda}{\epsilon} \| \beta \|_1 + 3\lambda \| \hat{\Delta} \|_1 - \frac{1}{n} \| X \hat{\Delta} \|_2^2
\]

\[
\leq \frac{1}{n} \| X (\beta - \beta^*) \|_2^2 + \frac{3\lambda}{\epsilon} \| \beta \|_1 + 3\lambda \| \hat{\Delta} \|_1 - \frac{1}{n} \| X \hat{\Delta} \|_2^2
\]

\[
\leq \frac{1}{n} \| X (\beta - \beta^*) \|_2^2 + \frac{3\lambda}{\kappa(3, s)} \frac{1}{\sqrt{n}} \| X \hat{\Delta} \|_2^2 - \frac{1}{n} \| X \hat{\Delta} \|_2^2 .
\]

If we set \( x = \| X \hat{\Delta} \|_2^2 / \sqrt{n} \), the last two terms are

\[
\sqrt{\frac{3\lambda}{\kappa(3, s)}} x - x^2 = \left( \sqrt{\frac{3\lambda}{2\kappa(3, s)}} \right)^2 - \left( \sqrt{\frac{3\lambda}{2\kappa(3, s)}} - x \right)^2 \leq \frac{9\lambda s}{4\kappa^2(3, s)} .
\]
and we get the desired result
\[
\frac{1}{n} \|X(\beta^* - \hat{\beta})\|_2^2 \leq \inf_{\beta \in \mathbb{R}^p} \min_{|J| \leq s} \left\{ \frac{1}{n} \|X(\beta - \beta^*)\|_2^2 + \frac{9\lambda^2 s}{4\kappa^2(3,s)} + \ldots \right\}.
\]

Lemma 4. The identity \( D_{\ell_1}^\rho (z, w) = \|z\|_1 - \langle \rho, z \rangle \) immediately follows from the definition of the Bregman divergence and the fact that \( \langle \rho, w \rangle = \|w\|_1 \). To prove convexity, we consider \( z_1, z_2, w \in \mathbb{R}^p \) arbitrary vectors and \( t \in [0, 1] \), hence by the definition of the Bregman divergence we can write
\[
D_{\ell_1}^\rho (t z_1 + (1-t)z_2, w) = \|t z_1 + (1-t)z_2\|_1 - \|w\|_1 - \langle \rho, t z_1 + (1-t)z_2 - w \rangle \\
\leq (\|z_1\|_1 - \langle \rho, z_1 \rangle) + (t - 1) (\|z_2\|_1 - \langle \rho, z_2 \rangle) + \langle \rho, w \rangle - \|w\|_1 \\
= t D_{\ell_1}^\rho (z_1, w) + (1-t) D_{\ell_1}^\rho (z_2, w),
\]
where we used the triangle inequality and the identity \( D_{\ell_1}^\rho (z, w) = \|z\|_1 - \langle \rho, z \rangle \). The bound \( 0 \leq D_{\ell_1}^\rho (z, w) \leq 2 \|z\|_1 \), follows from Hölder’s inequality and the fact that \( \|\rho\|_\infty \leq 1 \). Separability \( D_{\ell_1}^\rho (z, w) = \sum_{i=1}^{p} |z_i| - z_i \rho_i = \sum_{i=1}^{p} D_{\ell_1}^\rho (z_i, w_i) \) is a consequence of the separability of the \( \ell_1 \)-norm. Finally, the last property follows from the definition of the subgradient \( \rho \) and identity \( D_{\ell_1}^\rho (z, w) = \|z\|_1 - \langle \rho, z \rangle \).

**B Additional experiments**

Here we provide with additional experimental results

Figure 10: Leukemia dataset with \( p = 200, \text{SNR} = 2, s = 5 \), normal correlations scenario.

- Estimation error (a)
- Prediction error (b)
- Hamming error (c)
- True Positive (d)
- False Positive (e)
- Predicted sparsity (f)
Figure 11: *Leukemia* dataset with $p = 200$, $SNR = 2$, $s = 20$, high correlations scenario.

Figure 12: *Leukemia* dataset with $p = 1000$, $SNR = 8$, $s = 5$, normal correlations scenario.