Solving the Elastic Net and Lasso Regression Problems

Jean Gallier and Jocelyn Quaintance

Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
jean@seas.upenn.edu

May 30, 2022

Abstract. In these notes we investigate methods for solving the elastic net regression problem. The lasso regression problem is the special case of the elastic net regression problem where only the ℓ^1-norm of the weight vector is penalized. The elastic net regression problem has a unique minimal solution. This solution can be found directly from the primal program using ADMM, using a simple extension of the method used to solve lasso. For all of these problems we seek a minimal solution (w, b) where w is the weight vector and b is the intercept (we deal with the intercept directly, rather than by "centering" of the data).

Using the trick where we replace the ℓ^1-norm $\|w\|_1$ of the weight vector w occurring in the objective function by the sum $\mathbf{1}_n^\top \epsilon$ of the components of a vector ϵ and adding the inequalites $w - \epsilon \leq 0$ and $-w - \epsilon \leq 0$, which are equivalent to $|w_i| \leq \epsilon_i$ for $i = 1, \ldots, n$, we avoid having to use subgradients and we obtain an explicit form of the dual program in terms of the Lagrange multipliers associated with the above inequalities. This dual program can be solved using ADMM, and also yields the solution of the primal. This is because the matrix occurring in the quadratic part of the objective function is symmetric positive definite.

However, in the special case of lasso, this matrix is singular if the data matrix X does not have full rank. In this case, it is not possible to find an explicit formula for the dual function in terms of the Lagrange multipliers (associated with the inequalities $w - \epsilon \leq 0$ and $-w - \epsilon \leq 0$). We elucidate the relationship between the Lagrange multipliers and the subgradients of the ℓ^1-norm function occurring as a penalty term in lasso. We investigate the uniqueness of the solutions of lasso. We find sufficient conditions involving a notion of affine independence weaker than the standard notion of affine independence. We also characterize the space of minimal solutions. It is a polytope arising as the intersection of various hyperplanes related to the kernel of the matrix $[X\ \mathbf{1}_m]$ with a simplex contained in a special hyperplane.

1 Solving the Problem Directly

The first formulation of elastic net regression consists in expressing the objective function directly in terms of w, b and $\|w\|_1$ as

Program (**elastic net V1**):

$$\text{minimize} \quad J(w,b) = \frac{1}{2}(Xw + b\mathbf{1}_m - y)^\top (Xw + b\mathbf{1}_m - y) + \frac{1}{2}Kw^\top w + \tau \|w\|_1 \,,$$

where $K > 0$ and $\tau > 0$ are two constants controlling the influence of the ℓ^2-regularization and the ℓ^1-regularization. Recall that X is an $m \times n$ matrix, $w \in \mathbb{R}^n$, $y \in \mathbb{R}^m$, and $b \in \mathbb{R}$, and that X and y are given. Observe that as in the case of ridge regression, minimization is performed over w and b, but b is not penalized in the objective function. Lasso regression can be viewed as the special case of elastic net regression in which $K = 0$ and $\tau > 0$, and ridge regression as the special case where $K > 0$ and $\tau = 0$.

Another formulation of *elastic net regression* uses a trick to get rid of the term $\tau \|w\|_1$, as we explain in Section 19.4 of our book Gallier and Qaintance [6] (Vol II.)

Program (**elastic net V2**):

$$\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\xi^\top \xi + \frac{1}{2}Kw^\top w + \tau \mathbf{1}_n^\top \epsilon \\
&\text{subject to} \\
&\qquad y - Xw - b\mathbf{1}_m = \xi \\
&\qquad w \leq \epsilon \\
&\qquad -w \leq \epsilon,
\end{aligned}$$

In this formulation $\xi \in \mathbb{R}^m$, $\epsilon \in \mathbb{R}^n$, and minimization is performed over ξ, w, ϵ and b, but b is not penalized. The variant of Version (V2) in which $\xi = y - Xw - b\mathbf{1}_m$ is not a constraint but instead is incoporated into the objective function as in (V1) is convenient to derive the dual program; see Section 2.

In this section we use ADMM to solve Version (V1) of elastic net regression.

It is easy to show (see Section 7) that by expanding the first term in Version (V1) we get

$$J(w,b) = \frac{1}{2}\begin{pmatrix} w^\top & b \end{pmatrix} \begin{pmatrix} X^\top X + KI_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} - \begin{pmatrix} w^\top & b \end{pmatrix} \begin{pmatrix} X^\top y \\ \mathbf{1}_m^\top y \end{pmatrix} + \frac{1}{2}y^\top y + \tau \|w\|_1 \,.$$

Thus the elastic net regression problem is to minimize $J(w,b)$ with respect to w and b, without penalizing b. Since the term $\frac{1}{2}y^\top y$ is constant, it is equivalent to minimize $J(w,b) - \frac{1}{2}y^\top y$, so we will drop the term $\frac{1}{2}y^\top y$ from the objective function.

It can be shown that the matrix

$$B = \begin{pmatrix} X^\top X + KI_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix}$$

is symmetric positive definite, and that as a consequence, J is strictly convex; see Section 7. Since it is nonnegative, it has a unique minimum (see Gallier and Qaintance [6], Theorem 4.13(1,2)), a fact that is not obvious at first glance.

We derive a method for solving the above problem using ADMM directly on this version of the primal. When $K = 0$ (and $\tau > 0$) elastic net regression is actually lasso, and indeed the method that we describe reduces to ADMM for lasso. However it should be noted that in this case there may be more that one minimal solution because the matrix

$$B = \begin{pmatrix} X^\top X & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix}$$

may no longer be positive definite. As we explain in Section 3, this will happen if the $m \times (n+1)$ matrix

$$\widetilde{X} = \begin{pmatrix} X & \mathbf{1}_m \end{pmatrix}$$

does not have full rank, because $(\widetilde{X})^\top X = B$. The situation is actually quite subtle, see Sections 3, 4 and 5.

As in Sections 16.3 and 16.8 of our book Gallier and Qaintance [6] (Vol II) we need to figure out the x-minimization step and the z-minimization step of the scaled version of ADMM. In our situation,

$$x = \begin{pmatrix} w \\ b \end{pmatrix},$$

$A = I_{n+1}, B = -I_{n+1}$, $c = 0_{n+1}$, $f(w, b) = J(w, b) - \tau \left\|z\right\|_1$, and $g(z, z_b) = \tau \left\|z\right\|_1$. Observe that g is *independent of* z_b, since we are not penalizing b. We introduce the variables

$$\begin{pmatrix} w^k \\ b^k \end{pmatrix}, \quad \begin{pmatrix} z \\ z_b \end{pmatrix}, \quad \begin{pmatrix} z^k \\ z_b^k \end{pmatrix}, \quad \begin{pmatrix} \mu^k \\ \mu_b^k \end{pmatrix}.$$

Note that the superscript k in $w^k, b^k, z^k, z_b^k, \mu^k, \mu_b^k$ denotes the iteration stage, and not the kth power. Since $A = I_{n+1}, B = -I_{n+1}$, $c = 0_{n+1}$, we have the unique equation

$$\begin{pmatrix} w \\ b \end{pmatrix} - \begin{pmatrix} z \\ z_b \end{pmatrix} = \begin{pmatrix} 0_n \\ 0 \end{pmatrix}.$$

We start with some initial values $(z^0, z_b^0, \mu^0, \mu_b^0)$ and find $(w^{k+1}, b^{k+1}, z^{k+1}, z_b^{k+1}, \mu^{k+1}, \mu_b^{k+1})$ for all $k \geq 0$ using the following x-minimization steps and z-minimization steps in alternation.

The x-minimization step is to minimize

$$\frac{1}{2} \begin{pmatrix} w^\top & b \end{pmatrix} \begin{pmatrix} X^\top X + K I_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} - \begin{pmatrix} w^\top & b \end{pmatrix} \begin{pmatrix} X^\top y \\ \mathbf{1}_m^\top y \end{pmatrix} + \frac{\rho}{2} \left\| \begin{pmatrix} w \\ b \end{pmatrix} - \begin{pmatrix} z^k - \mu^k \\ z_b^k - \mu_b^k \end{pmatrix} \right\|_2^2$$

with respect to w and b. We immediately find that the gradient is given by

$$\nabla J_{w,b} = \left(\begin{pmatrix} X^\top X + KI_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix} + \rho I_{n+1} \right) \begin{pmatrix} w \\ b \end{pmatrix} - \begin{pmatrix} X^\top y \\ \mathbf{1}_m^\top y \end{pmatrix} - \rho \begin{pmatrix} z^k - \mu^k \\ z_b^k - \mu_b^k \end{pmatrix}.$$

Since the x-update is obtained by setting the above gradient to zero, we find that the x-update is given by

$$\begin{pmatrix} w^{k+1} \\ b^{k+1} \end{pmatrix} = \left(\begin{pmatrix} X^\top X + KI_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix} + \rho I_{n+1} \right)^{-1} \left(\begin{pmatrix} X^\top y \\ \mathbf{1}_m^\top y \end{pmatrix} + \rho \begin{pmatrix} z^k - \mu^k \\ z_b^k - \mu_b^k \end{pmatrix} \right).$$

The z-minimization step is to minimize

$$\tau \left\| z \right\|_1 + \frac{\rho}{2} \left\| \begin{pmatrix} w^{k+1} \\ b^{k+1} \end{pmatrix} - \begin{pmatrix} z - \mu^k \\ z_b - \mu_b^k \end{pmatrix} \right\|_2^2 = \tau \left\| z \right\|_1 + \frac{\rho}{2} \left\| \begin{pmatrix} w^{k+1} + \mu^k \\ b^{k+1} + \mu_b^k \end{pmatrix} - \begin{pmatrix} z \\ z_b \end{pmatrix} \right\|_2^2$$

with respect to z and z_b. Since $\left\| z \right\|_1$ does not depend on z_b, as in the case of lasso in terms of z, we see that the z-update is

$$\begin{aligned} z^{k+1} &= S_{\tau/\rho}(w^{k+1} + \mu^k) \\ z_b^{k+1} &= b^{k+1} + \mu_b^k, \end{aligned}$$

where the function S_c given by

$$S_c(v) = \begin{cases} v - c & \text{if } v > c \\ 0 & \text{if } |v| \leq c \\ v + c & \text{if } v < -c \end{cases}$$

is the soft thresholding operator; see Gallier and Qaintance [6] (Section 16.8, Vol II) The μ-update is simply

$$\begin{pmatrix} \mu^{k+1} \\ \mu_b^{k+1} \end{pmatrix} = \begin{pmatrix} \mu^k \\ \mu_b^k \end{pmatrix} + \begin{pmatrix} w^{k+1} \\ b^{k+1} \end{pmatrix} - \begin{pmatrix} z^{k+1} \\ z_b^{k+1} \end{pmatrix}.$$

In summary, the ADMM steps to solve elastic net regression are

$$\begin{aligned} \begin{pmatrix} w^{k+1} \\ b^{k+1} \end{pmatrix} &= \begin{pmatrix} X^\top X + (K + \rho)I_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m + \rho \end{pmatrix}^{-1} \left(\begin{pmatrix} X^\top y \\ \mathbf{1}_m^\top y \end{pmatrix} + \rho \begin{pmatrix} z^k - \mu^k \\ z_b^k - \mu_b^k \end{pmatrix} \right) \\ z^{k+1} &= S_{\tau/\rho}(w^{k+1} + \mu^k) \\ z_b^{k+1} &= b^{k+1} + \mu_b^k \\ \mu^{k+1} &= \mu^k + w^{k+1} - z^{k+1} \\ \mu_b^{k+1} &= \mu_b^k + b^{k+1} - z_b^{k+1}. \end{aligned}$$

Observe that the equation $z_b^{k+1} = b^{k+1} + \mu_b^k$ implies that $\mu_b^{k+1} = 0$. But then $\mu_b^1 = 0$, which implies $z_b^2 = b^2$ and $\mu_b^2 = 0$, and so we deduce that $\mu_b^k = 0$ and $z_b^{k+1} = b^{k+1}$ for all $k \geq 1$.

A nice feature of this method is that it needs very minor changes to deal with the special cases where $\tau = 0$ or $K = 0$.

If $\tau = 0$ and $K > 0$, Problem (V1) reduces to standard ridge regression with intercept. In this case, since $K > 0$, the matrix

$$B = \begin{pmatrix} X^\top X + KI_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix}$$

is symmetric positive definite, so the function $J(w, b)$ has a unique minimum obtained by setting its gradient to zero, which yields the system

$$\begin{pmatrix} X^\top X + KI_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} = \begin{pmatrix} X^\top y \\ \mathbf{1}_m^\top y \end{pmatrix}, \tag{$*_1$}$$

so we have the unique solution

$$\begin{pmatrix} w \\ b \end{pmatrix} = \begin{pmatrix} X^\top X + KI_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix}^{-1} \begin{pmatrix} X^\top y \\ \mathbf{1}_m^\top y \end{pmatrix}. \tag{$*_2$}$$

This amounts to setting $\rho = 0$ in the equation

$$\begin{pmatrix} w^{k+1} \\ b^{k+1} \end{pmatrix} = \begin{pmatrix} X^\top X + (K+\rho)I_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m + \rho \end{pmatrix}^{-1} \left(\begin{pmatrix} X^\top y \\ \mathbf{1}_m^\top y \end{pmatrix} + \rho \begin{pmatrix} z^k - \mu^k \\ z_b^k - \mu_b^k \end{pmatrix} \right). \tag{$*_3$}$$

In this case no iteration is necessary. In writing a computer program implementing this method we simply need to have a test for $\tau = 0$. If $\tau > 0$, we go through the ADMM steps, else if $\tau = 0$ and $K > 0$, we compute

$$\begin{pmatrix} w \\ b \end{pmatrix}$$

using ($*_2$), and we skip the loop of iteration steps.

If $K = 0$ and $\tau > 0$, then Problem (V1) reduces to lasso regression with intercept. Nothing needs to be changed.

The case $K = 0$ and $\tau = 0$ is interesting and can be handled. If $\tau = 0$, the function $S_{\tau/\rho} = S_0$ is the identity, so we have the update

$$z^{k+1} = S_{\tau/\rho}(w^{k+1} + \mu^k) = w^{k+1} + \mu^k,$$

and so

$$\begin{aligned} \mu^{k+1} &= 0 \\ \mu_b^{k+1} &= 0 \end{aligned}$$

for all $k \geq 0$, which implies that

$$\begin{aligned} z^{k+1} &= w^{k+1} \\ z_b^{k+1} &= b^{k+1} \end{aligned}$$

for all $k \geq 1$.

Equation $(*_3)$ still applies with $K = 0$. Altough we have no proof at this time, we conjecture that ADMM converges to the solution given by the pseudo-inverse of

$$B = \begin{pmatrix} X^\top X & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix},$$

namely

$$\begin{pmatrix} w \\ b \end{pmatrix} = \begin{pmatrix} X^\top X & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix}^+ \begin{pmatrix} X^\top y \\ \mathbf{1}_m^\top y \end{pmatrix}. \tag{$*_4$}$$

In writing a computer program that covers the case $\tau = K = 0$, it is better to test whether $\tau > 0$ or $K = 0$ (which covers the case $\tau = K = 0$) and run the ADMM steps as usual, or if $\tau = 0$ and $K > 0$, then to use $(*_2)$ and skip the loop of iteration steps. Another option when $\tau = 0$ and $K = 0$ is to compute w and b using the pseudo-inverse as in $(*_4)$ and skip the ADMM loop altogether. Our implemention shows that both options compute the same solution within an error smaller than 10^{-10}, confirming our conjecture that in this case, ADMM converges to the pseudo-inverse solution.

2 Solving the Problem Using the Dual of Version 3

We show that the dual program of Version 3 of elastic net regression (see below) can be explicitly derived and then solved using ADMM. It turns out that the values of the Lagrange multipliers α_+ and α_- are uniquely determined and then the minimal solution (w, b) can be computed in terms of α_+ and α_- because the matrix

$$B = \begin{pmatrix} X^\top X + KI_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix}$$

is positive definite if $K > 0$, and thus invertible.

However, in the case of lasso where $K = 0$, the matrix B is not necessarily invertible and it may not be possible to find an explicit expression for the dual function $G(\alpha_+, \alpha_-)$ in terms of the Lagrange multipliers α_+ and α_-. Even if we could, we have no way of finding (w, b) from α_+ and α_-. We tried to use the pseudo-inverse of B but this method does not yield a correct minimal solution in all cases.

The formulation of the primal problem Version 3 is obtained from Version 1 by replacing the term $\tau \left\|w\right\|_1$ by $\tau \mathbf{1}_n^\top \epsilon$ and adding two constraints as in Version 2.

Program (**elastic net V3**):

$$\text{minimize} \quad J(w, b, \epsilon) = \frac{1}{2}(Xw + b\mathbf{1}_m - y)^\top (Xw + b\mathbf{1}_m - y) + \frac{1}{2}Kw^\top w + \tau \mathbf{1}_n^\top \epsilon$$

subject to

$$\begin{aligned} w &\leq \epsilon \\ -w &\leq \epsilon, \end{aligned}$$

where $K > 0$ and $\tau \geq 0$ are two constants controlling the influence of the ℓ^2-regularization and the ℓ^1-regularization. Expanding $(Xw + b\mathbf{1}_m - y)^\top (Xw + b\mathbf{1}_m - y)$ as in Section 1 and subtracting the constant term $\frac{1}{2}y^\top y$, we obtain

Program (**elastic net V3**):

$$\text{minimize} \quad J(w, b, \epsilon) = \frac{1}{2} \begin{pmatrix} w^\top & b \end{pmatrix} \begin{pmatrix} X^\top X + KI_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} - \begin{pmatrix} w^\top & b \end{pmatrix} \begin{pmatrix} X^\top y \\ \mathbf{1}_m^\top y \end{pmatrix} + \tau \mathbf{1}_n^\top \epsilon$$

subject to

$$\begin{aligned} w &\leq \epsilon \\ -w &\leq \epsilon, \end{aligned}$$

Let $\alpha_+ \in \mathbb{R}^n_+$ be vectors of Lagrange multipliers associated with the inequalities $w - \epsilon \leq 0$ and let $\alpha_- \in \mathbb{R}^n_+$ be Lagrange multipliers associated with the inequalities $-w - \epsilon \leq 0$. Then the Lagrangian is

$$\begin{aligned} L(w, b, \epsilon, \alpha_+, \alpha_-) &= J(w, b, \epsilon) + (w - \epsilon)^\top \alpha_+ + (-w - \epsilon)^\top \alpha_- \\ &= \frac{1}{2} \begin{pmatrix} w^\top & b \end{pmatrix} \begin{pmatrix} X^\top X + KI_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} - \begin{pmatrix} w^\top & b \end{pmatrix} \begin{pmatrix} X^\top y \\ \mathbf{1}_m^\top y \end{pmatrix} \\ &\quad + w^\top (\alpha_+ - \alpha_-) + \epsilon^\top (\tau \mathbf{1}_n - \alpha_+ - \alpha_-) \\ &= \frac{1}{2} \begin{pmatrix} w^\top & b \end{pmatrix} \begin{pmatrix} X^\top X + KI_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} - \begin{pmatrix} w^\top & b \end{pmatrix} \begin{pmatrix} X^\top y - \alpha_+ + \alpha_- \\ \mathbf{1}_m^\top y \end{pmatrix} \\ &\quad + \epsilon^\top (\tau \mathbf{1}_n - \alpha_+ - \alpha_-). \end{aligned}$$

To find the dual function $G(\alpha_+, \alpha_-)$, we minimize $L(w, b, \epsilon, \alpha_+, \alpha_-)$ with respect to w, b and ϵ. Since $L(w, b, \epsilon, \alpha_+, \alpha_-)$ is a convex function (the quadratic term involves a symmetric positive semidefinite matrix) defined on a vector space, a minimum exists iff the gradient $\nabla L_{w,b,\epsilon}$ vanishes; see Theorem 4.13(4) of Gallier and Quaintance [6]. We have

$$\nabla L_{w,b,\epsilon} = \begin{pmatrix} \begin{pmatrix} X^\top X + KI_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} - \begin{pmatrix} X^\top y - \alpha_+ + \alpha_- \\ \mathbf{1}_m^\top y \end{pmatrix} \\ \tau \mathbf{1}_n - \alpha_+ - \alpha_- \end{pmatrix},$$

so we get the equations

$$\begin{aligned} \begin{pmatrix} X^\top X + KI_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} &= \begin{pmatrix} X^\top y - \alpha_+ + \alpha_- \\ \mathbf{1}_m^\top y \end{pmatrix} \\ \alpha_+ + \alpha_- &= \tau \mathbf{1}_n. \end{aligned}$$

This suggests to express the right-hand side of the first equation in terms of the vector of dimension $2n$ whose first n components are α_+ and whose last n components are α_-. We can write

$$\begin{aligned}\begin{pmatrix} X^\top y - \alpha_+ + \alpha_- \\ \mathbf{1}_m^\top y \end{pmatrix} &= \begin{pmatrix} -\alpha_+ + \alpha_- \\ 0 \end{pmatrix} + \begin{pmatrix} X^\top y \\ \mathbf{1}_m^\top y \end{pmatrix} \\ &= \begin{pmatrix} -I_n & I_n \\ 0_n^\top & 0_n^\top \end{pmatrix} \begin{pmatrix} \alpha_+ \\ \alpha_- \end{pmatrix} + \begin{pmatrix} X^\top y \\ \mathbf{1}_m^\top y \end{pmatrix}.\end{aligned}$$

Therefore we can write

$$\begin{aligned}\begin{pmatrix} X^\top X + KI_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} &= \begin{pmatrix} X^\top y - \alpha_+ + \alpha_- \\ \mathbf{1}_m^\top y \end{pmatrix} \\ &= \begin{pmatrix} -I_n & I_n \\ 0_n^\top & 0_n^\top \end{pmatrix} \begin{pmatrix} \alpha_+ \\ \alpha_- \end{pmatrix} + \begin{pmatrix} X^\top y \\ \mathbf{1}_m^\top y \end{pmatrix},\end{aligned}$$

and we have the equations

$$\begin{pmatrix} X^\top X + KI_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} = \begin{pmatrix} -I_n & I_n \\ 0_n^\top & 0_n^\top \end{pmatrix} \begin{pmatrix} \alpha_+ \\ \alpha_- \end{pmatrix} + \begin{pmatrix} X^\top y \\ \mathbf{1}_m^\top y \end{pmatrix} \tag{eq1}$$

$$\alpha_+ + \alpha_- = \tau \mathbf{1}_n. \tag{eq2}$$

It is convenient to define the $(n+1) \times (n+1)$ matrix B, the $(n+1) \times 2n$ matrix B_1, and the $(n+1) \times 1$ matrix B_2 as

$$\begin{aligned} B &= \begin{pmatrix} X^\top X + KI_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix} \\ B_1 &= \begin{pmatrix} -I_n & I_n \\ 0_n^\top & 0_n^\top \end{pmatrix} \\ B_2 &= \begin{pmatrix} X^\top y \\ \mathbf{1}_m^\top y \end{pmatrix}, \end{aligned}$$

so that

$$\begin{pmatrix} X^\top y - \alpha_+ + \alpha_- \\ \mathbf{1}_m^\top y \end{pmatrix} = B_1 \begin{pmatrix} \alpha_+ \\ \alpha_- \end{pmatrix} + B_2, \tag{eq3}$$

and (eq1) can be written more concisely as

$$B \begin{pmatrix} w \\ b \end{pmatrix} = B_1 \begin{pmatrix} \alpha_+ \\ \alpha_- \end{pmatrix} + B_2. \tag{eq4}$$

Up to this point we never used the fact that B is SPD (since in elastic net regression, $K > 0$). This allows us to express the dual function $G(\alpha_+, \alpha_-)$ explicitly in terms of α_+ and α_- and thus to obtain the dual program. If $K = 0$ and B is not invertible, we seem to be stuck. In order to proceed we assume that $K > 0$ (and $\tau > 0$).

Since B is SPD, it is invertible so we obtain

$$\begin{pmatrix} w \\ b \end{pmatrix} = B^{-1} B_1 \begin{pmatrix} \alpha_+ \\ \alpha_- \end{pmatrix} + B^{-1} B_2. \tag{eq5}$$

If (w, b, ϵ) is a minimizer of the Lagrangian $L(w, b, \epsilon, \alpha_+, \alpha_-)$ (holding α_+, α_- fixed), since the Lagrangian is

$$\begin{aligned} L(w, b, \epsilon, \alpha_+, \alpha_-) = \frac{1}{2} \begin{pmatrix} w^\top & b \end{pmatrix} \begin{pmatrix} X^\top X + K I_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} - \begin{pmatrix} w^\top & b \end{pmatrix} \begin{pmatrix} X^\top y - \alpha_+ + \alpha_- \\ \mathbf{1}_m^\top y \end{pmatrix} \\ + \epsilon^\top (\tau \mathbf{1}_n - \alpha_+ - \alpha_-), \end{aligned}$$

by (eq2), (eq3), and the definition of the matrices B, B_1, B_2, we obtain

$$G(\alpha_+, \alpha_-) = \frac{1}{2} \begin{pmatrix} w^\top & b \end{pmatrix} B \begin{pmatrix} w \\ b \end{pmatrix} - \begin{pmatrix} w^\top & b \end{pmatrix} \left(B_1 \begin{pmatrix} \alpha_+ \\ \alpha_- \end{pmatrix} + B_2 \right).$$

Since B is symmetric, Equation (eq5) yields

$$\begin{pmatrix} w^\top & b \end{pmatrix} = \begin{pmatrix} \alpha_+^\top & \alpha_-^\top \end{pmatrix} B_1^\top B^{-1} + B_2^\top B^{-1}, \tag{eq6}$$

so using (eq4) and (eq6) we obtain

$$\begin{aligned} G(\alpha_+, \alpha_-) &= \frac{1}{2} \left(\begin{pmatrix} \alpha_+^\top & \alpha_-^\top \end{pmatrix} B_1^\top B^{-1} + B_2^\top B^{-1} \right) \left(B_1 \begin{pmatrix} \alpha_+ \\ \alpha_- \end{pmatrix} + B_2 \right) \\ &\quad - \left(\begin{pmatrix} \alpha_+^\top & \alpha_-^\top \end{pmatrix} B_1^\top B^{-1} + B_2^\top B^{-1} \right) \left(B_1 \begin{pmatrix} \alpha_+ \\ \alpha_- \end{pmatrix} + B_2 \right) \\ &= -\frac{1}{2} \left(\begin{pmatrix} \alpha_+^\top & \alpha_-^\top \end{pmatrix} B_1^\top B^{-1} + B_2^\top B^{-1} \right) \left(B_1 \begin{pmatrix} \alpha_+ \\ \alpha_- \end{pmatrix} + B_2 \right) \\ &= -\frac{1}{2} \begin{pmatrix} \alpha_+^\top & \alpha_-^\top \end{pmatrix} B_1^\top B^{-1} B_1 \begin{pmatrix} \alpha_+ \\ \alpha_- \end{pmatrix} - \begin{pmatrix} \alpha_+^\top & \alpha_-^\top \end{pmatrix} B_1^\top B^{-1} B_2 - \frac{1}{2} B_2^\top B^{-1} B_2. \end{aligned}$$

The dual problem, which is to maximize $G(\alpha_+, \alpha_-)$ subject to Equation (eq2), is equivalent to minimizing $-G(\alpha_+, \alpha_-)$ subject to Equation (eq2), and since the constant term $-\frac{1}{2} B_2^\top B^{-1} B_2$ is irrelevant, we obtain the following dual program:

Program (dual elastic net V3):

$$\text{minimize} \quad \frac{1}{2} \begin{pmatrix} \alpha_+^\top & \alpha_-^\top \end{pmatrix} B_1^\top B^{-1} B_1 \begin{pmatrix} \alpha_+ \\ \alpha_- \end{pmatrix} + \begin{pmatrix} \alpha_+^\top & \alpha_-^\top \end{pmatrix} B_1^\top B^{-1} B_2$$

subject to

$$\begin{aligned} &\alpha_+ + \alpha_- = \tau \mathbf{1}_n \\ &\alpha_+ \geq 0, \ \alpha_- \geq 0. \end{aligned}$$

The equation

$$\alpha_+ + \alpha_- = \tau \mathbf{1}_n$$

is written in matrix form as

$$\begin{pmatrix} I_n & I_n \end{pmatrix} \begin{pmatrix} \alpha_+ \\ \alpha_- \end{pmatrix} = \tau \mathbf{1}_n,$$

and the matrix of the system obviously has rank n. Since B is SPD, so is B^{-1}, and it is well known that $B_1^\top B^{-1} B_1$ is symmetric positive semidefinite, thus ADMM applies. Once α_+ and α_- are computed, w and b are given by Equation (eq5).

Observe that Equation (eq1) implies that

$$\alpha_+ - \alpha_- = -(X^\top X + K I_n) w - (X^\top \mathbf{1}_m) b + X^\top y, \tag{eq1a}$$

and since Equation (eq2) is

$$\alpha_+ + \alpha_- = \tau \mathbf{1}_n, \tag{eq2}$$

we see that for any minimal solution (w, b), the Lagrange multipliers α_+ and α_- are uniquely determined by

$$\alpha_+ = \frac{1}{2} \left(-(X^\top X + K I_n) w - (X^\top \mathbf{1}_m) b + X^\top y + \tau \mathbf{1}_n \right) \tag{$*_{\alpha_+}$}$$

$$\alpha_- = \frac{1}{2} \left((X^\top X + K I_n) w + (X^\top \mathbf{1}_m) b - X^\top y + \tau \mathbf{1}_n \right). \tag{$*_{\alpha_-}$}$$

Actually, the components of α_+ and α_- corresponding to nonzero components w_i of w in a minimal solution (w, b) are either τ or 0 depending on the sign of w_i.

Proposition 2.1. *For any minimal solution (w, b), we have $|w_i| = \epsilon_i$ for $i = 1, \ldots, n$. Furthermore, if $w_i = \epsilon_i > 0$, then $(\alpha_+)_i = \tau$ and $(\alpha_-)_i = 0$, and if $w_i = -\epsilon_i < 0$, then $(\alpha_+)_i = 0$ and $(\alpha_-)_i = \tau$.*

Proof. For a minimal solution (w, b), we have $\alpha_+, \alpha_- \geq 0$,

$$\begin{aligned} w - \epsilon &\leq 0 \\ -w - \epsilon &\leq 0 \\ \alpha_+ + \alpha_- &= \tau \mathbf{1}_n, \end{aligned}$$

and we have the KKT equations

$$\alpha_+^\top (w - \epsilon) = 0 \tag{KKT1}$$

$$\alpha_-^\top (-w - \epsilon) = 0. \tag{KKT2}$$

Observe that the inequalities

$$\begin{aligned} w - \epsilon &\leq 0 \\ -w - \epsilon &\leq 0 \end{aligned}$$

are equivalent to $|w_i| \leq \epsilon_i$ for $i = 1, \ldots, n$, so $\epsilon \geq 0$. Also, since $\alpha_+, \alpha_- \geq 0$ and $\alpha_+ + \alpha_- = \tau \mathbf{1}_n$, either $(\alpha_+)_i > 0$ or $(\alpha_-)_i > 0$. We proceed by contradiction.

First assume that $w_i > 0$. If $(\alpha_-)_i > 0$, then by (KKT2)

$$-w_i = \epsilon_i,$$

and since $w_i > 0$ and $\epsilon_i \geq 0$, this is a contradiction, so we must have $(\alpha_-)_i = 0$. Since $\alpha_+ + \alpha_- = \tau \mathbf{1}_n$, we deduce that $(\alpha_+)_i = \tau$, and by (KKT1), $w_i = \epsilon_i$.

Next if $w_i < 0$ and $(\alpha_+)_i > 0$, then by (KKT1)

$$w_i = \epsilon_i,$$

and since $w_i < 0$ and $\epsilon_i \geq 0$, this is a contradiction so we must have $(\alpha_+)_i = 0$. Since $\alpha_+ + \alpha_- = \tau \mathbf{1}_n$, we deduce that $(\alpha_-)_i = \tau$, and by (KKT2), $w_i = -\epsilon_i$.

If $w_i = 0$, since either $(\alpha_+)_i > 0$ or $(\alpha_-)_i > 0$, by (KKT1) or (KKT2) we must have $\epsilon_i = w_i = 0$. Thus we confirm that for any minimal solution (w, b), we have $|w_i| = \epsilon_i$. $\square$

Corollary 2.2. *Since $\epsilon \geq 0$, the KKT equations imply that if $(\alpha_+)_i = \tau$, then $w_i \geq 0$, if $(\alpha_+)_i = 0$, then $w_i \leq 0$, and if $0 < (\alpha_+)_i < \tau$, then $w_i = 0$.*

For $w_i = 0$, any values of $(\alpha_+)_i, (\alpha_-)_i$ with $(\alpha_+)_i, (\alpha_-)_i > 0$ and $(\alpha_+)_i + (\alpha_-)_i = \tau$ are possible. In this case, $-\tau \leq (\alpha_+)_i - (\alpha_-)_i \leq \tau$. The readers familiar with subgradients in convex analysis will recognize that $(1/\tau)(\alpha_+ - \alpha_-)$ is a subgradient of the non-differentiable function $w \mapsto \|w\|_1$. This fact is also noted by Hastie, Tibshirani and Wainwright [7].

The previous discusion also explains why if we pick τ big, it becomes more difficult to satisfy simultaneously the conditions $(\alpha_+)_i + (\alpha_-)_i = \tau$, $(\alpha_+)_i > 0$ and $(\alpha_-)_i > 0$, so many w_i are driven to zero.

Inspired by Section 1, we can attempt to deal with the special cases $\tau = 0$ or $K = 0$ as follows.

If $\tau > 0$ and $K = 0$, then Problem (V3) reduces to lasso. The matrix

$$B = \begin{pmatrix} X^\top X & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix}$$

in Equation (eq4) is not necessarily invertible. This happens when B has rank strictly smaller than $n+1$. To overcome this problem we can try using the pseudo-inverse B^+ of B, and we use the matrices $B^+ B_1$ and $B^+ B_2$ in Equation (eq5). Similarly, to solve the dual program we can try using the matrices $B_1^\top B^+ B_1$ and $B_1^\top B^+ B_2$ instead of $B_1^\top B^{-1} B_1$ and $B_1^\top B^{-1} B_2$. Unfortunately this "obvious" solution does not work. We have an example of a data set with two linearly independent column vectors x_1 and x_2 of dimension 8 to which we add the third vector $x_3 = x_1 + x_2$. For $\tau = 0.1$, ADMM applied to the primal formulation of Section 2 yields the solution

$$w = \begin{pmatrix} 0.034 \\ 0 \\ 1.1384 \end{pmatrix},$$

but the solution obtained by using the pseudo-inverse of B to solve Equation (eq4) with the same values of the Lagrange multipliers (computed using $*_{\alpha_+}$ and $*_{\alpha_+}$) whose values are $\alpha_+ = (0.1, 0.05, 0.1)$ and $\alpha_- = (0, 0.05, 0)$, is

$$wps = \begin{pmatrix} 0.3998 \\ 0.3693 \\ 1.7691 \end{pmatrix},$$

which is incorrect since its second component should be zero. In fact, $\|w\|_1 = 1.1688$ and $\|wps\|_1 = 1.5382$, so wps is not a solution of minimal 1-norm. We will discuss this issue more extensively in Section 3.

If $\tau = 0$ and $K > 0$, Problem (V3) reduce to ridge regression. In this case there is no need to use ADMM to solve the dual because $\alpha_+ = \alpha_- = 0_n$, so the unique minimal solution is given by (eq4) with $\alpha_+ = \alpha_- = 0_n$, which is identical to $(*_2)$.

Finally, if $\tau = 0$ and $K = 0$, then this is the least-squares solution given by a pseudo-inverse, so we skip the ADMM steps and compute w and b using the pseudo-inverse of B using $(*_4)$.

Our implementations of the primal version of Section 1 and of the dual version of Section 2 show excellent agreement of the solutions in all the cases where B has full rank, typically with a numerical difference smaller than 10^{-10}. The use of the pseudo-inverse of B when $K = 0$ and $\tau = 0$ appears to be correct, although we have no formal proof of this fact.

We finish this section by proving that the difference $(\alpha_+ - \alpha_-)/\tau$ is indeed a subgradient of the ℓ^1-norm function $w \mapsto \|w\|_1$ at w, where (w, b) is any minimal solution.

Recall that if $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a proper convex function[1] then for any $x \in \mathbb{R}^n$, the set $\partial f(x)$ of *subgradients* of f at x is the set (possibly empty) of vectors $u \in \mathbb{R}^n$ such that

$$f(z) \geq f(x) + \langle z - x, u\rangle, \quad \text{for all } z \in \mathbb{R}^n.$$

See Gallier and Qaintance [6], Section 15.3, or Rockafellar [8], Section 23.

Proposition 2.3. *For any $w = (w_1, \ldots, w_n) \in \mathbb{R}^n$, the set $\partial f(w)$ of subgradients $u = (u_1, \ldots, u_n) \in \mathbb{R}^n$ of the ℓ^1-norm function $f(w) = \|w\|_1$ at w is given by*

$$u_i = \begin{cases} 1 & \text{if } w_i > 0 \\ -1 & \text{if } w_i < 0 \\ v_i \in [-1, +1] & \text{if } w_i = 0, \end{cases}$$

for $i = 1, \ldots, n$.

[1] To avoid technicalities we assume that f is proper. Norms are proper and convex functions.

Proof. For $i = 1, \ldots, n$ define the function $f_i \colon \mathbb{R}^n \to \mathbb{R}$ by

$$f_i(w_1, \ldots, w_n) = |w_i|.$$

Obviously

$$f(w_1, \ldots, w_n) = \|(w_1, \ldots, w_n)\|_1 = \sum_{i=1}^{n} |w_i| = \sum_{i=1}^{n} f_i(w_1, \ldots, w_n),$$

and the functions f_i are proper and convex. By Proposition 15.23 of Gallier and Qaintance [6], or Theorem 23.8 of Rockafellar [8], since the the f_i are total functions,

$$\partial f(w) = \partial f_1(w) + \cdots + \partial f_n(w).$$

Thus we are reduced to finding $\partial f_i(w)$. If $w_i \neq 0$, then $f_i(w) = w_i$ if $w_i > 0$ and $f_i(w) = -w_i$ if $w_i < 0$, so f_i is differentiable at w and its gradient is obviously given by

$$\nabla f_i(w) = \begin{cases} e_i & \text{if } w_i > 0 \\ -e_i & \text{if } w_i < 0, \end{cases}$$

where e_i is the ith canonical basis vector in $\mathbb{R}^n$. In this case $\partial f_i(w) = \{\nabla f_i(w)\}$.

If $w_i = 0$, then u is a subgradient at w iff

$$f_i(z) \geq \langle z, u \rangle \quad \text{for all } z \in \mathbb{R}^n$$

iff

$$|z_i| \geq \sum_{j=1}^{n} z_j u_j \quad \text{for all } z = (z_1, \ldots, z_n) \in \mathbb{R}^n.$$

For $z = e_j$, $j \neq i$, we obtain

$$0 \geq u_j$$

and for $z = -e_j$ $j \neq i$, we obtain

$$0 \geq -u_j,$$

so $u_j = 0$ for all $j \neq i$. Therefore u_i must satisfy the condition

$$|z_i| \geq z_i u_i \quad \text{for all } z_i \in \mathbb{R}.$$

If $z_i > 0$, then

$$z_i \geq z_i u_i,$$

which is equivalent to $z_i(1 - u_i) \geq 0$ for all $z_i > 0$, and thus we must have $1 - u_i \geq 0$, namely $u_i \leq 1$.

If $z_i < 0$, then

$$-z_i \geq z_i u_i,$$

which is equivalent to $-z_i(1+u_i) \geq 0$ for all $z_i < 0$, and thus we must have $1 + u_i \geq 0$, namely $u_i \geq -1$.

Obviously

$$|z_i| \geq z_i u_i \quad \text{for all } z_i \in \mathbb{R}$$

for all u_i such that $-1 \leq u_i \leq +1$, so we proved that if $w_i = 0$, then

$$\partial f_i(w) = \{(0, \ldots, 0, u_i, 0, \ldots, 0) \mid -1 \leq u_i \leq +1\},$$

which finishes the proof. $\square$

The objective function of the version (V1) of lasso is

$$J(w,b) = \frac{1}{2} \begin{pmatrix} w^\top & b \end{pmatrix} \begin{pmatrix} X^\top X + & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} - \begin{pmatrix} w^\top & b \end{pmatrix} \begin{pmatrix} X^\top y \\ \mathbf{1}_m^\top y \end{pmatrix} + \tau \left\| w \right\|_1$$

and we obtain

$$\begin{aligned} \partial J(w,b) &= \begin{pmatrix} X^\top X + & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} - \begin{pmatrix} X^\top y \\ \mathbf{1}_m^\top y \end{pmatrix} + \begin{pmatrix} \tau \partial f(w) \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} X^\top X + & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} - \begin{pmatrix} X^\top y - \tau \partial f(w) \\ \mathbf{1}_m^\top y \end{pmatrix}. \end{aligned}$$

By definition of a subgradient, the function J has a minimum at (w,b) iff $0_{n+1} \in \partial J(w,b)$ (see Proposition 15.34 of Gallier and Qaintance [6] or Section 27 in Rockafellar [8]) iff there is some $u \in \partial f(w)$ such that

$$\begin{pmatrix} X^\top X + & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} = \begin{pmatrix} X^\top y - \tau u \\ \mathbf{1}_m^\top y \end{pmatrix}.$$

Earlier using Version (V3) in terms of the Lagrange multipliers we found the equations

$$\begin{aligned} \begin{pmatrix} X^\top X + K I_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} &= \begin{pmatrix} X^\top y - \alpha_+ + \alpha_- \\ \mathbf{1}_m^\top y \end{pmatrix} \\ \alpha_+ + \alpha_- &= \tau \mathbf{1}_n. \end{aligned}$$

So now we see the connection between the subgradients of the ℓ^1-norm function $f(w) = \left\| w \right\|_1$ and the Lagrange multipliers α_+ and α_-: for a minimal solution (w,b), the difference $\alpha_+ - \alpha_-$ is equal to τu, where $u \in \partial f(w)$ is a subgradient the ℓ^1-norm function f at w. It turns out that this subgradient is the same for all minimal solutions (w,b), because the Lagrange multipliers α_+ and α_- are the same for all minimal solutions, a fact that is not obvious at first glance and will be established in Section 3.

3 Uniqueness of Minimal Solutions for Lasso

When $\tau > 0$ and $K = 0$, which corresponds to lasso minimization, if X does not have full rank, it is possible that infinitely many minimizers exist. It is actually possible to figure out exactly when this happens. To avoid subgradients, we use the formulations (V2) and (V3) with $K = 0$. A crucial property is that the Lagrange multipliers α_+ and α_- *have the same value for all minimal solutions*; see Proposition 3.2. This follows from the fact that error vectors $\xi = y - Xw - b\mathbf{1}_m$ have have the same value ξ^* for all minimal solution (w, b); see Proposition 3.1. Since the Lagrange multipliers α_+ and α_- are uniquely determined, it turns out that the set of minimal solutions depends heavily on the sets of indices K_0, K_+ and K_- given by

$$\begin{aligned}
K_0 &= \{i \in \{1, \ldots, n\} \mid 0 < (\alpha_+)_i < \tau\} \\
K_+ &= \{i \in \{1, \ldots, n\} \mid (\alpha_+)_i = \tau\} \\
K_- &= \{i \in \{1, \ldots, n\} \mid (\alpha_+)_i = 0\}.
\end{aligned}$$

Let

$$s_k = \begin{cases} 0 & \text{if } k \in K_+ \\ +1 & \text{if } k \in K_-, \end{cases}$$

We will show that for any two minimal solutions (w_1, b_1) and (w_2, b_2), if we write $\delta = w_2 - w_1$ and $\eta = b_2 - b_1$, then the following equations hold:

$$\sum_{k \in K_+ \cup K_-} \delta_k X^k + \eta \mathbf{1}_m = 0, \quad \sum_{k \in K_+ \cup K_-} (-1)^{s_k} \delta_k = 0, \;\text{ and }\; \delta_k \neq 0 \;\text{ for some } k.$$

See Proposition 3.4. The above equations place a heavy constraint on the vectors (δ, η), which belong to the kernel of $\widetilde{X} = (X \; \mathbf{1}_m)$. An additional source of constraints comes from the KKT conditions; see Proposition 3.3. These propositions will allow us to describe the structure of the minimal solutions. They are convex sets obtained by intersecting a simplex and an affine space related to the kernel of $\widetilde{X}$; see Proposition 5.1, so they are polytopes.

Interestingly some notions of affine geometry arise.[2] In particular, a notion of affine dependence stronger that the usual notion comes up. We characterize this (new?) notion in Proposition 4.1. The relevant concepts of affine geometry are discussed in Section 4, and the reader may want to read it first.

We begin by proving that the error vector ξ has the same value for all minimal solutions, a fact that relies on the fact that the function $\xi \mapsto \frac{1}{2}\xi^\top \xi$ is strictly convex.

In the formulation (V2) given by

[2] This would not be surprising to Eugenio Calabi.

Program (**lasso V2**):

$$
\begin{aligned}
&\text{minimize} \quad \frac{1}{2}\xi^\top \xi + \tau \mathbf{1}_n^\top \epsilon \\
&\text{subject to} \\
&\qquad\qquad y - Xw - b\mathbf{1}_m = \xi \\
&\qquad\qquad w \leq \epsilon \\
&\qquad\qquad -w \leq \epsilon,
\end{aligned}
$$

we can either view the objective function $J(\xi, \epsilon, w, b)$ as a function of ξ, ϵ, w, b constrained to the convex set

$$U = \{(\xi, \epsilon, w, b) \mid y - Xw - b\mathbf{1}_m = \xi,\ w \leq \epsilon,\ -w \leq \epsilon\},$$

which is obviously nonempty. As such the objective function

$$J(\xi, \epsilon, w, b) = \frac{1}{2}\xi^\top \xi + \tau \mathbf{1}_n^\top \epsilon$$

is convex, but not strictly convex. However, we can also view the objective function as the sum

$$J_2(\xi, \epsilon) = J_{2,1}(\xi) + J_{2,2}(\epsilon)$$

of the two functions

$$J_{2,1}(\xi) = \frac{1}{2}\xi^\top \xi, \quad J_{2,2}(\epsilon) = \tau \mathbf{1}_n^\top \epsilon.$$

In this case J_2 is defined on the convex set $V_1 \times V_2$, where $J_{2,1}$ is defined on the convex set V_2 given by

$$V_1 = \{\xi \mid (\exists \epsilon \in \mathbb{R}^n)(\exists w \in \mathbb{R}^n)(\exists b \in \mathbb{R})(y - Xw - b\mathbf{1}_m = \xi,\ w \leq \epsilon,\ -w \leq \epsilon)\}$$

and $J_{2,2}$ is defined on the convex set V_2 given by

$$V_2 = \{\epsilon \mid (\exists \xi \in \mathbb{R}^m)(\exists w \in \mathbb{R}^n)(\exists b \in \mathbb{R})(y - Xw - b\mathbf{1}_m = \xi,\ w \leq \epsilon,\ -w \leq \epsilon)\}.$$

Observe that V_1 is the projection of U by the projection map $\pi_1 \colon \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^m$, given by

$$\pi_1(\xi, \epsilon, w, b) = \xi$$

and V_2 is the projection of U by the projection map $\pi_2 \colon \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$, given by

$$\pi_2(\xi, \epsilon, w, b) = \epsilon.$$

Since π_1 and π_2 are linear and since U is convex, $V_1 = \pi_1(U)$ and $V_2 = \pi_2(U)$ are also convex. We observed earlier that for any $\epsilon \in V_2$, we have $|w| \leq \epsilon$, so (since $\tau > 0$) the function $J_{2,2}(\epsilon) = \tau \mathbf{1}_n^\top \epsilon$ is nonnegative on V_2, and obviously the function $J_{2,1}(\xi) = \frac{1}{2}\xi^\top \xi$ is nonnegative on V_1. Therefore, the minimum of the function $J_2(\xi,\epsilon) = J_{2,1}(\xi) + J_{2,2}(\epsilon)$ on the convex set $V_1 \times V_2$ is the sum of the minima of the functions $J_{2,1}(\xi)$ on V_1 and $J_{2,2}(\epsilon)$ on V_2. Since the function $J_{2,1}(\xi) = \frac{1}{2}\xi^\top\xi$ is strictly convex and nonnegative on the nonempty convex set V_1, by Theorem 4.13(2) of Gallier and Quaintance [6], it has a minimum achieved for a unique ξ, say ξ^*. Therefore,

$$\xi^* = y - Xw - b\mathbf{1}_m$$

has the same value for all minimal solutions and so

$$\begin{pmatrix} X & \mathbf{1}_m \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} = Xw + b\mathbf{1}_m = y - \xi^*$$

has the same value for all minimal solutions. This key fact is worth recording as the following proposition.

Proposition 3.1. *Let $\widetilde{X}$ be the $m \times (n+1)$-matrix*

$$\widetilde{X} = \begin{pmatrix} X & \mathbf{1}_m \end{pmatrix}.$$

If ξ^* is the unique value of the first component of any minimal solution (ξ, ϵ) of the objective function of Problem lasso (V2) on the nonempty convex set U given by

$$U = \{(\xi, \epsilon, w, b) \mid y - Xw - b\mathbf{1}_m = \xi,\ w \leq \epsilon,\ -w \leq \epsilon\},$$

we have

$$\widetilde{X} \begin{pmatrix} w \\ b \end{pmatrix} = Xw + b\mathbf{1}_m = y - \xi^*$$

for all minimal solutions (w, b) of Problem lasso (V2).

It is well-known that the set of solutions of the linear equation

$$\widetilde{X} \begin{pmatrix} w \\ b \end{pmatrix} = y - \xi \tag{$*$}$$

is the affine subspace

$$\begin{pmatrix} w_1 \\ b_1 \end{pmatrix} + \operatorname{Ker} \widetilde{X},$$

where $\begin{pmatrix} w_1 \\ b_1 \end{pmatrix}$ is any solution of $(*)$, so the set of solutions does not depend on the particular solution chosen; see Gallier and Quaintance [5], Proposition 7.20. If $\begin{pmatrix} \delta \\ \eta \end{pmatrix}$ is a nonzero vector in $\operatorname{Ker} \widetilde{X}$, then $\delta \neq 0$, because

$$\widetilde{X} \begin{pmatrix} \delta \\ \eta \end{pmatrix} = X\delta + \eta\mathbf{1}_m = 0,$$

and $\delta = 0$ implies that $\eta = 0$.

Observe that

$$(\widetilde{X})^\top \widetilde{X} = \begin{pmatrix} X & \mathbf{1}_m \end{pmatrix}^\top \begin{pmatrix} X & \mathbf{1}_m \end{pmatrix} = \begin{pmatrix} X^\top \\ \mathbf{1}_m^\top \end{pmatrix} \begin{pmatrix} X & \mathbf{1}_m \end{pmatrix} = \begin{pmatrix} X^\top X & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix},$$

the matrix B introduced earlier, and so

$$B \begin{pmatrix} w \\ b \end{pmatrix}$$

has the same value for all minimal solutions.

Recall that Version (V3) of lasso is

Program (lasso V3):

$$\text{minimize} \quad J(w, b, \epsilon) = \frac{1}{2} \begin{pmatrix} w^\top & b \end{pmatrix} \begin{pmatrix} X^\top X & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} - \begin{pmatrix} w^\top & b \end{pmatrix} \begin{pmatrix} X^\top y \\ \mathbf{1}_m^\top y \end{pmatrix} + \tau \mathbf{1}_n^\top \epsilon$$

subject to

$$\begin{aligned} w &\leq \epsilon \\ -w &\leq \epsilon. \end{aligned}$$

Since the matrix

$$B = \begin{pmatrix} X^\top X & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix}$$

is symmetric positive semi-definite, the objective function is convex, and since the contraints are affine, by Theorem 14.6 of Gallier and Quaintance [6], there is a minimal solution (w, b, ϵ) iff the KKT equations hold, and using the notations of Section 2, they are expressed as

$$\begin{aligned} B \begin{pmatrix} w \\ b \end{pmatrix} &= B_1 \begin{pmatrix} \alpha_+ \\ \alpha_- \end{pmatrix} + B_2 \\ \alpha_+ + \alpha_- &= \tau \mathbf{1}_n \\ \alpha_+^\top (w - \epsilon) &= 0 && \text{(KKT1)} \\ \alpha_-^\top (-w - \epsilon) &= 0 && \text{(KKT2)} \\ \alpha_+, \, \alpha_- &\geq 0. \end{aligned}$$

If (w_1, b_1) and (w_2, b_2) are two minimal solutions, then in view of Proposition 3.1,

$$\widetilde{X} \begin{pmatrix} w_1 \\ b_1 \end{pmatrix} = \widetilde{X} \begin{pmatrix} w_2 \\ b_2 \end{pmatrix} \quad \text{and} \quad B \begin{pmatrix} w_1 \\ b_1 \end{pmatrix} = B \begin{pmatrix} w_2 \\ b_2 \end{pmatrix},$$

so

$$X^\top X w_1 + X^\top \mathbf{1}_m b_1 = X^\top X w_2 + X^\top \mathbf{1}_m b_2,$$

and Equations $(*_{\alpha_1})$ and $(*_{\alpha_2})$ show that α_+ and α_- have the same value for all minimal solutions. In summary we proved the following crucial and somewhat surprising result.

Proposition 3.2. *The Lagrange multipliers α_+ and α_- have the same value for all minimal solutions of Problem lasso (V3). They are given by*

$$\alpha_+ = \frac{1}{2}\left(-(X^\top X + KI_n)w - (X^\top \mathbf{1}_m)b + X^\top y + \tau \mathbf{1}_n\right) \qquad (*_{\alpha_+})$$
$$\alpha_- = \frac{1}{2}\left((X^\top X + KI_n)w + (X^\top \mathbf{1}_m)b - X^\top y + \tau \mathbf{1}_n\right), \qquad (*_{\alpha_-})$$

for any minimal solution (w, b).

We explained earlier that

$$\begin{pmatrix} w_1 \\ b_1 \end{pmatrix} - \begin{pmatrix} w_2 \\ b_2 \end{pmatrix}$$

is in the kernel of $\widetilde{X}$. Since $B = (\widetilde{X})^\top \widetilde{X}$ and it is well-known that $\widetilde{X}$ and B have the same kernel (see Proposition 20.4 of Gallier and Quaintance [5]),

$$\begin{pmatrix} w_1 \\ b_1 \end{pmatrix} - \begin{pmatrix} w_2 \\ b_2 \end{pmatrix}$$

is in the kernel of B.

Let (w_1, b_1) and (w_2, b_2) be two minimal solutions, and let

$$\begin{pmatrix} \delta \\ \eta \end{pmatrix} = \begin{pmatrix} w_2 \\ b_2 \end{pmatrix} - \begin{pmatrix} w_1 \\ b_1 \end{pmatrix} = \begin{pmatrix} w_2 - w_1 \\ b_2 - b_1 \end{pmatrix},$$

a vector in the kernel of $\widetilde{X}$. Consider any component δ_i of δ, with $1 \leq i \leq n$. By Corollary 2.2, we have the following classification according to the value of $(\alpha_+)_i$:

(1) $(\alpha_+)_i = \tau$. Then by (KKT1), $(w_1)_i \geq 0$ and $(w_2)_i = (w_1)_i + \delta_i \geq 0$, and so $|(w_1)_i| = (w_1)_i$ and $|(w_2)_i| = (w_1)_i + \delta_i$. We must also have $\delta_i \geq -(w_1)_i$ (where $-(w_1)_i \leq 0$).

(2) $(\alpha_+)_i = 0$. Then by (KKT2), $(w_1)_i \leq 0$ and $(w_2)_i = (w_1)_i + \delta_i \leq 0$, and so $|(w_1)_i| = -(w_1)_i$ and $|(w_2)_i| = -(w_1)_i - \delta_i$. We must also have $\delta_i \leq -(w_1)_i$ (where $-(w_1)_i \geq 0$).

(3) $0 < (\alpha_+)_i < \tau$. Then by (KKT1) and (KKT2), $(w_1)_i = 0$ and $(w_2)_i = (w_1)_i + \delta_i = 0$, so $\delta_i = 0$.

Thus we have established the following proposition.

Proposition 3.3. *Let (w_1, b_1) and (w_2, b_2) be two minimal solutions, and let*

$$\begin{pmatrix} \delta \\ \eta \end{pmatrix} = \begin{pmatrix} w_2 - w_1 \\ b_2 - b_1 \end{pmatrix},$$

a vector in the kernel of $\widetilde{X}$. For any i with $1 \leq i \leq n$, the following properties hold:

(1) If $(\alpha_+)_i = \tau$, then $\delta_i \geq -(w_1)_i$ and $(w_1)_i \geq 0$.

(2) If $(\alpha_+)_i = 0$, then $\delta_i \leq -(w_1)_i$ and $(w_1)_i \leq 0$.

(3) If $0 < (\alpha_+)_i < \tau$, then $(w_1)_i = 0$ and $\delta_i = 0$.

Let us now assume that $\delta = w_2 - w_1 \neq 0$. It will be convenient to define the subsets $K_0, K_+, K_-, K_+(\delta)$ and $K_-(\delta)$ of $\{1, \ldots, n\}$ as follows:

$$\begin{aligned}
K_0 &= \{i \in \{1, \ldots, n\} \mid 0 < (\alpha_+)_i < \tau\} \\
K_+ &= \{i \in \{1, \ldots, n\} \mid (\alpha_+)_i = \tau\} \\
K_- &= \{i \in \{1, \ldots, n\} \mid (\alpha_+)_i = 0\} \\
K_+(\delta) &= \{i \in \{1, \ldots, n\} \mid \delta_i \neq 0,\ (\alpha_+)_i = \tau\} \\
K_-(\delta) &= \{i \in \{1, \ldots, n\} \mid \delta_i \neq 0,\ (\alpha_+)_i = 0\}.
\end{aligned}$$

Note that $K_+(\delta) \subseteq K_+$ and $K_-(\delta) \subseteq K_-$. Since we are assuming that $\delta \neq 0$, we must have $K_+(\delta) \cup K_-(\delta) \neq \emptyset$. By definition, $\delta_i = 0$ if $i \in K_+ - K_+(\delta)$ or if $i \in K_- - K_-(\delta)$.

Since $(w_1)_i = (w_2)_i$ iff $\delta_i = 0$ and $\delta_i = 0$ iff $i \in K_0 \cup (K_+ - K_+(\delta)) \cup (K_- - K_-(\delta))$, the above analysis shows that

$$\begin{aligned}
\|w_1\|_1 &= \sum_{i \in K_0} |(w_1)_i| + \sum_{i \in K_+ - K_+(\delta)} |(w_1)_i| + \sum_{i \in K_- - K_-(\delta)} |(w_1)_i| + \sum_{i \in K_+(\delta)} (w_1)_i + \sum_{j \in K_-(\delta)} -(w_1)_j \\
\|w_2\|_1 &= \sum_{i \in K_0} |(w_1)_i| + \sum_{i \in K_+ - K_+(\delta)} |(w_1)_i| + \sum_{i \in K_- - K_-(\delta)} |(w_1)_i| + \sum_{i \in K_+(\delta)} ((w_1)_i + \delta_i) \\
&\quad + \sum_{j \in K_-(\delta)} (-(w_1)_j - \delta_j).
\end{aligned}$$

Since w_1 and w_2 are part of a minimal solution, they have the same 1-norm $\|w_1\|_1 = \|w_2\|_1$, so we deduce that

$$\sum_{i \in K_+(\delta)} \delta_i - \sum_{j \in K_-(\delta)} \delta_j = 0. \tag{$\dagger_1$}$$

Since $\delta_k \neq 0$ if $k \in K_+(\delta)$ or $k \in K_-(\delta)$, Equation ($\dagger_1$) implies that $K_+(\delta) \cup K_-(\delta)$ has at least two elements. Also, if $w_1 = 0$, Proposition 3.3 implies that $\delta_i > 0$ for all $i \in K_+(\delta)$ and $\delta_j < 0$ for all $j \in K_-(\delta)$, but then Equation ($\dagger_1$) does not hold. Therefore, we must have $w_1 \neq 0$, and similarly $w_2 \neq 0$ since $\|w_1\|_1 = \|w_2\|_1$.

Since $\begin{pmatrix} \delta \\ \eta \end{pmatrix}$ is in the kernel of $\widetilde{X} = [X \ \ \mathbf{1}_m]$, together with ($\dagger_1$), we conclude that there is a linear combination

$$\sum_{i \in K_+(\delta)} \delta_i X^i + \sum_{j \in K_-(\delta)} \delta_j X^j + \eta \mathbf{1}_m = 0$$

where the X^i and X^j are columns of X, with

$$\sum_{i \in K_+(\delta)} \delta_i - \sum_{j \in K_-(\delta)} \delta_j = 0.$$

Observe that if $\mathbf{1}_m$ is not a linear combination of the columns of X, then we must have $\eta = 0$. Actually, a condition weaker than linear independence implies that $\eta = 0$, but we postpone a discussion of this condition until later.

In summary we proved the following fact.

Proposition 3.4. *Let (w_1, b_1) and (w_2, b_2) be two distinct minimal solutions, and let*

$$\begin{pmatrix} \delta \\ \eta \end{pmatrix} = \begin{pmatrix} w_2 - w_1 \\ b_2 - b_1 \end{pmatrix},$$

a nonzero vector in the kernel of $\widetilde{X}$. If $\mathbf{1}_m$ is not a linear combination of the columns of X, then $\eta = 0$. We have $\delta = w_2 - w_1 \neq 0$, the set of indices $K_+(\delta) \cup K_-(\delta)$ has at least two elements, and $w_1, w_2 \neq 0$. For each $k \in K_+(\delta) \cup K_-(\delta)$, let

$$s_k = \begin{cases} 0 & \text{if } k \in K_+(\delta) \\ +1 & \text{if } k \in K_-(\delta). \end{cases}$$

Then we have

$$\sum_{k \in K_+(\delta) \cup K_-(\delta)} \delta_k X^k + \eta \mathbf{1}_m = 0, \quad \sum_{k \in K_+(\delta) \cup K_-(\delta)} (-1)^{s_k} \delta_k = 0, \quad \text{and } \delta_k \neq 0 \text{ for all } k.$$

Since $\delta_i = 0$ if $i \in K_+ - K_+(\delta)$ and $\delta_i = 0$ if $i \in K_- - K_-(\delta)$, we can extend the definition of the s_k to $K_+ \cup K_-$ by

$$s_k = \begin{cases} 0 & \text{if } k \in K_+ \\ +1 & \text{if } k \in K_-, \end{cases}$$

and Proposition 3.4 implies that

$$\sum_{k \in K_+ \cup K_-} \delta_k X^k + \eta \mathbf{1}_m = 0, \quad \sum_{k \in K_+ \cup K_-} (-1)^{s_k} \delta_k = 0, \quad \text{and } \delta_k \neq 0 \text{ for some } k,$$

in fact for at least two $k \in K_+ \cup K_-$. So we see that for any two distinct minimal solutions w_1, w_2, the vector $\delta = w_2 - w_1$ has the property that the subvector consisting of the components of index $i \notin K_0$ belongs to the hyperplane of equation

$$\sum_{k \in K_+ \cup K_-} (-1)^{s_k} \delta_k = 0.$$

If we let $d = n - |K_0| = |K_+ \cup K_-|$, then this is a hyperplane in $\mathbb{R}^d$ that we denote by H_{K_+,K_-}. We will use this fact later to characterize the set of minimal solutions.

It is illuminating to figure out what the space of solutions looks like when multiple solutions exist. We begin with the case of an $m \times 2$ matrix $X \neq 0$. We also assume that $\mathbf{1}_m$ is not a linear combination of X^1 and X^2, which implies vectors in the kernel of $\widetilde{X}$ are of the form $(\delta_1, \delta_2, 0)$.

Example 3.1. Suppose that (w_1, b_1) is a minimal solution. If $(\delta_1, \delta_2, 0)$ is some vector in the kernel of $\widetilde{X}$, then (δ_1, δ_2) in the kernel of X. By Proposition 3.4, $|K_+ \cup K_-| \geq 2$ and $w_1 \neq 0$.

(1) $(\alpha_+)_1 = \tau$, $(\alpha_+)_2 = \tau$. This means that $K_+ = \{1, 2\}$ and $K_- = \emptyset$. Then we must have $\delta_1 + \delta_2 = 0$, so that the vectors in the kernel are of the form $(\delta_1, -\delta_1)$. We have

$$\delta_1 X^1 + (-\delta_1) X^2 = 0,$$

and for $\delta_1 \neq 0$ we get $X^2 = X^1$. The classification implies that

$$(w_1)_1 \geq 0,\ (w_1)_2 \geq 0,\ \delta_1 \geq -(w_1)_1,\ \delta_2 \geq -(w_1)_2,$$

and so

$$\begin{aligned}
&\delta_1 + \delta_2 = 0, \quad \delta_1 \geq -(w_1)_1, \quad \delta_2 \geq -(w_1)_2, \\
&(w_1)_1 \geq 0,\ (w_1)_2 \geq 0,\ \delta_1 \neq 0,\ \delta_2 \neq 0.
\end{aligned}$$

Since $(w_1)_1 = (w_1)_2 = 0$ is impossible, we conclude that (δ_1, δ_2) belongs to the line segment on the line $\delta_1 + \delta_2 = 0$ delimited by the half space $\delta_1 \geq -(w_1)_1$ and $\delta_2 \geq -(w_1)_2$. The space of minimal solutions is obtained by translation by w_1, so we obtain the line segment on the line

$$\delta_1 + \delta_2 = \mu$$

and in the orthant defined by $\delta_1, \delta_2 \geq 0$, where $\mu = (w_1) + (w_1)_2 = \|w_1\|_1$ is the minimum of the 1-norm for all minimal solutions. This is a 1-simplex.

(2) $(\alpha_+)_1 = 0$, $(\alpha_+)_2 = \tau$. This means that $K_+ = \{2\}$ and $K_- = \{1\}$. Then we must have $-\delta_1 + \delta_2 = 0$, so that the vectors in the kernel are of the form (δ_1, δ_1). We have

$$(-\delta_1)(-X^1) + \delta_1 X^2 = 0,$$

so for $\delta_1 \neq 0$ we get $X^2 = -X^1$. The classification implies that

$$(w_1)_1 \leq 0,\ (w_1)_2 \geq 0,\ \delta_1 \leq -(w_1)_1,\ \delta_2 \geq -(w_1)_2,$$

and so

$$\begin{aligned}
&-\delta_1 + \delta_2 = 0, \quad \delta_1 \leq -(w_1)_1 \quad \delta_2 \geq -(w_1)_2, \\
&(w_1)_1 \leq 0, (w_1)_2 \geq 0,\ \delta_1 \neq 0,\ \delta_2 \neq 0.
\end{aligned}$$

Since $(w_1)_1 = (w_1)_2 = 0$ is impossible, we conclude that (δ_1, δ_2) belongs to the line segment on the line $-\delta_1 + \delta_2 = 0$ delimited by the half space $\delta_1 \leq -(w_1)_1$ and $\delta_2 \geq -(w_1)_2$. The space of minimal solutions is obtained by translation by w_1, so we obtain the line segment on the line

$$-\delta_1 + \delta_2 = \mu$$

and in the orthant defined by $\delta_1 \leq 0$ and $\delta_2 \geq 0$, where $\mu = -(w_1) + (w_1)_2 = \|w_1\|_1$ is the minimum of the 1-norm for all minimal solutions. This is a 1-simplex.

(3) $(\alpha_+)_1 = \tau$, $(\alpha_+)_2 = 0$. This means that $K_+ = \{1\}$ and $K_- = \{2\}$. Then we must have $\delta_1 - \delta_2 = 0$, so that the vectors in the kernel are of the form (δ_1, δ_1). We have

$$\delta_1 X^1 + (-\delta_1)(-X^2) = 0,$$

so for $\delta_1 \neq 0$ we get $X^2 = -X^1$. The classification implies that

$$(w_1)_1 \geq 0,\ (w_1)_2 \leq 0,\ \delta_1 \geq -(w_1)_1,\ \delta_2 \leq -(w_1)_2,$$

and so

$$\begin{aligned} &\delta_1 - \delta_2 = 0, \quad \delta_1 \geq -(w_1)_1, \quad \delta_2 \leq -(w_1)_2, \\ &(w_1)_1 \geq 0,\ (w_1)_2 \leq 0,\ \delta_1 \neq 0,\ \delta_2 \neq 0. \end{aligned}$$

Since $(w_1)_1 = (w_1)_2 = 0$ is impossible, we conclude that (δ_1, δ_2) belongs to the line segment on the line $\delta_1 - \delta_2 = 0$ delimited by the half space $\delta_1 \geq -(w_1)_1$ and $\delta_2 \leq -(w_1)_2$. The space of minimal solutions is obtained by translation by w_1, so we obtain the line segment on the line

$$\delta_1 - \delta_2 = \mu$$

and in the orthant defined by $\delta_1 \geq 0$ and $\delta_2 \leq 0$, where $\mu = (w_1) - (w_1)_2 = \|w_1\|_1$ is the minimum of the 1-norm for all minimal solutions. This is a 1-simplex.

(4) $(\alpha_+)_1 = 0$, $(\alpha_+)_2 = 0$. This means that $K_+ = \emptyset$ and $K_- = \{1, 2\}$. Then we must have $-\delta_1 - \delta_2 = 0$, so that the vectors in the kernel are of the form $(\delta_1, -\delta_1)$. We have

$$(-\delta_1)(-X^1) + \delta_1(-X^2) = 0,$$

so for $\delta_1 \neq 0$ we get $X^2 = X^1$. The classification implies that

$$(w_1)_1 \leq 0,\ (w_1)_2 \leq 0,\ \delta_1 \leq -(w_1)_1,\ \delta_2 \leq -(w_1)_2,$$

and so

$$\begin{aligned} &-\delta_1 - \delta_2 = 0, \quad \delta_1 \leq -(w_1)_1, \quad \delta_2 \leq (w_1)_2, \\ &(w_1)_1 \leq 0,\ (w_1)_2 \leq 0,\ \delta_1 \neq 0,\ \delta_2 \neq 0. \end{aligned}$$

Since $(w_1)_1 = (w_1)_2 = 0$ is impossible, we conclude that (δ_1, δ_2) belongs to the line segment on the line $\delta_1 + \delta_2 = 0$ delimited by the half space $\delta_1 \leq -(w_1)_1$ and $\delta_2 \leq -(w_1)_2$. The space of minimal solutions is obtained by translation by w_1, so we obtain the line segment on the line

$$-\delta_1 - \delta_2 = \mu$$

and in the orthant defined by $\delta_1 \leq 0$ and $\delta_2 \leq 0$, where $\mu = -(w_1) - (w_1)_2 = \|w_1\|_1$ is the minimum of the 1-norm for all minimal solutions. This is a 1-simplex.

We confirm that either X^1 and X^2 are affinely dependent, or $-X^1$ and X^2 are affinely dependent, or X^1 and $-X^2$ are affinely dependent, or $-X^1$ and $-X^2$ are affinely dependent, which means that either $X^2 = X^1$ or $X^2 = -X^1$. We also discovered that in all cases, the space of minimal solution is a line segment.

Example 3.2. Consider the $m \times 4$ matrix X consisting of the columns

$$X^1,\ X^2,\ \frac{1}{2}(X^1 + X^2),\ \frac{2}{3}X^1 + \frac{1}{3}X^2,$$

where X^1, X^2 and $\mathbf{1}_m$ are linearly independent. Here is an explicit example for $m = 8$ specified in `Matlab` as follows:

```
X = [-10, 11; -6, 5; -2, 4; 0, 0; 1, 2; 2, -5; 6, -4; 10, -6];
y = [0; -2.5; 0.5; -2; 2.5; -4.2; 1; 4];
```

The first column of the matrix X is the column vector $X^1 = [-10; -6; -2; 0; 1; 2; 6; 10]$, and the second column is the column vector $X^2 = [11; 5; 4; 0; 2; -5; -4; -5]$. For this example, by running ADMM with $\tau = 0.1$ we find that the Lagrange multipliers are $(\alpha_+)_1 = \tau$, $(\alpha_+)_2 = \tau$, $(\alpha_+)_3 = \tau$, $(\alpha_+)_4 = \tau$, and we obtain the minimal solution

$$w_1 = \begin{pmatrix} 0.1578 \\ 0.4322 \\ 0.7950 \\ 0.9160 \end{pmatrix}, \quad b_1 = -1.2265.$$

In this case $K_+ = \{1, 2, 3, 4\}$, $K_- = \emptyset$, and the hyperplane H_{K_+,K_-} is given by

$$\delta_1 + \delta_2 + \delta_3 + \delta_4 = 0,$$

For any minimal solution $w_1 = ((w_1)_1, (w_1)_2, (w_1)_3, (w_1)_4)$ we have $(w_1)_i \geq 0$, for $i = 1, 2, 3, 4$, and $\|w_1\|_1 = (w_1)_1 + (w_1)_2 + (w_1)_3 + (w_1)_4 = \mu > 0$, which is the minimal value of the 1-norm of all minimal solutions. A vector $(\delta_1, \delta_2, \delta_3, \delta_4)$ is in the kernel of X and satisfies the conclusion of Proposition 3.4 if

$$\begin{aligned} \delta_1 X^1 + \delta_2 X^2 + \delta_3 \frac{1}{2}(X^1 + X^2) + \delta_4(\frac{2}{3}X^1 + \frac{1}{3}X^2) &= 0 \\ \delta_1 + \delta_2 + \delta_3 + \delta_4 &= 0, \end{aligned}$$

which yields

$$\left(\delta_1 + \frac{1}{2}\delta_3 + \frac{2}{3}\delta_4\right) X^1 + \left(\delta_2 + \frac{1}{2}\delta_3 + \frac{1}{3}\delta_4\right) X^2 = 0$$
$$\delta_1 + \delta_2 + \delta_3 + \delta_4 = 0,$$

and since X^1 and X^2 are linearly independent, we obtain the linear system

$$\begin{aligned} \delta_1 + \frac{1}{2}\delta_3 + \frac{2}{3}\delta_4 &= 0 \\ \delta_2 + \frac{1}{2}\delta_3 + \frac{1}{3}\delta_4 &= 0 \\ \delta_1 + \delta_2 + \delta_3 + \delta_4 &= 0. \end{aligned}$$

The third equation is the sum of the first two, so for this example the kernel of X is contained in the hyperplane of equation

$$\delta_1 + \delta_2 + \delta_3 + \delta_4 = 0.$$

We see that the space of solutions is a two-dimensional subspace given by

$$\begin{aligned} \delta_1 &= -\frac{1}{2}\delta_3 - \frac{2}{3}\delta_4 \\ \delta_2 &= -\frac{1}{2}\delta_3 - \frac{1}{3}\delta_4, \end{aligned}$$

and where δ_3, δ_4 are arbitrary. For $\delta_3 = \delta_4 = -6$, we get the vector

$$(7,\ 5\ -6,\ -6),$$

and for $\delta_3 = 6, \delta_4 = -6$, we get the vector

$$(1,\ -1,\ 6,\ -6).$$

The above vectors form a basis of the kernel of X. According to Proposition 3.4, for any minimal solution w_1 we must have

$$\delta_1 \geq -(w_1)_1,\ \delta_2 \geq -(w_1)_2,\ \delta_3 \geq -(w_1)_3,\ \delta_4 \geq -(w_1)_4.$$

These inequalities define the polyhedral cone obtained by translating by $-w_1$ the positive orthant defined by $\delta_1, \delta_2, \delta_3, \delta_4 \geq 0$. So for any $\delta = w_2 - w_1$, the difference of two minimal solutions w_1 and w_2, we see that δ belongs to the intersection of

(1) The kernel $\mathrm{Ker}\, X$, a subspace of dimension 2.

(2) The hyperplane H_{K_+, K_-}, a subspace of dimension 3.

(3) The translate of the positive orthant by $-w_1$.

But then $w_2 = w_1 + \delta$ belongs to the intersection of

(1) The affine subspace $w_1 + \operatorname{Ker} X$, a subspace of dimension 2.

(2) The affine hyperplane $w_1 + H_{K_+,K_-}$, a subspace of dimension 3.

(3) The positive orthant.

We noted earlier that the affine subspace $w_1 + \operatorname{Ker} X$ does not depend on the choice of a minimal solution w_1. Also, since the hyperplane H_{K_+,K_-} is given by the equation

$$\delta_1 + \delta_2 + \delta_3 + \delta_4 = 0,$$

it is immediate to see that the affine hyperplane $w_1 + H_{K_+,K_-}$ is given by the equation

$$\delta_1 - (w_1)_1 + \delta_2 - (w_1)_2 + \delta_3 - (w_1)_3 + \delta_4 - (w_1)_3 = 0,$$

that is,

$$\delta_1 + \delta_2 + \delta_3 + \delta_4 = \mu,$$

since $\mu = \|w_1\|_1 = (w_1)_1 + (w_1)_2 + (w_1)_3 + (w_1)_4$. Thus this affine hyperplane $w_1 + H_{K_+,K_-}$ is *the same* for all minimal solutions. Let us denote this affine hyperplane by $H_{K_+,K_-,\mu}$.

Then the space of minimal solutions is contained in the intersection of

(1) The affine subspace $w_1 + \operatorname{Ker} X$, a subspace of dimension 2.

(2) The affine hyperplane $H_{K_+,K_-,\mu}$ of equation

$$\delta_1 + \delta_2 + \delta_3 + \delta_4 = \mu.$$

(3) The positive orthant.

But the intersection of the convex spaces in (2) and (3) is the 3-simplex whose vertices are the points $(\mu, 0, 0, 0), (0, \mu, 0, 0), (0, 0, \mu, 0), (0, 0, 0, \mu)$, and so the space of minimal solutions is contained in the intersection of this simplex with the affine plane $w_1 + \operatorname{Ker} X$, resulting in a polytope, in this case a convex poylygon since $w_1 + \operatorname{Ker} X$ is a subspace of $H_{K_+,K_-,\mu}$. It is also immediately verified that any vector w in this polygon is in fact a minimal solution. We will come back to this point in Proposition 5.1.

In order to find other minimal solutions we compute a basis of the kernel of X. Altough we already did this earlier, we describe a systematic approach to do so. Since $\mathbf{1}_4$ is not a linear combination of the first two columns of X, the vectors in $\operatorname{Ker} \widetilde{X}$ are of the form $(\delta, 0)$ where $\delta \in \operatorname{Ker} X$, and since $\widetilde{X}$ and $B = (\widetilde{X})^\top \widetilde{X}$ have the same kernel but B is symmetric, we computed an SVD $U\Sigma U^\top$ of B. Since B is a 5×5 matrix and $\widetilde{X}$ (and thus B) has rank 3, the last two columns of U form a basis of $\operatorname{Ker} B$. Our program confirms that these vectors belong to the hyperplane $\delta_1 + \delta_2 + \delta_3 + \delta_4 = 0$. In this particular case, if u_2 is the last column vector in U (with its fifth coordinate, which is 0, dropped) because the signs of the components of $w_2 = w_1 + u_2$ are the same as the signs of the components of w_1, the vector $w_2 = w_1 + u_2$ is also a minimal solution, with

$$w_2 = \begin{pmatrix} 0.2016 \\ 0.2653 \\ 1.5504 \\ 0.2838 \end{pmatrix}, \quad b_2 = b_1 = -1.2265.$$

Example 3.3. Consider the $m \times 4$ matrix X consisting of the columns

$$X^1,\ X^2,\ 2X^1 - 3X^2,\ -X^1 + 2X^2,$$

where X^1, X^2 and $\mathbf{1}_m$ are linearly independent as above. We also use the data set for $m = 8$ from Example 3.2 specified in `Matlab` as follows:

```
X = [-10, 11; -6, 5; -2, 4; 0, 0; 1, 2; 2, -5; 6, -4; 10, -6];
y = [0; -2.5; 0.5; -2; 2.5; -4.2; 1; 4];
```

For this example, by running ADMM with $\tau = 0.1$ we find that the Lagrange multipliers are $(\alpha_+)_1 = \tau$, $(\alpha_+)_2 = \tau$, $(\alpha_+)_3 = 0$, $(\alpha_+)_4 = \tau$, and we obtain the minimal solution

$$w_1 = \begin{pmatrix} 1.1870 \\ 1.0986 \\ -0.0054 \\ 0.0101 \end{pmatrix}, \quad b_1 = -1.2265.$$

In this case $K_+ = \{1, 2, 4\}$, $K_- = \{3\}$, and the hyperplane H_{K_+,K_-} is given by

$$\delta_1 + \delta_2 - \delta_3 + \delta_4 = 0.$$

For any minimal solution $w_1 = ((w_1)_1, (w_1)_2, (w_1)_3, (w_1)_4)$ we have $(w_1)_i \geq 0$, for $i = 1, 2, 4$, $(w_1)_3 \leq 0$, and $\|w_1\|_1 = (w_1)_1 + (w_1)_2 - (w_1)_3 + (w_1)_4 = \mu > 0$, which is the minimal value of the 1-norm of all minimal solutions. A vector $(\delta_1, \delta_2, \delta_3, \delta_4)$ is in the kernel of X and satisfies the conclusion of Proposition 3.4 if

$$\begin{aligned} \delta_1 X^1 + \delta_2 X^2 + \delta_3(2X^1 - 3X^2) + \delta_4(-X^1 + 2X^2) &= 0 \\ \delta_1 + \delta_2 - \delta_3 + \delta_4 &= 0, \end{aligned}$$

which yields

$$\begin{aligned} (\delta_1 + 2\delta_3 - \delta_4)\, X^1 + (\delta_2 - 3\delta_3 + 2\delta_4)\, X^2 &= 0 \\ \delta_1 + \delta_2 - \delta_3 + \delta_4 &= 0, \end{aligned}$$

and since X^1 and X^2 are linearly independent, we obtain the linear system

$$\begin{aligned}\delta_1 + 2\delta_3 - \delta_4 &= 0\\ \delta_2 - 3\delta_3 + 2\delta_4 &= 0\\ \delta_1 + \delta_2 - \delta_3 + \delta_4 &= 0.\end{aligned}$$

The third equation is the sum of the first two, so for this example the kernel of X is also contained in the hyperplane of equation

$$\delta_1 + \delta_2 - \delta_3 + \delta_4 = 0.$$

We see that the space of solutions is a two-dimensional subspace with

$$\begin{aligned}\delta_1 &= -2\delta_3 + \delta_4\\ \delta_2 &= 3\delta_3 - 2\delta_4,\end{aligned}$$

and where δ_3, δ_4 are arbitrary. For $\delta_3 = \delta_4 = 1$, we get the vector

$$(-1,\ 1,\ 1,\ 1)$$

and for $\delta_3 = 1$ and $\delta_4 = -1$ we get the vector

$$(-3,\ 5,\ 1,\ -1).$$

The above vectors form a basis of the space of solutions.

Next the discussion is very similar to the discussion in Example 3.2, except that we need to consider the hyperplane H_{K_+,K_-} of equation

$$\delta_1 + \delta_2 - \delta_3 + \delta_4 = 0,$$

and a different orthant. Thus we will not provide as much details.

According to Proposition 3.4, for any minimal solution w_1 we must have

$$\delta_1 \geq -(w_1)_1,\ \delta_2 \geq -(w_1)_2,\ \delta_3 \leq -(w_1)_3,\ \delta_4 \geq -(w_1)_4.$$

These inequalities define the polyhedral cone obtained by translating by $-w_1$ the orthant defined by $\delta_1, \delta_2, \delta_4 \geq 0$ and $\delta_3 \leq 0$. Let us denote this orthant as $O_{0,0,1,0}$. So for any $\delta = w_2 - w_1$, the difference of two minimal solutions w_1 and w_2, we see that δ belongs to the intersection of

(1) The kernel $\mathrm{Ker}\, X$, a subspace of dimension 2.

(2) The hyperplane H_{K_+,K_-}, a subspace of dimension 3.

(3) The translate of the orthant $O_{0,0,1,0}$ by $-w_1$.

But then $w_2 = w_1 + \delta$ belong to the intersection of

(1) The affine subspace $w_1 + \mathrm{Ker}\, X$, a subspace of dimension 2.

(2) The affine hyperplane $w_1 + H_{K_+,K_-}$, a subspace of dimension 3.

(3) The orthant $O_{0,0,1,0}$.

Since the hyperplane H_{K_+,K_-} is given by the equation

$$\delta_1 + \delta_2 - \delta_3 + \delta_4 = 0,$$

it is immediate to see that the affine hyperplane $w_1 + H_{K_+,K_-}$ is given by the equation

$$\delta_1 + \delta_2 - \delta_3 + \delta_4 = \mu,$$

since $\mu = \|w_1\|_1 = (w_1)_1 + (w_1)_2 - (w_1)_3 + (w_1)_4$. Thus this affine hyperplane $w_1 + H_{K_+,K_-}$ is *the same* for all minimal solutions. Let us denote this affine hyperplane by $H_{K_+,K_-,\mu}$.

Then the space of minimal solutions is contained in the intersection of

(1) The affine subspace $w_1 + \mathrm{Ker}\, X$, a subspace of dimension 2.

(2) The affine hyperplane $H_{H_+,H_-,\mu}$ of equation

$$\delta_1 + \delta_2 - \delta_3 + \delta_4 = \mu.$$

(3) The orthant $O_{0,0,1,0}$.

But the intersection of the convex spaces in (2) and (3) is the 3-simplex whose vertices are the points $(\mu, 0, 0, 0), (0, \mu, 0, 0), (0, 0, -\mu, 0), (0, 0, 0, \mu)$, and so the space of minimal solutions is contained in the intersection of this simplex with the affine plane $w_1 + \mathrm{Ker}\, X$, resulting in a polytope, in this case a convex polygon since $w_1 + \mathrm{Ker}\, X$ is a subspace of $H_{H_+,H_-,\mu}$. It is also immediately verified that any vector w in this polygon is in fact a minimal solution. We will come back to this point in Proposition 5.1.

As in Example 3.3, in order to find other minimal solutions we computed an SVD $U\Sigma U^\top$ of B. Again, B is a 5×5 matrix of rank 3, so the last two columns of U form a basis of $\mathrm{Ker}\, B$. Our program confirms that these vectors belong to the hyperplane $\delta_1 + \delta_2 - \delta_3 + \delta_4 = 0$. If u_2 is the last column vector in U (with its fifth coordinate, which is 0, dropped), it turns out that sign of the fourth component of $w_2 = w_1 + u_2$ is wrong. However, if we shrink u_2 by replacing it by $0.01u_2$, then the signs of the components of $w_2 = w_1 + 0.01u_2$ are the same as the signs of the components of w_1, so the vector $w_2 = w_1 + 0.01u_2$ is also a minimal solution, with

$$w_2 = \begin{pmatrix} 1.1870 \\ 1.1026 \\ -0.0095 \\ 0.0020 \end{pmatrix}, \quad b_2 = b_1 = -1.2265.$$

Example 3.4. Consider the $m \times 4$ matrix X consisting of the columns

$$X^1,\ X^2,\ 2X^1 - 3X^2,\ X^1 + 2X^2,$$

where X^1, X^2 and $\mathbf{1}_m$ are linearly independent as above. We also use the data set for $m = 8$ from Example 3.2 specified in Matlab as follows:

```
X = [-10, 11; -6, 5; -2, 4; 0, 0; 1, 2; 2, -5; 6, -4; 10, -6];
y = [0; -2.5; 0.5; -2; 2.5; -4.2; 1; 4];
```

For this example, by running ADMM we find that the Lagrange multipliers are $0 < (\alpha_+)_1 < \tau$, $0 < (\alpha_+)_2 < \tau$, $(\alpha_+)_3 = \tau$, $(\alpha_+)_4 = \tau$, and we obtain the minimal solution

$$w_1 = \begin{pmatrix} 0 \\ 0 \\ 0.1714 \\ 0.8264 \end{pmatrix}, \quad b_1 = -1.23.$$

In this case $K_+ = \{3, 4\}$, $K_- = \emptyset$, and the hyperplane H_{K_+,K_-} is given by

$$\delta_3 + \delta_4 = 0.$$

Every minimal solution must be of the form $(0, 0, (w_1)_3, (w_1)_4)$. A vector $(0, 0, \delta_3, \delta_4)$ is in the kernel of X and satisfies the conclusion of Proposition 3.4 if

$$\begin{aligned} \delta_3(2X^1 - 3X^2) + \delta_4(X^1 + 2X^2) &= 0 \\ \delta_3 + \delta_4 &= 0, \\ \delta_1 = 0,\ \delta_2 &= 0, \end{aligned}$$

which yields

$$\begin{aligned} (2\delta_3 + \delta_4)\, X^1 + (-3\delta_3 + 2\delta_4)\, X^2 &= 0 \\ \delta_3 + \delta_4 &= 0, \end{aligned}$$

and since X^1 and X^2 are linearly independent, we obtain the linear system

$$\begin{aligned} 2\delta_3 + \delta_4 &= 0 \\ -3\delta_3 + 2\delta_4 &= 0 \\ \delta_3 + \delta_4 &= 0. \end{aligned}$$

The first two equations have the unique solution $\delta_3 = \delta_4 = 0$, since from the first equation $\delta_4 = -2\delta_3$, and from the second equation $-3\delta_3 - 4\delta_3 = -7\delta_3 = 0$. This time the only solution is the trivial solution $(0, 0, 0, 0)$, so there is a unique minimal solution.

Observe that the vector $(-3, 1, 1, 1)$ is in the kernel of X and for this vector $\delta_1 + \delta_2 + \delta_3 + \delta_4 = 0$, but this vector has all its components nonzero so it does not satisfy the conditions imposed by the Lagrange multipliers.

Example 3.5. Consider the $m \times 3$ matrix X consisting of the columns

$$X^1,\ X^2,\ \beta X^1 + \gamma X^2,$$

where X^1, X^2 and $\mathbf{1}_m$ are linearly independent. In this case the vectors in the kernel of $\widetilde{X}$ are of the form $(\delta_1, \delta_2, \delta_3, 0)$, with

$$\delta_1 X^1 + \delta_2 X^2 + \delta_3(\beta X^1 + \gamma X^2) = 0,$$

that is,

$$(\delta_1 + \beta\delta_3) X^1 + (\delta_2 + \gamma\delta_3) X^2 = 0,$$

and since X^1 and X^2 are linearly independent, this is equivalent to

$$\begin{aligned} \delta_1 + \beta\delta_3 &= 0 \\ \delta_2 + \gamma\delta_3 &= 0. \end{aligned}$$

We deduce that we must have $\delta_3 \neq 0$ for all nonzero vectors in $\operatorname{Ker} X$. Thus if there are multiple solutions the case $K_0 = \{3\}$ is ruled out. Since we must have $|K_+ \cup K_-| \geq 2$ in order to have multiple solutions, the other two possibilies are $K_0 = \{1\}$ and $K_0 = \{2\}$.

If $K_0 = \{1\}$, then $\delta_1 = 0$, in which case, since $\delta_3 \neq 0$ if there are multiple solutions, $\beta\delta_3 = 0$ implies that $\beta = 0$. There are four cases for K_+ and K_-, and since $\delta_2 = -\gamma\delta_3$, one of the following four possibilities must hold:

$$\begin{aligned} \delta_2 + \delta_3 &= (-\gamma + 1)\delta_3 = 0 \\ \delta_2 - \delta_3 &= (-\gamma - 1)\delta_3 = 0 \\ -\delta_2 + \delta_3 &= (\gamma + 1)\delta_3 = 0 \\ -\delta_2 - \delta_3 &= (\gamma - 1)\delta_3 = 0, \end{aligned}$$

and since $\delta_3 \neq 0$, either $\gamma = 1$ or $\gamma = -1$. Thus the columns of X are X^1, X^2, X^2 or $X^1, X^2, -X^2$. Since $\delta_1 = 0$, we are back to Example 3.1, with X^2, X^2 or $X^2, -X^2$.

If $K_0 = \{2\}$, then $\delta_2 = 0$, in which case, since $\delta_3 \neq 0$ if there are multiple solutions, then $\gamma\delta_3 = 0$ implies $\gamma = 0$. There are four cases for K_+ and K_-, and since $\delta_1 = -\beta\delta_3$, one of the following four possibilities must hold:

$$\begin{aligned} \delta_1 + \delta_3 &= (-\beta + 1)\delta_3 = 0 \\ \delta_1 - \delta_3 &= (-\beta - 1)\delta_3 = 0 \\ -\delta_1 + \delta_3 &= (\beta + 1)\delta_3 = 0 \\ -\delta_1 - \delta_3 &= (\beta - 1)\delta_3 = 0, \end{aligned}$$

and since $\delta_3 \neq 0$, either $\beta = 1$ or $\beta = -1$. Thus the columns of X are X^1, X^2, X^1 or $X^1, X^2, -X^1$. Since $\delta_2 = 0$, we are back to Example 3.1, with X^1, X^1 or $X^1, -X^1$.

If $K_0 = \emptyset$, then there are eight cases depending of K_+ and K_-, but actually due to symmetries we only need to consider four of them, namely the four cases where $(\alpha_+)_1 = \tau$.

(1) $K_+ = \{1, 2, 3\}$, $K_- = \emptyset$, which means that $(\alpha_+)_1 = (\alpha_+)_2 = (\alpha_+)_2 = \tau$. Then $\delta_1, \delta_2, \delta_3$ must satisfy the equation

$$\delta_1 + \delta_2 + \delta_3 = 0,$$

and since $\delta_1 = -\beta\delta_3$ and $\delta_2 = -\gamma\delta_3$, we must have

$$(-\beta - \gamma + 1)\delta_3 = 0.$$

Since $\delta_3 \neq 0$ if there are multiple solutions, we must have

$$\beta + \gamma = 1.$$

In this case $\gamma = 1 - \beta$ and the kernel of X is a subspace of the plane given by

$$\delta_1 + \delta_2 + \delta_3 = 0,$$

which consists of the vectors of the form $(-\beta\delta_3, -(1-\beta)\delta_3, \delta_3)$. Geometrically, these vectors belong to the line which is the common intersection of the planes given by $\delta_1 + \beta\delta_3 = 0$, $\delta_2 + (1-\beta)\delta_3 = 0$, and $\delta_1 + \delta_2 + \delta_3 = 0$. If $\beta = 0$, then $\delta_1 = 0$, and if $\beta = 1$, we have $\delta_2 = 0$, and then we are back to the previous cases.

For any minimal solution $w_1 = ((w_1)_1, (w_1)_2, (w_1)_3)$ we have $(w_1)_i \geq 0$, for $i = 1, 2, 3$, and $\|w_1\|_1 = (w_1)_1 + (w_1)_2 + (w_1)_3 = \mu > 0$, which is the minimal value of the 1-norm of all minimal solutions. Since we already explained in great detail how to derive the set of minimal solutions, we simply describe the solution, leaving the details to the reader.

The set of minimal solutions is the line segment which is the intersection of the 2-simplex whose vertices are $(\mu, 0, 0), (0, \mu, 0), (0, 0, \mu)$ with both planes obtained by translating by w_1 the planes $\delta_1 + \beta\delta_3 = 0$ and $\delta_2 + (1-\beta)\delta_3 = 0$, which are the planes given by $\delta_1 + \beta\delta_3 - (w_1)_1 - \beta(w_1)_3 = 0$ and $\delta_2 + (1-\beta)\delta_3 - (w_1)_2 - (1-\beta)(w_1)_3 = 0$.

A concrete illustration is provided by the data set for $m = 8$ from Example 3.2 specified in `Matlab` as follows:

```
X = [-10, 11; -6, 5; -2, 4; 0, 0; 1, 2; 2, -5; 6, -4; 10, -6];
y = [0; -2.5; 0.5; -2; 2.5; -4.2; 1; 4];
```

and for $\beta = 2$, $\gamma = -1$. For this example, by running ADMM we obtain the minimal solution

$$w_1 = \begin{pmatrix} 0.1004 \\ 1.6679 \\ 0.5328 \end{pmatrix}, \quad b_1 = -1.2265.$$

As in Examples 3.3 and 3.4, in order to find other minimal solutions we compute an SVD $U\Sigma U^\top$ of B. Here B is a 4×4 matrix of rank 3. Our program confirms that the last vector in U belong to the hyperplane $\delta_1 + \delta_2 + \delta_3 = 0$. If u_1 is the last vector in U (with its fourth coordinate, which is 0, dropped), because the signs of the components of $w_2 = w_1 + u_1$ are the same as the signs of the components of w_1, the vector $w_2 = w_1 + u_1$ is also a minimal solution, with

$$w_2 = \begin{pmatrix} 0.9169 \\ 1.2597 \\ 0.1246 \end{pmatrix}, \quad b_2 = b_1 = -1.2265.$$

(2) $K_+ = \{1, 3\}$, $K_- = \{2\}$, which means that $(\alpha_+)_1 = (\alpha_+)_3 = \tau$ and $(\alpha_+)_2 = 0$. Then $\delta_1, \delta_2, \delta_3$ must satisfy the equation

$$\delta_1 - \delta_2 + \delta_3 = 0,$$

and since $\delta_1 = -\beta\delta_3$ and $\delta_2 = -\gamma\delta_3$, we must have

$$(-\beta + \gamma + 1)\delta_3 = 0.$$

Since $\delta_3 \neq 0$ if there are multiple solutions, we must have

$$\beta - \gamma = 1.$$

In this case $\gamma = \beta - 1$ and the kernel of X is a subspace of the plane given by

$$\delta_1 - \delta_2 + \delta_3 = 0,$$

which consists of the vectors of the form $(-\beta\delta_3, -(\beta - 1)\delta_3, \delta_3)$. Geometrically, these vectors belong to the line which is the common intersection of the planes given by $\delta_1 + \beta\delta_3 = 0$, $\delta_2 + (\beta - 1)\delta_3 = 0$, and $\delta_1 - \delta_2 + \delta_3 = 0$. If $\beta = 0$, then $\delta_1 = 0$, and if $\beta = 1$, we have $\delta_2 = 0$, and then we are back to the previous cases.

For any minimal solution $w_1 = ((w_1)_1, (w_1)_2, (w_1)_3)$ we have $(w_1)_i \geq 0$, for $i = 1, 3$, $(w_1)_2 \leq 0$, $\|w_1\|_1 = (w_1)_1 - (w_1)_2 + (w_1)_3 = \mu > 0$, which is the minimal value of the 1-norm of all minimal solutions.

The set of minimal solutions is the line segment which is the intersection of the 2-simplex whose vertices are $(\mu, 0, 0), (0, -\mu, 0), (0, 0, \mu)$ with both planes obtained by translating by w_1 the planes $\delta_1 + \beta\delta_3 = 0$ and $\delta_2 + (\beta - 1)\delta_3 = 0$, which are the planes given by $\delta_1 + \beta\delta_3 - (w_1)_1 - \beta(w_1)_3 = 0$ and $\delta_2 + (\beta - 1)\delta_3 - (w_1)_2 - (\beta - 1)(w_1)_3 = 0$.

A concrete illustration is provided by the data set for $m = 8$ specified in `Matlab` as follows:

```
Xb = [-8, 11.1; 6.1, 5; 2.2, 4; 0, 0; 1, 1.9; 1, -5; 6.2, -4; 5.1, -6];
y = [0; -2.5; 0.5; -2; 2.5; -4.2; 1; 4];
```

and for $\beta = 3$, $\gamma = 2$.

For this example, by running ADMM we obtain the minimal solution

$$w_1 = \begin{pmatrix} 0.0530 \\ -0.0258 \\ 0.0040 \end{pmatrix}, \quad b_1 = -0.1824.$$

As in previous examples, in order to find other minimal solutions we compute an SVD $U\Sigma U^\top$ of B. Here B is a 4×4 matrix of rank 3. Our program confirms that the last vector in U belong to the hyperplane $\delta_1 - \delta_2 + \delta_3 = 0$. If u_1 is the last column vector in U (with its fourth coordinate, which is 0, dropped), it turns out that sign of the third component of $w_2 = w_1 + u_1$ is wrong. However, if we shrink u_1 by replacing it by $0.01u_1$, then the signs of the components of $w_2 = w_1 + 0.01u_1$ are the same as the signs of the components of w_1, so the vector $w_2 = w_1 + 0.01u_1$ is also a minimal solution, with

$$w_2 = \begin{pmatrix} 0.0449 \\ -0.0311 \\ 0.0067 \end{pmatrix}, \quad b_2 = b_1 = -0.1824.$$

(3) $K_+ = \{1, 2\}$, $K_- = \{3\}$, which means that $(\alpha_+)_1 = (\alpha_+)_2 = \tau$ and $(\alpha_+)_3 = 0$. Then $\delta_1, \delta_2, \delta_3$ must satisfy the equation

$$\delta_1 + \delta_2 - \delta_3 = 0,$$

and since $\delta_1 = -\beta\delta_3$ and $\delta_2 = -\gamma\delta_3$, we must have

$$(-\beta - \gamma - 1)\delta_3 = 0.$$

Since $\delta_3 \neq 0$ if there are multiple solutions, we must have

$$\beta + \gamma = -1.$$

In this case $\gamma = -(\beta + 1)$ and the kernel of X is a subspace of the plane given by

$$\delta_1 + \delta_2 - \delta_3 = 0,$$

which consists of the vectors of the form $(-\beta\delta_3, (\beta + 1)\delta_3, \delta_3)$. Geometrically, these vectors belong to the line which is the common intersection of the planes given by $\delta_1 + \beta\delta_3 = 0$, $\delta_2 - (\beta + 1)\delta_3 = 0$, and $\delta_1 + \delta_2 - \delta_3 = 0$. If $\beta = 0$, then $\delta_1 = 0$, and if $\beta = -1$, we have $\delta_2 = 0$, and then we are back to the previous cases.

For any minimal solution $w_1 = ((w_1)_1, (w_1)_2, (w_1)_3)$ we have $(w_1)_i \geq 0$, for $i = 1, 2$, $(w_1)_3 \leq 0$, $\|w_1\|_1 = (w_1)_1 + (w_1)_2 - (w_1)_3 = \mu > 0$, which is the minimal value of the 1-norm of all minimal solutions.

The set of minimal solutions is the line segment which is the intersection of the 2-simplex whose vertices are $(\mu, 0, 0), (0, \mu, 0), (0, 0, -\mu)$ with both planes obtained by translating by w_1 the planes $\delta_1 + \beta\delta_3 = 0$ and $\delta_2 - (\beta + 1)\delta_3 = 0$, which are the planes given by $\delta_1 + \beta\delta_3 - (w_1)_1 - \beta(w_1)_3 = 0$ and $\delta_2 - (\beta+1)\delta_3 - (w_1)_2 + (\beta+1)(w_1)_3 = 0$.

A concrete illustration is also provided by the data set for $m = 8$ from Example 3.2 specified in `Matlab` as follows:

```
X = [-10, 11; -6, 5; -2, 4; 0, 0; 1, 2; 2, -5; 6, -4; 10, -6];
y = [0; -2.5; 0.5; -2; 2.5; -4.2; 1; 4];
```

and for $\beta = 2$, $\gamma = -3$. For this example, by running ADMM we obtain the minimal solution

$$w_1 = \begin{pmatrix} 1.1742 \\ 1.1228 \\ -0.0041 \end{pmatrix}, \quad b_1 = -1.2265.$$

As in previous examples, in order to find other minimal solutions we compute an SVD $U\Sigma U^\top$ of B. Here B is a 4×4 matrix of rank 3. Our program confirms that the last vector in U belong to the hyperplane $\delta_1 + \delta_2 - \delta_3 = 0$. If u_1 is the last column vector in U (with its fourth coordinate, which is 0, dropped), because the signs of the components of $w_2 = w_1 + u_1$ are the same as the signs of the components of w_1, the vector $w_2 = w_1 + u_1$ is also a minimal solution, with

$$w_2 = \begin{pmatrix} 1.7087 \\ 0.3211 \\ -0.2713 \end{pmatrix}, \quad b_2 = b_1 = -1.2265.$$

(4) $K_+ = \{1\}$, $K_- = \{2, 3\}$, which means that $(\alpha_+)_1 = \tau$, and $(\alpha_+)_2 = (\alpha_+)_3 = 0$. Then $\delta_1, \delta_2, \delta_3$ must satisfy the equation

$$\delta_1 - \delta_2 - \delta_3 = 0,$$

and since $\delta_1 = -\beta\delta_3$ and $\delta_2 = -\gamma\delta_3$, we must have

$$(-\beta + \gamma - 1)\delta_3 = 0.$$

Since $\delta_3 \neq 0$ if there are multiple solutions, we must have

$$-\beta + \gamma = 1.$$

In this case $\gamma = \beta + 1$ and the kernel of X is a subspace of the plane given by

$$\delta_1 - \delta_2 - \delta_3 = 0,$$

which consists of the vectors of the form $(-\beta\delta_3, -(\beta+1)\delta_3, \delta_3)$. Geometrically, these vectors belong to the line which is the common intersection of the planes given by $\delta_1 + \beta\delta_3 = 0$, $\delta_2 + (\beta+1)\delta_3 = 0$, and $\delta_1 - \delta_2 - \delta_3 = 0$. If $\beta = 0$, then $\delta_1 = 0$, and if $\beta = -1$, we have $\delta_2 = 0$, and then we are back to the previous cases.

For any minimal solution $w_1 = ((w_1)_1, (w_1)_2, (w_1)_3)$ we have $(w_1)_1 \geq 0$, $(w_1)_i \leq 0$ for $i = 2, 3$, $\|w_1\|_1 = (w_1)_1 - (w_1)_2 - (w_1)_3 = \mu > 0$, which is the minimal value of the 1-norm of all minimal solutions.

The set of minimal solutions is the line segment which is the intersection of the 2-simplex whose vertices are $(\mu, 0, 0), (0, -\mu, 0), (0, 0, -\mu)$ with both planes obtained by translating by w_1 the planes $\delta_1 + \beta\delta_3 = 0$ and $\delta_2 + (\beta+1)\delta_3 = 0$, which are the planes given by $\delta_1 + \beta\delta_3 - (w_1)_1 - \beta(w_1)_3 = 0$ and $\delta_2 + (\beta+1)\delta_3 - (w_1)_2 - (\beta+1)(w_1)_3 = 0$.

A concrete illustration is provided by the data set for $m = 8$ specified in `Matlab` as follows:

```
Xb2=[-10, 11.1; 6.1, 5; 2.2, 4; 0, 0; 1, 1.9; 1, -5; 6.2, -4; 4.1, -6];
y = [0; -2.5; 0.5; -2; 2.5; -4.2; 1; 4];
```

and for $\beta = 2$, $\gamma = 3$. For this example, by running ADMM we obtain the minimal solution

$$w_1 = \begin{pmatrix} 0.0358 \\ -0.0039 \\ -0.0141 \end{pmatrix}, \quad b_1 = -0.0573.$$

As in previous examples, in order to find other minimal solutions we compute an SVD $U\Sigma U^\top$ of B. Here B is a 4×4 matrix of rank 3. Our program confirms that the last vector in U belong to the hyperplane $\delta_1 - \delta_2 - \delta_3 = 0$. If u_1 is the last column vector in U (with its fourth coordinate, which is 0, dropped), it turns out that signs of the first and third components of $w_2 = w_1 + u_1$ are wrong. However, if we shrink u_1 by replacing it by $0.01u_1$, then the signs of the components of $w_2 = w_1 + 0.01u_1$ are the same as the signs of the components of w_1, so the vector $w_2 = w_1 + 0.01u_1$ is also a minimal solution, with

$$w_2 = \begin{pmatrix} 0.0305 \\ -0.0119 \\ -0.0114 \end{pmatrix}, \quad b_2 = b_1 = -0.0573.$$

The four cases in which $(\alpha_+)_1 = 0$ are symmetric to the previous cases. In these cases, the line segments are on the negative side of the the plane $x = 0$ (which means that $x \leq 0$) instead of the positive side (which means that $x \geq 0$).

4 Some Relevant Concepts of Affine Geometry

Proposition 3.4 suggests that some notions of affine geometry are relevant. In order to make this clear we rewrite the conclusion of Proposition 3.4 by introducing

$$\beta_k = (-1)^{s_k}\delta_k,$$

and then we have

$$\sum_{k\in K_+(\delta)\cup K_-(\delta)} \beta_k\big((-1)^{s_k}X^k\big) + \eta\mathbf{1}_m = 0, \quad \sum_{k\in K_+(\delta)\cup K_-(\delta)} \beta_k = 0, \text{ and } \beta_k \neq 0 \text{ for all } k.$$

The key point is that if $\eta = 0$, then the above conditions are reminiscent of the notion of affine dependence of the family $\big((-1)^{s_k}X^k\big)_{k\in K_+(\delta)\cup K_-(\delta)}$. In fact a stronger notion of affine dependence is needed, as we now explain in detail.

The reader may consult Gallier [4, 3], Berger [1], Boyd and Vandenberghe [2], Rockafellar [8], or Ziegler [9], for the basic notions of affine and convex geometrty involved. Technically an affine space is more than a vector space, since it consists of a transitive and faithful action of a vector space V on a set A of points, but in our situation $V = A$, and the action is just vector addition, so we can simplify our definitions a bit.

Given a family of vectors $(a_1, \ldots, a_n)$, recall that an *affine combination* is a linear combination

$$\lambda_1 a_1 + \cdots + \lambda_n a_n \quad \text{with} \quad \lambda_1 + \cdots + \lambda_n = 1,\ \lambda_i \in \mathbb{R}.$$

A *convex combination* is an affine combination such that the scalars $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ also satisfy the conditions $\lambda_i \geq 0$ for $i = 1, \ldots, n$.

A family of vectors $(a_1, \ldots, a_n)$ is *affinely independent* iff either $n = 1$ (the family consists of a single point) or $n \geq 2$ and the family $(a_2 - a_1, \ldots, a_n - a_1)$ is linearly independent. If $n = 1$, we see that any family (a_1) consisting of a single vector (even the zero vector) is affinely independent. A family $(a_1, \ldots, a_n)$ is *affinely dependent* iff it is not affinely independent. It can be shown that a family of vectors $(a_1, \ldots, a_n)$ is *affinely dependent* iff there is a family of scalars $(\lambda_1, \ldots, \lambda_n)$ such that

$$\begin{aligned}&\lambda_1 a_1 + \cdots + \lambda_n a_n = 0,\ \ \lambda_1 + \cdots + \lambda_n = 0,\\ &\text{and } \lambda_i \neq 0 \text{ for some } i, 1 \leq i \leq n.\end{aligned}$$

In this case we must have $n \geq 2$ and at least two scalars λ_i, λ_j are nonzero. Then we see that a family of vectors $(a_1, \ldots, a_n)$ is *affinely independent* iff for all $(\lambda_1, \ldots, \lambda_n)$, if

$$\begin{aligned}&\lambda_1 a_1 + \cdots + \lambda_n a_n = 0 \quad \text{and} \quad \lambda_1 + \cdots + \lambda_n = 0,\\ &\text{then } \lambda_i = 0 \text{ for all } i, 1 \leq i \leq n.\end{aligned}$$

Observe that affine independence is a weaker notion that linear independence since linear independence must apply to *all* n-tuples $(\lambda_1, \ldots, \lambda_n)$, not only to those for which $\lambda_1 + \cdots +$

$\lambda_n = 0$. For example, three vectors in $\mathbb{R}^2$ are affinely independent if they are the vertices of a nondegenerate triangle (a triangle not collapsed to three collinear points), but any three vectors in $\mathbb{R}^2$ are never linearly independent.

An *affine subspace* A is either the empty set or it is closed under affine combinations. It can be shown that if A is a nonempty affine subspace, then

$$\overrightarrow{A} = \{b - a \mid a, b \in A\}$$

is a vector space called the *direction of* A and that

$$A = a + \overrightarrow{A} = \{a + u \mid u \in \overrightarrow{A}\}$$

for all $a \in A$. The *dimension* of the affine subspace A is the dimension of the vector space $\overrightarrow{A}$.

A *convex set* C in $\mathbb{R}^n$ is either the empty set or it is closed under convex combinations.

Given a nonempty family of vectors S, the *affine subspace spanned by* S is the set $\mathrm{aff}(S)$ of all affine combinations of finite families of vectors $(a_1, \ldots, a_n)$ with $a_i \in S$.

The *convex hull* of the nonempty family S is the set $\mathrm{conv}(S)$ of all convex combinations of finite families $(a_1, \ldots, a_n)$ with $a_i \in S$.

The *dimension* of a convex set C is the dimension of the affine space $\mathrm{aff}(C)$ spanned by C.

Given any $d + 1$ vectors $(a_1, \ldots, a_{d+1})$ in $\mathbb{R}^n$, where $d \leq n$, if $(a_1, \ldots, a_{d+1})$ are affinely independent, then the convex hull $\mathrm{conv}(a_1, \ldots, a_{d+1})$ is called a *d-simplex*. Obviously any two d-simplices can be mapped into one another by a bijective affine map. The *standard d-simplex* Δ_d is the simplex in $\mathbb{R}^{n+1}$ spanned by the canonical basis vectors $e_1, \ldots, e_{d+1}$, with $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$ with 1 in the ith position. Obviously

$$\Delta_d = \{(x_1, \ldots, x_{d+1}) \in \mathbb{R}^{d+1} \mid x_1 + \cdots + x_{d+1} = 1,\ x_i \geq 0,\ i = 1, \ldots, d+1\}.$$

If $\eta = 0$ in Proposition 3.4, then

$$\begin{aligned} &\sum_{k \in K_+(\delta) \cup K_-(\delta)} \beta_k \big((-1)^{s_k} X^k\big) = 0, \qquad \sum_{k \in K_+(\delta) \cup K_-(\delta)} \beta_k = 0, \\ &\text{and } \beta_k \neq 0 \text{ for all } k \in K_+(\delta) \cup K_-(\delta), \end{aligned}$$

so the family $((-1)^{s_k} X^k)_{k \in K_+(\delta) \cup K_-(\delta)}$ is affinely dependent, but in a stronger way. To capture this notion we propose the following definition.

Definition 4.1. A family of vectors $(a_1, \ldots, a_n)$ is *strongly affinely dependent* if there is a family of scalar $(\lambda_1, \ldots, \lambda_n)$ such that

$$\begin{aligned} &\lambda_1 a_1 + \cdots + \lambda_n a_n = 0, \quad \lambda_1 + \cdots + \lambda_n = 0, \\ &\text{and } \lambda_i \neq 0 \text{ for all } i, 1 \leq i \leq n. \end{aligned}$$

A family $(a_1, \ldots, a_n)$ is *weakly affinely independent* iff it is not strongly affinely dependent iff for all $(\lambda_1, \ldots, \lambda_n)$, if

$$\begin{aligned}
&\lambda_1 a_1 + \cdots + \lambda_n a_n = 0 \quad \text{and} \quad \lambda_1 + \cdots + \lambda_n = 0, \\
&\text{then } \lambda_i = 0 \text{ for some } i, 1 \leq i \leq n.
\end{aligned}$$

Clearly strong affine dependence implies affine dependence and affine independence implies weak affine independence.

If the family $(a_1, \ldots, a_n)$ is strongly affinely dependent, since $\lambda_i \neq 0$ for all i and $\lambda_1 + \cdots + \lambda_n = 0$, we have

$$-\lambda_i a_i = \sum_{j \neq i} \lambda_j a_j, \quad -\lambda_i = \sum_{k \neq i} \lambda_k,$$

and so

$$a_i = \sum_{j \neq i} \frac{\lambda_j}{\sum_{k \neq i} \lambda_k} a_j,$$

which means that a_i belongs to the affine subspace spanned by the family $(a_j)_{j \neq i}$, and consequently the affine subspaces spanned by the $n+1$ families $(a_i)_{i=1}^n$ and $(a_j)_{j=1, j \neq i}^n$, for $i = 1, \ldots, n$ are all identical. We will prove the converse shortly using a trick involving roots of polynomials.

Example 4.1. For example if (a_1, a_2, a_3, a_4) are strongly affinely dependent, the affine subspaces spanned by (a_1, a_2, a_3, a_4), (a_2, a_3, a_4), (a_1, a_3, a_4), (a_1, a_2, a_4), and also (a_1, a_2, a_3) are all identical. If the dimension of this subspace if 2, then (a_2, a_3, a_4), (a_1, a_3, a_4), (a_1, a_2, a_4) all span the same (affine) plane, so they are affinely independent.

On the other hand, if (a_1, a_2, a_3) belong to a line (an affine subspace of dimension 1) and a_4 does not belong to this line, then (a_1, a_2, a_3, a_4) are affinely dependent but not strongly affinely dependent, since otherwise a_4 would belong to this line.

Proposition 4.1. *A family of vectors $(a_1, \ldots, a_n)$ is strongly affinely dependent iff the affine subspaces spanned by the $n+1$ families $(a_i)_{i=1}^n$ and $(a_j)_{j=1, j \neq i}^n$, for $i = 1, \ldots, n$ are all identical.*

Proof. We already proved that if the family $(a_1, \ldots, a_n)$ is strongly affinely dependent, then the affine subspaces spanned by the $n+1$ families $(a_i)_{i=1}^n$ and $(a_j)_{j=1, j \neq i}^n$, for $i = 1, \ldots, n$ are all identical.

Conversely assume that the affine subspaces spanned by the $n+1$ families $(a_i)_{i=1}^n$ and $(a_j)_{j=1, j \neq i}^n$, are all identical. Let A be this common subspace and say A has dimension d. Since A is spanned by each of the families $(a_j)_{j=1, j \neq i}^n$ that has $n-1$ vectors, we must have $d \leq n-2$. There is a subsequence of $d+1 \leq n-1$ vectors that are affinely independent and A is spanned by these vectors. Reindexing our vectors if needed we may assume that these vectors are $a_1, \ldots, a_{d+1}$. The remaining $n-(d+1) \geq 1$ vectors a_j are affine combinations of $a_1, \ldots, a_{d+1}$, so there are unique scalars λ_j^k $(1 \leq j \leq d+1, 1 \leq k \leq n-(d+1))$ such that

$$\begin{aligned} \lambda_1^1 a_1 + \cdots + \lambda_j^1 a_j + \cdots + \lambda_{d+1}^1 a_{d+1} &= a_{d+2}, \\ &\vdots \\ \lambda_1^{n-(d+1)} a_1 + \cdots + \lambda_j^{n-(d+1)} a_j + \cdots + \lambda^{n-(d+1)} a_{d+1} &= a_n, \end{aligned}$$

with

$$\lambda_1^k + \cdots + \lambda_j^k + \cdots + \lambda_{d+1}^k = 1, \quad 1 \leq k \leq n-(d+1).$$

Observe for every $j = 1, \ldots, d+1$, there is some k with $1 \leq k \leq n-(d+1)$ such that $\lambda_j^k \neq 0$, in other words, every column of the above linear system has some nonzero entry. Otherwise, for some j such that $1 \leq j \leq d+1$, we would have $\lambda_j^k = 0$ for all k with $1 \leq k \leq n-(d+1)$, and then the space A spanned by the family $(a_j)_{j=1, j\neq i}^n$ would also be spanned by $(a_1, \ldots, a_{j-1}, a_{j+1}, \ldots, a_{d+1})$, and so A would have dimension stricty less than d, a contradiction. If we can find $n-1$ scalars $\mu_j \neq 0$ so that

$$a_n = \mu_1 a_1 + \cdots + \mu_{n-1} a_{n-1}, \quad \mu_1 + \cdots + \mu_{n-1} = 1,$$

we are done because

$$\mu_1 a_1 + \cdots + \mu_{n-1} a_{n-1} - a_n = 0$$

is a strong affine dependence since $\mu_j \neq 0$ for $j = 1, \ldots, n-1$ and $\mu_1 + \cdots + \mu_{n-1} - 1 = 0$.

If $d = n-2$, there is a single equation

$$\lambda_1^1 a_1 + \cdots + \lambda_{n-1}^1 a_{n-1} = a_n$$

with $\lambda_1^1 + \cdots + \lambda_{n-1}^1 = 1$, and by the previous remark we must have $\lambda_j^1 \neq 0$ for $j = 1, \ldots, n-1$. In this case we are done. Let us now asssume that $d \leq n-3$.

Consider the system

$$\begin{aligned} \lambda_1^1 a_1 + \cdots + \lambda_j^1 a_j + \cdots + \lambda_{d+1}^1 a_{d+1} &= a_{d+2} \\ &\vdots \\ \lambda_1^{n-(d+2)} a_1 + \cdots + \lambda_j^{n-(d+2)} a_j + \cdots + \lambda^{n-(d+2)} a_{d+1} &= a_{n-1} \\ \lambda_1^{n-(d+1)} a_1 + \cdots + \lambda_j^{n-(d+1)} a_j + \cdots + \lambda^{n-(d+1)} a_{d+1} &= a_n. \end{aligned}$$

If we multiply the first equation of the system above by $\rho^{n-(d+2)}$ for some $\rho \neq 0$ to be determined later, the second equation by $\rho^{n-(n+3)}$, and so on, with the next to last equation multiplied by ρ, by adding up these $n-(d+1)$ equations we get the equation

$$\nu_1 a_1 + \cdots + \nu_{d+1} a_{d+1} = \rho^{n-(d+2)} a_{d+2} + \cdots + \rho a_{n-1} + a_n,$$

with

$$\nu_j = \rho^{n-(d+2)}\lambda_j^1 + \cdots + \rho\lambda_j^{n-(d+2)} + \lambda_j^{n-(d+1)}, \quad 1 \le j \le d+1.$$

We also have

$$\begin{aligned}
\nu_1 + \cdots + \nu_{d+1} &= \rho^{n-(d+2)}\lambda_1^1 + \cdots + \rho\lambda_1^{n-(d+2)} + \lambda_1^{n-(d+1)} + \cdots \\
&\quad + \rho^{n-(d+2)}\lambda_{d+1}^1 + \cdots + \rho\lambda_{d+1}^{n-(d+2)} + \lambda_{d+1}^{n-(d+1)} \\
&= \rho^{n-(d+2)}(\lambda_1^1 + \cdots + \lambda_{d+1}^1) + \cdots + \rho(\lambda_1^{n-(d+2)} + \cdots + \lambda_{d+1}^{n-(d+2)}) \\
&\quad + \lambda_1^{n-(d+1)} + \cdots + \lambda_{d+1}^{n-(d+1)} \\
&= \rho^{n-(d+2)} + \cdots + \rho + 1,
\end{aligned}$$

since

$$\lambda_1^k + \cdots + \lambda_j^k + \cdots + \lambda_{d+1}^k = 1, \quad 1 \le k \le n-(d+1).$$

Consequently,

$$a_n = \nu_1 a_1 + \cdots + \nu_{d+1}a_{d+1} + (-\rho^{n-(d+2)})a_{d+2} + \cdots + (-\rho)a_{n-1},$$

with

$$\nu_1 + \cdots + \nu_{d+1} - (\rho^{n-(d+2)} + \cdots + \rho) = \rho^{n-(d+2)} + \cdots + \rho + 1 - (\rho^{n-(d+2)} + \cdots + \rho) = 1.$$

We claim that can pick some $\rho \neq 0$ so that

$$\nu_j = \rho^{n-(d+2)}\lambda_j^1 + \cdots + \rho\lambda_j^{n-(d+2)} + \lambda_j^{n-(d+1)} \neq 0, \quad 1 \le j \le d+1.$$

Observe that for $j = 1, \ldots, d+1$, $(\lambda_j^1 \ldots, \lambda_j^{n-(d+2)}, \lambda_j^{n-(d+1)})$ constitute the elements of the jth column of the system

$$\begin{aligned}
\lambda_1^1 a_1 + \cdots + \lambda_j^1 a_j + \cdots + \lambda_{d+1}^1 a_{d+1} &= a_{d+2} \\
&\ \ \vdots \\
\lambda_1^{n-(d+2)} a_1 + \cdots + \lambda_j^{n-(d+2)} a_j + \cdots + \lambda^{n-(d+2)} a_{d+1} &= a_{n-1} \\
\lambda_1^{n-(d+1)} a_1 + \cdots + \lambda_j^{n-(d+1)} a_j + \cdots + \lambda^{n-(d+1)} a_{d+1} &= a_n,
\end{aligned}$$

so by a previous remark, $\lambda_j^k \neq 0$ for some k with $1 \le k \le n-(d+1)$. Therefore, viewing ρ as a variable, the $d+1$ polynomials

$$\rho^{n-(d+2)}\lambda_j^1 + \cdots + \rho\lambda_j^{n-(d+2)} + \lambda_j^{n-(d+1)}$$

are nonzero polynomials of degree at most $n-(d+2)$. Since a real polynomial of degree $n-(d+2)$ has at most $n-(d+1)$ distinct roots in $\mathbb{C}$, we can pick $\rho \neq 0$ (in $\mathbb{R}$) distinct from all these roots and ensure that

$$\rho^{n-(d+2)}\lambda_j^1 + \cdots + \rho\lambda_j^{n-(d+2)} + \lambda_j^{n-(d+1)} \neq 0, \quad 1 \le j \le d+1,$$

which finishes the proof. $\square$

Going back to Proposition 3.4, if $\eta \neq 0$, by dividing by η we see that either $\mathbf{1}_m$ is a linear combination

$$\mathbf{1}_m = \sum_{k \in K_+(\delta) \cup K_-(\delta)} \gamma_k \big((-1)^{s_k} X^k\big)$$

for some $\gamma_k \neq 0$ (in fact, $\gamma_k = -\beta_k/\eta$) such that

$$\sum_{k \in K_+(\delta) \cup K_-(\delta)} \gamma_k = 0,$$

or else $\eta = 0$ and the family $((-1)^{s_k} X^k)_{k \in K_+(\delta) \cup K_-(\delta)}$ is strongly affinely dependent.

Using the above remark and taking the negation of Proposition 3.4, we obtain the following result which gives sufficient conditions for lasso to have a unique minimal solution.

Proposition 4.2. *For every family $((-1)^{s_k} X^k)_{k \in K}$ where $K \subseteq \{1, \ldots, n\}$ and the X^k are columns of X, with $s_k = 0, 1$, for every scalar η, for every sequence $(\beta_k)_{k \in K}$, if $|K| \geq 2$ and*

$$\sum_{k \in K} \beta_k = 0 \ \ \text{and} \ \ \eta \mathbf{1}_m + \sum_{k \in K} \beta_k \big((-1)^{s_k} X^k\big) = 0$$

implies that $\beta_k = 0$ for some $k \in K$, then the minimal solution (w, b) is unique.

Equivalently, for every family $((-1)^{s_k} X^k)_{k \in K}$ where $K \subseteq \{1, \ldots, n\}$ and the X^k are columns of X, with $s_k = 0, 1$, if

(1) for every sequence $(\beta_k)_{k \in K}$, if $|K| \geq 2$ and

$$\sum_{k \in K} \beta_k = 0 \ \ \text{and} \ \ \mathbf{1}_m = \sum_{k \in K} \beta_k \big((-1)^{s_k} X^k\big)$$

implies that $\beta_k = 0$ for some $k \in K$, and

(2) the family $((-1)^{s_k} X^k)_{k \in K}$ is weakly affinely independent,

then the minimal solution (w, b) is unique.

Hastie, Tibshirani and Wainwright define what it means for the columns of the matrix X to be in *general position*; see [7], Section 2.6. In our terminology this is basically the condition that the families $((-1)^{s_k} X^k)_{k \in K}$ are affinely independent. This condition is stronger that weak affine independence so our Proposition 4.2 is a stronger result. In practice this really does not make any difference because data sets, unless they are rather pathological, satisfy the stronger assumption (being in general position). Hastie, Tibshirani and Wainwright also do not deal with the intercept b.

Observe that if X is the zero matrix, then Proposition 4.2 does not apply because Condition (2) is not satisfied, but by going back to Version (V3), we easily see that there is a unique solution. In fact, we see immediately that $w = 0$ and $b = (\mathbf{1}_m^\top y)/m = \overline{y}$. Also, all the Lagrange multiplers are equal to $\tau/2$.

A sufficient condition for (1) is that $\mathrm{rank}(X) < \mathrm{rank}(\widetilde{X})$, namely that $\mathbf{1}_m$ is not a linear combination of the columns of X. A sufficient condition for (2) is that the families $((-1)^{s_k} X^k)_{k \in K}$ are affinely independent. Most "real" data sets X satisfy this condition unless they are pathological.

5 The Space of Minimal Solutions for lasso

This situation in Examples 3.1, 3.2, 3.3 3.4, and 3.5 generalizes to the minimal solutions associated with any pair of sets K_+ and K_-.

For any d with $2 \leq d \leq n$, for any sequence $s = (s_1, \ldots, s_d)$ with $s_i = 0$ or $s_i = 1$, let O_s be the polyhedral cone in $\mathbb{R}^d$, called the *s-orthant*, defined by the inequalities

$$\begin{cases} \delta_i \geq 0 & \text{if } s_i = 0 \\ \delta_i \leq 0 & \text{if } s_i = 1. \end{cases}$$

These can be written concisely as

$$(-1)^{s_i} \delta_i \geq 0, \quad 1 \leq i \leq d.$$

There are 2^d s-orthants.

We also have the hyperplanes H_s and $H_{s,\mu}$ in $\mathbb{R}^d$ given by the equations

$$(-1)^{s_1} \delta_1 + \cdots + (-1)^{s_d} \delta_d = 0.$$

and

$$(-1)^{s_1} \delta_1 + \cdots + (-1)^{s_d} \delta_d = \mu,$$

with $\mu > 0$. Earlier we used the notations H_{K_+,K_-} and H_{K_+,K_-}, μ but the notations H_s and $H_{s,\mu}$ are more concise.

The intersection $\Delta_{d-1}^{s,\mu} = H_{s,\mu} \cap O_s$ of $H_{s,\mu}$ and O_s is the convex set defined by the affine constraints

$$\begin{aligned} (-1)^{s_1} \delta_1 + \cdots + (-1)^{s_d} \delta_d = \mu & \\ (-1)^{s_i} \delta_i \geq 0 \quad 1 \leq i \leq d. & \end{aligned}$$

Using the change of variable $z_i = (-1)^{s_i} \delta_i$, we see that $\Delta_{d-1}^{s,\mu}$ is the $(d-1)$-simplex whose vertices are the d points $(0, \ldots, 0, (-1)^{s_i} \mu, 0, \ldots, 0)$, with $(-1)^{s_i} \mu$ in the ith slot.

To deal with the case where vectors $\begin{pmatrix} \delta \\ \eta \end{pmatrix}$ in $\mathrm{Ker}\, \widetilde{X}$ have a nonzero η component, we also define the polyhedral cylinder $C\Delta_{d-1}^{s,\mu} \subseteq \mathbb{R}^{d+1}$ above the simplex $\Delta_{d-1}^{s,\mu}$ as

$$C\Delta_{d-1}^{s,\mu} = \{(\delta, \eta) \in \mathbb{R}^{d+1} \mid \delta \in \Delta_{d-1}^{s,\mu},\ \eta \in \mathbb{R}\}.$$

The cylinder $C\Delta_{d-1}^{s,\mu} \subseteq \mathbb{R}^{d+1}$ has dimension d. Observe that

$$C\Delta_{d-1}^{s,\mu} = \Delta_{d-1}^{s,\mu} \times \mathbb{R} = (H_{s,\mu} \times \mathbb{R}) \cap (O_s \times \mathbb{R}).$$

Proposition 5.1. *Let K_0, K_+ and K_- be the subsets given by*

$$\begin{aligned} K_0 &= \{i \in \{1, \ldots, n\} \mid 0 < (\alpha_+)_i < \tau\} \\ K_+ &= \{i \in \{1, \ldots, n\} \mid (\alpha_+)_i = \tau\} \\ K_- &= \{i \in \{1, \ldots, n\} \mid (\alpha_+)_i = 0\} \end{aligned}$$

uniquely associated with all minimal solutions of an instance of lasso, let $d = n - |K_0| \geq 2$, and let s be the sequence in which $s_i = 0$ if $i \in K_+$, and $s_i = 1$ if $i \in K_-$. Also let (w_1, b_1) be any minimal solution and let

$$\mu = \|w_1\|_1 = \sum_{i \in K_+ \cup K_-}^{d} (-1)^{s_i} (w_1)_i$$

(with $\mu > 0$), the minimal value of the 1-norm of all minimal solutions. Let $(\mathrm{Ker}\, \widetilde{X})_{K_+,K_-}$ be the subspace of $\mathbb{R}^{d+1}$ consisting of all vectors $(\delta_{i \in K_+ \cup K_-}, \eta)$ such that $(\delta, \eta) \in \mathrm{Ker}\, \widetilde{X}$ and $\delta_i = 0$ if $i \in K_0$, or equivalently

(1) $\delta_i = 0$ if $i \in K_0$.

(2) $\sum_{i \in K_+ \cup K_-} \delta_i X^i + \eta \mathbf{1}_m = 0$.

The subspace $(\mathrm{Ker}\, \widetilde{X})_{K_+,K_-}$ is the projection onto $\mathbb{R}^{d+1}$ of the subspace of $\mathrm{Ker}\, \widetilde{X}$ consisting of the vectors $(\delta, \eta) \in \mathrm{Ker}\, \widetilde{X}$ such that $\delta_i = 0$ if $i \in K_0$. The affine subspace $(w_1, b_1) + (\mathrm{Ker}\, \widetilde{X})_{K_+,K_-}$ has dimension at most $n + 1 - r$ where r is the rank of $\widetilde{X}$, and is not parallel to the η-axis. Then the space of minimal solution is the intersection of the polyhedral cylinder $C\Delta_{d-1}^{s,\mu}$ with the affine subspace $(w_1, b_1) + (\mathrm{Ker}\, \widetilde{X})_{K_+,K_-}$. It is a polytope (a bounded polyhedron) of dimension at most $d - 1$.

Proof. Let (w_2, b_2) be any minimal solution and let $(\delta, \eta) = (w_2 - w_1, b_2 - b_1)$. The classification implies that

$$\begin{cases} \delta_i \geq -(w_1)_i, \ (w_1)_i \geq 0 & \text{if } s_i = 0 \\ \delta_i \leq -(w_1)_i, \ (w_1)_i \leq 0 & \text{if } s_i = 1. \end{cases}$$

Observe that this convex set is the translate $O_s - w_1$ of the s-orthant O_s. Then by Proposition 3.4, the vector $(\delta_{i \in K_+ \cup K_-}, \eta) \in \mathbb{R}^{d+1}$ must belong to the intersection of the convex sets

(1) The kernel subspace $(\mathrm{Ker}\, \widetilde{X})_{K_+,K_-}$.

(2) The hyperplane $H_s \times \mathbb{R}$ in $\mathbb{R}^{d+1}$, where H_s is given by the equation

$$\sum_{i \in K_+ \cup K_-} (-1)^{s_i} \delta_i = 0,$$

since η is not yet constrained.

(3) The polyhedral cylinder $(O_s - w_1) \times \mathbb{R}$, since η is not yet constrained.

But then the minimal solution $(w_2, b_2) = (w_1, b_1) + (\delta, \eta)$ belongs to the intersection of the convex sets

(1) The affine subspace $(w_1, b_1) + (\mathrm{Ker}\, \widetilde{X})_{K_+,K_-}$.

(2) The translate $(w_1, b_1) + (H_s \times \mathbb{R})$, which is the affine hyperplane $H_{s,\mu} \times \mathbb{R}$ in $\mathbb{R}^{d+1}$, where $H_{s,\mu}$ is given by the equation

$$\sum_{i \in K_+ \cup K_-} (-1)^{s_i} \delta_i = \mu,$$

since η is not yet constrained.

(3) The polyhedral cylinder $(w_1, b_1) + ((O_s - w_1) \times \mathbb{R}) = O_s \times \mathbb{R}$, since η is not yet constrained.

Since $(\mathrm{Ker}\, \widetilde{X})_{K_+,K_-}$ is the projection of a subspace of $\mathrm{Ker}\, \widetilde{X}$, which has dimension $n+1-r$ where r is the rank of $\widetilde{X}$, it has dimension at most $n+1-r$. Since the vectors in $(\mathrm{Ker}\, \widetilde{X})_{K_+,K_-}$ satisfy the equation

$$\sum_{i \in K_+ \cup K_-} \delta_i X^i + \eta \mathbf{1}_m = 0,$$

the m equations with scalar coefficients corresponding to the above system all have 1 as the coefficient of η, which means that the vectors normal to these hyperplanes are not orthogonal to the η-axis which is the line spanned by the vector $(0, \ldots, 0, 1) \in \mathbb{R}^{d+1}$. Therefore none of the hyperplanes defined by these equations are parallel to the η-axis, and so the affine subspace $(w_1, b_1) + (\mathrm{Ker}\, \widetilde{X})_{K_+,K_-}$ is not parallel to the η-axis.

The intersection of the two convex sets $H_{s,\mu} \times \mathbb{R}$ and $O_s \times \mathbb{R}$ is the polyhedral cylinder $C\Delta^{s,\mu}_{d-1}$. Therefore the minimal solution (w_2, b_2) belongs to the convex set which is the intersection of the polyhedral cylinder $C\Delta^{s,\mu}_{d-1}$ with the affine subspace $(w_1, b_1) + (\mathrm{Ker}\, \widetilde{X})_{K_+,K_-}$.

Because the affine subspace $(w_1, b_1) + (\mathrm{Ker}\, \widetilde{X})_{K_+,K_-}$ is not parallel to the η-axis, we claim that it intersects the polyhedral cylinder $C\Delta^{s,\mu}_{d-1}$ in a bounded set, which is a polytope.

We prove this fact as follows. A hyperplane in $\mathbb{R}^{d+1}$ not parallel to the η-axis (the x_{d+1}-axis) has an equation of the form

$$a_1 x_1 + \cdots + a_d x_d + \eta = c.$$

The boundary of the polyhedral cylinder $C\Delta^{s,\mu}_{d-1}$ consists of the lines parallel to the η-axis passing though the boundary of the simplex $\Delta^{s,\mu}_{d-1}$ and consists of faces whose boundaries are the lines parallel to the η-axis passing though the d vertices of the simplex $\Delta^{s,\mu}_{d-1}$. The intersection the vertical line passing though $(0, \ldots, 0, (-1)^{s_i}\mu, 0, \ldots, 0, 0)$ with $(-1)^{s_i}\mu$ in the ith slot ($1 \leq i \leq d$) and the hyperplane

$$a_1 x_1 + \cdots + a_d x_d + \eta = c$$

has η-coordinate given by

$$\eta_i = c - (-1)^{s_i} a_i \mu.$$

The d points $(0, \ldots, 0, (-1)^{s_i}\mu, 0, \ldots, 0, c - (-1)^{s_i} a_i \mu)$ are obviously all distinct, so there is a unique bijective affine map sending the simplex $\Delta^{s,\mu}_{d-1}$ to the intersection of the polyhedral cylinder $C\Delta^{s,\mu}_{d-1}$ with the hyperplane

$$a_1 x_1 + \cdots + a_d x_d + \eta = c,$$

so this intersection is a simplex. Since any affine subspace is the intersection of affine hyperplanes, by intersecting the previous simplex with more affine hyperplanes we obtain a polytope.

Finally, by construction, any vector (w, b) in the intersection of the polyhedral cylinder $C\Delta^{s,\mu}_{d-1}$ with the affine subspace $(w_1, b_1) + (\mathrm{Ker}\, \widetilde{X})_{K_+,K_-}$ has the same minimal norm μ as the minimal solution (w_1, b_1) and satisfies all the KKT conditions, so it is also a minimal solution. $\square$

Observe that if all the minimal solutions have the same b-component, which is the case if $\mathbf{1}_m$ is not a linear combinations of the columns of X, then the space of minimal solutions is a polytope obtained by intersecting the simplex $\Delta^{s,\mu}_{d-1}$ with some affine subspace.

In conclusion, note that for every polytope P, there is an affine bijection between P and the intersection of a simplex with some affine space (Ziegler [9], Exercise 2.2). Indeed, if P is a polytope in $\mathbb{R}^n$, it is defined as the bounded polyhedron which is the intersection of m half spaces, so it is defined as the set of solutions of a system of inequalities of the form

$$Ax \leq b.$$

Since P is bounded, each variable x_i is bounded below (and above) so by some suitable translation we may also assume that $x_i \geq 0$ for $i = 1, \ldots, n$. Using slack variables $y_1, \ldots, y_m$, we have the polytope Q in $\mathbb{R}^{n+m}$ defined by

$$\begin{aligned} \begin{pmatrix} A & I_m \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} - b &= 0 \\ x, y &\geq 0. \end{aligned}$$

We check immediately that the map $x \mapsto (x, b - Ax)$ from $\mathbb{R}^n$ to $\mathbb{R}^{n+m}$ is an afffine bijection between P and Q. The convex set Q specified by the above system is the intersection of the positive orthant in $\mathbb{R}^{n+m}$ with the affine subspace specified by the above system of equations. But because P is bounded, Q is also bounded, so we can find a $(n+m)$-simplex in $\mathbb{R}^{n+m}$ in the positive orthant that contains Q, and then Q is the intersection of a simplex with an affine subspace.

6 An Attempt Using Pseudo-Inverses

We now investigate how to use the pseudo-inverse of B to find minimal solutions in lasso when B has rank r less than $n+1$. Since B is symmetric positive semi-definite, it can be diagonalized as $B = Q\Sigma Q^\top$, where Q is an orthogonal matrix and Σ is a diagonal matrix of the form

$$\Sigma = \begin{pmatrix} \Sigma_r & 0_{r,n+1-r} \\ 0_{n+1-r,r} & 0_{n+1-r,n+1-r} \end{pmatrix},$$

where Σ_r is a diagonal matrix with positive entries (the strictly positiive eigenvalues of B). Our goal is to solve Equation (eq4) from Section 2 assuming that α_+ and α_- are known, namely

$$B \begin{pmatrix} w \\ b \end{pmatrix} = B_1 \begin{pmatrix} \alpha_+ \\ \alpha_- \end{pmatrix} + B_2.$$

If we write

$$C = B_1 \begin{pmatrix} \alpha_+ \\ \alpha_- \end{pmatrix} + B_2,$$

then we have the system

$$Q \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} Q^\top \begin{pmatrix} w \\ b \end{pmatrix} = C,$$

and so

$$\begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} Q^\top \begin{pmatrix} w \\ b \end{pmatrix} = Q^\top C.$$

If we define the new variable z_1 ranging over $\mathbb{R}^r$ and z_2 ranging over $\mathbb{R}^{n+1-r}$,

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = Q^\top \begin{pmatrix} w \\ b \end{pmatrix},$$

then we have

$$\begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = Q^\top C,$$

which implies that

$$\begin{pmatrix} \Sigma_r z_1 \\ 0_{n+1-r,1} \end{pmatrix} = Q^\top C.$$

Therefore $Q^\top C$ must be of the form

$$Q^\top C = \begin{pmatrix} c \\ 0_{n+1-r,1} \end{pmatrix}$$

for some $c \in \mathbb{R}^r$, so we obtain

$$z_1 = \Sigma_r^{-1} c,$$

and $z_2 \in \mathbb{R}^{n+1-r}$ is arbitrary. Since

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = Q^\top \begin{pmatrix} w \\ b \end{pmatrix},$$

the general solution is given by

$$\begin{aligned}
\begin{pmatrix} w \\ b \end{pmatrix} &= Q \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = Q \begin{pmatrix} \Sigma_r^{-1} c \\ 0_{n+1-r,1} \end{pmatrix} + Q \begin{pmatrix} 0_{r,1} \\ z_2 \end{pmatrix} \\
&= Q \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} c \\ 0_{n+1-r,1} \end{pmatrix} + Q \begin{pmatrix} 0_{r,1} \\ z_2 \end{pmatrix} \\
&= Q \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} Q^\top C + Q \begin{pmatrix} 0_{r,1} \\ z_2 \end{pmatrix}.
\end{aligned}$$

We recognize the pseudo-inverse B^+ of B given by

$$B^+ = Q \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} Q^\top,$$

so the general solution is given by

$$\begin{pmatrix} w \\ b \end{pmatrix} = B^+ C + Q \begin{pmatrix} 0_{r,1} \\ z_2 \end{pmatrix} = B^+ B_1 \begin{pmatrix} \alpha_+ \\ \alpha_- \end{pmatrix} + B^+ B_2 + Q \begin{pmatrix} 0_{r,1} \\ z_2 \end{pmatrix},$$

namely

$$\begin{pmatrix} w \\ b \end{pmatrix} = B^+ \begin{pmatrix} -\alpha_+ + \alpha_- \\ 0 \end{pmatrix} + B^+ B_2 + Q \begin{pmatrix} 0_{r,1} \\ z_2 \end{pmatrix}.$$

Given a minimal solution (w_1, b_1), say given by running ADMM on the primal problem, we can compute α_+ and α_- using the formulae of Proposition 3.2. We did this in a `Matlab` program. Then we can compute

$$\begin{pmatrix} w_2 \\ b_2 \end{pmatrix} = B^+ \begin{pmatrix} -\alpha_+ + \alpha_- \\ 0 \end{pmatrix} + B^+ B_2$$

using the pseudo-inverse B^+. If the signs of the components of w_2 satisfy the conditions of Proposition 3.3, then every time we ran our program we found that (w_2, b_2) is also a minimal solution, and in particular, $\|w_2\|_1 = \|w_1\|_1$. However we have no explanation for this fact. If the signs of the components of w_2 do not satisfy the conditions of Proposition 3.3, then (w_2, b_2) is not a minimal solution. We observed this behavior in Examples 3.3 and 3.4. If in Case (4) of Example 3.5 we replace Xb2 by

```
Xb3 = [20, -18; 6.1, 5; 2.2, 4; 0, 0; 4, 1.9; 1, -5; 6.2, -8; 4.1, -6];
```

then the solution computed using the pseudo-inverse is also wrong.

7 Filling in the Missing Steps

The missing steps that we mentioned in Section 1 are worked out as follows.

(1) We have

$$\begin{aligned}J_2(w,b) = \xi^\top\xi + Kw^\top w &= (w^\top X^\top + b\mathbf{1}_m^\top - y^\top)(Xw + b\mathbf{1}_m - y) + Kw^\top w\\ &= w^\top X^\top Xw + w^\top X^\top \mathbf{1}_m b - w^\top X^\top y\\ &\quad + b\mathbf{1}_m^\top Xw + b^2\mathbf{1}_m^\top\mathbf{1}_m - b\mathbf{1}_m^\top y\\ &\quad - y^\top Xw - by^\top\mathbf{1}_m + y^\top y + Kw^\top w\\ &= w^\top(X^\top X + KI_n)w + 2w^\top X^\top\mathbf{1}_m b - 2w^\top X^\top y - 2b\mathbf{1}_m^\top y\\ &\quad + b^2\mathbf{1}_m^\top\mathbf{1}_m + y^\top y.\end{aligned}$$

We can rewrite the above expression in matrix form (using the fact that $\mathbf{1}_m^\top\mathbf{1}_m = m$) as

$$J_2(w,b) = \begin{pmatrix} w^\top & b\end{pmatrix}\begin{pmatrix} X^\top X + KI_n & X^\top\mathbf{1}_m\\ \mathbf{1}_m^\top X & m\end{pmatrix}\begin{pmatrix} w\\ b\end{pmatrix} - 2\begin{pmatrix} w^\top & b\end{pmatrix}\begin{pmatrix} X^\top y\\ \mathbf{1}_m^\top y\end{pmatrix} + y^\top y.$$

(2) The issue now is to prove that the matrix

$$B = \begin{pmatrix} X^\top X + KI_n & X^\top\mathbf{1}_m\\ \mathbf{1}_m^\top X & m\end{pmatrix}$$

is symmetric positive definite. Since $m > 0$ and $X^\top X + KI_n$ is symmetric positive definite we can use Schur complements (Gallier and Qaintance [6], Proposition 7.3) to prove that B is symmetric positive definite. The Schur complement of m is

$$S = X^\top X + KI_n - m^{-1}X^\top\mathbf{1}_m\mathbf{1}_m^\top X,$$

and the Schur complement of $X^\top X + KI_n$ is

$$T = m - \mathbf{1}_m^\top X(X^\top X + KI_n)^{-1}X^\top\mathbf{1}_m = \mathbf{1}_m^\top(I_m - X(X^\top X + KI_n)^{-1}X^\top)\mathbf{1}_m. \tag{T}$$

But

$$(X^\top X + KI_n)^{-1}X^\top = X^\top(XX^\top + KI_m)^{-1},$$

so

$$T = \mathbf{1}_m^\top(I_m - XX^\top(XX^\top + KI_m)^{-1})\mathbf{1}_m.$$

Then

$$\begin{aligned}I_m - XX^\top(XX^\top + KI_m)^{-1} &= (XX^\top + KI_m)(XX^\top + KI_m)^{-1} - XX^\top(XX^\top + KI_m)^{-1}\\ &= K(XX^\top + KI_m)^{-1},\end{aligned}$$

which is SPD, since $XX^\top + KI_m$ is SPD, so

$$T = \mathbf{1}_m^\top K(XX^\top + KI_m)^{-1}\mathbf{1}_m > 0. \tag{$*_T$}$$

Since $X^\top X + KI_n$ is SPD and $T > 0$, by Proposition 7.3, B is also SPD.

Here is another proof not using the Schur complement. Let g be the function given by

$$\begin{aligned} g(w,b) &= \begin{pmatrix} w^\top & b \end{pmatrix} \begin{pmatrix} X^\top X + KI_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} \\ &= w^\top (X^\top X + KI_n) w + 2 w^\top X^\top \mathbf{1}_m b + \mathbf{1}_m^\top \mathbf{1}_m b^2. \end{aligned}$$

We need to prove that if $(w,b) \neq 0$, then $g(w,b) > 0$. If $(w,b) \neq 0$ and $b = 0$, then $w \neq 0$, and in this case $g(w,0) = w^\top (X^\top X + KI_n) w > 0$, since $X^\top X + KI_n$ is SPD.

Let us now assume that $b \neq 0$. For b fixed, the function $w \mapsto g(w,b)$ is strictly convex because $X^\top X + KI_n$ is SPD, so it has a unique minimum obtained by setting the gradient $\nabla_w g$ to 0; see Gallier and Qaintance [6], Theorem 4.13(4)). This yields

$$2(X^\top X + KI_n) w + 2 X^\top \mathbf{1}_m b = 0,$$

and we obtain

$$w^* = -(X^\top X + KI_n)^{-1} X^\top \mathbf{1}_m b.$$

The minimum of g with respect to w is obtained by substituting the above value of w into g and we get

$$\begin{aligned} g(w^*,b) &= b \mathbf{1}_m^\top X (X^\top X + KI_n)^{-1} X^\top \mathbf{1}_m b - 2b \mathbf{1}_m^\top X (X^\top X + KI_n)^{-1} X^\top \mathbf{1}_m b + \mathbf{1}_m^\top \mathbf{1}_m b^2 \\ &= (\mathbf{1}_m^\top \mathbf{1}_m - \mathbf{1}_m^\top X (X^\top X + KI_n)^{-1} X^\top \mathbf{1}_m) b^2 \\ &= \mathbf{1}_m^\top (I_m - X (X^\top X + KI_n)^{-1} X^\top) \mathbf{1}_m b^2 = T b^2, \end{aligned}$$

where T is defined in (T). We proved above in $(*_T)$ that

$$T = \mathbf{1}_m^\top (I_m - X (X^\top X + KI_n)^{-1} X^\top) \mathbf{1}_m = \mathbf{1}_m^\top K (X X^\top + KI_m)^{-1} \mathbf{1}_m > 0.$$

Consequently, if $b \neq 0$, we see that $g(w,b) \geq g(w^*,b) = Tb^2 > 0$, which shows that B is SPD. In summary, we proved that if $(w,b) \neq 0$, then $g(w,b) > 0$, which means that B is SPD.

Since the matrix

$$B = \begin{pmatrix} X^\top X + KI_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix}$$

is symmetric positive definite, the function $J_2(w,b)$ has a unique minimum obtained by setting its gradient to zero, which yields the system

$$\begin{pmatrix} X^\top X + KI_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} = \begin{pmatrix} X^\top y \\ \mathbf{1}_m^\top y \end{pmatrix}. \qquad (*_1)$$

Acknowledgement. We thank the students in CIS 515, Spring 2022, for prompting us to clarify the nature or the dual in elastic net and lasso. We also thank Gaetano Donizetti whose extraordinary music inspired us not to give up during our sometimes sluggish investigations.

References

[1] Marcel Berger. *Géométrie 1*. Nathan, 1990. English edition: Geometry 1, Universitext, Springer Verlag.

[2] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, first edition, 2004.

[3] Jean H. Gallier. *Curves and Surfaces In Geometric Modeling: Theory And Algorithms*. Morgan Kaufmann, 1999.

[4] Jean H. Gallier. *Geometric Methods and Applications, For Computer Science and Engineering*. TAM, Vol. 38. Springer, second edition, 2011.

[5] Jean H. Gallier and Jocelyn Quaintance. *Linear Algebra and Optimization with Applications to Machine Learning, Volume I*. World Scientific, first edition, 2020.

[6] Jean H. Gallier and Jocelyn Quaintance. *Linear Algebra and Optimization with Applications to Machine Learning, Volume II*. World Scientific, first edition, 2020.

[7] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity. The Lasso and Generalizations*. CRC Press, first edition, 2015.

[8] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics. Princeton University Press, 1970.

[9] Gunter Ziegler. *Lectures on Polytopes*. GTM No. 152. Springer Verlag, first edition, 1997.