OPT2022: 14th Annual Workshop on Optimization for Machine Learning

The Solution Path of the Group Lasso

Aaron Mishkin AMISHKIN@CS.STANFORD.EDU
Mert Pilanci PILANCI@STANFORD.EDU
Stanford University, USA

Abstract

We prove continuity of the solution path for the group lasso, a popular method of computing group-sparse models. Unlike the more classical lasso method, the group lasso solution path is non-linear and cannot be evaluated algorithmically. To circumvent this, we first characterize the group lasso solution set and then show how to construct an implicit function for the min-norm path. We prove our implicit representation is continuous almost everywhere and extend this to continuity everywhere when the group lasso solution is unique. These results can be viewed as extending solution path analyses from the lasso to the group lasso and imply that grid-search is a sensible approach to hyper-parameter selection. Our work applies to linear models as well as convex reformulations of neural networks and provides new tools for understanding solution paths of shallow ReLU models.

1. Introduction

The group lasso extends the classical lasso algorithm (R. Tibshirani, 1996) for feature-sparse regression to problems with group structure. Similar to the lasso, the group lasso is an *embedded method*; it achieves group sparsity by augmenting the training objective with a sparsity-inducing penalty function, in this case the sum of the ℓ_2 norms of each feature group (Bakin et al., 1999; Lin and H. H. Zhang, 2003; Yuan and Lin, 2006). Solving the resulting non-smooth optimization problem gives a group-sparse solution, with the degree of sparsity controlled by a regularization parameter.

Group-sparsity is desirable in learning problems for both computational and statistical reasons. Computationally, group-sparse models require less memory and can speed-up prediction by leveraging fast algorithms for sparse linear algebra (Wen et al., 2016). They may also generalize better at test time than dense models (Huang and T. Zhang, 2010; Mitra and C.-H. Zhang, 2016). Finally, group-sparse models like the group lasso are consistent in certain settings, meaning they recover the support of the ground-truth model (Bach, 2008; Liu and J. Zhang, 2009; Nardi and Rinaldo, 2008).

The group lasso also has deep connections to neural network optimization. Recent work by Pilanci and Ergen (2020) shows that two-layer neural networks with ReLU activations can be reformulated as a group lasso problem with additional constraints. Dropping these constraints yields a simpler "gated ReLU" neural network with the same expressive power (Fiat et al., 2019; Mishkin et al., 2022). In both cases, the degree of group sparsity corresponds directly to the number of hidden units in the final network. Thus, properties of the group lasso translate into properties of neural nets.

This paper studies the solution function—the mapping from regularization parameter to optimal model—of the linear group lasso problem. Our main contribution is a proof that the solution function is continuous when the group lasso solution is unique. Although technical, continuity of the solution

function is necessary for efficient tuning of the regularization parameter (see, e.g. Nesterov et al. (2018, Sec. 1.1)) and our result is an important "sanity check" that parameter tuning is possible for the group lasso. As part of our proof, we characterize the set of optimal models as well as the min-norm model, and provide new sufficient conditions for the group lasso solution to be unique. Finally, our work lays the foundation for analyzing the solution functions of (convex) neural networks.

1.1. Related Work

The Lasso Solution Path: Efron et al. (2004) and Osborne et al. (2000) develop homotopy methods for directly computing the lasso solution path, which is piecewise linear. These works establish continuity of the lasso solution as a side affect, but also require the solution to be unique. Hastie et al. (2007) connect the lasso path to forward stage-wise regression, while R. J. Tibshirani (2013) extend many results, including path continuity, to the min-norm solution.

Beyond the Lasso: Limited results exist outside of the lasso setting. R. J. Tibshirani and Taylor (2011) analyze the generalized lasso and provide path-computation algorithms in this setting. Vaiter et al. (2012) establish almost-everywhere continuity of the group lasso solution function with respect to the targets, rather than the regularization parameter.

2. Notation

We use lower-case a to denote vectors and upper-case A for matrices. Calligraphic letters $\mathcal{C}$ denote sets. The boundary of a set is written $\text{bd}(\mathcal{C})$. For a block of indices $b_i \subseteq \{1, \dots, d\}$, we write A_{b_i} for the sub-matrix of columns indexed by b_i. Similarly, a_{b_i} is the sub-vector indexed by b_i. If $\mathcal{M}$ is a collection of blocks, then $A_{\mathcal{M}}$ is the submatrix and $a_{\mathcal{M}}$ the sub-vector with columns/elements indexed by blocks in the collection. Finally, $|\mathcal{M}|$ is cardinality of the union of blocks in $\mathcal{M}$.

3. The Group Lasso

Let $X \in \mathbb{R}^{n \times d}$ be a data matrix, $y \in \mathbb{R}^n$ the associated vector of targets, and $\mathcal{B} = \{b_1, \dots, b_m\}$ a disjoint partition of the feature indices $\{1, \dots, d\}$. Given a regularization parameter $\lambda \geq 0$, the linear group lasso solves the following regularized regression problem:

$$\min_w f_\lambda(w) := \frac{1}{2}\|Xw - y\|_2^2 + \lambda \sum_{b_i \in \mathcal{B}} \|w_{b_i}\|_2. \tag{1}$$

Solutions to Eq. (1) are block sparse when λ is sufficiently large, meaning $w_{b_i} = 0$ for some subset of b_i. This is similar to the feature sparsity given by the lasso, to which the group lasso naturally reduces when $b_i = \{i\}$ for each $b_i \in \mathcal{B}$.

Our primary interest in this work is the solution function or "regularization path",

$$\mathcal{W}^*(\lambda) := \arg\min_w f_\lambda(w).$$

For a general data matrix, f_λ is not strictly convex and the linear group lasso problem may admit multiple solutions. As such, $\mathcal{W}^*(\lambda) \subset \mathbb{R}^d$ is set-valued and only the min-norm solution mapping

$$w^*(\lambda) = \arg\min \{\|w\|_2 : w \in \mathcal{W}^*(\lambda)\},$$

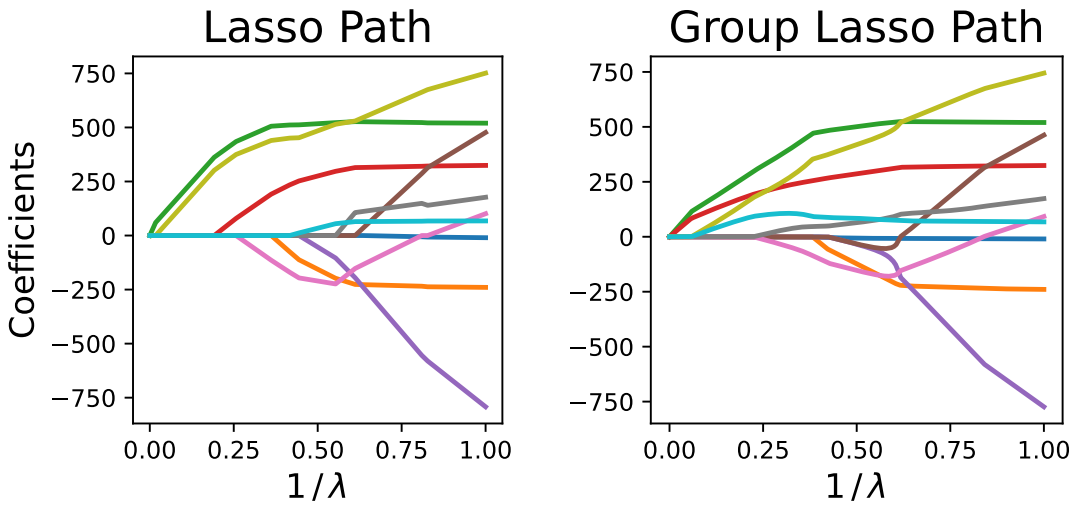

Figure 1: Comparison of solution paths for the lasso and group lasso models on the diabetes dataset (Dua and Graff, 2017). The lasso path was generated using the lasso modification of the LARS algorithm via Scikit-Learn (Pedregosa et al., 2011). Note the obvious non-linearity in the group lasso path, while the lasso is piece-wise linear. Such non-linearity prevents an explicit characterization of the path between breakpoints and significantly complicates our study of continuity.

defines a function. Even in the non-unique case, the model fit $\hat{y}(\lambda) = Xw$ is the same for all $w \in \mathcal{W}^*(\lambda)$ (Vaiter et al., 2012, Lemma 2). Moreover, if $\lambda > 0$, then it is straightforward to deduce that the sum of group norms $\sum_{i \in \mathcal{B}} \|w_{b_i}\|_2$ is also constant over $\mathcal{W}^*(\lambda)$. Uniqueness of $\hat{y}(\lambda)$ extends to the residual vector $r(\lambda) = y - \hat{y}(\lambda)$ and the block correlation vectors $v_{b_i}(\lambda) = X_{b_i}^\top r(\lambda)$. We write $v \in \mathbb{R}^d$ as the concatenation of the v_{b_i} blocks; v plays a critical role in defining $\mathcal{W}^*(\lambda)$.

Unlike the lasso, the group lasso solution function is not piece-wise linear (see Fig. 1). This prevents us from obtaining an explicit expression for the solution path between breakpoints, where features enter or exit the active set. We overcome this difficulty by relying only on an *implicit* characterization of the path provided by the implicit function theorem. We start towards this goal by developing optimality conditions for the group lasso problem.

Since Eq. (1) is a convex optimization problem, first-order (FO) optimality conditions are both necessary and sufficient for weights w to be globally optimal. Sub-differentiating the optimization problem gives the following FO conditions:

$$0 \in \partial f_\lambda(w) \iff v_{b_i}(\lambda) := X_{b_i}^\top (y - Xw) \in \begin{cases} \left\{ \lambda \frac{w_{b_i}}{\|w_{b_i}\|_2} \right\} & \text{if } w_{b_i} \neq 0 \\ \{v : \|v\|_2 \leq \lambda\} & \text{otherwise,} \end{cases} \tag{2}$$

showing that $w_{b_i} = 0$ when $\|v_{b_i}\|_2 < \lambda$. As a result, the equicorrelation set

$$\mathcal{E}_\lambda := \left\{ b_i \in \mathcal{B} : \|X_{b_i}^\top (y - Xw)\|_2 = \lambda \right\}, \tag{3}$$

contains all blocks which may be active for fixed λ. The complement of $\mathcal{E}_\lambda$ in $\mathcal{B}$ is the inactive set, denoted $\mathcal{I}_\lambda$. We combine FO conditions with uniqueness of $\hat{y}(\lambda)$ to characterize the solution function in terms of $\text{Null}(X_{\mathcal{E}_\lambda})$.

Proposition 1 *Let $\lambda > 0$ and $\mathcal{N}_\lambda = Null(X_{\mathcal{E}_\lambda}) \bigcap \{z : z_{b_i} \in Span(v_{b_i}), i \in \mathcal{E}_\lambda\}$. The optimal set is*

$$\mathcal{W}^*(\lambda) = \left\{ w \in \mathbb{R}^d : w_{\mathcal{E}_\lambda} = w^*_{\mathcal{E}_\lambda}(\lambda) + z, z \in \mathcal{N}_\lambda,\ w_{\mathcal{I}_\lambda} = 0,\ w_{b_i} \neq 0 \implies \frac{\lambda w_{b_i}}{\|w_{b_i}\|_2} = v_{b_i}(\lambda) \right\}.$$

Proposition 1 extends a similar result for the lasso solution to the group lasso (R. J. Tibshirani, 2013, Eq. 9). It implies the group lasso solution is unique when the columns of X are linearly independent; we will use it later to obtain a more refined condition for uniqueness. See Appendix A for proof.

4. The Minimum-norm Solution

Now we turn our attention to the min-norm solution and its path. Unlike $\mathcal{W}^*$, the min-norm solution is a single-valued function of λ and can be analyzed precisely. Our focus is on developing new characterizations of w^*. All proofs are deferred to Appendix B.

First, we introduce additional notation which will be necessary throughout our discussion. Divide the equicorrelation set into active and transitioning blocks, respectively,

$$\mathcal{A}_\lambda(w) := \{b_i \in \mathcal{B} : w_{b_i} \neq 0\}, \quad \mathcal{T}_\lambda(w) := \{b_i \in \mathcal{B} : w_{b_i} = 0, \|v_{b_i}\|_2 = \lambda\}.$$

Intuitively, $\mathcal{T}_\lambda(w)$ is the set of blocks which may become active (i.e. enter $\mathcal{A}_\lambda$) or cease to be equicorrelated in a neighbourhood of λ. We write $\mathcal{A}^*_\lambda$ and $\mathcal{T}^*_\lambda$ for the active and transiting blocks of the min-norm solution. By manipulating FO conditions, we now give a concise expression for $w^*_{\mathcal{A}^*_\lambda}$ in terms of the correlation vector v.

Lemma 2 *For $\lambda > 0$, the active blocks of the min-norm solution are given by*

$$w^*_{\mathcal{A}^*_\lambda}(\lambda) = (X^\top_{\mathcal{A}^*_\lambda} X_{\mathcal{A}^*_\lambda})^+ X^\top_{\mathcal{A}^*_\lambda} \left(y - (X^\top_{\mathcal{A}^*_\lambda})^+ v_{\mathcal{A}^*_\lambda}(\lambda) \right) \tag{4}$$

As a corollary, $w^*_{\mathcal{A}^*_\lambda}$ is the unique solution to the group lasso which is orthogonal to $\text{Null}(X_{\mathcal{A}^*_\lambda})$. Although we state this lemma in terms of $X_{\mathcal{A}^*_\lambda}$, it is easily extended to $X_{\mathcal{E}_\lambda}$ by including FO conditions for the blocks $\mathcal{T}^*_\lambda$ in the same argument. Thus, Lemma 2 also has computational consequences: given any solution $w \in \mathcal{W}^*(\lambda)$, the unique correlation vector $v(\lambda)$ can be computed and used to find w^* as the solution to a linear system. Since $w^*_{\mathcal{A}^*_\lambda}$ has no component of $\text{Null}(X_{\mathcal{A}^*_\lambda})$, we can also compute it via a *reduced-block* optimization problem constrained to $\text{Row}(X_{\mathcal{A}^*_\lambda})$.

Lemma 3 *Suppose the support of the min-norm solution is $\mathcal{A}^*_\lambda$. Then the active blocks are the unique solution to the following optimization problem:*

$$w^*_{\mathcal{A}^*_\lambda} = \underset{w_{\mathcal{A}^*_\lambda}}{\arg\min}\ \frac{1}{2}\|X_{\mathcal{A}^*_\lambda} w_{\mathcal{A}^*_\lambda} - y\|_2^2 + \lambda \sum_{i \in \mathcal{A}^*_\lambda} \|w_{b_i}\|_2 \quad \textit{s.t.} \quad (I - X^+_{\mathcal{A}^*_\lambda} X_{\mathcal{A}^*_\lambda}) w_{\mathcal{A}^*_\lambda} = 0. \tag{5}$$

Introducing dual variables $\eta \in \mathbb{R}^{|\mathcal{A}^*_\lambda|}$, the KKT system for Eq. (5) is

$$\begin{aligned} X_{b_i}^\top X_{\mathcal{A}^*_\lambda} w_{\mathcal{A}^*_\lambda} - X_{b_i}^\top y + \lambda \frac{w_{b_i}}{\|w_{b_i}\|_2} + (I - X^+_{\mathcal{A}^*_\lambda} X_{\mathcal{A}^*_\lambda})_{b_i} \eta = 0 \quad \text{for all } b_i \in \mathcal{A}^*_\lambda \\ (I - X^+_{\mathcal{A}^*_\lambda} X_{\mathcal{A}^*_\lambda}) w_{\mathcal{A}^*_\lambda} = 0. \end{aligned} \tag{6}$$

Satisfying Eq. (6) is both necessary and sufficient for a primal-dual pair $(w_{\mathcal{A}^*_\lambda}, \eta)$ to be optimal because the reduced-block problem is convex with linear constraints. It turns out that the space of dual solutions has a simple structure which is easily incorporated into the KKT system.

Lemma 4 *Fix $\lambda \geq 0$ and suppose the support of the min-norm solution is $\mathcal{A}^*_\lambda$. Then the space of dual solutions to the KKT system is exactly Row$(X_{\mathcal{A}^*_\lambda})$.*

As a result, the min-norm dual solution is simply $\eta^* = 0$ and is attained by adding the constraint $X^+_{\mathcal{A}^*_\lambda} X_{\mathcal{A}^*_\lambda} \eta = 0$. Since $(I - X^+_{\mathcal{A}^*_\lambda} X_{\mathcal{A}^*_\lambda})$ and $X^+_{\mathcal{A}^*_\lambda} X_{\mathcal{A}^*_\lambda}$ project onto orthogonal spaces, we can incorporate this constraint into the KKT system as follows:

$$\begin{aligned} X_{b_i}^\top X_{\mathcal{A}^*_\lambda} w_{\mathcal{A}^*_\lambda} - X_{b_i}^\top y + \lambda \frac{w_{b_i}}{\|w_{b_i}\|_2} + (I - X^+_{\mathcal{A}^*_\lambda} X_{\mathcal{A}^*_\lambda})_{b_i} \eta = 0 \quad \text{for all } b_i \in \mathcal{A}^*_\lambda \\ (I - X^+_{\mathcal{A}^*_\lambda} X_{\mathcal{A}^*_\lambda}) w_{\mathcal{A}^*_\lambda} + X^+_{\mathcal{A}^*_\lambda} X_{\mathcal{A}^*_\lambda} \eta = 0. \end{aligned} \tag{7}$$

We call Eq. (7) the *unique* KKT system. The following proposition uses the unique KKT system to obtain a local, implicit solution function for the reduced-block problem. We then extend to this to the full min-norm solution and obtain local continuity of the solution path as a consequence.

Proposition 5 *Let Λ be an open interval on which $\mathcal{A}^*_\lambda$ is constant. For every $\lambda \in \Lambda$, there exists a neighbourhood of $\mathcal{O}$ of λ on which w^* is continuously differentiable. Moreover, the path for $w^*_{\mathcal{A}^*_\lambda}$ is the same on $\mathcal{O}$ as that for the reduced-block problem in Eq. (5).*

Proposition 5 is the key technical result from which we derive continuity of the solution path.

4.1. Continuity of the Solution Path

Now we consider the case where the solution to the group Lasso is unique. In this setting, the solution map $\mathcal{W}^*$ is equal to the min-norm solution w^* and the analysis is considerably eased. To facilitate the analysis, we introduce a sufficient condition for uniqueness of the group lasso solution.

Assumption 6 (Group General Position) *For every $\mathcal{E} \subseteq \mathcal{B}$, $|\mathcal{E}| \leq n + 1$, there do not exist unit vectors $z_{b_i} \in \mathbb{R}^{|b_i|}$ such that for any $j \in \mathcal{E}$,*

$$X_{b_j} z_{b_j} \in \textit{affine}(\{X_{b_i} z_{b_i} : b_i \in \mathcal{E} \setminus b_j\}).$$

We call Assumption 6 *group general position* because it is a natural extension of general position to groups of column vectors. General position itself is an extension of affine independence and is sufficient for the lasso solution to be unique (R. J. Tibshirani, 2013). Group general position is strictly weaker than linear independence of the columns of X, but neither implies nor is implied by general position (Proposition 11). See Proposition 12 for a formal proof of sufficiency.

Following Vaiter et al. (2012), define the *transition space* to be

$$\mathcal{H} = \bigcup_{b_i \in \mathcal{B}} \text{bd}(\{\lambda : b_i \in \mathcal{A}_\lambda^*\}), \tag{8}$$

By construction, $\mathcal{H}$ is the set of regularization parameters at which a block b_i transitions from being active to inactive or vice-versa. Note that, as a direct consequence of Proposition 5, $\mathcal{H}$ is the exactly set of potential discontinuities of $\mathcal{W}^*$. Crucially, we show $\mathcal{H}$ contains no intervals.

Lemma 7 *The transition space $\mathcal{H}$ is discrete.*

Since $\mathcal{H}$ is discrete and Proposition 5 implies $\mathcal{W}^*$ is continuous away from $\mathcal{H}$, proving continuity of the path now reduces to showing that the left and right limits of the unique solution are equal at each "break point" in $\mathcal{H}$. Doing so yields our main result.

Theorem 8 *Suppose group general position holds. Then the unique group lasso solution path is continuous.*

The proof of Theorem 8 relies on uniqueness of the solution; without uniqueness, we can only show the right- and left-hand limits of the min-norm solution at $\bar{\lambda} \in \mathcal{H}$ are also solutions to the group lasso problem. Since we do not recover their active blocks from the argument, we cannot show they are the min-norm solution. Thus, we develop machinery for characterizing the min-norm solution path, but additional work is required to extend continuity to this more general setting.

5. Conclusion

This paper analyzes the path of the group lasso solution set as the regularization parameter is varied. We prove that the path, called the *solution function*, is continuous when the group lasso admits a unique solution and provide a new sufficient condition for such uniqueness to hold. Outside of the unique-solution setting, we characterize the optimal solution set and show how to compute the min-norm solution as a linear system of the block correlation vectors. In several cases, our research extends results for the lasso to the more general setting of the group lasso.

The main technical difficulty in our analysis is that the group lasso path cannot be computed algorithmically, unlike the piecewise linear lasso. In particular, non-linearity makes computing the active set of the min-norm solution challenging when the group lasso admits multiple solutions. As such, we leave proving continuity of the min-norm path to future work.

A consequence of our results is that convex reformulations of neural networks with gated ReLU activations also have continuous solution paths under mild assumptions. We believe it is possible to extend continuity to convex reformulations of ReLU neural networks by incorporating the necessary constraints into our analysis. We leave this exciting direction of research to future work.

References

[Bac08] Francis R Bach. "Consistency of the group lasso and multiple kernel learning." In: *Journal of Machine Learning Research* 9.6 (2008).

[Bak+99] Sergey Bakin et al. "Adaptive regression and model selection in data mining problems". In: (1999).

[DG17] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml.

[Efr+04] Bradley Efron et al. "Least angle regression". In: *The Annals of statistics* 32.2 (2004), pp. 407–499.

[FMS19] Jonathan Fiat, Eran Malach, and Shai Shalev-Shwartz. "Decoupling gating from linearity". In: *arXiv preprint arXiv:1906.05032* (2019).

[Has+07] Trevor Hastie et al. "Forward stagewise regression and the monotone lasso". In: *Electronic Journal of Statistics* 1 (2007), pp. 1–29.

[HZ10] Junzhou Huang and Tong Zhang. "The benefit of group sparsity". In: *The Annals of Statistics* 38.4 (2010), pp. 1978–2004.

[LZ03] Yi Lin and Hao Helen Zhang. *Component selection and smoothing in smoothing spline analysis of variance models–COSSO*. Tech. rep. North Carolina State University. Dept. of Statistics, 2003.

[LZ09] Han Liu and Jian Zhang. "Estimation consistency of the group lasso and its applications". In: *Artificial Intelligence and Statistics*. PMLR. 2009, pp. 376–383.

[MSP22] Aaron Mishkin, Arda Sahiner, and Mert Pilanci. "Fast Convex Optimization for Two-Layer ReLU Networks: Equivalent Model Classes and Cone Decompositions". In: *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*. Vol. 162. Proceedings of Machine Learning Research. 2022, pp. 15770–15816.

[MZ16] Ritwik Mitra and Cun-Hui Zhang. "The benefit of group sparsity in group inference with de-biased scaled group Lasso". In: *Electronic Journal of Statistics* 10.2 (2016), pp. 1829–1873.

[Nes+18] Yurii Nesterov et al. *Lectures on convex optimization*. Vol. 137. Springer, 2018.

[NR08] Yuval Nardi and Alessandro Rinaldo. "On the asymptotic properties of the group lasso estimator for linear models". In: *Electronic Journal of Statistics* 2 (2008), pp. 605–633.

[OPT00] Michael R Osborne, Brett Presnell, and Berwin A Turlach. "A new approach to variable selection in least squares problems". In: *IMA journal of numerical analysis* 20.3 (2000), pp. 389–403.

[PE20] Mert Pilanci and Tolga Ergen. "Neural Networks are Convex Regularizers: Exact Polynomial-time Convex Optimization Formulations for Two-layer Networks". In: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*. Vol. 119. Proceedings of Machine Learning Research. 2020, pp. 7695–7705.

[Ped+11] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[Tib13] Ryan J Tibshirani. "The lasso problem and uniqueness". In: *Electronic Journal of statistics* 7 (2013), pp. 1456–1490.

[Tib96] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[TT11] Ryan J Tibshirani and Jonathan Taylor. "The solution path of the generalized lasso". In: *The annals of statistics* 39.3 (2011), pp. 1335–1371.

[Vai+12] Samuel Vaiter et al. "The degrees of freedom of the Group Lasso for a General Design". In: *CoRR* abs/1212.6478 (2012).

[Wen+16] Wei Wen et al. "Learning structured sparsity in deep neural networks". In: *Advances in neural information processing systems* 29 (2016).

[YL06] Ming Yuan and Yi Lin. "Model selection and estimation in regression with grouped variables". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.1 (2006), pp. 49–67.

Appendix A. Group Lasso: Proofs

Proposition 1 *Let $\lambda > 0$ and $\mathcal{N}_\lambda = Null(X_{\mathcal{E}_\lambda}) \bigcap \{z : z_{b_i} \in Span(v_{b_i}), i \in \mathcal{E}_\lambda\}$. The optimal set is*

$$\mathcal{W}^*(\lambda) = \left\{ w \in \mathbb{R}^d : w_{\mathcal{E}_\lambda} = w^*_{\mathcal{E}_\lambda}(\lambda) + z, z \in \mathcal{N}_\lambda,\ w_{\mathcal{I}_\lambda} = 0,\ w_{b_i} \neq 0 \implies \frac{\lambda w_{b_i}}{\|w_{b_i}\|_2} = v_{b_i}(\lambda) \right\}.$$

Proof For ease of notation, let

$$\mathcal{X} = \left\{ w \in \mathbb{R}^d : w_{\mathcal{E}_\lambda} = w^*_{\mathcal{E}_\lambda}(\lambda) + z, z \in \mathcal{N}_\lambda,\ w_{\mathcal{I}_\lambda} = 0,\ w_{b_i} \neq 0 \implies \frac{\lambda w_{b_i}}{\|w_{b_i}\|_2} = v_{b_i}(\lambda) \right\}.$$

Recall that $\hat{y} = Xw$ is constant over $\mathcal{W}^*(\lambda)$ and each $w \in \mathcal{W}^*(\lambda)$ can only have support on $\mathcal{E}_\lambda$ by FO conditions. Dropping the zero entries of w, $\hat{y} = X_{\mathcal{E}_\lambda} w_{\mathcal{E}_\lambda}$ and we deduce that

$$z = w_{\mathcal{E}_\lambda} - w^*_{\mathcal{E}_\lambda} \in \text{Null}(X_{\mathcal{E}_\lambda}),$$

for every $w \in \mathcal{W}^*(\lambda)$. Let us show that $z \in \mathcal{N}_\lambda$.

If $w^*_{b_i} \neq 0$, then first order optimality conditions require

$$\frac{w^*_{b_i}}{\|w^*_{b_i}\|_2} = \lambda \cdot v_{b_i},$$

from which we have $w^*_{b_i} \in \text{Span}(v_{b_i})$. Thus, $w_{b_i} = 0$ implies $z_{b_i} = -w^*_{b_i} \in \text{Span}(v_{b_i})$ (this also trivially holds when $w^*_{b_i} = 0$).

Otherwise, if $w_{b_i} \neq 0$, then first-order optimality again implies

$$\begin{aligned} \frac{w_{b_i}}{\|w_{b_i}\|_2} &= \lambda \cdot v_{b_i} \\ \implies w^*_{b_i} + z_{b_i} &\in \text{Span}(v_{b_i}) \\ \implies z_{b_i} &\in \text{Span}(v_{b_i}), \end{aligned}$$

since $w^*_{b_i} \in \text{Span}(v_{b_i})$. As a result, $z \in \mathcal{N}_\lambda$. It is straightforward to check that first-order optimality implies $w_{\mathcal{I}_\lambda} = 0$ so that $\mathcal{W}^* \subseteq \mathcal{X}$.

For the reverse inclusion, let $w \in \mathcal{X}$. Clearly

$$Xw = X_{\mathcal{E}_\lambda} w_{\mathcal{E}_\lambda} = X_{\mathcal{E}_\lambda} w^*_{\mathcal{E}_\lambda} = \hat{y}(\lambda),$$

so that w has the correct model fit. If $w_{b_i} = 0$ and $i \in \mathcal{I}_\lambda$, then

$$\|X^\top_{b_i}(y - Xw)\|_2 = \|v_{b_i}\|_2 \leq \lambda,$$

shows FO conditions are satisfied. Alternatively, if $i \in \mathcal{E}_\lambda$, then $X^\top_{b_i}(y - Xw) = v_{b_i}$ and FO conditions are satisfied. Finally, we check blocks for which $w_{b_i} \neq 0$. Membership in $\mathcal{X}$ implies

$$\frac{w_{b_i}}{\|w_{b_i}\|_2} = v_{b_i}$$

and w_{b_i} again satisfies first-order optimality conditions. Putting these conditions together implies $w \in \mathcal{W}^*$. This completes the proof. ■

Appendix B. The Minimum-norm Solution: Proofs

Lemma 2 *For $\lambda > 0$, the active blocks of the min-norm solution are given by*

$$w^*_{\mathcal{A}^*_\lambda}(\lambda) = (X^\top_{\mathcal{A}^*_\lambda} X_{\mathcal{A}^*_\lambda})^+ X^\top_{\mathcal{A}^*_\lambda} \left(y - (X^\top_{\mathcal{A}^*_\lambda})^+ v_{\mathcal{A}^*_\lambda}(\lambda) \right) \tag{4}$$

Proof Recall the definition of the block correlation vector,

$$v_{b_i} = X^\top_{b_i} r(\lambda),$$

which shows $v_{b_i} \in \text{Row}(X_{b_i})$ for each $b_i \in \mathcal{A}_\lambda$ and $v_{\mathcal{A}^*_\lambda} \in \text{Row}(X_{\mathcal{A}^*_\lambda})$. As a result, $v_{\mathcal{A}^*_\lambda}$ is unchanged by the projection onto $\text{Row}(X_{\mathcal{A}^*_\lambda})$, i.e.

$$v_{\mathcal{A}^*_\lambda} = X^\top_{\mathcal{A}^*_\lambda} (X^\top_{\mathcal{A}^*_\lambda})^+ v_{\mathcal{A}^*_\lambda}.$$

Combining this with the definition of v_{b_i} and summing over the non-zero blocks of the min-norm solution, we obtain

$$\begin{aligned} X^\top_{\mathcal{A}^*_\lambda} X_{\mathcal{A}^*_\lambda} w^*_{\mathcal{A}^*_\lambda} &= X^\top_{\mathcal{A}^*_\lambda} y - X^\top_{\mathcal{A}^*_\lambda} (X^\top_{\mathcal{A}^*_\lambda})^+ v_{\mathcal{A}^*_\lambda} \\ &= X^\top_{\mathcal{A}^*_\lambda} \left[y - (X^\top_{\mathcal{A}^*_\lambda})^+ v_{\mathcal{A}^*_\lambda} \right], \end{aligned}$$

which implies that w' satisfying

$$w'_{\mathcal{A}^*_\lambda} = (X^\top_{\mathcal{A}^*_\lambda} X_{\mathcal{A}^*_\lambda})^+ X^\top_{\mathcal{A}^*_\lambda} \left[y - (X^\top_{\mathcal{A}^*_\lambda})^+ v_{\mathcal{A}^*_\lambda} \right],$$

and $w'_{\mathcal{B} \setminus \mathcal{A}^*_\lambda} = 0$ defines one solution to the group Lasso. Moreover, $w'_{\mathcal{A}^*_\lambda}$ is orthogonal to $\text{Null}(X_{\mathcal{A}^*_\lambda})$.

Now, suppose that the min-norm solution is not orthogonal to $\text{Null}(X_{\mathcal{A}^*_\lambda})$. Then $w^*_{\mathcal{A}^*_\lambda} = w'_{\mathcal{A}^*_\lambda} + a$, where $z \in \text{Null}(X_{\mathcal{A}^*_\lambda})$, $z \neq 0$ and

$$\|w^*_{\mathcal{A}^*_\lambda}\|_2^2 = \|w'_{\mathcal{A}^*_\lambda}\|_2^2 + \|b\|_2^2 > \|w'_{\mathcal{A}^*_\lambda}\|_2^2,$$

which is a contradiction. We conclude $w^*_{\mathcal{A}^*_\lambda} = w'_{\mathcal{A}^*_\lambda}$ as this is only solution which is orthogonal to $\text{Null}(X_{\mathcal{A}^*_\lambda})$. ■

Lemma 3 *Suppose the support of the min-norm solution is $\mathcal{A}^*_\lambda$. Then the active blocks are the unique solution to the following optimization problem:*

$$w^*_{\mathcal{A}^*_\lambda} = \underset{w_{\mathcal{A}^*_\lambda}}{\arg\min} \ \frac{1}{2} \|X_{\mathcal{A}^*_\lambda} w_{\mathcal{A}^*_\lambda} - y\|_2^2 + \lambda \sum_{i \in \mathcal{A}^*_\lambda} \|w_{b_i}\|_2 \quad \textit{s.t.} \quad (I - X^+_{\mathcal{A}^*_\lambda} X_{\mathcal{A}^*_\lambda}) w_{\mathcal{A}^*_\lambda} = 0. \tag{5}$$

Proof Recall that $w^*_{\mathcal{A}^*_\lambda}$ is orthogonal to $\text{Null}\left(X_{\mathcal{A}^*_\lambda}\right)$. Thus,

$$(I - X^+_{\mathcal{A}^*_\lambda} X_{\mathcal{A}^*_\lambda}) w^*_{\mathcal{A}^*_\lambda} = 0,$$

and $w^*_{\mathcal{A}^*_\lambda}$ is a feasible point for the reduced-block problem. Since $\mathcal{A}^*_\lambda$ contains all active blocks of w^*, we know

$$\begin{aligned}\min_{w_{\mathcal{A}^*_\lambda}} \|X_{\mathcal{A}^*_\lambda} w_{\mathcal{A}^*_\lambda} - y\|_2^2 + \lambda \sum_{i \in \mathcal{A}^*_\lambda} \|w_{b_i}\|_2 &\geq \min_w \|Xw - y\|_2^2 + \lambda \sum_{i \in \mathcal{B}} \|w_{b_i}\|_2 \\ &= \|X_{\mathcal{A}^*_\lambda} w_{\mathcal{A}^*_\lambda} - y\|_2^2 + \lambda \sum_{i \in \mathcal{A}^*_\lambda} \|w_{b_i}\|_2,\end{aligned}$$

so that $w^*_{\mathcal{A}^*_\lambda}$ is optimal for the reduced-block problem.

Let $w'_{\mathcal{A}^*_\lambda} \neq w^*_{\mathcal{A}^*_\lambda}$ be another solution to the reduced-block problem. Vaiter et al. (2012, Lemma 2) implies $w'_{\mathcal{A}^*_\lambda}$ must have the same model fit $\hat{y} = X_{\mathcal{A}^*_\lambda} w'_{\mathcal{A}^*_\lambda}$ as $w^*_{\mathcal{A}^*_\lambda}$. Thus, $z = w'_{\mathcal{A}^*_\lambda} - w^*_{\mathcal{A}^*_\lambda} \in \text{Null}(\mathcal{A}^*_\lambda)$, which implies

$$0 = (I - X^+_{\mathcal{A}^*_\lambda} X_{\mathcal{A}^*_\lambda}) w'_{\mathcal{A}^*_\lambda} = (I - X^+_{\mathcal{A}^*_\lambda} X_{\mathcal{A}^*_\lambda}) z.$$

We deduce $z = 0$ and $w^*_{\mathcal{A}^*_\lambda}$ is the unique solution. ■

Lemma 4 *Fix $\lambda \geq 0$ and suppose the support of the min-norm solution is $\mathcal{A}^*_\lambda$. Then the space of dual solutions to the KKT system is exactly Row$(X_{\mathcal{A}^*_\lambda})$.*

Proof Let $\eta \in \text{Row}(X_{\mathcal{A}^*_\lambda})$. The min-norm solution w^* is member of $\mathcal{W}^*(\lambda)$. Thus, FO conditions imply

$$X_{b_i}^\top X_{\mathcal{A}^*_\lambda} w_{\mathcal{A}^*_\lambda} - X_{b_i}^\top y + \lambda \frac{w_{b_i}}{\|w_{b_i}\|_2} = 0,$$

for all $b_i \in \mathcal{A}^*_\lambda$. We also have

$$(I - X^+_{\mathcal{A}^*_\lambda} X_{\mathcal{A}^*_\lambda}) \eta = 0,$$

by choice of η, which implies any $\eta \in \text{Row}(X_{\mathcal{A}^*_\lambda})$ satisfies the first $|\mathcal{A}^*_\lambda|$ constraints of the KKT system. Since w^* is the min-norm solution, it is orthogonal to Null$(X_{\mathcal{A}^*_\lambda})$, implying

$$(I - X^+_{\mathcal{A}^*_\lambda} X_{\mathcal{A}^*_\lambda}) w_{\mathcal{A}^*_\lambda} = 0.$$

We conclude that (w^*, η) is a KKT point. ■

Lemma 9 *Fix $\lambda \geq 0$ and $\bar{w}, \bar{\eta} \in \mathbb{R}^{|\mathcal{A}^*_\lambda|}$. The Jacobian of the unique KKT system is full rank at $(\bar{w}, \bar{\eta})$.*

Proof Let $M(w)$ be the block-diagonal projection matrix given by

$$M(w)_{b_i} = \frac{1}{\|w_{b_i}\|_2} \left(I - \frac{w_{b_i}}{\|w_{b_i}\|_2} \frac{w_{b_i}^\top}{\|w_{b_i}\|_2} \right).$$

The Jacobian of the unique KKT system with respect to $(w_{\mathcal{A}^*_\lambda}, \eta)$ is

$$J_\lambda = \begin{bmatrix} X^\top_{\mathcal{A}^*_\lambda} X_{\mathcal{A}^*_\lambda} + \lambda M(w) & (I - X^+_{\mathcal{A}^*_\lambda} X_{\mathcal{A}^*_\lambda}) \\ (I - X^+_{\mathcal{A}^*_\lambda} X_{\mathcal{A}^*_\lambda}) & X^+_{\mathcal{A}^*_\lambda} X_{\mathcal{A}^*_\lambda} \end{bmatrix} = \begin{bmatrix} X^\top_{\mathcal{A}^*_\lambda} X_{\mathcal{A}^*_\lambda} + \lambda M(w) & P^\perp_{\mathcal{A}^*_\lambda} \\ P^\perp_{\mathcal{A}^*_\lambda} & P_{\mathcal{A}^*_\lambda} \end{bmatrix},$$

where $P_{\mathcal{A}^*_\lambda}$ and $P^\perp_{\mathcal{A}^*_\lambda}$ are the projection matrices for $\text{Row}(X_{\mathcal{A}^*_\lambda})$ and $\text{Null}(X_{\mathcal{A}^*_\lambda})$, respectively.

Let $w, \eta \in \mathbb{R}^{|\mathcal{A}^*_\lambda|}$ such that $(w, \eta) \neq 0$ and assume

$$J_\lambda \begin{bmatrix} w \\ \eta \end{bmatrix} = \begin{bmatrix} X^\top_{\mathcal{A}^*_\lambda} X_{\mathcal{A}^*_\lambda} w + \lambda M(\bar{w}) w + P^\perp_{\mathcal{A}^*_\lambda} \eta \\ P^\perp_{\mathcal{A}^*_\lambda} w + P_{\mathcal{A}^*_\lambda} \eta \end{bmatrix} = 0.$$

If $P_{\mathcal{A}^*_\lambda} \eta \neq 0$ or $P^\perp_{\mathcal{A}^*_\lambda} w \neq 0$ then the second block cannot be zero. Thus, $\eta \in \text{Null}(X_{\mathcal{A}^*_\lambda})$ and $w \in \text{Row}(X_{\mathcal{A}^*_\lambda})$ must hold. We deduce

$$J_\lambda \begin{bmatrix} w \\ \eta \end{bmatrix} = \begin{bmatrix} X^\top_{\mathcal{A}^*_\lambda} X_{\mathcal{A}^*_\lambda} w + \lambda M(\bar{w}) w + \eta \\ 0 \end{bmatrix} = 0.$$

Since $w \in \text{Row}(X_{\mathcal{A}^*_\lambda})$, it holds that $X_{\mathcal{A}^*_\lambda} w \neq 0$ and thus

$$w^\top X^\top_{\mathcal{A}^*_\lambda} X_{\mathcal{A}^*_\lambda} w + \lambda w^\top M(\bar{w}) w + w^\top \eta = \|X^\top_{\mathcal{A}^*_\lambda} w\|_2^2 + w^\top \lambda M(\bar{w}) w > 0,$$

since $M(\bar{w})$ is positive semi-definite and $w \perp \eta$. But this is a contradiction. We conclude $\text{Null}(J_\lambda) = \{0\}$. ■

Proposition 5 *Let Λ be an open interval on which $\mathcal{A}^*_\lambda$ is constant. For every $\lambda \in \Lambda$, there exists a neighbourhood of $\mathcal{O}$ of λ on which w^* is continuously differentiable. Moreover, the path for $w^*_{\mathcal{A}^*_\lambda}$ is the same on $\mathcal{O}$ as that for the reduced-block problem in Eq. (5).*

Proof Let $\lambda \in \Lambda$ Then, there exists a unique η such that Eq. (7) is satisfied at $(w^*_{\mathcal{A}^*_\lambda}, \eta)$. (In fact, Lemma 4 implies $\eta = 0$ is exactly the optimal dual parameter.) Furthermore, Lemma 9 shows that the Jacobian of the unique KKT system at $(w^*_{\mathcal{A}^*_\lambda}, \eta)$ is invertible.

The implicit function theorem implies there exists an open interval $\tilde{\mathcal{O}} \subseteq \Lambda$ containing λ such that $\lambda \mapsto (\tilde{g}(\lambda), h(\lambda))$, where $\tilde{g} : \tilde{\mathcal{O}} \to \mathbb{R}^{|\mathcal{A}^*_\lambda|}$ and $h : \tilde{\mathcal{O}} \to \mathbb{R}^{|\mathcal{A}^*_\lambda|}$, is a unique, continuously differentiable function satisfying $(\tilde{g}(\lambda), h(\lambda)) = (w^*_{\mathcal{A}^*_\lambda}, \eta)$ and $(\tilde{g}(\lambda'), h(\lambda'))$ is a zero of the unique KKT system for all $\lambda' \in \tilde{\mathcal{O}}$. That is, $(\tilde{g}, h)$ is locally the primal-dual solution function for the reduced problem (Eq. (5)); we will show that an extension of $\tilde{g}$ to $\mathbb{R}^d$ locally provides the min-norm solution function for the full group Lasso problem.

Define the extension of $\tilde{g}$ to be $g : \tilde{\mathcal{O}} \to \mathbb{R}^d$ such that $g(\lambda')_{\mathcal{A}^*_\lambda} = \tilde{g}(\lambda')$ and $g(\lambda')_{\mathcal{B} \setminus \mathcal{A}^*_\lambda} = 0$. For $i \in \mathcal{I}_\lambda$, first-order optimality says

$$\|X^\top_{b_i}(y - X w^*)\|_2 < \lambda.$$

Continuity of g guarantees there exists $\mathcal{O} \subseteq \tilde{\mathcal{O}}$ on which

$$\|X^\top_{b_i}(y - X g(\lambda'))\|_2 \leq \lambda',$$

This implies $g_{\mathcal{I}_\lambda}$ satisfies FO conditions on $\mathcal{O}$.

By assumption, $\mathcal{A}_\lambda^*$ is constant on Λ and thus on $\mathcal{O}$. Applying Lemma 3, we deduce that $w^*_{\mathcal{A}_\lambda^*}$ is given by the unique solution to the KKT system on $\mathcal{O}$. Thus, $g_{\mathcal{A}_\lambda^*}$ also satisfies FO conditions on $\mathcal{O}$ and $\hat{y}(\lambda') = Xg(\lambda')$ provides the unique model fit on $\mathcal{O}$.

It remains only to check $g_{\mathcal{T}_\lambda^*}(\lambda')$. Suppose there exists $b_i \in \mathcal{T}_\lambda^*$ such that

$$\|X_{b_i}^\top (y - Xg(\lambda'))\|_2 > \lambda.$$

Since $Xg(\lambda') = Xw^*(\lambda')$ by uniqueness of the model fit, it must be that

$$\|X_{b_i}^\top (y - Xw^*(\lambda'))\|_2 > \lambda,$$

which is a contradiction. So FO conditions are also satisfied for all blocks in $\mathcal{T}_\lambda^*$.

Putting these results together, we have shown g satisfies the (sufficient) FO conditions for the group lasso on $\mathcal{O}$. We conclude that $w^*(\lambda)$ is continuously differentiable on $\mathcal{O}$ and corresponds with the min-norm solution for the reduced-block problem. ■

B.1. Continuity of the Min-norm Path: Proofs

Lemma 10 *The minimum-norm solution $w^*(\lambda)$ is bounded for all $\lambda \geq 0$.*

Proof For $\lambda = 0$, the min-norm solution reduces to the min-norm least-squares solution, which is bounded. For $\lambda > 0$, we show boundedness directly. Optimality of w^* implies

$$\frac{1}{2}\|Xw^* - y\|_2^2 + \lambda \sum_{b_i \in \mathcal{B}} \|w^*_{b_i}\|_2 \leq \frac{1}{2}\|y\|_2^2,$$

from which we deduce the residual vector $r(\lambda)$ is bounded for any choice of λ. Thus, the block correlation vectors satisfy,

$$\begin{aligned} \|v_{b_i}(\lambda)\|_2^2 &= \|X_{b_i}^\top r(\lambda)\| \\ &\leq \|X_{b_i}\|_2^2 \|y\|_2^2 \\ \implies \|v(\lambda)\|_2 &\leq \left[\sum_{b_i \in \mathcal{B}} \{\|X_{b_i}\|_2\} \|y\|_2\right]^{1/2}. \end{aligned}$$

The right-hand quantity is independent of λ; therefore, $v(\lambda)$ is also bounded for all λ. By Lemma 2, $w^*_{\mathcal{A}_\lambda^*}$ is given by

$$\begin{aligned} w^*_{\mathcal{A}_\lambda^*}(\lambda) &= (X_{\mathcal{A}_\lambda^*}^\top X_{\mathcal{A}_\lambda^*})^+ X_{\mathcal{A}_\lambda^*}^\top \left(y - (X_{\mathcal{A}_\lambda^*}^\top)^+ v_{\mathcal{A}_\lambda^*}(\lambda)\right) \\ \implies \|w^*(\lambda)\|_2 &\leq \|(X_{\mathcal{A}_\lambda^*}^\top X_{\mathcal{A}_\lambda^*})^+ X_{\mathcal{A}_\lambda^*}^\top\|_2 \left(\|y\|_2 + \|(X_{\mathcal{A}_\lambda^*}^\top)^+\|_2 \|v_{\mathcal{A}_\lambda^*}(\lambda)\|_2\right), \end{aligned}$$

where the right-hand side is bounded. This completes the proof. ■

Lemma 7 *The transition space $\mathcal{H}$ is discrete.*

Proof To see that that $\mathcal{H}$ is discrete, define

$$\mathcal{H}_i = \text{bd}\left(\{\lambda : b_i \in \mathcal{A}_\lambda\}\right),$$

so that $\mathcal{H} = \bigcup_{i \in \mathcal{B}} \mathcal{H}_i$. The (Lebesgue) measure of $\mathcal{H}_i$ is zero because it is the boundary of a one-dimensional set and thus is discrete. By properties of the Lebesgue measure (denoted μ),

$$\mu(\mathcal{H}) \leq \bigcup_{i \in \mathcal{B}} \mu(\mathcal{H}_i) = 0,$$

since $\mathcal{B}$ is finite. We deduce that $\mathcal{H}$ is discrete. ■

Proposition 11 *Group general position (Assumption 6) does not imply the columns of X are in general position. Similarly, general position of the columns of X does not imply group general position.*

Proof Consider the simple case where we have two groups: $b_1 = \{1\}$ and $b_2 = \{2, \ldots, d\}$. Group general position is violated if there exists a unit vector z_{b_2} such that

$$\begin{aligned} & x_1 = X_{b_2} z_{b_2}. \\ \iff & x_1 \in X_{b_2} B_{d-1}, \end{aligned}$$

where $B_{n-1} = \left\{z \in \mathbb{R}^{d-1} : \|z\| \leq 1\right\}$. In contrast, general position is violated if

$$\begin{aligned} & x_1 \in \text{affine}(x_2, \ldots, x_d) \\ \iff & x_1 \in X \left\{z : \langle z, 1 \rangle = 1\right\}. \end{aligned}$$

Taking $X_{b_2} = I$, it is trivial to see that group general position can hold when general position is violated and vice-versa.

■

Proposition 12 *Suppose Assumption 6 holds and $\lambda > 0$. Then the group Lasso solution is unique.*

Proof Suppose the group Lasso solution is not unique. Then, Proposition 1 implies

$$\mathcal{N}_\lambda = \text{Null}(X_{\mathcal{E}_\lambda}) \bigcap \left\{z : z_{b_i} \propto v_{b_i}, i \in \mathcal{E}_\lambda\right\},$$

is non-empty. That is, there exist $s_{b_i} \in \{+1, -1\}$ and $\alpha_{b_i} \geq 0$ such that

$$\begin{aligned} s_{b_j} X_{b_j} v_{b_j} &= \sum_{b_i \in \mathcal{E}_\lambda \setminus j} \alpha_{b_i} s_{b_i} X_{b_i} v_{b_i} \\ \implies X_{b_j} v_{b_j} &= \sum_{b_i \in \mathcal{E}_\lambda \setminus j} \alpha_{b_i} s_{b_j} s_{b_i} X_{b_i} v_{b_i}. \end{aligned}$$

Taking inner-products on both sides with the residual r,

$$\begin{aligned} \implies \lambda^2 &= \sum_{b_i \in \mathcal{E}_\lambda \setminus j} \alpha_{b_i} s_{b_j} s_{b_i} \lambda^2 \\ \implies 1 &= \sum_{b_i \in \mathcal{E}_\lambda \setminus j} \alpha_{b_i} s_{b_j} s_{b_i}. \end{aligned}$$

Thus, we deduce that

$$X_{b_j} v_{b_j} = \sum_{b_i \in \mathcal{E}_\lambda \setminus j} \beta_{b_i} X_{b_i} v_{b_i}, \tag{9}$$

where $\sum_{b_i \in \mathcal{E}_\lambda \setminus j} \beta_{b_i} = 1$. Now, suppose that $|\mathcal{E}_\lambda| > n+1$. Then, $\{X_{b_i} v_{b_i} : b_i \in \mathcal{E}_\lambda \setminus j\}$ are linearly dependent and, by eliminating dependent vectors $X_{b_i} v_{b_i}$, we can repeat the above proof with a subset $\mathcal{E}'$ of at most $n+1$ blocks. Noting $\|v_{b_i}\|_2 = \lambda$ for each $b_i \in \mathcal{E}_\lambda$ and rescaling both sides of Eq. (9) by λ implies the existence of unit vectors z_{b_i} which contradict Assumption 6. This completes the proof. ■

Theorem 8 *Suppose group general position holds. Then the unique group lasso solution path is continuous.*

Proof Let $\bar{\lambda} \in \mathcal{H}$. By Lemma 7, $\mathcal{H}$ is discrete so that there exists some open interval $\mathcal{O} = (\bar{\lambda}, \lambda^0)$ containing no points of $\mathcal{H}$. Let g be the continuous solution function on $\mathcal{O}$, which exists by Proposition 5 and recall that $\mathcal{A}^*_\lambda$ is constant on this set.

Let $\lambda_k \downarrow \bar{\lambda}$ so that $\lambda_k \in \mathcal{O}$ for all sufficiently large k. By Lemma 10, the min-norm solution $g(\lambda_k)$ is bounded. Thus, dropping to a convergent subsequence if necessary, the limit $\bar{w} = \lim_k g(\lambda_k)$ exists. It is trivial to argue in the same fashion for the dual problem, where $h(\lambda_k) = 0$ for all k implies $\bar{\eta} = 0$ is the limit.

For each $b_i \in \mathcal{A}^*_\lambda$, FO conditions give

$$X_{b_i}^\top (y - X g(\lambda_k)) = \lambda_k \frac{g(\lambda_k)}{\|g(\lambda_k)\|_2},$$

which, by taking limits on both sides, yields

$$X_{b_i}^\top (y - X\bar{w}) = \lambda_k u,$$

with $\|u_{b_i}\|_2 = 1$. If $\bar{w}_{b_i} \neq 0$, then

$$u_{b_i} = \frac{\bar{w}_{b_i}}{\|\bar{w}_{b_i}\|_2},$$

and FO conditions are satisfied. On the other hand, if $\bar{w}_{b_i} = 0$, then taking norms shows $b_i \in \mathcal{E}_{\bar{\lambda}}$ and FO conditions are also satisfied.

Thus, $\bar{w}_{\mathcal{A}^*_\lambda}$ satisfies FO conditions at $\bar{\lambda}$. Similarly, taking limits for the blocks in $b_i \in \mathcal{I}_\lambda \cup \mathcal{T}^*_\lambda$ gives

$$\begin{aligned} &\|X_{b_i}^\top (y - X g(\lambda_k))\|_2 \leq \lambda_k \\ \implies &\|X_{b_i}^\top (y - X\bar{w})\|_2 \leq \bar{\lambda}, \end{aligned}$$

showing the extension $\bar{w}_{b_i} = 0$ satisfies FO conditions for the group Lasso. We conclude that $\bar{w}$ is a solution to the group Lasso problem.

Since the solution to the group Lasso is unique under Assumption 6, it must be that $\bar{w}$ is the unique solution at $\bar{\lambda}$. Since this argument holds for every limit point of $g(\lambda_k)$, we have shown that the limit from the right exists and $\mathcal{W}^*$ is continuous from the right. Arguing identically for the left-hand limit establishes continuity of the solution function at $\bar{\lambda}$. ■