Pattern Recovery in Penalized and Thresholded Estimation and its Geometry

Piotr Graczyk, Ulrike Schneider, Tomasz Skalski, Patrick J C Tardivel

Pattern Recovery in Penalized and Thresholded Estimation and its Geometry

Piotr Graczyk[1], Ulrike Schneider[2], Tomasz Skalski[3], and Patrick Tardivel[4]

[1]Université d'Angers, Angers, France
[2]TU Vienna, Vienna, Austria
[3]Politechnika Wrocławska, Wrocław, Poland
[4]Institut de Mathématiques de Bourgogne, UMR 5584 CNRS, Dijon, France

Abstract

We consider the framework of penalized estimation where the penalty term is given by a real-valued polyhedral gauge, which encompasses methods such as LASSO (and many variants thereof such as the generalized LASSO), SLOPE, OSCAR, PACS and others. Each of these estimators can uncover a different structure or "pattern" of the unknown parameter vector. We define a general notion of patterns based on subdifferentials and formalize an approach to measure their complexity. For pattern recovery, we provide a minimal condition for a particular pattern to be detected by the procedure with positive probability, the so-called accessibility condition. Using our approach, we also introduce the stronger noiseless recovery condition. For the LASSO, it is well known that the irrepresentability condition is necessary for pattern recovery with probability larger than 1/2 and we show that the noiseless recovery plays exactly the same role, thereby extending and unifying the irrepresentability condition of the LASSO to a broad class of penalized estimators. We show that the noiseless recovery condition can be relaxed when turning to thresholded penalized estimators, extending the idea of the thresholded LASSO: we prove that the accessibility condition is already sufficient (and necessary) for sure pattern recovery by thresholded penalized estimation provided that the signal of the pattern is large enough. Throughout the article, we demonstrate how our findings can be interpreted through a geometrical lens.

Keywords: penalized estimation, regularization, gauge, pattern recovery, polytope, geometry, LASSO, generalized LASSO, SLOPE, irrepresentability condition, uniqueness.

1 Introduction

Consider the linear regression model

$$Y = X\beta + \varepsilon,$$

where $X \in \mathbb{R}^{n\times p}$ is a design matrix, $\varepsilon \in \mathbb{R}^n$ represents random noise having a symmetric distribution with a positive density on $\mathbb{R}^n$ and $\beta \in \mathbb{R}^p$ is the vector of unknown regression coefficients. Penalized

estimation of β has been studied extensively in the literature, of particular interest the case where the penalization is polyhedral so that the estimator may detect particular features of β. Depending on the penalization term, different features can be recovered by the procedure. The most prominent example is of course the LASSO (Tibshirani, 1996) with its ability to perform model selection, i.e., potentially uncovering zero components of β. In addition to this sparsity property, the fused LASSO (Tibshirani et al., 2005) may set adjacent components to be equal. Using the supremum norm in the penalty term promotes clustering of components that are maximal in absolute value (Jégou et al., 2012). SLOPE (Bogdan et al., 2015) as well as OSCAR (Bondell & Reich, 2008) display further clustering phenomena where certain components may be equal in absolute value, to name just a few. When looking closer at these phenomena under a geometric lens, Schneider & Tardivel (2022) show that for the LASSO, the naturally arising pattern structure not only carries information about zero components, but also about the signs of the non-zero components. This natural pattern structure of the LASSO has appeared many times in the literature, such as in the so-called sign-consistency of the LASSO (see e.g. Zhao & Yu, 2006; Tardivel & Bogdan, 2022) and also the conditioning event in the selective inference approach of Lee et al. (2016). For SLOPE, the natural pattern structure describes not only signs (zero components as well as the signs of non-zero components) and clustering (components may be equal in absolute value, a well-known phenomenon for this estimator), but also conveys information about the ordering of the coefficients, see Schneider & Tardivel (2022) for details.

In this article, we provide a general approach to characterize the pattern structure naturally arising for a particular method. We do so by introducing the notion of patterns inherent to a method as equivalence classes of elements in $\mathbb{R}^p$ exhibiting the same subdifferential with respect to the penalizing term. We assume that the penalty term is given by a polyhedral gauge, a concept slightly more general than a polyhedral norm which allows to also treat methods such as the generalized LASSO (Ali & Tibshirani, 2019). We show that the pattern equivalence classes coincide with the normal cones of the polytope B^*, where B^* is the subdifferential of the penalizing gauge at zero, and that each equivalence class can be identified with a face of B^*. We also introduce the concept of complexity of a pattern, defined to be the dimension of the linear span of the corresponding equivalence class, and prove that this complexity measure coincides with the codimension of the associated face of B^*.

Given this general notion of patterns, we turn to the question of when an estimation procedure may recover a specific pattern. A minimal condition is the so-called accessibility condition of a pattern of β which gives equivalent criteria for the existence of point $y \in \mathbb{R}^n$ such that the resulting estimator exhibits the pattern under consideration. We express this criterion both in an analytic manner and through a geometric criterion involving how the row span of X intersects the polytope B^*. This extends the geometric condition given for LASSO and SLOPE in Schneider & Tardivel (2022) to the general framework of gauge-penalized estimation. Note that a different approach for an accessibility criterion for the LASSO under a uniqueness assumption was also considered in Sepehri & Harris (2017). Under uniqueness, we prove that this minimal condition already ensures pattern detection with positive probability, provided that the response vector follows a continuous distribution on $\mathbb{R}^n$.

A stronger condition is given by the noiseless recovery condition, where the estimator determined by the noiseless signal $y = X\beta$ is required to possess the same pattern as β. This condition can be proven to be equivalent to the irrepresentability condition in case of the LASSO (see e.g. Bühlmann

& Van de Geer, 2011) which is a necessary condition for pattern recovery with probability of at least 1/2 (Wainwright, 2009). In fact, the noiseless recovery condition is shown to play exactly the same role as the irrepresentability condition in the general gauge-penalized estimation framework: it is a a necessary condition for pattern recovery with probability of at least 1/2 and allows to unify and extend the concept of an irrepresentability condition to entire class of gauge-penalized estimators.

Inspired by the fact that the thresholded LASSO (where small non-zero components may be set to zero additionally to existing zeros) can alleviate the recovery of LASSO patterns (Tardivel & Bogdan, 2022), we define a general concept of thresholded estimators that alter the penalized estimator by in some sense moving to the closest, less complex patterns. We show that for this thresholded penalized estimation, the noiseless recovery condition which is necessary for pattern recovery without thresholding, the much weaker accessibility condition is already sufficient for sure pattern recovery under a uniqueness assumption, provided the signal of the pattern is large enough. For completeness, we also extend the necessary and sufficient condition for uniform uniqueness from Schneider & Tardivel (2022) to gauge-penalized estimation which again relies on the connection between patterns and the faces of B^* and essentially shows that uniqueness occurs if no pattern of complexity exceeding the rank of X is accessible. Finally, we illustrate some pattern recovery properties with numerical experiments.

The paper is organized as follows. In Section 2, we introduce the given setting and notation. Section 3 treats defining and illustrating pattern structures. Pattern recovery by penalized estimation is investigated in Section 4, whereas we turn to pattern recovery by thresholded penalized estimation in Section 5. Uniform uniqueness is proven in Section 6 and Section 7 gives some numerical illustrations. All proofs are relegated to Appendix B, before which Appendix A provides some definitions and results on polytopes and gauges. Finally, Appendix C contains additional results referred to throughout, including a result on solution existence of the optimization problem treated in the article.

2 Setting and notation

The optimization problem we consider throughout the article is the gauge-penalized least-squares problem described in the following. Let $X \in \mathbb{R}^{n\times p}$ be completely arbitrary. Given $y \in \mathbb{R}^n$ and $\lambda > 0$, we define the set $S_{X,\lambda\text{pen}}(y)$ of minimizers to be given by

$$S_{X,\lambda\text{pen}}(y) = \operatorname*{Arg\,min}_{b\in\mathbb{R}^p} \frac{1}{2}\|y - Xb\|_2^2 + \lambda\text{pen}(b), \tag{1}$$

where "pen" is a real-valued polyhedral gauge and $\|.\|_2$ denotes the Euclidean norm. A gauge is any non-negative and positively homogeneous convex function that vanishes at 0, and it is polyhedral if its unit ball is given by a (possibly unbounded) polyhedron. A polyhedral gauge $b \in \mathbb{R}^p \mapsto \text{pen}(b) \in [0,\infty)$ can always be written as the maximum of finitely many linear functions (Rockafellar, 1997; Mousavi & Shen, 2019), so that we can assume that

$$\text{pen}(b) = \max\{u_1'b, \ldots, u_k'b\}, \text{ for some } u_1, \ldots, u_k \in \mathbb{R}^p \text{ with } u_1 = 0.$$

Note that a polyhedral gauge whose unit ball $B = \{b \in \mathbb{R}^p : \text{pen}(x) \leq 1\}$ is a bounded and symmetric polyhedron is in fact a polyhedral norm. Examples of polyhedral norms and gauges are discussed

in more detail in Section 3. For our geometric considerations, a central object of study will be the polytope B^* defined as

$$B^* = \text{conv}(u_1, \ldots, u_k),$$

where conv(.) denotes the convex hull. In case pen is a norm, B^* coincides with the unit ball of the dual norm. The optimization problem in (1) always possesses a solution, as we show in Proposition C.1 in Appendix C[1], but it does not have to be unique. We treat uniqueness by giving a necessary and sufficient condition in Section 6.

The following additional notation will be used throughout the article. By $[p]$, we denote the set $\{1, \ldots, p\}$. For a set $I \subseteq [p]$, the symbol I^c denotes its complement $I^c = [p] \setminus I$. Given a matrix X and an index set I, X_I is the matrix with columns corresponding to indices in I only, with analogous notation for a vector b, so that b_I denotes the vector with components corresponding to indices in I only. The column and row space of X are $\text{col}(X)$ and $\text{row}(X)$, respectively, and $\text{rk}(X)$ refers to the rank of X. For a set $S \subseteq \mathbb{R}^p$, $\text{lin}(S)$ is the linear span of S, i.e., the smallest vector space containing S and $\text{aff}(S)$ is the affine hull of S, i.e., the smallest affine space containing S, whereas $\vec{\text{aff}}(S)$ refers to the vector space parallel to $\text{aff}(S)$ given by $\{u - s : u \in \text{aff}(S)\}$ for a fixed, but arbitrary $s \in \text{aff}(S)$. The relative interior of S is denoted by $\text{ri}(S)$. The symbol $V^\perp$ is used for the orthogonal complement of the vector space V and $\mathbb{1}(.)$ stands for the indicator function. For a convex function $\phi : \mathbb{R}^p \to \mathbb{R}$, a vector $s \in \mathbb{R}^p$ is a *subgradient of ϕ at $\beta \in \mathbb{R}^p$* if

$$f(b) \geq f(\beta) + s'(b - \beta) \;\; \forall b \in \mathbb{R}^p.$$

The convex, non-empty set of all subgradients of ϕ at β is called the *subdifferential of ϕ at β*, denoted by $\partial_\phi(\beta)$. Finally, for a closed and convex set $K \subseteq \mathbb{R}^p$ and $\beta \in K$, the normal cone of K at β is given by

$$N_K(\beta) = \{s \in \mathbb{R}^p : s'(b - \beta) \leq 0 \; \forall b \in K\},$$

see e.g. (Hiriart-Urruty & Lemarechal, 2001, p.65).

3 The notion of patterns

For a gauge-penalized estimation method, we implicitly define its canonical pattern structure and complexity in the following definition.

Definition 3.1 (Pattern equivalence class). Let pen be a real-valued polyhedral gauge on $\mathbb{R}^p$. We say that β and $\tilde{\beta} \in \mathbb{R}^p$ have the same *pattern with respect to* pen if

$$\partial_{\text{pen}}(\beta) = \partial_{\text{pen}}(\tilde{\beta}),$$

i.e., if their subdifferentials of pen coincide. We then write $\beta \overset{\text{pen}}{\sim} \tilde{\beta}$. The set of all elements of $\mathbb{R}^p$ sharing the same pattern as β is called the pattern equivalence class C_β. Furthermore, we define the

[1]The existence of a minimizer is clear when pen is a norm. For the special case of the generalized LASSO (in which pen is not a norm), existence is shown in Ali & Tibshirani (2019) or Dupuis & Vaiter (2019). However, these proofs cannot be generalized to arbitrary polyhedral gauges.

complexity of the pattern of β to be the dimension of $\text{lin}(C_\beta)$.

From Lemma A.2 in Appendix A, it can be learned that the faces of B^*, which is in fact the subdifferential of pen at 0, are made up of the subdifferentials $\partial_{\text{pen}}(\beta)$, so that there is a one-to-one relationship between the the pattern equivalence classes and the faces of B^*. This relationship can be made fully concrete by the following proposition which states that the pattern equivalence classes are, in fact, given by the (relative interior) of the normal cones of the corresponding subdifferential.

Theorem 3.2. *Let* pen *be a real-valued polyhedral gauge on* $\mathbb{R}^p$ *and let* $\beta \in \mathbb{R}^p$. *Then* $C_\beta = \text{ri}(N_{B^*}(b))$ *where* b *is an arbitrary element of* $\text{ri}(\partial_{\text{pen}}(\beta))$ *and* $\text{lin}(C_\beta) = \vec{\text{aff}}(\partial_{\text{pen}}(\beta))^\perp$.

It is known that the relative interior of the normal cones of a polytope form a partition $\mathbb{R}^p$ (see Ewald, 1996, p. 17, Theorem 4.13), so the first part of Theorem 3.2 shows that this partition is the same as partitioning the space by the pattern equivalence classes C_β.

Note that the second statement proves that $\text{lin}(C_\beta)$ matches the notion of model subspace as defined in Vaiter et al. (2015, 2018). This statement also demonstrates that the measure of complexity of the pattern of β introduced in Definition 3.1 coincides with the codimension of the face $\partial_{\text{pen}}(\beta)$ of B^* which is given by $p - \dim(\partial_{\text{pen}}(\beta))$[2] as summarized in the corollary below. Note that this quantity that is also relevant for uniform uniqueness characterized in Theorem 6.1.

Corollary 3.3. *Let* pen *be a real-valued polyhedral gauge on* $\mathbb{R}^p$ *and let* $\beta \in \mathbb{R}^p$. *Then the complexity of the pattern of* β *with respect to* pen *is given by the codimension of* $\partial_{\text{pen}}(\beta)$.

We illustrate the notion of patterns and their complexity as well as the above theorem for several examples of gauges in the following.

Example (Different penalizations and their patterns).

ℓ_1-norm: The subdifferential of the ℓ_1-norm at 0 is given by $B^* = \partial_{\|.\|_1}(0) = [-1, 1]^p$. The pattern of $\beta \in \mathbb{R}^p$ can be represented by its sign vector, where $\text{sign}(\beta) \in \{-1, 0, 1\}^p$ is defined as

$$\text{sign}(\beta) = (\text{sign}(\beta_1), \ldots, \text{sign}(\beta_p))' \text{ with } \text{sign}(\beta_j) = \mathbb{1}\{\beta_j \geq 0\} - \mathbb{1}\{\beta_j \leq 0\}.$$

Indeed, the subdifferentials $\partial_{\|.\|_1}(.)$ at two points in $\mathbb{R}^p$ will be the same if and only if their sign vectors coincide so that $C_\beta = \{b \in \mathbb{R}^p : \text{sign}(b) = \text{sign}(\beta)\}$ and the pattern structure of the LASSO carries not only information about zero components, but also the signs of the non-zero coefficients. Note that the complexity of the LASSO pattern of β, which coincides with the codimension of $\partial_{\|.\|_1}(\beta)$ by Corollary 3.3, is given by $\|\text{sign}(\beta)\|_1$, the number of non-null components of β.

SLOPE-norm: For $b \in \mathbb{R}^p$, the sorted-ℓ_1 or SLOPE norm is defined as $\|b\|_w = \sum_{j=1}^p w_j |b|_{(j)}$, where $|b|_{(1)} \geq \cdots \geq |b|_{(p)}$ and $w_1 \geq \cdots \geq w_p \geq 0$ with $w_1 > 0$ are pre-defined weights. It can be shown that $B^* = \partial_{\|.\|_w}(0) = \text{conv}\{(w_{\pi(1)}, \ldots, w_{\pi(p)})' : \pi \in \mathcal{S}_p\}$ with $\mathcal{S}_p$ denoting the set of all permutations on $[p]$. The polytope B^* is the so-called signed permutahedron, see Negrinho &

[2]The dimension of a face is defined as the dimension of its affine hull, see Appendix A for details.

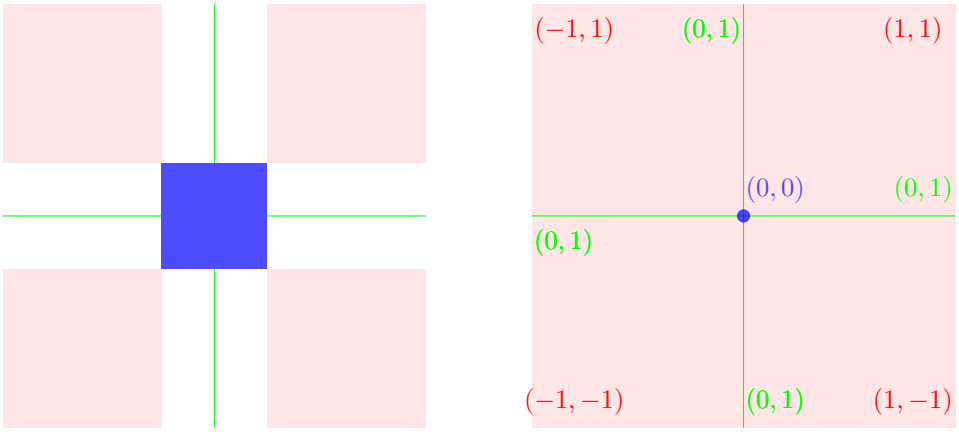

Figure 1: Pattern equivalence classes for the LASSO in $p = 2$ dimensions: On the left, the blue polytope is $B^* = \partial_{\|.\|_1}(0) = \text{conv}\{\pm(1,1), \pm(1,-1)\}$, together with the (uncentered) normal cones of the faces of B^* in pink and green. The picture on the right provides the actual normal cones which, by Theorem 3.2, coincide with the pattern equivalence classes $C_\beta = \{b \in \mathbb{R}^p : \text{sign}(b) = \text{sign}(\beta)\}$ for the patterns $\text{sign}(\beta) \in \{(0,0), \pm(0,1), \pm(1,0), \pm(1,1), \pm(1,-1)\}$.

Martins (2014) and Schneider & Tardivel (2022) for details. The SLOPE pattern of $\beta \in \mathbb{R}^p$ is represented by $\text{patt}_{\text{slope}}(\beta) \in \mathbb{Z}^p$ with each component given by

$$\text{patt}_{\text{slope}}(\beta)_j = \text{sign}(\beta_j)\,\text{rank}(|\beta|)_j,$$

where $\text{rank}(|\beta|)_j \in \{0, 1, \dots, m\}$ with m the number of non-zero values in $\{|\beta_1|, \dots, |\beta_p|\}$ is defined as follows: $\text{rank}(|\beta|)_j = 0$ if $\beta_j = 0$, $\text{rank}(|\beta|)_j > 0$ if $|\beta_j| > 0$ and $\text{rank}(|\beta|)_i < \text{rank}(|\beta|)_j$ if $|\beta_i| < |\beta_j|$, as can be learned in Schneider & Tardivel (2022). For example, the SLOPE pattern of $\beta = (3.1, -1.2, 0.5, 0, 1.2, -3.1)$ is given by $\text{patt}_{\text{slope}}(\beta) = (3, -2, 1, 0, 2, -3)$. Indeed, if $w \in \mathbb{R}^p$ satisfies $w_1 > \cdots > w_p > 0$, the subdifferentials $\partial_{\|.\|_w}(.)$ at two points in $\mathbb{R}^p$ will be the same if and only if their SLOPE patterns coincide so that $C_\beta = \{b \in \mathbb{R}^p : \text{patt}_{\text{slope}}(b) = \text{patt}_{\text{slope}}(\beta)\}$. This shows that the SLOPE patterns do not only carry information about zeros, signs and clustering, but also about the order of the clusters. SLOPE patterns are also treated in Hejný et al. (2023). Note that the complexity of the SLOPE pattern of β, which coincides with the codimension of $\partial_{\|.\|_w}(\beta)$ by Corollary 3.3, is given by $\|\text{patt}_{\text{slope}}(\beta)\|_\infty$, the number of non-zero clusters in β, see Schneider & Tardivel (2022).

ℓ_∞-norm: The subdifferential of the ℓ_∞-norm at 0 is the unit ball of the ℓ_1-norm, $B^* = \partial_{\|.\|_\infty}(0) = \{s : \|s\|_1 \leq 1\}$. The pattern of $\beta \in \mathbb{R}^p$ can be represented by $\text{patt}_\infty(\beta) \in \{-1, 0, 1\}^p$ where each component is defined as

$$\text{patt}_\infty(\beta)_j = \mathbb{1}\{\beta_j = \|\beta\|_\infty\} - \mathbb{1}\{\beta_j = -\|\beta\|_\infty\}.$$

Note that a zero component of $\text{patt}_\infty(\beta)$ represents a component of β that is not maximal in absolute value or a component of the zero vector. For instance, for $\beta = (1.45, 1.45, 0.56, 0, -1.45)'$, the pattern is given by $\text{patt}_\infty(\beta) = (1, 1, 0, 0, -1)$. Indeed, the subdifferentials at two points in

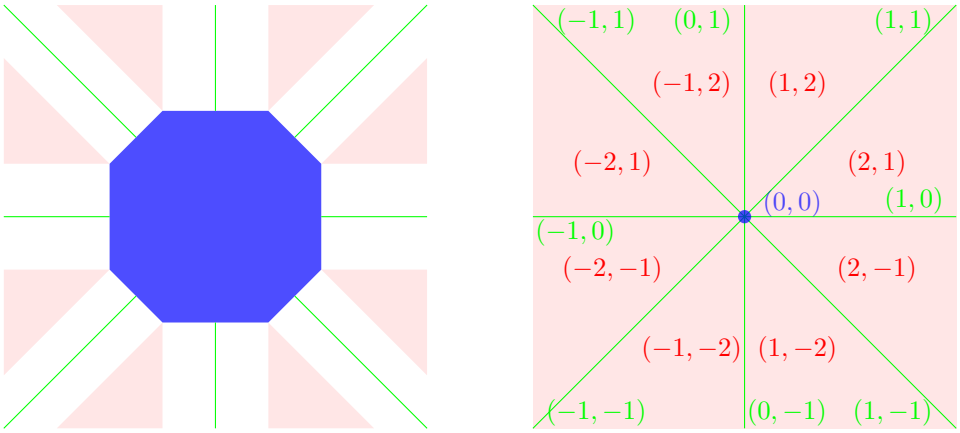

Figure 2: Pattern equivalence classes for SLOPE in $p = 2$ dimensions: On the left, the blue polytope is $B^* = \partial_{\|.\|_w}(0) = \text{conv}\{\pm(w_1, w_2), \pm(w_1, -w_2), \pm(w_2, w_1), \pm(w_2, -w_1)\}$, the signed permutahedron for the SLOPE weights $w_1 > w_2 > 0$, together with the (uncentered) normal cones of the faces of B^* in pink and green. The picture on the right provides the actual normal cones which, by Theorem 3.2, coincide with the pattern equivalence classes $C_\beta = \{b \in \mathbb{R}^p : \text{patt}_{\text{slope}}(b) = \text{patt}_{\text{slope}}(\beta)\}$ for the patterns $\{(0,0), \pm(1,0), \pm(0,1), \pm(1,1), \pm(1,-1), \pm(1,2), \pm(1,-2), \pm(2,1), \pm(2,-1)\}$.

$\beta, \tilde{\beta} \in \mathbb{R}^p$ will be the same if and only if $\text{patt}_\infty(\beta) = \text{patt}_\infty(\tilde{\beta})$ so that $C_\beta = \{b \in \mathbb{R}^p : \text{patt}_\infty(b) = \text{patt}_\infty(\beta)\}$. This shows that the sup-norm patterns carry information about maximal (in absolute value) and non-maximal components, as well as the sign information of the maximal coefficients. Note that the complexity of $\text{patt}_\infty(\beta)$, which coincides with the codimension of $\partial_{\|.\|_\infty}(\beta)$[3] by Corollary 3.3, is given by $\mathbb{1}\{\beta \neq 0\}(\sum_{j=1}^p \mathbb{1}\{|\beta_j| < \|\beta\|_\infty\} + 1)$, the number of non-maximal components in case $\beta \neq 0$ and 0 otherwise.

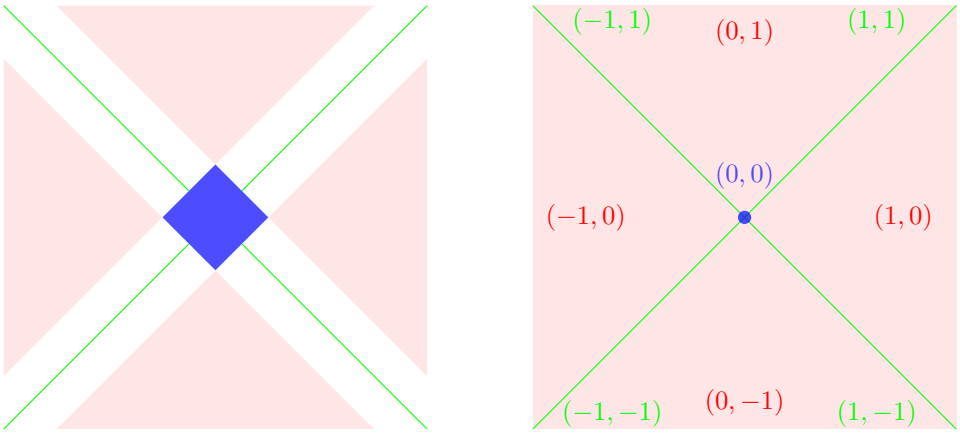

Figure 3: Pattern equivalence classes for the sup-norm in $p = 2$ dimensions: On the left, the blue polytope is $B^* = \partial_{\|.\|_\infty}(0) = \text{conv}\{\pm(1,0), \pm(0,1)\}$, together with the (uncentered) normal cones of the faces of B^* in pink and green. The picture on the right provides the actual normal cones which, by Theorem 3.2, coincide with the pattern equivalence classes $C_\beta = \{b \in \mathbb{R}^p : \text{patt}_\infty(b) = \text{patt}_\infty(\beta)\}$ for the patterns $\text{patt}_\infty(\beta) \in \{(0,0), \pm(0,1), \pm(1,0), \pm(1,1), \pm(1,-1)\}$.

[3] An explicit expression for $\partial_{\|.\|_\infty}(\beta)$ can be found in Appendix C.2.

Generalized LASSO: For the generalized Lasso, the penalty term is given by $\text{pen}(b) = \|Db\|_1$ where $D \in \mathbb{R}^{m\times p}$. Note that, when $\ker(D) \neq \{0\}$, pen is only a semi-norm. We list two common choices of D. For the subdifferential at 0, we have $\partial_{\|D.\|_1}(0) = D'[-1,1]^m$, see Hiriart-Urruty & Lemarechal (2001, p.184).

1. Let $p \geq 2$ and let $D^{\text{tv}} \in \mathbb{R}^{(p-1)\times p}$ be the first-order difference matrix defined as

$$D^{\text{tv}} = \begin{pmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & -1 & 1 \end{pmatrix}.$$

   The subdifferentials $\partial_{\|D^{\text{tv}}.\|_1}(\beta)$ and $\partial_{\|D^{\text{tv}}.\|_1}(\tilde{\beta})$ are equal if and only if $\text{sign}(D^{\text{tv}}\beta) = \text{sign}(D^{\text{tv}}\tilde{\beta})$, so that we can represent the pattern by this expression. Note that $\text{sign}(D^{\text{tv}}\beta)_j = 0$ if $\beta_{j+1} = \beta_j$. Moreover, $\text{sign}(D^{\text{tv}}\beta)_j = 1$ or $\text{sign}(D^{\text{tv}}\beta)_j = -1$ if $\beta_{j+1} > \beta_j$ or $\beta_{j+1} < \beta_j$, respectively. For example, the pattern of $\beta = (1.45, 1.45, 0.56, 0.56, -0.45, 0.35)'$ is given by $\text{patt}_{\text{tv}}(\beta) = \text{sign}(D^{\text{tv}}\beta) = (0, -1, 0, -1, 1)'$. Clearly, $C_\beta = \{b \in \mathbb{R}^p : \text{patt}_{\text{tv}}(b) = \text{patt}_{\text{tv}}(\beta)\}$. Note that the complexity of $\text{patt}_{\text{tv}}(\beta)$, which coincides with the codimension of $\partial_{\|D^{\text{tv}}.\|_1}(\beta)$ by Corollary 3.3, is given by $1 + \|\text{sign}(D^{\text{tv}}\beta)\|_1$, the number of equal adjacent components plus 1.

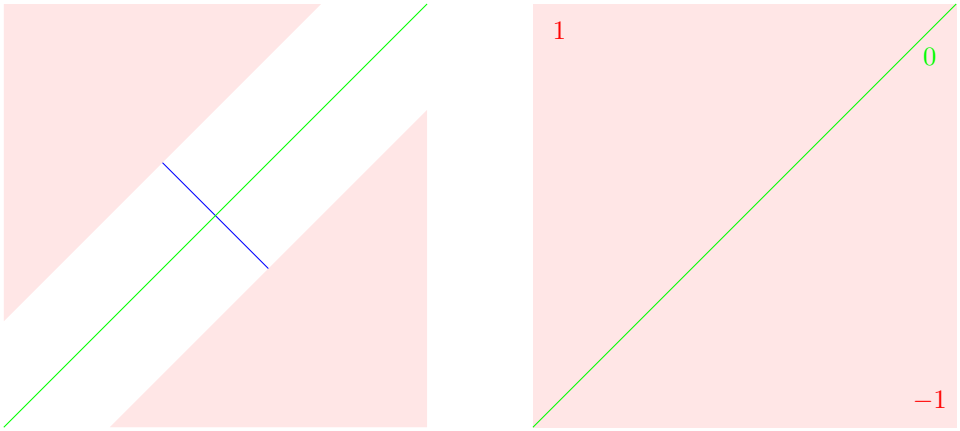

Figure 4: Pattern equivalence classes for the generalized LASSO with penalizing first-order differences ($D = D^{\text{tv}}$) in $p = 2$ dimensions: On the left, the blue polytope is $B^* = \partial_{\text{pen}}(0) = \text{conv}\{\pm(1, -1)\}$ together with the (uncentered) normal cones of the faces of B^* in pink and green. The picture on the right provides the actual normal cones which, by Theorem 3.2, coincide with the pattern equivalence classes $C_\beta = \{b \in \mathbb{R}^p : \text{patt}_{\text{tv}}(b) = \text{patt}_{\text{tv}}(\beta)\}$ for the patterns $\text{patt}_{\text{tv}}(\beta) \in \{-1, 0, 1\}$.

2. Let $p \geq 3$ and let $D^{\mathrm{tf}} \in \mathbb{R}^{(p-2)\times p}$ be the second-order difference matrix defined as

$$D^{\mathrm{tf}} = \begin{pmatrix} 1 & -2 & 1 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & -2 & 1 \end{pmatrix}.$$

The resulting method is called ℓ_1-trend filtering (Kim et al., 2009) which in this context can be viewed as a special case of the generalized LASSO. The subdifferentials $\partial_{\|D^{\mathrm{tf}}.\|_1}(\beta)$ and $\partial_{\|D^{\mathrm{tf}}.\|_1}(\tilde{\beta})$ are equal if and only if $\mathrm{sign}(D^{\mathrm{tf}}\beta) = \mathrm{sign}(D^{\mathrm{tf}}\tilde{\beta})$, so that we can represent the pattern by this expression. Subdifferentials $\partial_{\|D^{\mathrm{tf}}.\|_1}(\beta) = \partial_{\|D^{\mathrm{tf}}.\|_1}(\tilde{\beta})$ are equal if and only if $\mathrm{sign}(D^{\mathrm{tf}}\beta) = \mathrm{sign}(D^{\mathrm{tv}}\tilde{\beta})$. To illustrate this pattern structure, consider the piecewise linear curve $G_\beta = \cup_{j=1}^{p-1}[(j, \beta_j), (j+1, \beta_{j+1})]$. Note that $\mathrm{sign}(D^{\mathrm{tf}}\beta)_j = 0$ if, in a neighborhood of the point (j, β_j), the curve G_β is linear. Moreover, $\mathrm{sign}(D^{\mathrm{tf}}\beta)_j = 1$ or $\mathrm{sign}(D^{\mathrm{tf}}\beta)_j = -1$ if, in a neighborhood of the point (j, β_j), the curve G_β convex or concave, respectively. For instance, Figure 5 provides an illustration of $\mathrm{sign}(D^{\mathrm{tf}}(x))$ for a particular $x \in \mathbb{R}^9$. Finally, note that the complexity of $\mathrm{patt}_{\mathrm{tf}}(\beta)$, which coincides with the codimension of $\partial_{\|D^{\mathrm{tf}}.\|_1}(\beta)$ by Corollary 3.3, is given by is given by $2 + \|\mathrm{sign}(D^{\mathrm{tf}}\beta)\|_1$, the number "non-linear points" plus 2.

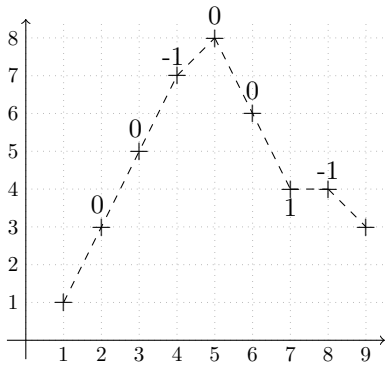

Figure 5: In this figure the dotted curve represents C described above for $\beta = (1, 3, 5, 7, 8, 6, 4, 4, 3)'$. Here, $\mathrm{sign}(D^{\mathrm{tv}}\beta) = (0, 0, -1, 0, 0, 1, -1)'$.

Note that for the ℓ_1-norm, the sorted-ℓ_1-norm and the ℓ_∞-norm, the pattern of β itself is a canonical representative of the equivalence class C_β. On the other hand, for generalized LASSO, $\mathrm{sign}(D^{\mathrm{tv}}\beta)$ and $\mathrm{sign}(D^{\mathrm{tf}}\beta)$, respectively, characterize the pattern but are not an element of C_β as there seems to be no natural way to represent the pattern as such.

4 Pattern recovery in penalized estimation

We now turn to the question of when a pattern can be recovered by a penalized estimation procedure. For this, we introduce the notion of accessible patterns in the following definition which requires

the existence of *a* response vector such that the resulting estimator exhibits the required pattern. Accessibility clearly is a minimal condition for possible pattern recovery. This definition generalizes the notion of accessible sign vectors for LASSO (Sepehri & Harris, 2017; Schneider & Tardivel, 2022) and accessible patterns for SLOPE (Schneider & Tardivel, 2022) to the general class of penalized estimators considered in this paper.

Definition 4.1 (Accessible pattern). Let $X \in \mathbb{R}^{n\times p}$, $\lambda > 0$ and pen be a real-valued polyhedral gauge. We say that $\beta \in \mathbb{R}^p$ has an accessible pattern with respect to X and λpen, if there exists $y \in \mathbb{R}^n$ and $\hat{\beta} \in S_{X,\lambda\text{pen}}(y)$ such that $\hat{\beta} \overset{\text{pen}}{\sim} \beta$.

When pen is the ℓ_1-norm scaled by a tuning parameter $\lambda > 0$, i.e., pen $= \lambda\|.\|_1$ the above definition coincides with the notion of accessibility of sign vectors with respect to X. When pen is the sorted ℓ_1-norm, i.e., pen $= \|.\|_w$ for some $w \in \mathbb{R}^p$ with $w_1 > \cdots > w_p > 0$, the above definition coincides with the notion of accessible SLOPE patterns with respect to X. Proposition 4.2 provides both a geometric and an analytic characterization for the general notion of accessible patterns.

Proposition 4.2 (Characterization of accessible patterns)**.** *Let $X \in \mathbb{R}^{n\times p}$ and* pen $: \mathbb{R}^p \to \mathbb{R}$ *be a real-valued polyhedral gauge.*

1. *Geometric characterization: The pattern of $\beta \in \mathbb{R}^p$ is accessible with respect to X and λ*pen *if and only if*
$$\text{row}(X) \cap \partial_{\text{pen}}(\beta) \neq \varnothing.$$

2. *Analytic characterization: The pattern of $\beta \in \mathbb{R}^p$ is accessible with respect to X and λ*pen *if and only if for any $b \in \mathbb{R}^p$ the implication*
$$X\beta = Xb \implies \text{pen}(\beta) \leq \text{pen}(b)$$
holds.

Based on Proposition 4.2, it is clear that the notion of accessibility does not depend on the value of the tuning parameter λ. We therefore also say that the pattern of β is accessible with respect to X and pen. The geometric characterization shows that we have accessibility for the pattern of β if and only if row(X) intersects the face of B^* that corresponds to the pattern of β. The following proposition strengthens the notion of accessibility, showing that under uniform uniqueness, accessibility already implies the existence a set of $y's$ in $\mathbb{R}^n$ with non-empty interior that lead the pattern of interest:

Proposition 4.3. *Let $X \in \mathbb{R}^{n\times p}$, $\lambda > 0$ and* pen $: \mathbb{R}^p \to \mathbb{R}$ *be a real-valued polyhedral gauge. Assume that uniform uniqueness holds, i.e. for any $y \in \mathbb{R}^n$, the set $S_{X,\lambda\text{pen}}(y)$ contains the unique minimizer $\hat{\beta}(y)$. Let $\beta \in \mathbb{R}^p$. If the pattern of β is accessible with respect to X and* pen*, the set*
$$A_\beta = \{y : \hat{\beta}(y) \overset{\text{pen}}{\sim} \beta\} \subseteq \mathbb{R}^n$$
has non-empty interior.

Clearly, Proposition 4.3 demonstrates that under a uniqueness assumption, accessibility of a pattern already implies that the pattern can be detected by the penalized procedure with positive probability,

provided that y is generated by a continuous distribution taking on all values in $\mathbb{R}^n$. Hejný et al. (2023) call this concept attainability which they view in an asymptotic setting for SLOPE.

Corollary 4.4. *Let $X \in \mathbb{R}^{n\times p}$, $\lambda > 0$ and $\mathrm{pen} : \mathbb{R}^p \to \mathbb{R}$ be a real-valued polyhedral gauge. Let $\beta \in \mathbb{R}^p$ have an accessible pattern and assume that uniform uniqueness holds. If Y follows a distribution with a positive Lebesgue-density on $\mathbb{R}^n$, then*

$$\mathbb{P}(\hat{\beta}(Y) \overset{\text{pen}}{\sim} \beta) > 0.$$

We now turn to a stronger requirement for pattern recovery. For this, we consider the solution path of a penalized estimator which is given by the curve $0 < \lambda \mapsto \hat{\beta}_\lambda$, where $\hat{\beta}_\lambda$ is the (assumed to be unique) element of $S_{X,\lambda\mathrm{pen}}(y)$ for fixed $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$. The solution path of the generalized LASSO or OSCAR and Clustered LASSO is studied in Tibshirani & Taylor (2011) or Takahashi & Nomura (2020), respectively. Definition 4.5 below alludes to the notion of a solution path. Note, however, that Definition 4.5 does not require uniqueness of estimator.

Definition 4.5 (Noiseless recovery condition). Let pen be a real-valued polyhedral gauge, $X \in \mathbb{R}^{n\times p}$ and $\beta \in \mathbb{R}^p$. We say that the pattern of β satisfies the noiseless recovery condition with respect to X and pen if

$$\exists \lambda > 0, \exists \hat{\beta} \in S_{X,\lambda\mathrm{pen}}(X\beta) \text{ such that } \hat{\beta} \overset{\text{pen}}{\sim} \beta.$$

For instance, $\beta = 0$ satisfies the noiseless recovery condition with respect to X and pen since then $X\beta = 0$ and $0 \in S_{X,\lambda\mathrm{pen}}(0)$. Another way of stating the noiseless recovery condition is to require that in the noiseless case $Y = X\beta$, the solution path contains a minimizer having the same pattern as β. The noiseless recovery condition is illustrated for the supremum norm in Figure 6 for the particular case where X and β are given by

$$X = \begin{pmatrix} 1 & 0 & 2 \\ 0 & 1 & 1 \end{pmatrix} \text{ and } \beta = (0, 2, 2)'.$$

In Theorem C.5 in Appendix C.3, we prove that the noiseless recovery condition occurs if and only if $X'X\mathrm{lin}(C_\beta) \cap \partial_{\mathrm{pen}}(\beta) \neq \varnothing$. Based on this characterization, it is clear that the condition depends on β only through its pattern. For an analytic expression for checking the noiseless recovery condition, some formulas are given in the literature. For example, when $\mathrm{pen} = \|.\|_1$, the noiseless recovery condition can be shown to be equivalent to

$$\|X'(X_I')^+\mathrm{sign}(\beta_I)\|_\infty \leq 1 \text{ and } \mathrm{sign}(\beta_I) \in \mathrm{row}(X_I), \tag{2}$$

where $I = \{j \in [p] : \beta_j \neq 0\}$. Note that if $\ker(X_I) = \{0\}$, we have $\mathrm{sign}(\beta_I) \in \mathrm{row}(X_I)$ and the expression (2) coincides with the well-known *irrepresentability condition* for the LASSO given by $\|X_{I^c}'X_I(X_I'X_I)^{-1}\mathrm{sign}(\beta_I)\|_\infty \leq 1$, (Bühlmann & Van de Geer, 2011; Wainwright, 2009; Zou, 2006; Zhao & Yu, 2006). Thus, the irrepresentability condition for the LASSO can be thought of as an analytical shortcut for checking the noiseless recovery condition. For the sorted-ℓ_1-norm, when $m =$

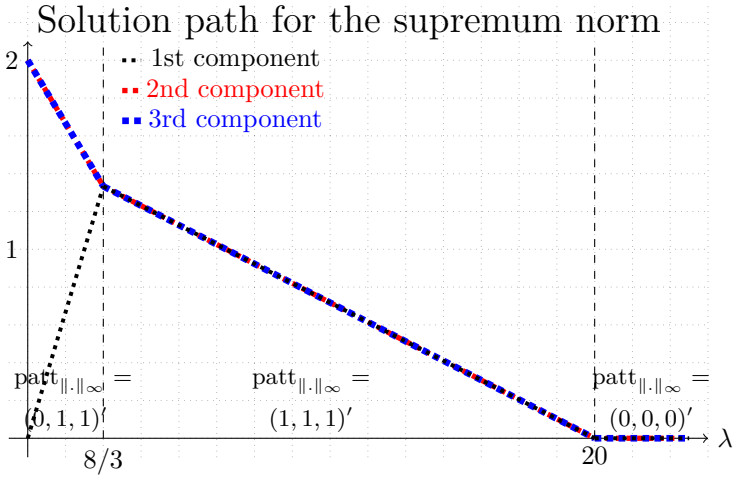

Figure 6: Shown are the curves of the three component functions $\lambda \mapsto \hat{\beta}_{\lambda,1}$ (black dotted curve), $\lambda \mapsto \hat{\beta}_{\lambda,2}$ (red dotted curve) and $\lambda \mapsto \hat{\beta}_{\lambda,3}$ (blue dotted curve) for $\lambda > 0$, where $\{\hat{\beta}_\lambda\} = S_{X,\lambda\|.\|_\infty}(X\beta)$. Note that $\text{patt}_\infty(\beta)$ satisfies the noiseless recovery condition. Indeed, $\text{patt}_\infty(\hat{\beta}_\lambda) = (0,1,1)'$ for $\lambda \in (0, 8/3)$.

$\text{patt}_{\text{slope}}(\beta)$, the noiseless recovery condition is equivalent to

$$\|X'(\tilde{X}'_m)^+ \tilde{W}_m\|_w^* \leq 1 \text{ and } \tilde{W}_m \in \text{row}(\tilde{X}_m),$$

where $\|.\|_w^*$ is the dual sorted-ℓ_1-norm, $\tilde{X}_m$ is the so-called clustered matrix and $\tilde{W}_m$ is the clustered parameter, see Bogdan et al. (2022) for details. In Proposition C.4 in Appendix C.2, we also provide an analytic characterization of the noiseless recovery condition for the supremum norm: Let $I = \{j \in [p] : |\beta_j| < \|\beta\|_\infty\}$, $\tilde{X} = (\tilde{X}_1 | X_I)$ where $\tilde{X}_1 = X_{I^c}\text{sign}(\beta_{\bar{I}})$. The noiseless recovery condition holds if and only if

$$\|X'(\tilde{X}')^+ e_1\|_1 \leq 1 \text{ and } e_1 \in \text{row}(\tilde{X}), \text{ where } e_1 = (1, 0, \ldots, 0)'.$$

Figure 6 confirms this characterization. Indeed, in the above example we have

$$\tilde{X} = \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix}, \; e_1 = (1,0)' \text{ and } X'(\tilde{X}')^+ e_1 = (0, 1/2, 1/2)'$$

and based on Figure 6 one may observe that the noiseless recovery condition holds for β. In the following, we show that

1. The noiseless recovery condition is a necessary condition for pattern recovery with a probability larger than 1/2, see Theorem 4.6.

2. Thresholded penalized estimators recover the pattern of β under much weaker condition than the noiseless recovery condition, see Section 5.

Theorem 4.6. *Let $Y = X\beta + \varepsilon$ where $X \in \mathbb{R}^{n\times p}$ is a fixed matrix, $\beta \in \mathbb{R}^p$ and ε follows a symmetric distribution. Let* pen *be a real-valued polyhedral gauge. If β does not satisfy the noiseless recovery*

condition with respect to X *and* pen*, then*

$$\mathbb{P}\left(\exists \lambda > 0 \; \exists \hat{\beta} \in S_{X,\lambda \mathrm{pen}}(Y) \text{ such that } \hat{\beta} \overset{\mathrm{pen}}{\sim} \beta\right) \leq 1/2.$$

By Theorem 4.6, if the noiseless recovery condition does not hold for the LASSO (for example, when $\|X_I' X_I (X_I' X_I)^{-1} \mathrm{sign}(\beta_I)\|_\infty > 1$), then

$$\mathbb{P}(\exists \lambda > 0 \; \exists \hat{\beta} \in S_{X,\lambda\|\cdot\|_1}(Y) \text{ such that } \mathrm{sign}(\hat{\beta}) = \mathrm{sign}(\beta)) \leq 1/2.$$

This above result is stronger than the one given in Theorem 2 in Wainwright (2009) which shows that $\mathbb{P}(\mathrm{sign}(\hat{\beta}^{\mathrm{LASSO}}(\lambda)) = \mathrm{sign}(\beta)) \leq 1/2$ for fixed $\lambda > 0$. Theorem 4.6 demonstrates that the noiseless recovery condition can be viewed as a unified irrepresentability condition in a general penalized estimation framework.

Clearly, if β satisfies the noiseless recovery condition with respect to X and pen, β is accessible with respect to X and pen by taking $y = X\beta$ in the definition of accessibility. In the following section, we show that thresholded penalized least-squares estimators recover the pattern of β under the accessibility condition only provided that the signal is strong enough.

5 Pattern recovery by thresholded penalized estimators

In practice, some additional information about β may be known a priori, e.g. its sparsity. Therefore it can be quite natural to threshold small components of $\hat{\beta}^{\mathrm{LASSO}}$ and so consider the thresholded LASSO estimator $\hat{\beta}^{\mathrm{LASSO},\tau}$ for some threshold $\tau \geq 0$. In fact, if the threshold is appropriately selected, the estimator allows to recover $\mathrm{sign}(\beta)$ under weaker conditions than LASSO itself (Tardivel & Bogdan, 2022). We aim at generalizing this property to the broader class of penalized estimators considered in this article. Before introducing this general notion of thresholded estimation, recall that for any threshold $\tau \geq 0$, the inclusion $\partial_{\|\cdot\|_1}(\hat{\beta}^{\mathrm{LASSO}}) \subseteq \partial_{\|\cdot\|_1}(\hat{\beta}^{\mathrm{LASSO},\tau})$ occurs. This observation is essential to formally define the notion of a thresholded estimator as defined in Definition 5.1 below. To motivate the formal definition, we list the following heuristic examples.

1. The penalty term $\|\cdot\|_\infty$ promotes clustering of components that are maximal in absolute value: Once $|\hat{\beta}_j| < \|\hat{\beta}\|_\infty$ but $|\hat{\beta}_j| \approx \|\hat{\beta}\|_\infty$, it is quite natural to set $|\hat{\beta}_j| = \|\hat{\beta}\|_\infty$. Let $\hat{\beta}^{\mathrm{thr}}$ be the estimator taking into account this approximation, obtained after slightly modifying $\hat{\beta}$. Then $\partial_{\|\cdot\|_\infty}(\hat{\beta}) \subseteq \partial_{\|\cdot\|_\infty}(\hat{\beta}^{\mathrm{thr}})$.

2. The sorted-ℓ_1-norm penalty promotes clustering of components equal in absolute value: Once $|\hat{\beta}_j^{\mathrm{SLOPE}}| \approx |\hat{\beta}_i^{\mathrm{SLOPE}}|$, it is quite natural to set $|\hat{\beta}_i^{\mathrm{SLOPE}}| = |\hat{\beta}_j^{\mathrm{SLOPE}}|$. Let $\hat{\beta}^{\mathrm{thr}}$ be the estimator taking into account this approximation and obtained after slightly modifying $\hat{\beta}^{\mathrm{SLOPE}}$. Then, $\partial_{\|\cdot\|_w}(\hat{\beta}^{\mathrm{SLOPE}}) \subseteq \partial_{\|\cdot\|_w}(\hat{\beta}^{\mathrm{thr}})$.

3. The penalty term $\|D^{tv} \cdot\|$ promotes neighboring components to be equal: Once $\hat{\beta}_j \approx \hat{\beta}_{j+1}$, it is quite natural to set $\hat{\beta}_j = \hat{\beta}_{j+1}$. Let $\hat{\beta}^{\mathrm{thr}}$ be the estimator taking into account this approximation and obtained after slightly modifying $\hat{\beta}$. Then, $\partial_{\|D^{\mathrm{tv}}\cdot\|_1}(\hat{\beta}) \subseteq \partial_{\|D^{\mathrm{tv}}\cdot\|_1}(\hat{\beta}^{\mathrm{thr}})$.

Motivated by the above examples, we define the concept of a thresholded estimator below.

Definition 5.1 (τ-thresholded penalized estimator). Let $X \in \mathbb{R}^{n\times p}$, $\lambda > 0$, and pen be a real-valued polyhedral gauge. Moreover, let $y \in \mathbb{R}^n$. Given $\hat{\beta} \in S_{X,\lambda\text{pen}}(y)$, we say that $\hat{\beta}^{\text{thr},\tau}$ is a τ-thresholded penalized estimator of $\hat{\beta}$ if

1. $\|\hat{\beta} - \hat{\beta}^{\text{thr},\tau}\|_\infty \leq \tau$,
2. $\partial_{\text{pen}}(\hat{\beta}) \subseteq \partial_{\text{pen}}(\hat{\beta}^{\text{thr},\tau})$,
3. $\dim(\partial_{\text{pen}}(b)) \leq \dim(\partial_{\text{pen}}(\hat{\beta}^{\text{thr},\tau}))$ for all b with $\|\hat{\beta} - b\|_\infty \leq \tau$.

Note that the thresholded LASSO is, in fact, an example of τ-thresholded estimator with threshold τ in the sense of the above definition. Another example of a τ-thresholded estimator when the penalty term is the supremum norm can be found in Algorithm 1 in Appendix C.4. Generally, we require the thresholded penalized $\hat{\beta}^{\text{thr},\tau}$ estimator to be close to the penalized estimator $\hat{\beta}$ (1.), to exhibit a pattern structure that is embedded in the pattern structure of $\hat{\beta}$ (2.), and to have a pattern of minimal complexity in that neighborhood of $\hat{\beta}$ (3.).

The notion of accessibility introduced for penalized estimators in Section 4 also covers thresholded estimators as can be learned from the proposition below.

Proposition 5.2. *Let* pen *be a real-valued polyhedral gauge, $X \in \mathbb{R}^{n\times p}$, $\lambda > 0$ and $\beta \in \mathbb{R}^p$. We have*

$$\begin{aligned} &\exists y \in \mathbb{R}^n,\, \exists \hat{\beta} \in S_{X,\lambda\text{pen}}(y) \text{ such that } \hat{\beta} \overset{\text{pen}}{\sim} \beta \\ &\quad \iff \exists y \in \mathbb{R}^n, \exists \hat{\beta} \in S_{X,\lambda\text{pen}}(y) \text{ such that } \partial_{\text{pen}}(\hat{\beta}) \subseteq \partial_{\text{pen}}(\beta). \end{aligned}$$

According to Propositions 4.2 and 5.2, if there exists $b \in \mathbb{R}^p$ such that $Xb = X\beta$ and $\text{pen}(b) < \text{pen}(\beta)$, then for any $y \in \mathbb{R}^n$, $\lambda > 0$, and $\hat{\beta} \in S_{X,\lambda\text{pen}}(y)$ we have $\partial_{\text{pen}}(\hat{\beta}) \not\subseteq \partial_{\text{pen}}(\beta)$. Consequently, no penalized *nor* thresholded penalized estimator can recover the pattern of β. On the other hand, if $\text{pen}(b) \geq \text{pen}(\beta)$ for all $b \in \mathbb{R}^p$ with $Xb = X\beta$, then both penalized and thresholded penalized estimator can recover the pattern of β albeit with different "choices" of y. Of course, in practice, a statistician does not aim at picking the appropriate y to recover the pattern of β, but instead uses the response of a linear regression model as a particular y to infer this pattern. Along these lines, by Theorem 4.6, if $Y = X\beta + \varepsilon$ under a symmetric distribution for ε, the noiseless recovery condition (a stronger condition than accessibility) is necessary for recovering the pattern of β via a penalized estimator with probability larger than 1/2. In Theorem 5.3, we show how the noiseless recovery condition can be relaxed when turning to thresholded estimators. More concretely, the minimal condition of accessibility is already sufficient for *sure* pattern recovery by thresholded estimation, provided that the signal of the pattern is "large enough", as is formalized in the following theorem.

Theorem 5.3. *Let* pen *be a real-valued polyhedral gauge, $X \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^p$, and $\lambda > 0$. Assume that uniform uniqueness holds, i.e. for any $y \in \mathbb{R}^n$, the set $S_{X,\lambda\text{pen}}(y)$ contains the unique minimizer $\hat{\beta}(y)$. For arbitrary $\varepsilon \in \mathbb{R}^n$ and for $r \in \mathbb{N}$, set $y^{(r)} = X(r\beta) + \varepsilon$. If* $\text{pen}(b) \geq \text{pen}(b)$ *for any $b \in \mathbb{R}^p$*

with $Xb = X\beta$, then there exists $r_0 \in \mathbb{N}$ and $\tau \geq 0$ such that for all $r \geq r_0$

$$\begin{cases} \partial_{\text{pen}}(b) \subseteq \partial_{\text{pen}}(\beta) \text{ for any } b \text{ with } \|b - \beta\|_\infty \leq \tau \\ \exists b_0 \text{ with } \|b_0 - \beta\|_\infty \leq \tau \text{ such that } b_0 \overset{\text{pen}}{\sim} \beta. \end{cases}$$

Consequently, a τ-thresholded penalized estimator $\hat{\beta}^{\text{thr},\tau}(y^{(r)})$ recovers the pattern of β.

Similar results for the LASSO (in which non-null components are large enough, i.e., $r \geq r_0$ in Theorem 5.3) are given in Tardivel & Bogdan (2022) and Descloux et al. (2022). In particular, Theorem 5.3 corroborates Theorem 1 in Tardivel & Bogdan (2022), which proves that the thresholded LASSO estimator recovers the sign of β once accessibility (termed identifiability in that reference) holds and non-null components of β are large enough.

6 A necessary and sufficient condition for uniform uniqueness

In Proposition 4.3, Corollary 4.4 and Theorem 5.3 we require uniform uniqueness, i.e., uniqueness of the penalized optimization problem (1) for a given $X \in \mathbb{R}^{n\times p}$ for all $\lambda > 0$ and all $y \in \mathbb{R}^n$. We provide a necessary and sufficient condition for this kind of uniqueness in Theorem 6.1 below. This theorem relaxes the coercivity condition for the penalty term needed in Theorem 1 in Schneider & Tardivel (2022) and extends the result to encompass methods such as the generalized LASSO.

Theorem 6.1 (Necessary and sufficient condition for uniform uniqueness)**.** *Let* pen *be a real-valued polyhedral gauge, $X \in \mathbb{R}^{n\times p}$, and $\lambda > 0$. Then the solution set $S_{X,\lambda\text{pen}}(y)$ from* (1) *is a singleton for all $y \in \mathbb{R}^n$ if and only if* $\text{row}(X)$ *does not intersect a face of B^* whose dimension*[4] *is strictly less than* $\text{def}(X) = \dim(\ker(X))$.

Note that a face F of B^* satisfies

$$\dim(F) < \text{def}(X) \iff \text{codim}(F) > \text{rk}(X),$$

where $\text{codim}(F) = p - \dim(F)$. Using Corollary 3.3 and Proposition 4.2, we may therefore conclude the following result from Theorem 6.1.

Corollary 6.2. *Let* pen *be a real-valued polyhedral gauge, $X \in \mathbb{R}^{n\times p}$, and $\lambda > 0$. Then the optimization problem in* (1) *is uniquely solvable for all $y \in \mathbb{R}^n$ if and only if no pattern with complexity exceeding* $\text{rk}(X)$ *is accessible.*

We now illustrate some cases of non-uniqueness occurring for the generalized LASSO with $\text{pen}(b) = \|Db\|_1$ for some $D \in \mathbb{R}^{m\times p}$. Clearly, the set of generalized LASSO minimizers $S_{X,\lambda\|D.\|_1}(y)$ is unbounded for every $y \in \mathbb{R}^n$ once $\ker(X) \cap \ker(D) \neq \{0\}$. Consequently, $\ker(X) \cap \ker(D) = \{0\}$ is a necessary condition for uniform uniqueness, yet, it is not sufficient, as illustrated in the example below.

[4]The dimension of a face is defined as the dimension of its affine hull, see Appendix A for details.

Example. An example of generalized LASSO optimization problem for which the set of minimizers is not restricted to a singleton is given in Barbara et al. (2019):

$$\operatorname*{Arg\,min}_{b\in\mathbb{R}^p} \frac{1}{2}\|y - Xb\|_2^2 + \frac{1}{2}\|Db\|_1 \text{ where } X = \begin{pmatrix} 1 & 1 & 1 \\ 3 & 1 & 1 \\ \sqrt{2} & 0 & 0 \end{pmatrix},\ D = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 2 & 1 & 1 \end{pmatrix} \text{ and } y = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}.$$

Note that $S_{X,\frac{1}{2}\|D.\|_1}(y) = \operatorname{conv}\{(0,1/2,0)', (0,0,1/2)'\}$ (Barbara et al., 2019). Since

$$\|Db\|_1 = \max\{\pm(4b_1 + 2b_2 + 2b_3), \pm(2b_1 + 2b_2), \pm(2b_1 + 2b_3)\}$$

we have $B^* = \operatorname{conv}\{\pm(4,2,2)', \pm(2,2,0)', \pm(2,0,2)'\}$. Because the vertex $F = (4,2,2)'$ is an element of $\operatorname{row}(X)$ and satisfies $\operatorname{codim}(F) = 3 - \dim(F) = 3 - 0 > 2 = \operatorname{rk}(X)$, uniform uniqueness cannot hold. This complies of course with the fact that $S_{X,\frac{1}{2}\|D.\|_1}(y)$ is not a singleton.

When $\ker(X) \cap \ker(D) = \{0\}$, in broad generality, the set of generalized LASSO minimizers is a polytope, i.e., a bounded polyhedron (Barbara et al., 2019), and extremal points can be computed explicitly (Dupuis & Vaiter, 2019). This description is relevant when the set of minimizers is not a singleton.

7 Numerical experiments

We illustrate the accessibility and the noiseless recovery condition in numerical experiments for the case when the penalty term is given by the supremum norm. More concretely, we start by illustrating the relationship between the probability of either condition holding and the particular pattern under consideration. For these simulations, we consider $n = 100$, $p = 150$ and the matrix $X = (X_1 \dots X_{150}) \in \mathbb{R}^{100\times 150}$ having iid $\mathcal{N}(0, 1/100)$ entries.

Clearly, both the noiseless recovery as well as the accessibility condition for the supremum norm depend on β through $\operatorname{patt}_\infty(\beta)$, providing information about maximal and non-maximal components (both understood in absolute value), as well as the signs of the maximal components in absolute value, see the example in Section 3 for details. Furthermore, since the distribution of X we consider here is invariant under changing signs of and permuting columns, the probability that a non-zero vector $\beta \in \mathbb{R}^p$ having k non-maximal components satisfies the noiseless recovery condition is given by

$$\mathbb{P}_X(\|X'(\tilde{X}')^+ e_1\|_\infty \le 1 \text{ and } e_1 \in \operatorname{row}(\tilde{X})),$$

where $\tilde{X} = (\tilde{X}_1 | X_I)$ with $\tilde{X}_1 = \sum_{j=1}^{p-k} X_j$ and X_j denoting the j-th column of X, $e_1 = (1, 0, \dots, 0)'$ and $I = \{p-k+1, \dots, p\}$, see also Proposition C.4 in Appendix C.2. This shows that the probability of satisfying the noiseless recovery condition for a non-zero $\beta \in \mathbb{R}^p$ depends only on the number of non-maximal components of β. Additionally, the accessibility condition is satisfied with probability

$$\mathbb{P}_X(\min\{\|\gamma\|_\infty : X\gamma = \tilde{X}_1\} = 1).$$

Note that asymptotically, when both n and p are large, the accessibility condition has probability

almost 1 when $k < 2n - p$ and probability almost 0 when $k > 2n - p$, see (Amelunxen et al., 2014). Figure 7 now provides both the probability of the accessibility condition and the noiseless recovery condition as a function of k, the number of non-maximal components. We see that the noiseless recovery condition essentially never holds for our choice of X, implying that the probability of pattern recovery is always bounded by $1/2$ in this setting.

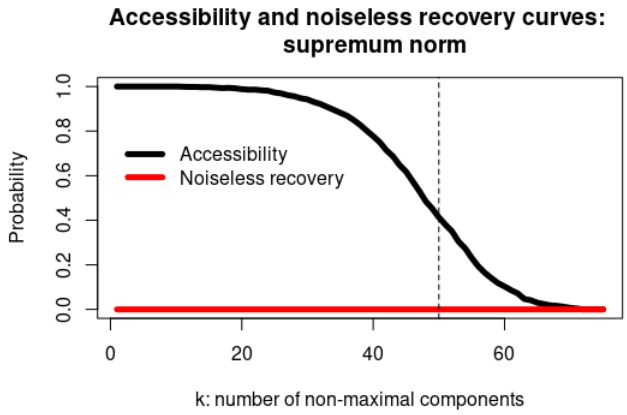

Figure 7: The probability of the accessibility and the noiseless recovery holding for (the pattern of) β, plotted as a function of the number of non-maximal components k of the β. For $k = 2n - p = 50$ (see the dotted line), the probability of the accessibility condition holding is roughly 0.5.

Given the insights from Figure 7 and the fact that thresholding the penalized estimator alleviates pattern recovery, as learned from Theorem 5.3, we illustrate the pattern recovery properties for penalized and thresholded penalized estimators in a second numerical study. For this, we consider the linear model $Y = X\beta + \varepsilon$ for a fixed $X \in \mathbb{R}^{100\times 150}$ generated once according to the distribution mentioned above and for $\varepsilon \in \mathbb{R}^n$ with iid $\mathcal{N}(0,1)$ entries. For $\beta \in \mathbb{R}^{150}$, we choose $\beta_1 = \cdots = \beta_{60} = 20$, $\beta_{61} = \cdots = \beta_{120} = -20$ and $\beta_{121} = \cdots = \beta_{150} = 0$. The tuning parameter is selected by the SURE formula which, for a given X and y, minimizes the function $0 < \lambda \mapsto \frac{1}{2}\|y - X\hat{\beta}_\lambda\|_2^2 + \mathrm{card}(\{j \in [p] : |\hat{\beta}_{\lambda,j}| < \|\hat{\beta}_\lambda\|_\infty\})$, see for example Minami (2020) or Vaiter et al. (2017). Note that, up to the additive constant 1, the second term is indeed the complexity of the pattern of $\hat{\beta}$ according to Definition 3.1. Figure 8 now shows how the penalized estimator fails to recover the pattern and how a thresholded estimator could detect the correct pattern given an appropriate threshold.

Acknowledgments

We would like to thank Samuel Vaiter, Mathurin Massias and Abderrahim Jourani for their insightful comments on the paper. Patrick Tardivel's institute is supported by the EIPHI Graduate School (contract ANR-17-EURE-0002) and his work was supported by the region Bourgogne-Franche-Comté (EPADM project). The work of Tomasz Skalski was supported by a French Government Scholarship.

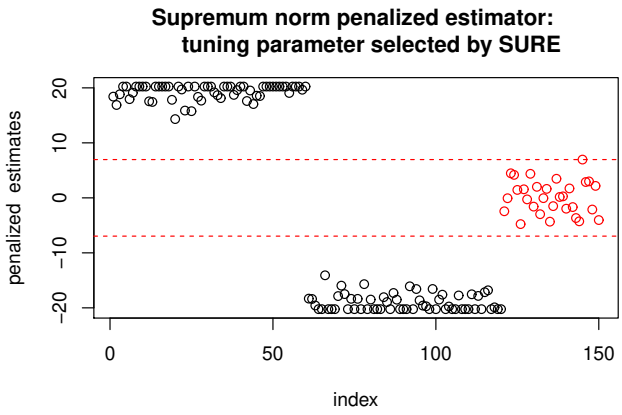

Figure 8: Illustration of pattern recovery by sup-norm penalized estimation and the potential of thresholded estimation: The figure shows a plot of $\hat{\beta}_\lambda \in \mathbb{R}^{150}$, by plotting each component $(j, \hat{\beta}_{\lambda,j})$. The true parameter β has $k = 30$ non-maximal components. Here, the noiseless recovery condition does not hold and we see that $\hat{\beta}_\lambda$ cannot recover the pattern of β, in particular maximal components of β are not estimated at a similar value by $\hat{\beta}_\lambda$. On the other hand, the accessibility condition does hold thus a thresholded penalized estimator may potentially recover the pattern of β. In particular, components of β equal to 20 (-20 or 0) are *approximately* estimated at 20 (-20 or 0, respectively). Thus – with an appropriate threshold – a thresholded penalized estimator can recover the pattern of β.

A Facts about polytopes and polyhedral gauges

We recall some basic definitions and facts about polytopes which we will use throughout the proofs. The following can be found in textbooks such as Gruber (2007) and Ziegler (2012).

A set $P \subseteq \mathbb{R}^p$ is called a *polytope* if it is the convex hull of a finite set of points $\{v_1, \ldots, v_k\} \subseteq \mathbb{R}^p$, namely,

$$P = \mathrm{conv}\{v_1, \ldots, v_k\}.$$

The *dimension* $\dim(P)$ of a polytope is defined as the dimension of $\mathrm{aff}(P)$, the affine subspace spanned by P. An inequality $a'x \leq c$ is called a *valid inequality* of P if $P \subseteq \{x \in \mathbb{R}^p : a'x \leq c\}$. A *face* F of P is any subset $F \subseteq P$ that satisfies

$$F = \{x \in P : a'x = c\} \text{ for some } a \in \mathbb{R}^p \text{ and } c \in \mathbb{R} \text{ with } P \subseteq \{x \in \mathbb{R}^p : a'x \leq c\}.$$

Note that $F = \varnothing$ and $F = P$ are faces of P and that any face F is again a polytope. A non-empty face F with $F \neq P$ is called *proper*. A point $x_0 \in P$ lies in $\mathrm{ri}(P)$, the *relative interior* of P, if x_0 is not contained in a proper face of P. We state two useful properties of faces in the following lemma.

Lemma A.1. *Let $P \subseteq \mathbb{R}^p$ be a polytope given by $P = \mathrm{conv}\{v_1, \ldots, v_k\}$, where $v_1, \ldots, v_k \in \mathbb{R}^p$. The following properties hold.*

1. *If F and $\tilde{F}$ are faces of P, then so is $F \cap \tilde{F}$.*

2. *Let L be an affine line contained in the affine span of P. If $L \cap \mathrm{ri}(P) \neq \varnothing$, then L intersects a*

proper face of P.

Lemma A.2 characterizes the connection between a certain class of convex functions (which encompasses polyhedral gauges) and the faces of a related polytope. The lemma is needed to prove Theorem 6.1.

Lemma A.2. *Let $v_1, \ldots, v_k \in \mathbb{R}^p$, P be the polytope $P = \operatorname{conv}\{v_1, \ldots, v_k\}$ and ϕ be the convex function defined by*

$$\phi(x) = \max\{v_1'x, \ldots, v_k'x\} \text{ for } x \in \mathbb{R}^p.$$

Then the subdifferential of ϕ at x is a face of P and is given by

$$\partial_\phi(x) = \operatorname{conv}\{v_l : l \in I_\phi(x)\} = \{s \in P : s'x = \phi(x)\}, \text{ where } I_\phi(x) = \{l \in [k] : v_l'x = \phi(x)\}.$$

Conversely, let F be a non-empty face of P. Then $F = \partial_\phi(x)$ for some $x \in \mathbb{R}^p$.

Proof. The fact that $\partial_\phi(x) = \operatorname{conv}\{u_l : l \in I_\phi(x)\}$ can be found in (Hiriart-Urruty & Lemarechal, 2001, p. 183). To prove the second equality, we consider the following. If $l \in I_\phi(x)$, by definition of $I_\phi(x)$, $v_l'x = \phi(x)$ and thus $v_l \in \{s \in P : s'x = \phi(x)\}$. Since $\{s \in P : s'x = \phi(x)\}$ is a convex set, one may deduce that

$$\operatorname{conv}\{v_l : l \in I_\phi(x)\} \subseteq \{s \in P : s'x = \phi(x)\}.$$

Conversely, assume $s \in P$ is such that $s \notin \operatorname{conv}\{v_l : l \in I_\phi(x)\}$. We then have $s = \sum_{l=1}^k \alpha_l v_l$ where $\alpha_1, \ldots, \alpha_k \geq 0$, $\sum_{l=1}^k \alpha_l = 1$ and $\alpha_{l_0} > 0$ for some $l_0 \notin I_\phi(x)$. Since $v_l'x \leq \phi(x)$ for all $l \in [k]$ and $u_{l_0}'x < \phi(x)$, we also get

$$s'x = \sum_{l=1}^k \alpha_l v_l'x < \phi(x).$$

Consequently, $s \notin \{s \in P : s'x = \phi(x)\}$ and thus

$$\{s \in P : s'x = \phi(x)\} \subseteq \operatorname{conv}\{v_l : l \in I_\phi(x)\}.$$

Therefore, $\partial_\phi(x) = \operatorname{conv}\{v_l : l \in I_\phi(x)\} = \{s \in P : s'x = \phi(x)\}$.

Now we show that the subdifferentials of ϕ are the (non-empty) faces of P. Let $x \in \mathbb{R}^p$. By definition of ϕ, $v_l'x \leq \phi(x)$ for every $l \in [k]$ so that the inequality $x's \leq \phi(x)$ is valid for all $s \in P$. This implies that $\partial_\phi(x)$ is a non-empty face of P. Conversely, let $F = \{s \in P : a's = c\}$ be a non-empty face of P where $a \in \mathbb{R}^p$, $c \in \mathbb{R}$ and $a's \leq c$ is a valid inequality for all $s \in P$. We prove that $F = \partial_\phi(a)$. For this, take any $s \in F$. We get $a's = c$ as well as $a's \leq \phi(a)$ as shown above, implying that $c \leq \phi(a)$. Analogously, for any $s \in \partial_\phi(a)$, $a's = \phi(a)$ as well as $a's \leq c$ since $\partial_\phi(a) \subseteq P$, yielding $\phi(a) \leq c$. Therefore we may deduce that $\phi(a) = c$ and thus $F = \partial_\phi(a)$. □

B Appendix – Proofs

We use the following additional notation for the remainder of the appendix. Given a matrix $A \in \mathbb{R}^{n \times p}$, $\operatorname{col}(A)$ represents the vector space spanned by columns of A: $\operatorname{col}(A) = \{Ab : b \in \mathbb{R}^p\}$ and A^+ stands

for the Moore-Penrose inverse of A. For a vector v, $v^\perp$ represents $\text{lin}(\{v\})^\perp$, the hyperplane orthogonal to v. We denote the convex cone or positive hull generated by $v_1, \ldots, v_k$ with $\text{cone}(v_1, \ldots, v_k)$.

B.1 Proof of Theorem 3.2 from Section 3

We now prove Theorem 3.2, the first part stating that the equivalence classes C_β with respect to pen coincide with the relative interior of the normal cones of the faces of B^*. Note that, since B^* is a polytope, the normal cones of B^* are the same at all points in the relative interior of a particular face of B^*, (see e.g. Ewald, 1996, p.16). Since $\partial_{\text{pen}}(\beta)$ is a face of B^* by Lemma A.2 in Appendix A, this means that $N_{B^*}(b) = N_{B^*}(\tilde{b})$ for all $b, \tilde{b} \in \text{ri}(\partial_{\text{pen}}(\beta))$. For simplicity, we write $N_{B^*}(\partial_{\text{pen}}(\beta))$ for the normal cone of B^* at (any) $b \in \text{ri}(\partial_{\text{pen}}(\beta))$, so that Theorem 3.2 states that

$$C_\beta = \text{ri}(N_{B^*}(\partial_{\text{pen}}(\beta))).$$

For the second part of Theorem 3.2, namely $\text{lin}(C_\beta) = \vec{\text{aff}}(\partial_{\text{pen}}(\beta))^\perp$, we will need the following lemma.

Lemma B.1. *Let P be the polyhedron $\{b \in \mathbb{R}^p : s_1' b \leq r_1, \ldots, s_m' b \leq r_m\}$, $\bar{b} \in P$ and $\bar{I} = \{l \in [m] : s_l' \bar{b} = r_l\}$. We then have*

$$\text{lin}(N_P(\bar{b})) = \vec{\text{aff}}(F)^\perp,$$

where F is the smallest face of P containing $\bar{b}$, i.e., $F = \{b \in P : s_l' b = r_l \quad \forall l \in \bar{I}\}$.

Proof. According to (Gruber, 2007, Proposition 14.1, p. 250), we have $N_P(\bar{b}) = \text{cone}(\{s_l\}_{l \in \bar{I}})$ and therefore $\text{lin}(N_P(\bar{b})) = \text{lin}(\{s_l\}_{l \in \bar{I}})$. Clearly, $\vec{\text{aff}}(F) \subseteq \text{lin}(\{s_l\}_{l \in \bar{I}})^\perp$ and conversely, if $h \in \text{lin}(\{s_l\}_{l \in \bar{I}})^\perp$ then, for $\eta > 0$ small enough, $s_l'(\bar{b} + \eta h) = r_l$ for all $l \in \bar{I}$ and $s_l'(\bar{b} + \eta h) < r_l$ for all $l \notin \bar{I}$. Therefore $\bar{b} + \eta h \in F$ and thus $h \in \vec{\text{aff}}(F)$ and consequently $\text{lin}(N_P(\bar{b})) = \text{lin}(\{s_l\}_{l \in \bar{I}}) = \vec{\text{aff}}(F)^\perp$. □

Proof of Theorem 3.2: $C_\beta = \text{ri}(N_{B^*}(\partial_{\text{pen}}(\beta)))$. We split the proof into four steps.

1) We first show that $C_\beta \subseteq N_{B^*}(\partial_{\text{pen}}(\beta))$. For this, take $b \in C_\beta$ and $v \in \text{ri}(\partial_{\text{pen}}(b))$. Since for any $z \in B^*$ we have

$$b'(z - v) = \underbrace{b'z}_{\leq \text{pen}(b)} - \underbrace{b'z}_{= \text{pen}(b)} \leq \text{pen}(b) - \text{pen}(b) = 0,$$

we may conclude that $b \in N_{B^*}(v) = N_{B^*}(\partial_{\text{pen}}(b)) = N_{B^*}(\partial_{\text{pen}}(\beta))$.

2) Next, we show that $\text{lin}(N_{B^*}(\partial_{\text{pen}}(\beta))) \subseteq \vec{\text{aff}}(\partial_{\text{pen}}(\beta))^\perp$: take $s \in N_{B^*}(\partial_{\text{pen}}(\beta))$ and note that this implies

$$s'(z - v) \leq 0 \quad \forall z \in B^*, \forall v \in \text{ri}(\partial_{\text{pen}}(\beta)).$$

Since $\partial_{\text{pen}}(\beta) \subseteq B^*$, we can conclude for any $v, w \in \text{ri}(\partial_{\text{pen}}(\beta))$ that both

$$s'(w - v) \leq 0 \text{ and } s'(v - w) \leq 0$$

hold so that $s \perp v - w$ for any $v, w \in \text{ri}(\partial_{\text{pen}}(\beta))$. Consequently $s \perp v - w$ for any $v, w \in \partial_{\text{pen}}(\beta)$[5].

[5]There exists sequences $(v_n)_{n \in \mathbb{N}}$ and $(w_n)_{n \in \mathbb{N}}$ in $\text{ri}(\partial_{\text{pen}}(\beta))$ such that $\lim_{n \to \infty} v_n = v$ and $\lim_{n \to \infty} w_n = w$. Since $s'(v_n - w_n) = 0$ one may deduce that $s'(v - w) = 0$.

Therefore, the following holds.

$$\operatorname{lin}(N_{B^*}(\partial_{\text{pen}}(\beta))) \subseteq \operatorname{lin}\{v - w : v, w \in \operatorname{ri}(\partial_{\text{pen}}(\beta))\}^{\perp} = \vec{\operatorname{aff}}(\partial_{\text{pen}}(\beta))^{\perp}.$$

3) By 1), we have $C_\beta \subseteq N_{B^*}(\partial_{\text{pen}}(\beta))$. We now establish the stronger result $C_\beta \subseteq \operatorname{ri}(N_{B^*}(\partial_{\text{pen}}(\beta)))$. For this, let $b \in C_\beta$. We show that $B(b, \varepsilon) \cap \operatorname{aff}(N_{B^*}(\partial_{\text{pen}}(\beta))) \subseteq N_{B^*}(\partial_{\text{pen}}(\beta))$ for small enough $\varepsilon > 0$, implying the desired claim. Take any $s \in B(b, \varepsilon) \cap \operatorname{aff}(N_{B^*}(\partial_{\text{pen}}(\beta)))$. By Lemma B.7, we know that $s \in B(b, \varepsilon)$ implies $\partial_{\text{pen}}(s) \subseteq \partial_{\text{pen}}(b) = \partial_{\text{pen}}(\beta)$ for small enough $\varepsilon > 0$. If $\partial_{\text{pen}}(s) \subsetneq \partial_{\text{pen}}(\beta)$, pick $v \in \partial_{\text{pen}}(s)$ and $w \in \partial_{\text{pen}}(\beta) \setminus \partial_{\text{pen}}(s)$. Since $v - w \in \vec{\operatorname{aff}}(\partial_{\text{pen}}(\beta))$ and $s \in \operatorname{aff}(N_{B^*}(\partial_{\text{pen}}(\beta))) \subseteq \operatorname{lin}(N_{B^*}(\partial_{\text{pen}}(\beta)))$ then, by 2), we have $s \in \vec{\operatorname{aff}}(\partial_{\text{pen}}(\beta))^{\perp}$ and therefore $s'(v - w) = 0$. Finally, since $s'v = \text{pen}(s)$, we may deduce that $s'w = \text{pen}(s)$ and thus $w \in \partial_{\text{pen}}(s)$ which leads to a contradiction. Consequently, $s \in C_\beta$ so that $B(b, \varepsilon) \cap \operatorname{aff}(N_{B^*}(\partial_{\text{pen}}(\beta))) \subseteq C_\beta \subseteq N_{B^*}(\partial_{\text{pen}}(\beta))$.

4) So far, we have shown that $C_\beta \subseteq \operatorname{ri}(N_{B^*}(\partial_{\text{pen}}(\beta)))$. We now argue that equality holds. For this, note that it is known that the relative interior of the normal cones provide a partition of the underlying space (see e.g. Ewald, 1996, p.17), so that the sets $\operatorname{ri}(N_{B^*}(\partial_{\text{pen}}(\beta)))$ form a partition of $\mathbb{R}^p$. Since the sets C_β also form a partition one may deduce that $C_\beta = \operatorname{ri}(N_{B^*}(\partial_{\text{pen}}(\beta)))$.

We now show the second part $\operatorname{lin}(C_\beta) = \vec{\operatorname{aff}}(\partial_{\text{pen}}(\beta))^{\perp}$. Because $C_\beta = \operatorname{ri}(N_{B^*}(\partial_{\text{pen}}(\beta)))$ and linear subspaces are closed, one may deduce that $N_{B^*}(\partial_{\text{pen}}(\beta))) \subseteq \operatorname{lin}(\operatorname{ri}(N_{B^*}(\partial_{\text{pen}}(\beta)))$. Consequently, $\operatorname{lin}(\operatorname{ri}(N_{B^*}(\partial_{\text{pen}}(\beta))) = \operatorname{lin}(N_{B^*}(\partial_{\text{pen}}(\beta)))$. Let $s \in \operatorname{ri}(\partial_{\text{pen}}(\beta))$. Because $\operatorname{lin}(N_{B^*}(\partial_{\text{pen}}(\beta))) = \operatorname{lin}(N_{B^*}(s))$ and since $\partial_{\text{pen}}(\beta)$ is the smallest face of B^* containing s, we may deduce by Lemma B.1 that $\operatorname{lin}(N_{B^*}(s)) = \operatorname{lin}(C_\beta) = \vec{\operatorname{aff}}(\partial_{\text{pen}}(\beta))^{\perp}$. □

B.2 Proofs for Section 4

Proof of Proposition 4.2

The following lemma can be seen as generalizing Proposition 4.1 from Gilbert (2017) from the ℓ_1-norm to all convex functions.

Lemma B.2. *Let $\beta \in \mathbb{R}^p$ and ϕ be a convex function on $\mathbb{R}^p$. Then* $\operatorname{row}(X)$ *intersects $\partial_\phi(\beta)$ if and only if, for any $b \in \mathbb{R}^p$, the following implication holds*

$$X\beta = Xb \implies \phi(\beta) \leq \phi(b). \tag{3}$$

Proof. Consider the function $\iota_\beta : \mathbb{R}^p \to \{0, \infty\}$ given by

$$\iota_\beta(b) = \begin{cases} 0 & \text{when } Xb = X\beta \\ \infty & \text{else.} \end{cases}$$

Then (3) holds for any $b \in \mathbb{R}^p$ if and only if β is a minimizer of the function $b \mapsto \phi(b) + \iota_\beta(b)$. It is straightforward to show that $\partial_{\iota_\beta}(\beta) = \operatorname{row}(X)$. We can therefore deduce that the implication (3) holds for any $b \in \mathbb{R}^p$ if and only if

$$0 \in \operatorname{row}(X) + \partial_\phi(\beta) \iff \operatorname{row}(X) \cap \partial_\phi(\beta) \neq \varnothing.$$

□

Proof of Proposition 4.2. By Lemma B.2, the geometric characterization of accessible patterns is equivalent to the analytic one. We show the geometric characterization.

($\Longrightarrow$) When the pattern of β is accessible with respect to X and λpen, there exists $y \in \mathbb{R}^n$ and $\hat{\beta} \in S_{X,\lambda\text{pen}}(y)$ such that $\hat{\beta} \overset{\text{pen}}{\sim} \beta$. Because $\hat{\beta}$ is a minimizer, $\frac{1}{\lambda}X'(y - X\hat{\beta}) \in \partial_{\text{pen}}(\hat{\beta}) = \partial_{\text{pen}}(\beta)$, so that, clearly, $\text{row}(X)$ intersects $\partial_{\text{pen}}(\beta)$.

($\Longleftarrow$) If $\text{row}(X)$ intersects the face $\partial_{\text{pen}}(\beta)$, then there exists $z \in \mathbb{R}^n$ such that $X'z \in \partial_{\text{pen}}(\beta)$. For $y = X\beta + \lambda z$, we have $\frac{1}{\lambda}X'(y - X\beta) = X'z$, so that $\beta \in S_{X,\lambda\text{pen}}(y)$, and the pattern of β is accessible with respect to X and λpen. □

Proof of Proposition 4.3

Proof. Assume that the pattern of β is accessible. Using Gilbert (2017, Proposition 5.2,(35))[6], we may conclude that there exists $z \in \mathbb{R}^n$ such that $X'z \in \text{ri}(\partial_{\text{pen}}(\beta))$. We set $y = \lambda z + X\beta$ and note that

$$\frac{1}{\lambda}X'(y - X\beta) = X'z \in \text{ri}(\partial_{\text{pen}}(\beta)),$$

so that $y \in A_\beta$. We now show that for small, but otherwise arbitrary $\varepsilon \in \mathbb{R}^n$, $y + \varepsilon$ still lies in A_β. For this, we decompose $\mathbb{R}^n$ into

$$\mathbb{R}^n = \text{col}(XU_\beta) \oplus \text{col}(XU_\beta)^\perp = \text{col}(XU_\beta) \oplus \ker(U_\beta' X'),$$

where $U_\beta \in \mathbb{R}^{p \times m}$ contains a basis of $\text{lin}(C_\beta)$ as columns. (Note that m is the complexity of the pattern β.) We accordingly decompose $\varepsilon = \tilde{\varepsilon} + \check{\varepsilon}$, where $\tilde{\varepsilon} \in \text{col}(XU_\beta)$ and $\check{\varepsilon} \in \ker(U_\beta' X')$ which satisfy $\|\check{\varepsilon}\|_2 \leq \|\varepsilon\|_2$ and $\|\tilde{\varepsilon}\|_2 \leq \|\varepsilon\|_2$. By construction, we have $\tilde{\varepsilon} = XU_\beta(XU_\beta)^+\tilde{\varepsilon}$. We set $\tilde{\beta} = \beta + U_\beta(XU_\beta)^+\tilde{\varepsilon}$. Note that $\beta \in C_\beta$ and $U_\beta(XU_\beta)^+\tilde{\varepsilon} \in \text{lin}(C_\beta)$. By Theorem 3.2, C_β is relatively open. Moreover, we have $\text{lin}(C_\beta) = \text{aff}(C_\beta) = \vec{\text{aff}}(C_\beta)$, which holds since 0 lies in the relative boundary of C_β by Theorem 3.2 and $\text{aff}(C_\beta)$ is closed, so that $0 \in \text{aff}(C_\beta)$. Therefore, there exists $r_0 > 0$ such that $\|\varepsilon\|_2 \leq r_0$ implies $\tilde{\beta} \in C_\beta$. Moreover,

$$\frac{1}{\lambda}X'(y + \varepsilon - X\tilde{\beta}) = \frac{1}{\lambda}X'\left(y + XU_\beta(XU_\beta)^+\tilde{\varepsilon} + \check{\varepsilon} - X(\beta + U_\beta(XU_\beta)^+\tilde{\varepsilon})\right) = X'z + \frac{1}{\lambda}X'\check{\varepsilon}.$$

Since $X'z \in \text{ri}(\partial_{\text{pen}}(\beta))$ and $X'\check{\varepsilon}/\lambda \in \text{col}(U_\beta)^\perp = \text{lin}(C_\beta)^\perp = \vec{\text{aff}}(\partial_{\text{pen}}(\beta))$, by Theorem 3.2, there exists $r_1 > 0$ such that $\|\varepsilon\|_2 \leq r_1$ implies $X'(y + \varepsilon - X\tilde{\beta})/\lambda \in \partial_{\text{pen}}(\beta)$. Finally, when $\|\varepsilon\|_2 \leq \min\{r_0, r_1\}$ then $\partial_{\text{pen}}(\tilde{\beta}) = \partial_{\text{pen}}(\beta)$ proving that $S_{X,\lambda\text{pen}}(y + \varepsilon) = \{\tilde{\beta}\}$, where $\tilde{\beta} \overset{\text{pen}}{\sim} \beta$. □

Proof of Theorem 4.6

Lemma B.3. *Let $\phi : \mathbb{R}^p \to \mathbb{R}$ be the polyhedral gauge defined as*

$$\phi(x) = \max\{u_1'x, \ldots, u_k'x\} \text{ for some } u_1, \ldots, u_k \in \mathbb{R}^p$$

[6] To make the connection to the constrained problem treated in this reference, set $A = X$ and $b = X\hat{\beta}(y)$.

If $\partial_\phi(x) = \partial_\phi(v)$, we have $\partial_\phi(x) = \partial_\phi(\alpha x + (1-\alpha)v) = \partial_\phi(v)$ for all $\alpha \in [0,1]$.

Proof. Let $s \in \partial_\phi(x) = \partial_\phi(v)$. Since s is a subgradient at x and at v, the following two inequalities hold

$$\phi(\alpha x + (1-\alpha)v) \geq \phi(x) - (1-\alpha)s'(x-v)$$
$$\phi(\alpha x + (1-\alpha)v) \geq \phi(v) + \alpha s'(x-v).$$

Multiplying the first inequality by α, the second by $(1-\alpha)$ and adding them, we get

$$\phi(\alpha x + (1-\alpha)v) \geq \alpha\phi(x) + (1-\alpha)\phi(v).$$

Using the convexity of ϕ, we arrive at

$$\phi(\alpha x + (1-\alpha)v) = \alpha\phi(x) + (1-\alpha)\phi(v).$$

By Lemma A.2 we have $\partial_\phi(x) = \operatorname{conv}\{u_l : l \in I\}$, where $I_\phi(x) = \{l \in [k] : u_l'x = \phi(x)\}$. Therefore, if $u_l \in \partial_\phi(x) = \partial_\phi(v)$, then $u_l'x = \phi(x)$ and $u_l'v = \phi(v)$, thus

$$u_l'(\alpha x + (1-\alpha)v) = \alpha\phi(x) + (1-\alpha)\phi(v) = \phi(\alpha x + (1-\alpha)v).$$

Consequently, $u_l \in \partial_\phi(\alpha x + (1-\alpha)v)$. On the other hand, if $u_l \notin \partial_\phi(x)$, then $u_l'x < \phi(x)$ and $u_l'v < \phi(v)$, thus

$$u_l'(\alpha x + (1-\alpha)v) < \alpha\phi(x) + (1-\alpha)\phi(v) = \phi(\alpha x + (1-\alpha)v).$$

Consequently, $u_l \notin \partial_\phi(\alpha x + (1-\alpha)v)$ and the claim follows. □

Lemma B.4. *Let $X \in \mathbb{R}^{n\times p}$ and $\beta \in \mathbb{R}^p$. The following set is convex*

$$V_\beta = \{y \in \mathbb{R}^n : \exists\lambda > 0\, \exists\hat{\beta} \in S_{X,\lambda\text{pen}}(y) \text{ such that } \hat{\beta} \overset{\text{pen}}{\sim} \beta\}.$$

Note that V_β may be empty.

Proof. Assume that $V_\beta \neq \varnothing$. Let $y, \tilde{y} \in V_\beta$. Then there exist $\lambda > 0$ and $\tilde{\lambda} > 0$ such that $\hat{\beta} \in S_{X,\lambda\text{pen}}(y)$ and $\tilde{\beta} \in S_{X,\tilde{\lambda}\text{pen}}(\tilde{y})$ with $\partial_{\text{pen}}(\hat{\beta}) = \partial_{\text{pen}}(\tilde{\beta}) = \partial_{\text{pen}}(\beta)$. Consequently,

$$X'(y - X\hat{\beta}) \in \lambda\partial_{\text{pen}}(\beta) \text{ and } X'(\tilde{y} - X\tilde{\beta}) \in \tilde{\lambda}\partial_{\text{pen}}(\beta).$$

Let $\alpha \in (0,1)$ and $\breve{y} = \alpha y + (1-\alpha)\tilde{y}$. Define $\breve{\lambda} = \alpha\lambda + (1-\alpha)\tilde{\lambda}$ and $\breve{\beta} = \alpha\hat{\beta} + (1-\alpha)\tilde{\beta}$. We show that $\breve{y} \in V_\beta$. Indeed, observe that

$$X'\left(\breve{y} - X\breve{\beta}\right) = \alpha X'(y - X\hat{\beta}) + (1-\alpha)X'(\tilde{y} - X\tilde{\beta}) \in \alpha\lambda\partial_{\text{pen}}(\beta) + (1-\alpha)\tilde{\lambda}\partial_{\text{pen}}(\beta) = \breve{\lambda}\partial_{\text{pen}}(\beta).$$

By Lemma B.3, $\partial_{\text{pen}}(\breve{\beta}) = \partial_{\text{pen}}(\alpha\hat{\beta} + (1-\alpha)\tilde{\beta}) = \partial_{\text{pen}}(\beta)$, so that $\breve{\beta} \in S_{X,\breve{\lambda}\text{pen}}(\breve{y})$ also, which proves the claim. □

Proof of Theorem 4.6. Assume that the noiseless recovery condition does not hold for β. Then $X\beta \notin V_\beta$, where V_β is defined as in Lemma B.4. Consequently, by convexity of V_β, we have $X\beta + \varepsilon \notin V_\beta$ or $X\beta - \varepsilon \notin V_\beta$ for any realization of $\varepsilon \in \mathbb{R}^n$. Therefore

$$\begin{aligned} 1 &= \mathbb{P}_\varepsilon(\{X\beta + \varepsilon \notin V_\beta\} \cup \{X\beta - \varepsilon \notin V_\beta\}) \\ &\leq \mathbb{P}_\varepsilon(\{X\beta + \varepsilon \notin V_\beta\}) + \mathbb{P}(\{X\beta - \varepsilon \notin V_\beta\}) = 2\mathbb{P}_\varepsilon(\{X\beta + \varepsilon \notin V_\beta\}). \end{aligned}$$

Consequently,

$$\frac{1}{2} \geq \mathbb{P}_\varepsilon(\{X\beta + \varepsilon \in V_\beta\}) = \mathbb{P}_\varepsilon(\exists \lambda > 0 \; \exists \hat{\beta} \in S_{X,\lambda\text{pen}}(Y) \text{ such that } \hat{\beta} \overset{\text{pen}}{\sim} \beta).$$

□

B.3 Proofs for Section 5

Proof of Proposition 5.2. We only need to prove the implication ($\Longleftarrow$), as the other direction is obvious. Assume that $\partial_{\text{pen}}(\hat{\beta}) \subseteq \partial_{\text{pen}}(\beta)$. Since $\hat{\beta} \in S_{X,\lambda\text{pen}}(y)$, we have $\frac{1}{\lambda}X'(y - X\hat{\beta}) \in \partial_{\text{pen}}(\hat{\beta}) \subseteq \partial_{\text{pen}}(\beta)$. Consequently, $\text{row}(X)$ intersects $\partial_{\text{pen}}(\beta)$ which implies that the pattern of β is accessible with respect to X and pen by Proposition 4.2. Consequently, there exists $y \in \mathbb{R}^n$ and there exists $\hat{\beta} \in S_{X,\lambda\text{pen}}(y)$ for which $\hat{\beta} \overset{\text{pen}}{\sim} \beta$. □

Proof of Theorem 5.3

Lemmas B.5 and B.6 are used to prove Theorem 5.3 which claims that, asymptotically, $\hat{\beta}(y^{(r)})/r$ converges to β when r tends to ∞.

Before stating these lemmas, note that for a non-empty closed and convex set $K \subseteq \mathbb{R}^p$ and $x \in K$, the asymptotic cone is defined as (cf. Hiriart-Urruty & Lemarechal, 2001)

$$K_\infty = \{d \in \mathbb{R}^p : x + td \in K \; \forall t > 0\}.$$

Moreover, the following statements hold.

- The set K_∞ does not depend on the choice of $x \in K$.
- A non-empty closed and convex set K is compact if and only if $K_\infty = \{0\}$.

Lemma B.5. *Let* pen *be a real-valued polyhedral gauge on* $\mathbb{R}^p$*,* $X \in \mathbb{R}^{n\times p}$*,* $v \in \text{col}(X)$*. Let* $K_1 \geq 0$*,* $K_2 \geq 0$ *be large enough such that* $K = \{b \in \mathbb{R}^p : \text{pen}(b) \leq K_1, \|Xb - v\|_2 \leq K_2\}$ *is non-empty. If* $\ker(X) \cap \ker(\text{pen}) = \{0\}$ *then, the set* K *is compact.*

Proof. Clearly, K is closed and convex. If $\text{pen}(d) > 0$ or if $Xd \neq 0$ then $d \notin K_\infty$. Consequently, $K_\infty \subset \ker(X) \cap \ker(\text{pen}) = \{0\}$ and thus K is compact. □

Lemma B.6. *Let* $X \in \mathbb{R}^{n\times p}$*,* $\lambda > 0$*,* pen *be a real-valued polyhedral gauge on* $\mathbb{R}^p$ *and assume that uniform uniqueness holds for* (1)*. Let* $\beta \in \mathbb{R}^p$*,* $\varepsilon \in \mathbb{R}^n$ *and set* $y^{(r)} = X(r\beta) + \varepsilon$*. If* β *is accessible with*

respect to X *and* pen*, then*

$$\lim_{r\to\infty} \hat{\beta}(y^{(r)})/r = \beta.$$

Proof. Since $\hat{\beta}(y^{(r)})$ is a minimizer of $S_{X,\lambda\text{pen}}(y^{(r)})$, the following inequality holds

$$\frac{1}{2}\|y^{(r)} - X\hat{\beta}(y^{(r)})\|_2^2 + \lambda\text{pen}(\hat{\beta}(y^{(r)})) \leq \frac{1}{2}\|y^{(r)} - X(r\beta)\|_2^2 + \lambda\text{pen}(r\beta).$$

Since $y^{(r)} - X(r\beta) = \varepsilon$, one may deduce that

$$\begin{aligned} \lambda\text{pen}(\hat{\beta}(y^{(r)})) &\leq \frac{1}{2}\|\varepsilon\|_2^2 + \lambda\text{pen}(r\beta) \\ \implies \text{pen}(\hat{\beta}(y^{(r)})/r) &\leq \frac{\|\varepsilon\|_2^2}{2\lambda r} + \text{pen}(\beta) \\ \implies \limsup_{r\to\infty} \text{pen}(\hat{\beta}(y^{(r)})/r) &\leq \text{pen}(\beta). \end{aligned} \tag{4}$$

Consequently, the sequence $\left(\text{pen}(\hat{\beta}(y^{(r)})/r)\right)_{r\in\mathbb{N}}$ is bounded. In addition, the Cauchy-Schwarz inequality gives the following implications

$$\begin{aligned} &\frac{1}{2}\|\varepsilon + X(r\beta) - X\hat{\beta}(y^{(r)})\|_2^2 + \lambda\text{pen}(\hat{\beta}(y^{(r)})) \leq \frac{1}{2}\|\varepsilon\|_2^2 + \lambda\text{pen}(r\beta) \\ &\quad\implies -\|\varepsilon\|_2 \|X(r\beta) - X\hat{\beta}(y^{(r)})\|_2 + \frac{1}{2}\|X(r\beta) - X\hat{\beta}(y^{(r)})\|_2^2 \leq \lambda\text{pen}(r\beta) - \lambda\text{pen}(\hat{\beta}(y^{(r)})) \\ &\quad\implies -\|\varepsilon\|_2/r \left\|X\left(\hat{\beta}(y^{(r)})/r - \beta\right)\right\|_2 + \frac{1}{2}\left\|X\left(\hat{\beta}(y^{(r)})/r - \beta\right)\right\|_2^2 \leq \lambda\text{pen}(\beta)/r - \lambda/r\,\text{pen}\left(\hat{\beta}(y^{(r)})/r\right). \end{aligned} \tag{5}$$

Let $\alpha \in [0,\infty]$ be the limes superior of the sequence

$$\left(\left\|X\left(\hat{\beta}(y^{(r)})/r - \beta\right)\right\|_2\right)_{r\in\mathbb{N}}. \tag{6}$$

By (5) we get

$$\limsup_{r\to\infty} \frac{\lambda\text{pen}(\beta) - \lambda\text{pen}(\hat{\beta}(y^{(r)})/r)}{r} \;\geq\; \begin{cases} \alpha^2/2 & \text{if } \alpha < \infty \\ \infty & \text{if } \alpha = \infty. \end{cases}$$

Moreover, by (4) we get

$$\limsup_{r\to\infty} \frac{\lambda\text{pen}(\beta) - \lambda\text{pen}(\hat{\beta}(y^{(r)})/r)}{r} = 0$$

We can conclude that $\alpha = 0$ and that the sequence (6) converges to 0.

Due to uniform uniqueness, we have $\ker(\text{pen})\cap\ker(X) = \{0\}$ and thus, by Lemma B.5, the sequence $(\hat{\beta}(y^{(r)})/r)_{r\in\mathbb{N}}$ is bounded. Therefore, to prove that $\lim_{r\to\infty} \hat{\beta}(y^{(r)})/r = \beta$, it suffices to show that β is the unique accumulation point of this sequence. We extract a subsequence $(\hat{\beta}(y^{\phi(r)})/\phi(r))_{r\in\mathbb{N}}$ converging to $\gamma \in \mathbb{R}^p$ (where $\phi : \mathbb{N} \to \mathbb{N}$ is an increasing function). By (4), one may deduce that

$\text{pen}(\gamma) \leq \text{pen}(\beta)$. Moreover, we get that

$$0 = \lim_{r\to\infty} \left\| X\left(\hat{\beta}(y^{(\phi(r))}/\phi(r) - \beta\right)\right\|_2^2 = \|X(\gamma-\beta)\|_2^2.$$

Finally, γ satisfies

$$X\gamma = X\beta \text{ and } \text{pen}(\gamma) \leq \text{pen}(\beta),$$

and we show that the only $\gamma \in \mathbb{R}^p$ satisfying the above is $\gamma = \beta$. Because the pattern of β is accessible, there exists $z \in \mathbb{R}^n$ such that $X'z \in \partial_{\text{pen}}(\beta)$. Let $y = X\beta + \lambda z$, then $\beta \in S_{X,\lambda\text{pen}}(y)$. Consequently, if there exists $\gamma \neq \beta$ such that $X\beta = X\gamma$ and $\text{pen}(\gamma) \leq \text{pen}(\beta)$, one may deduce that $\gamma \in S_{X,\lambda\text{pen}}(y)$ also, contradicting the uniform uniqueness assumption. Consequently, $\gamma = \beta$ and

$$\lim_{r\to\infty} \frac{\hat{\beta}(y^{(r)})}{r} = \beta.$$

□

Finally, the proof of the sufficient condition in Theorem 5.3 is based on Lemma B.6 and on Lemma B.7 given below.

Lemma B.7. *Let* pen *be a real-valued polyhedral gauge on* $\mathbb{R}^p$*. Then, there exists* $\tau > 0$ *depending on* β *such that*

$$\partial_{\text{pen}}(b) \subseteq \partial_{\text{pen}}(\beta) \text{ for all } b \in \overline{B}_\infty(\beta, \tau).$$

Proof. Let $I = \{l \in [k] : u_l'\beta = \text{pen}(\beta)\}$. By Lemma A.2, $\partial_{\text{pen}}(\beta) = \text{conv}\{u_l\}_{l\in I}$. Since

$$u_l'\beta < \text{pen}(\beta)\ \forall l \notin I,$$

and since linear functions and the gauge pen are continuous, one may pick $\tau > 0$ small enough such that

$$u_l'b < \text{pen}(b)\ \forall l \notin I,\ \forall b \in \overline{B}_\infty(\beta, \tau).$$

Consequently, for any $b \in \overline{B}_\infty(\beta, \tau)$, we have $J = \{l \in [k] : u_l'b = \text{pen}(b)\} \subseteq I$ and thus

$$\partial_{\text{pen}}(b) = \text{conv}\{u_l\}_{l\in J} \subseteq \text{conv}\{u_l\}_{l\in I} = \partial_{\text{pen}}(\beta).$$

□

Proof of Theorem 5.3. By Lemma B.7, there exists $\tau_0 > 0$ such that for any $b \in \overline{B}_\infty(\beta, \tau_0)$ we have $\partial_{\text{pen}}(b) \subseteq \partial_{\text{pen}}(\beta)$. By Lemma B.6, $\hat{\beta}(y^{(r)})/r$ converges to β when r tends to ∞. Consequently, we have

$$\exists r_0 \in \mathbb{N} \text{ such that } \forall r \geq r_0, \|\hat{\beta}(y^{(r)})/r - \beta\|_\infty \leq \tau_0/2.$$

Consequently, for $r \geq r_0$ we have

$$\begin{aligned} &\forall b \in \overline{B}_\infty(\hat{\beta}(y^{(r)})/r, \tau_0/2),\ \partial_{\text{pen}}(b) \subseteq \partial_{\text{pen}}(\beta) \text{ and} \\ &\exists \tilde{b} \in \overline{B}_\infty(\hat{\beta}(y^{(r)})/r, \tau_0/2), \partial_{\text{pen}}(\tilde{b}) = \partial_{\text{pen}}(\beta). \end{aligned}$$

Since for any $t > 0$ and any $x \in \mathbb{R}^p$, we have $\partial_{\text{pen}}(x) = \partial_{\text{pen}}(tx)$, one may deduce that

$$\forall b \in \overline{B}_\infty(\hat{\beta}(y^{(r)}), r\tau_0/2), \partial_{\text{pen}}(b) \subseteq \partial_{\text{pen}}(\beta)$$
$$\exists \tilde{b} \in \overline{B}_\infty(\hat{\beta}(y^{(r)}), r\tau_0/2), \partial_{\text{pen}}(\tilde{b}) = \partial_{\text{pen}}(\beta)$$

Consequently, the claim follows by taking $\tau = r\tau_0/2$. □

B.4 Proof of Theorem 6.1 from Section 6

The following lemma – needed to show Theorem 6.1 – states that the fitted values are unique over all non-unique solutions of the penalized problem for a given y. It is a generalization of Lemma 1 in Tibshirani (2013), which shows this fact for the special case of the LASSO.

Lemma B.8. *Let $X \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$, $\lambda > 0$ and* pen *be a polyhedral gauge. Then $X\hat{\beta} = X\tilde{\beta}$ and* $\text{pen}(\hat{\beta}) = \text{pen}(\tilde{\beta})$ *for all $\hat{\beta}, \tilde{\beta} \in S_{X,\text{pen}}(y)$.*

Proof. Assume that $X\hat{\beta} \neq X\tilde{\beta}$ for some $\hat{\beta}, \tilde{\beta} \in S_{X,\lambda\text{pen}}(y)$ and let $\check{\beta} = (\hat{\beta}+\tilde{\beta})/2$. Because the function $\mu \in \mathbb{R}^n \mapsto \|y-\mu\|_2^2$ is strictly convex, one may deduce that

$$\|y - X\check{\beta}\|_2^2 < \frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \frac{1}{2}\|y - X\tilde{\beta}\|_2^2.$$

Consequently,

$$\frac{1}{2}\|y - X\check{\beta}\|_2^2 + \lambda\text{pen}(\check{\beta}) < \frac{1}{2}\left(\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\text{pen}(\hat{\beta}) + \frac{1}{2}\|y - X\tilde{\beta}\|_2^2 + \lambda\text{pen}(\tilde{\beta})\right),$$

which contradicts both $\hat{\beta}$ and $\tilde{\beta}$ being minimizers. Finally, $X\hat{\beta} = X\tilde{\beta}$ clearly implies $\text{pen}(\hat{\beta}) = \text{pen}(\tilde{\beta})$. □

Proof of Theorem 6.1. ($\implies$) Assume that there exists a face F of $B^* = \text{conv}\{u_1, \ldots, u_k\}$ that intersects $\text{row}(X)$ and satisfies $\dim(F) < \text{def}(X)$. By Lemma A.2, $F = \partial_{\text{pen}}(\hat{\beta})$ for some $\hat{\beta} \in \mathbb{R}^p$. Let $z \in \mathbb{R}^n$ with $X'z \in F$, which exists by assumption. Now let $y = X\hat{\beta} + \lambda z$. Note that $\hat{\beta} \in S_{X,\lambda\text{pen}}(y)$ since

$$0 \in X'X\hat{\beta} - X'y + \lambda\partial_{\text{pen}}(\hat{\beta}) \iff \frac{1}{\lambda}X'(y - X\hat{\beta}) = X'z \in \partial_{\text{pen}}(\hat{\beta}).$$

We now construct $\tilde{\beta} \in S_{X,\lambda\text{pen}}(y)$ with $\tilde{\beta} \neq \hat{\beta}$. According to Lemma A.2, $\partial_{\text{pen}}(\hat{\beta}) = \text{conv}\{u_l : l \in I\}$ where $I = I_{\text{pen}}(\hat{\beta}) = \{l \in [k] : u_l'\hat{\beta} = \text{pen}(\hat{\beta})\}$ and thus $u_l'\hat{\beta} < \text{pen}(\hat{\beta})$ whenever $l \notin I$. We now show that it is possible to pick $h \in \ker(X)$ with $h \neq 0$ but $u_l'h = 0$ for all $l \in I$. We then make h small enough such that $u_l'(\hat{\beta} + h) \leq \text{pen}(\hat{\beta})$ still holds for all $l \notin I$, which in turn implies that $\text{pen}(\hat{\beta} + h) = \max\{u_l'\hat{\beta} : l \in I\} = \text{pen}(\hat{\beta})$. This, together with $X\hat{\beta} = X(\hat{\beta} + h)$, yields $\hat{\beta} \neq \tilde{\beta} = \hat{\beta} + h \in S_{X,\lambda\text{pen}}(y)$ also. We now show that $\ker(X) \cap \text{col}(U)^\perp \neq \{0\}$, where $U = (u_l)_{l\in I} \in \mathbb{R}^{p\times|I|}$. For this, we distinguish two cases:

1) Assume that $0 \in \text{aff}\{u_l : l \in I\}$. Then $\text{aff}\{u_l : l \in I\} = \text{col}(U)$ and $\dim(F) = \text{rk}(U) < \text{def}(X)$. This implies that

$$\dim(\ker(X)) + \dim(\text{col}(U)^\perp) > p,$$

which proves what was claimed.

2) Assume that $0 \notin \text{aff}\{u_l : l \in I\}$. Note that this implies that $v = X'z \in \text{row}(X) \cap \text{conv}\{u_l : l \in I\}$ satisfies $X'z \neq 0$. We also have $\text{rk}(U) = \dim(\text{aff}\{u_l : l \in I\}) + 1 = \dim(F) + 1 \leq \text{def}(X)$ which implies that

$$\dim(\ker(X)) + \dim(\text{col}(U)^\perp) \geq p.$$

If $\ker(X) \cap \text{col}(U)^\perp = \{0\}$, then $\mathbb{R}^p = \ker(X) \oplus \text{col}(U)^\perp$. But we also have $\ker(X) \subseteq v^\perp$ as well as $\text{col}(U)^\perp \subseteq v^\perp$, yielding a contradiction and proving the claim.

($\Longleftarrow$) Assume that there exists $y \in \mathbb{R}^n$ such that $\hat{\beta}, \tilde{\beta} \in S_{X,\lambda\text{pen}}(y)$ with $\hat{\beta} \neq \tilde{\beta}$. We then have

$$\frac{1}{\lambda}X'(y - X\hat{\beta}) \in \partial_{\text{pen}}(\hat{\beta}) \quad \text{and} \quad \frac{1}{\lambda}X'(y - X\tilde{\beta}) \in \partial_{\text{pen}}(\tilde{\beta}).$$

According to Lemma B.8, $X\hat{\beta} = X\tilde{\beta}$, thus $\frac{1}{\lambda}X'(y - X\hat{\beta}) = \frac{1}{\lambda}X'(y - X\tilde{\beta})$. Consequently, one may deduce that $\text{row}(X)$ intersects the face $\partial_{\text{pen}}(\hat{\beta}) \cap \partial_{\text{pen}}(\tilde{\beta})$. Let $F^* = \text{conv}\{u_l : l \in I^*\}$ be a face of $\partial_{\text{pen}}(\hat{\beta}) \cap \partial_{\text{pen}}(\tilde{\beta})$ of smallest dimension among all faces of $\partial_{\text{pen}}(\hat{\beta}) \cap \partial_{\text{pen}}(\tilde{\beta})$ intersecting $\text{row}(X)$. By minimality of $\dim(F^*)$, $\text{row}(X)$ intersects the relative interior of F^*, namely, there exists $z \in \mathbb{R}^n$ such that $v = X'z$ lies in F^*, but not on a proper face of F^*. We will now show that if $\dim(F^*) \geq \text{def}(X)$, then $\text{row}(X)$ intersects a proper face of F^*, yielding a contradiction.

For this, first observe that $\dim(F^*) = \dim(\text{aff}\{u_l : l \in I^*\})$ and that we can write the affine space $\text{aff}\{u_l : l \in I^*\} = u_{l_0} + \text{col}(\tilde{U}^*)$ where $l_0 \in I^*$ and $\tilde{U}^* = (u_l - u_{l_0})_{l \in I^* \setminus \{l_0\}} \in \mathbb{R}^{p \times |I^*|-1}$, implying that $\dim(F^*) = \text{rk}(\tilde{U}^*)$.

Now let $h = \hat{\beta} - \tilde{\beta} \neq 0$. Clearly, $h \in \ker(X)$. Moreover, since $\text{pen}(\hat{\beta}) = \text{pen}(\tilde{\beta})$ by Lemma B.8, and since $u_l \in \partial_{\text{pen}}(\hat{\beta}) \cap \partial_{\text{pen}}(\tilde{\beta})$ for all $l \in I^*$, by Lemma A.2, we get

$$u_l'h = u_l'\hat{\beta} - u_l'\tilde{\beta} = \text{pen}(\hat{\beta}) - \text{pen}(\tilde{\beta}) = 0 \; \forall l \in I^*.$$

Therefore, $h \in \ker(X) \cap \text{col}(U^*)^\perp$, where $U^* = (u_l)_{l \in I^*} \in \mathbb{R}^{p \times |I^*|}$. Assume that $\dim(F^*) \geq \text{def}(X)$. Then

$$\dim(\text{row}(X)) + \dim(\text{col}(\tilde{U}^*)) \geq \text{rk}(X) + \text{def}(X) = p.$$

If $\text{row}(X) \cap \text{col}(\tilde{U}^*) = \{0\}$, then $\mathbb{R}^p = \text{row}(X) \oplus \text{col}(\tilde{U}^*)$. However, the last relationship cannot hold since $\text{row}(X) = \ker(X)^\perp \subseteq h^\perp$ as well as $\text{col}(\tilde{U}^*) \subseteq \text{col}(U^*) \subseteq h^\perp$, where $h \neq 0$. Consequently, there exists $0 \neq \tilde{v} \in \text{row}(X) \cap \text{col}(\tilde{U}^*)$. The affine line $L = \{X'z + t\tilde{v} : t \in \mathbb{R}\} \subseteq \text{row}(X)$ intersects the relative interior of F^* at $t = 0$ and clearly lies in $\text{aff}(F^*) = u_{l_0} + \text{col}(\tilde{U}^*)$, since $X'z \in F^*$ and $\tilde{v} \in \text{col}(\tilde{U}^*)$. Therefore, L must intersect a proper face of F^* by Lemma A.1. But then also $\text{row}(X)$ intersects a proper face of F^*, which yields the required contradiction. □

C Appendix – Additional results

C.1 Existence of a minimizer

We show that the optimization problem of interest in this article always has a minimizer.

Proposition C.1. *Let $X \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$, $\mathrm{pen}(x) = \max\{u_1'x, \dots, u_l'x\}$ where $u_1, \dots, u_l \in \mathbb{R}^p$ with $u_1 = 0$. For*

$$f : b \in \mathbb{R}^p \mapsto \frac{1}{2}\|y - Xb\|_2^2 + \lambda \mathrm{pen}(b),$$

the optimization problem $\min_{b\in\mathbb{R}^p} f(b)$ has at least one minimizer.

For the remainder of this section, without loss of generality, we set $\lambda = 1$ since otherwise, this parameter can be absorbed into the penalty function. The proof of Proposition C.1 relies on the following two lemmas.

Lemma C.2. *Let the assumptions of Proposition C.1 hold and let $(\beta_m)_{m\in\mathbb{N}}$ be a minimizing sequence of f:*

$$\lim_{m\to\infty} f(\beta_m) = \inf_{b\in\mathbb{R}^p} f(b).$$

Then also $(X\beta_m)_{m\in\mathbb{N}}$ and $(\mathrm{pen}(\beta_m))_{m\in\mathbb{N}}$ converge. Moreover, these limits do not depend on the minimizing sequence.

Proof. The sequence $(X\beta_m)_{n\in\mathbb{N}}$ is bounded. Otherwise, $\|y - X\beta_m\|_2^2$ would be unbounded also, contradicting $\inf\{f(b) : b \in \mathbb{R}^p\} \le f(0) < \infty$. Let $\tilde{\beta}_m$ be another minimizing sequence. Note that also $X\tilde{\beta}_m$ is bounded. Now extract arbitrary converging subsequences $(X\beta_{n_m})_{m\in\mathbb{N}}$ and $(X\tilde{\beta}_{\tilde{n}_m})_{m\in\mathbb{N}}$ with limits l and $\tilde{l}$, respectively. Note that $(\beta_{n_m})_{m\in\mathbb{N}}$ and $(\tilde{\beta}_{\tilde{n}_m})_{m\in\mathbb{N}}$ are still minimizing sequences so that also $\mathrm{pen}(\beta_{n_m})$ and $\mathrm{pen}(\tilde{\beta}_{\tilde{n}_m})$ must converge. We now show that $l = \tilde{l}$. If $l \neq \tilde{l}$, set $\bar{\beta}_m = (\beta_{n_m} + \tilde{\beta}_{\tilde{n}_m})/2$. By the above considerations, $(f(\bar{\beta}_m))_{m\in\mathbb{N}}$ is convergent. Since the function $z \in \mathbb{R}^n \mapsto \|y - z\|_2^2$ is strictly convex and pen is convex, we may deduce that

$$\begin{aligned}
\limsup_{m\to\infty} f(\bar{\beta}_m) &\le \frac{1}{2}\|y - (l+\tilde{l})/2\|_2^2 + \limsup_{m\to\infty} \mathrm{pen}(\bar{\beta}_m) \\
&< \frac{1}{2}\left(\|y - l\|_2^2/2 + \|y - \tilde{l}\|_2^2/2\right) + \lim_{m\to\infty} \mathrm{pen}(\beta_{n_m})/2 + \lim_{m\to\infty} \mathrm{pen}(\tilde{\beta}_{\tilde{n}_m})/2 \\
&= \frac{1}{2}\lim_{m\to\infty} f(\beta_{n_m}) + \frac{1}{2}\lim_{m\to\infty} f(\tilde{\beta}_{\tilde{n}_m}) = \inf_{b\in\mathbb{R}^p} f(b),
\end{aligned}$$

yielding a contradiction. Since the selection of convergent subsequences was arbitrary, this implies that $(X\beta_m)_{m\in\mathbb{N}}$ and $(X\tilde{\beta}_m)_{m\in\mathbb{N}}$ share a unique limit point and that the sequences $(\mathrm{pen}(\beta_m))_{m\in\mathbb{N}}$ and $(\mathrm{pen}(\tilde{\beta}_m))_{m\in\mathbb{N}}$ converges as well. □

We remark that Lemma C.2 also holds for any non-negative, convex function in place of the polyhedral gauge pen.

Lemma C.3. *Let the assumptions of Proposition C.1 hold and let $\gamma \ge 0$. The optimization problem*

$$\min_{b\in\mathbb{R}^p} \|y - Xb\|_2^2 \quad \text{subject to} \quad \mathrm{pen}(b) \le \gamma \tag{7}$$

has at least one minimizer.

Proof. Let $P_\gamma = \{b \in \mathbb{R}^p : \mathrm{pen}(b) \le \gamma\}$ be the closed and convex feasible region of (7). We set $z = Xb$ and note that the linearly transformed set XP_γ is still closed and convex. Therefore, the minimization problem

$$\min \ \|y - z\|_2^2 \quad \text{subject to} \quad z \in XP_\gamma$$

has a unique solution $\hat{z} \in XP_\gamma$, namely, the projection of y onto XP_γ. Consequently, $\hat{z} = X\hat{b}$ for some $\hat{b} \in P_\gamma$, where $\hat{b}$ is not necessarily unique. Finally, $\hat{b}$ clearly is a solution of the optimization problem (7). □

Before we turn to the proof of Proposition C.1, we make the following observations. Note that we can decompose the polyhedron $P_\gamma = \{b \in \mathbb{R}^p : \text{pen}(b) \leq \gamma\} = \{b \in \mathbb{R}^p : u_1'b, \ldots, u_l'b \leq \gamma\}$, where $\gamma \geq 0$, into the sum of a polyhedral cone (the so-called recession cone of P_γ) and a polytope, (see, e.g., Ziegler, 2012, Theorem 1.2 and Proposition 1.12). For $\gamma = 1$, we can therefore write

$$P_1 = \{b \in \mathbb{R}^p : u_1'b \leq 0, \ldots, u_l'b \leq 0\} + E,$$

where E is a polytope and therefore bounded. For arbitrary $\gamma \geq 0$, we then write

$$P_\gamma = P_0 + \gamma E. \tag{8}$$

Proof of Proposition C.1. Let $(\beta_m)_{m \in N}$ be a minimizing sequence of f. By Lemma C.2, both sequences $(X\beta_m)_{m\in\mathbb{N}}$ and $(\text{pen}(\beta_m))_{m\in\mathbb{N}}$ converge to, say, l and γ, respectively. This implies that

$$\frac{1}{2}\|y - l\|_2^2 + \gamma = \inf_{b\in\mathbb{R}^p} f(b).$$

Let $\hat{\beta}$ be an arbitrary solution of (7). We prove that $f(\hat{\beta}) = \|y - l\|_2^2 + \gamma$. For this, we distinguish the following two cases.

1) Assume that $\gamma > 0$. For n large enough so that $\text{pen}(\beta_m) > 0$, we set u_n as

$$u_m = \frac{\gamma}{\text{pen}(\beta_m)}\beta_m.$$

Clearly, $\text{pen}(u_m) = \gamma$ so that $u_m \in P_\gamma$. Consequently, by definition of $\hat{\beta}$, we have $\|y - X\hat{\beta}\|_2^2 \leq \|y - Xu_m\|_2^2$ and $\text{pen}(\hat{\beta}) \leq \gamma$, so that

$$f(\hat{\beta}) = \frac{1}{2}\left\|y - X\hat{\beta}\right\|_2^2 + \text{pen}(\hat{\beta}) \leq \frac{1}{2}\|y - Xu_m\|_2^2 + \gamma \longrightarrow \frac{1}{2}\|y - l\|_2^2 + \gamma$$

as $m \to \infty$, implying $f(\hat{\beta}) = \inf\{f(b) : b \in \mathbb{R}^p\}$.

2) Assume that $\gamma = 0$. Using (8), we can write $\beta_m = u_m + \text{pen}(\beta_m)v_m$ with $u_m \in P_0$ and $v_m \in E$, where E is bounded. Since $X\beta_m \to l$ and $\text{pen}(\beta_m)v_n \to 0$ one may deduce that also $Xu_m \to l$, yielding

$$f(\hat{\beta}) = \frac{1}{2}\left\|y - X\hat{\beta}\right\|_2^2 \leq \frac{1}{2}\|y - Xu_m\|_2^2 \longrightarrow \frac{1}{2}\|y - l\|_2^2$$

as $m \to \infty$ implying again that $f(\hat{\beta}) = \inf\{f(b) : b \in \mathbb{R}^p\}$ which completes the proof. □

C.2 A characterization of the noiseless recovery condition for the supremum norm

Note that the noiseless recovery condition is always satisfied for $\beta = 0$. We give a characterization for $\beta \neq 0$ when the penalty term is given by the supremum norm.

Proposition C.4. *Let $X \in \mathbb{R}^{n\times p}$ and $\beta \in \mathbb{R}^p$ where $\beta \neq 0$ and $I = \{j \in [p] : |\beta_j| < \|\beta\|_\infty\}$. Furthermore, let $\tilde{X} = (\tilde{X}_1|X_I)$ where*

$$\tilde{X}_1 = X_{I^c}\text{sign}(\beta_{I^c}).$$

Then

$$\exists \lambda > 0, \exists \hat{\beta} \in S_{X,\lambda\|.\|_\infty}(X\beta) \text{ with } \hat{\beta} \overset{\|.\|_\infty}{\sim} \beta \iff e_1 \in \text{row}(\tilde{X}) \text{ and } \|X'(\tilde{X}')^+ e_1\|_1 \leq 1,$$

where $e_1 = (1, 0, \ldots, 0)' \in \mathbb{R}^p$.

Before presenting the proof, recall that the subdifferential of the ℓ_∞-norm at 0 is the unit ball of the ℓ_1-norm, and for $\beta \neq 0$, this subdifferential is equal to

$$\begin{aligned}\partial_{\|.\|_\infty}(\beta) &= \{s \in \mathbb{R}^p : \|s\|_1 \leq 1 \text{ and } s'\beta = \|\beta\|_\infty\} \\ &= \left\{ s \in \mathbb{R}^p : \|s\|_1 = 1 \text{ and } \forall j \in [p] \begin{cases} s_j\beta_j \geq 0 & \text{if } |\beta_j| = \|\beta\|_\infty \\ s_j = 0 & \text{otherwise} \end{cases} \right\}. \end{aligned} \tag{9}$$

Proof. ($\implies$) Assume there exists $\lambda > 0$ and $\hat{\beta} \in S_{X,\lambda\|.\|_\infty}(X\beta)$ such that $\hat{\beta} \overset{\|.\|_\infty}{\sim} \beta$. Then

$$\frac{1}{\lambda}X'(X\beta - X\hat{\beta}) \in \partial_{\|.\|_\infty}(\hat{\beta}) = \partial_{\|.\|_\infty}(\beta). \tag{10}$$

We set $c = (\|\beta\|_\infty, \beta_I')'$ and $\hat{c} = (\|\hat{\beta}\|_\infty, \hat{\beta}_I')'$. By construction, $\tilde{X}c = X\beta$. Moreover, since $\partial_{\|.\|_\infty}(\beta) = \partial_{\|.\|_\infty}(\hat{\beta})$, we also have $\tilde{X}\hat{c} = X\hat{\beta}$. Consequently, by (10), we get

$$\frac{1}{\lambda}X'\tilde{X}(c - \hat{c}) \in \partial_{\|.\|_\infty}(\beta).$$

Therefore, using (9), we get that

$$X_I'X(c - \hat{c}) = 0,$$

as well as

$$\beta'\frac{1}{\lambda}X'\tilde{X}(c - \hat{c}) = \frac{1}{\lambda}\beta_{I^c}'X_{I^c}'\tilde{X}(c - \hat{c}) = \frac{1}{\lambda}\|\beta\|_\infty\text{sign}(\beta_{I^c})'X_{I^c}'\tilde{X}(c - \hat{c}) = \frac{1}{\lambda}\|\beta\|_\infty\tilde{X}_1'\tilde{X}(c - \hat{c}) = \|\beta\|_\infty,$$

so that

$$\frac{1}{\lambda}\tilde{X}_1'\tilde{X}(c - \hat{c}) = 1.$$

Therefore, we may conclude

$$\frac{1}{\lambda}\tilde{X}'\tilde{X}(c-\hat{c}) = e_1 \implies \tilde{X}(c-\hat{c}) = \lambda(\tilde{X}')^+ e_1,$$

which also yields

$$\frac{1}{\lambda}X'(X\beta - X\hat{\beta}) = \frac{1}{\lambda}X'\tilde{X}(c-\hat{c}) = X'(\tilde{X}')^+ e_1 \in \partial_{\|.\|_\infty}(\beta).$$

We therefore immediately get $\|X'(\tilde{X}')^+ e_1\|_1 \leq 1$. It remains to show that and $e_1 \in \mathrm{row}(\tilde{X})$. Note that. analogously to above, $X'(\tilde{X}')^+ e_1 \in \partial_{\|.\|_\infty}(\beta)$ implies that

$$\tilde{X}'(\tilde{X}')^+ e_1 = e_1.$$

Since $\tilde{X}'(\tilde{X}')^+$ is the orthogonal projection onto $\mathrm{row}(\tilde{X})$, we may deduce that $e_1 \in \mathrm{row}(\tilde{X})$.

($\Longleftarrow$) As above, let $c = (\|\beta\|_\infty, \beta_I')'$ and set $\hat{c} = c - \lambda\tilde{X}^+(\tilde{X}')^+ e_1 = (\hat{c}_1, \hat{c}_{-1})'$, where the first component is $\hat{c}_1$ and remaining components are $\hat{c}_{-1}$. We define $\hat{\beta}$ through

$$\hat{\beta}_{I^c} = \hat{c}_1 \mathrm{sign}(\beta_{I^c}) \text{ and } \hat{\beta}_I = \hat{c}_{-1}.$$

Since $c_1 = \|\beta\|_\infty$ for small enough $\lambda > 0$, we have $\{j \in [p] : |\hat{\beta}_j| < \|\hat{\beta}\|_\infty\} = \{j \in [p] : |\beta_j| < \|\beta\|_\infty\} = I$ as well as $\beta_j\hat{\beta}_j > 0$ for $j \notin I$. Therefore, for small enough λ, $\partial_{\|.\|_\infty}(\hat{\beta}) = \partial_{\|.\|_\infty}(\beta)$ holds. To conclude the proof, it suffices to show that $\hat{\beta} \in S_{X,\lambda\|.\|_\infty}(X\beta)$, i.e., $\frac{1}{\lambda}X'(X\beta - X\hat{\beta}) \in \partial_{\|.\|_\infty}(\hat{\beta})$. Since $\mathrm{col}((\tilde{X}')^+) = \mathrm{col}(\tilde{X})$ and $\tilde{X}\tilde{X}^+$ is the orthogonal projection onto $\mathrm{col}(\tilde{X})$, we get

$$\frac{1}{\lambda}X'(X\beta - X\hat{\beta}) = \frac{1}{\lambda}X'(\tilde{X}c - \tilde{X}\hat{c}) = X'\tilde{X}\tilde{X}^+(\tilde{X}')^+ e_1 = X'(\tilde{X}')^+ e_1,$$

so that left to show is $X'(\tilde{X}')^+ e_1 \in \partial_{\|.\|_\infty}(\beta)$, which holds if both $\|X'(\tilde{X}')^+ e_1\|_1 \leq 1$ and $\beta' X'(\tilde{X}')^+ e_1 = \|\beta\|_\infty$ are true. The first inequality holds by assumption. To show the latter, note that the assumption $\tilde{X}'(\tilde{X}')^+ e_1 = e_1$ implies that

$$\|\beta\|_\infty = c' e_1 = c'\tilde{X}'(\tilde{X}')^+ e_1 = \beta' X'(\tilde{X}')^+ e_1.$$

Consequently, for $\lambda > 0$ small enough, $\hat{\beta} \in S_{X,\lambda\|.\|_\infty}(X\beta)$. □

C.3 A geometric characterization of the noiseless recovery condition

We provide a geometric criterion for the noiseless recovery condition in the theorem below. In particular, this characterization shows that the noiseless recovery condition depends on β only through its pattern.

Theorem C.5. *Let $X \in \mathbb{R}^{n\times p}$ and $\mathrm{pen} : \mathbb{R}^p \to \mathbb{R}$ be a real-valued polyhedral gauge. Let $\beta \in \mathbb{R}^p$. Then*

$$\exists \lambda > 0, \exists \hat{\beta} \in S_{X,\lambda\mathrm{pen}}(X\beta) \text{ with } \hat{\beta} \overset{\mathrm{pen}}{\sim} \beta \iff X'X\mathrm{lin}(C_\beta) \cap \partial_{\mathrm{pen}}(\beta) \neq \varnothing$$
$$\iff X'X\vec{\mathrm{aff}}(\partial_{\mathrm{pen}}(\beta))^\perp \cap \partial_{\mathrm{pen}}(\beta) \neq \varnothing$$

Proof. First note that by Theorem 3.2, $\mathrm{lin}(C_\beta) = \vec{\mathrm{aff}}(\partial_{\mathrm{pen}}(\beta))^\perp$. Therefore $X'X\mathrm{lin}(C_\beta) \cap \partial_{\mathrm{pen}}(\beta) \neq \varnothing \iff X'X\vec{\mathrm{aff}}(\partial_{\mathrm{pen}}(\beta))^\perp \cap \partial_{\mathrm{pen}}(\beta) \neq \varnothing$ and the last equivalence clearly holds. We now prove the first equivalence.

($\Longrightarrow$) Let $\hat{\beta} \in S_{X,\lambda\mathrm{pen}}(X\beta)$ with $\hat{\beta} \overset{\mathrm{pen}}{\sim} \beta$. Then

$$\frac{1}{\lambda}X'X(\beta - \hat{\beta}) \in \partial_{\mathrm{pen}}(\beta).$$

Since $\hat{\beta} \in C_\beta$, we get $(\beta - \hat{\beta})/\lambda \in \mathrm{lin}(C_\beta)$ which yields the desired implication.

($\Longleftarrow$) We assume that $X'X\mathrm{lin}(C_\beta) \cap \partial_{\mathrm{pen}}(\beta) \neq \varnothing$, i.e., there exists such $b \in \mathrm{lin}(C_\beta)$ such that $X'Xb \in \partial_{\mathrm{pen}}(\beta)$. We set $\hat{\beta} = \beta - \lambda b$. Again, by Theorem 3.2, C_β is relatively open. Note that $b \in \mathrm{lin}(C_\beta) = \mathrm{aff}(C_\beta) = \vec{\mathrm{aff}}(C_\beta)$, which holds since 0 lies in the relative boundary of C_β by Theorem 3.2 and $\mathrm{aff}(C_\beta)$ is closed, so that $0 \in \mathrm{aff}(C_\beta)$. Therefore, for small enough λ, we have $\hat{\beta} \in C_\beta$ and $\hat{\beta} \overset{\mathrm{pen}}{\sim} \beta$. Consequently,

$$\frac{1}{\lambda}X'(X\beta - X\hat{\beta}) = X'Xb \in \partial_{\mathrm{pen}}(\hat{\beta}),$$

so that $\hat{\beta} \in S_{X,\lambda\mathrm{pen}}(X\beta)$, which finishes the proof. □

C.4 An algorithm for a τ-thresholded estimator for the supremum norm

Algorithm 1 Thresholded penalized least-squares estimator when the penalty term is the ℓ_∞-norm:

Require: estimation: $\hat{\beta}$, threshold $\tau \geq 0$.
  if $\|\hat{\beta}\|_\infty \leq \tau$ **then**
    $\hat{\beta}^{\mathrm{thr},\tau} \leftarrow 0$.
  else

$$\forall j \in [p] : \hat{\beta}_j^{\mathrm{thr},\tau} \leftarrow \begin{cases} \|\hat{\beta}\|_\infty - \tau & \text{if } \hat{\beta}_j \geq \|\hat{\beta}\|_\infty - 2\tau \text{ and } \hat{\beta}_j \geq 0, \\ -\|\hat{\beta}\|_\infty + \tau & \text{if } \hat{\beta}_j \leq -\|\hat{\beta}\|_\infty + 2\tau \text{ and } \hat{\beta}_j < 0, \\ \hat{\beta}_j & \text{otherwise.} \end{cases}$$

  end if
  return $\hat{\beta}^{\mathrm{thr},\tau}$

References

Ali, A. & Tibshirani, R. J. (2019). The generalized lasso problem and uniqueness. *Electronic Journal of Statistics* **13**, 2307–2347.

Amelunxen, D., Lotz, M., McCoy, M. B. & Tropp, J. A. (2014). Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA* **3**, 224–294.

Barbara, A., Abderrahim, J. & Vaiter, S. (2019). Maximal solutions of sparse analysis regularization. *Journal of Optimization Theory and Applications* **180**, 374–396.

Bogdan, M., Dupuis, X., Graczyk, P., Kołodziejek, B., Skalski, T., Tardivel, P. & Wilczyński, M. (2022). Pattern recovery by SLOPE. Preprint 2203.12086, arXiv.

Bogdan, M., van den Berg, E., C. Sabatti, W. S. & Candès, E. J. (2015). SLOPE – adaptive variable selection via convex optimization. *Annals of Applied Statistics* **9**, 1103–1140.

Bondell, H. D. & Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics* **64**, 115–123.

Bühlmann, P. & Van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Heidelberg: Springer.

Descloux, P., Boyer, C., Josse, J., Sportisse, A. & Sardy, S. (2022). Robust Lasso-zero for sparse corruption and model selection with missing covariates. *Scandinavian Journal of Statistics* **49**, 1605–1635.

Dupuis, X. & Vaiter, S. (2019). The geometry of sparse analysis regularization. Preprint 1907.01769, arXiv.

Ewald, G. (1996). *Combinatorial Convexity and Algebraic Geometry*. Springer.

Gilbert, J. C. (2017). On the solution uniqueness characterization in the l1 norm and polyhedral gauge recovery. *Journal of Optimization Theory and Applications* **172**, 70–101.

Gruber, P. (2007). *Convex and Discrete Geometry*. Heidelberg: Springer.

Hejný, I., Wallin, J. & Bogdan, M. (2023). Weak pattern convergence for slope and its robust versions. Preprint 2303.10970, arxiv.

Hiriart-Urruty, J.-B. & Lemarechal, C. (2001). *Fundamentals of Convex Analysis*. Heidelberg: Springer.

Jégou, H., Furon, T. & Fuchs, J. J. (2012). Anti-sparse coding for approximate nearest neighbor search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Kim, S.-J., Koh, K., Boyd, S. & Gorinevsky, D. (2009). ℓ_1 trend filtering. *SIAM Review* **51**, 339–360.

Lee, J. D., Sun, D. L., Sun, Y. & Taylor, J. E. (2016). Exact post-selection inference with an application to the Lasso. *Annals of Statistics* **44**, 907–927.

Minami, K. (2020). Degrees of freedom in submodular regularization: A computational perspective of Stein's unbiased risk estimate. *Journal of Multivariate Analysis* **175**, 104546.

Mousavi, S. & Shen, J. (2019). Solution uniqueness of convex piecewise affine functions based optimization with applications to constrained l_1 minimization. *ESAIM: Control, Optimisation and Calculus of Variations* **25**, 1–56.

Negrinho, R. & Martins, A. (2014). Orbit regularization. In *Advances in Neural Information Processing Systems*, vol. 27.

Rockafellar, R. (1997). *Convex Analysis*. Princeton University Press.

Schneider, U. & Tardivel, P. (2022). The geometry of uniqueness, sparsity and clustering in penalized estimation. *Journal of Machine Learning Research* **23**, 1–36.

Sepehri, A. & Harris, N. (2017). The accessible lasso models. *Statistics* **51**, 711–721.

Takahashi, A. & Nomura, S. (2020). Efficient path algorithms for clustered Lasso and OSCAR. Preprint 2006.08965, arxiv.

Tardivel, P. & Bogdan, M. (2022). On the sign recovery by least absolute shrinkage and selection operator, thresholded least absolute shrinkage and selection operator and thresholded basis pursuit denoising. *Scandinavian Journal of Statistics* , to appear.

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

Tibshirani, R. J. (2013). The Lasso problem and uniqueness. *Electronic Journal of Statistics* **7**, 1456–1490.

Tibshirani, R. J., Sanders, M., Rosset, S., Zhu, J. & Knight, K. (2005). Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society Series B* **67**, 91–108.

Tibshirani, R. J. & Taylor, J. (2011). The solution path of the generalized Lasso. *Annals of Statistics* **39**, 1335–1371.

Vaiter, S., Deledalle, C., Fadili, J., Peyré, G. & Dossal, C. (2017). The degrees of freedom of partly smooth regularizers. *Annals of the Institute of Statistical Mathematics* **69**, 791–832.

Vaiter, S., Goldabaee, M., Fadili, J. & Peyré, G. (2015). Model selection with low complexity priors. *Information and Inference: A Journal of the IMA* **4**, 230–287.

Vaiter, S., Peyré, G. & Fadili, J. (2018). Model consistency of partly smooth regularizers. *IEEE Transactions on Information Theory* **64**, 1725–1737.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory* **55**, 2183–2202.

Zhao, P. & Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research* **7**, 2541–2563.

Ziegler, G. (2012). *Lectures on Polytopes*, vol. 152. New York: Springer.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.