10-702 Statistical Machine Learning: Assignment 2 Solutions

1. Convex sets and functions

(a)

- First, we prove that if $p \geq 1$, then C is convex set. Let $g(x) = |x|$, where $x \in \mathbb{R}$. $g(x)$ is convex function. Let $h(y) = y^p$, where $y > 0$ and $p \geq 1$. Since $h'(y) = py^{p-1} \geq 0$ and $h''(y) = p(p-1)y^{p-2} \geq 0$ for $y > 0$ and $p \geq 1$, $h(y)$ is a non-decreasing convex function. Therefore $h(g(x)) = |x|^p$ is a convex function. $\forall x, y \in C$, $\|x\|_p^p = \sum_{j=1}^n |x_j|^p \leq 1$ and $\|y\|_p^p = \sum_{j=1}^n |y_j|^p \leq 1$. According to Jensen's inequality, $\forall \lambda \in [0, 1]$:

$$
\begin{aligned}
\|\lambda x + (1-\lambda)y\|_p^p &= \sum_{j=1}^n |\lambda x_j + (1-\lambda)y_j|^p \\
&\leq \lambda \sum_{j=1}^n |x_j|^p + (1-\lambda)\sum_{j=1}^n |y_j|^p \\
&\leq \lambda + (1-\lambda) = 1
\end{aligned}
$$

Since $\|\lambda x + (1-\lambda)y\|_p \leq 1$, $\lambda x + (1-\lambda)y \in C$. C is convex set

- Next for $p < 1$, let $x = [1, 0, \ldots, 0]^T$ and $y = [0, 1, \ldots, 0]^T$ and $\lambda = 0.5$:

$$
\|\lambda x + (1-\lambda)y\|_p^p = \sum_{j=1}^n |\lambda x_j + (1-\lambda)y_j|^p = 2 \cdot 0.5^p = 2^{1-p} > 1
$$

Therefore, $\lambda x + (1-\lambda)y \notin C$. C is not convex set

(b)

i. Convex function

The function can be separated into the two terms of the non-negative weighted sum. Consider the individual arguments inside the max term. $|x|$ is convex in x, and $1 - y$ is affine, so $1 + |x| - y$ is convex. Also, $\frac{1}{\sqrt{z}}$ is a negative-power function, so it is convex in z. The max term is convex, since each of its arguments is convex. The second term $\frac{(x-z)^2}{y+1}$ is the composition of the quadratic-over-linear function $\frac{s^2}{t}$ with the affine function that maps (x, y, z) to $(x - z, y + 1)$ – so this term is convex. The function $f(x, y, z)$ is the nonnegative weighted sum of convex functions, and thus is convex.

ii. Concave function

Let $g(t) = f(X + tV)$ with $\text{dom}(t) = \{t | X + tV \succ 0, X \in S_{++}^n, V \in S^n\}$.

$$
g(t) = (\det(X + tV))^{1/n} = (\det(X))^{1/n} \left(\det(I + tX^{-1/2}VX^{-1/2})\right)^{1/n}
$$

Let $\lambda_1, \ldots, \lambda_n$ as the eigenvalues of $X^{-1/2}VX^{-1/2}$, then

$$
\left(\det(I + tX^{-1/2}VX^{-1/2})\right)^{1/n} = (\Pi_{i=1}^n (1 + t\lambda_i))^{1/n}
$$

Define $h(t) = \left(\det(I + tX^{-1/2}VX^{-1/2})\right)^{1/n}$ which can be written in the form of vector composition:

$$h(t) = h(x_1, \ldots, x_n) = (\Pi_{i=1}^n x_i)^{1/n},$$

where $x_i = \varphi_i(t) = 1 + t\lambda_i$.
$h(x_1, \ldots, x_n)$ is the geometric mean which is a concave function in $x_1, \ldots, x_n$ and each x_i is a linear function in t. So

$$h''(t) = \phi(t)^T \nabla^2 h(x) \phi(t) + \nabla h(x)^T \phi''(t) = \phi(t)^T \nabla^2 h(x) \phi(t) \leq 0$$

Therefore, $h(t)$ is a concave function in t and hence $g(t)$ is concave in t. $f(X) = (\det(X))^{1/n}$ is concave in X.

(c) Since f is a differential convex function, $\forall x, y \in C$:

$$\begin{array}{rcl} f(y) & \geq & f(x) + \nabla f(x)^T (y - x) \\ f(x) & \geq & f(y) + \nabla f(y)^T (x - y) \end{array}$$

By adding them up, we obtain the result.

2. (Convex Conjugacy)

(a)

$$\begin{aligned} f^*(a,z) &= \sup_{x,y} \{xa + yz - f(x,y)\} \\ &= \sup_{x,y} \{xa + yz - f_1(x) - f_2(y)\} \\ &= \sup_{x,y} \; \{\{xa - f_1(x)\} + \{yz - f_2(y)\}\} \end{aligned}$$

Since the two bracketed terms do not share any variables, we can maximize over each separately

$$\begin{aligned} f^*(a,z) &= \sup_x\{xa - f_1(x)\} + \sup_y\{yz - f_2(y)\} \\ &= f_1^*(a) + f_2^*(z) \end{aligned}$$

Since we did not assume any convexity, this will hold even if f_1 and f_2 are not convex.

(b) For a convex mapping $f : \mathbb{R} \mapsto \mathbb{R}$, the convex conjugate is defined as

$$f^*(z) := \sup_x\{xz - f(x)\}.$$

This implies that for any $x, y \in \mathbb{R}$,

$$f^*(y) \geq xy - f(x),$$

and, in particular,

$$\lambda_i f^*(y) + \lambda_i f(x_i) \geq \lambda_i x_i y, \quad i = \{1, \ldots, n\}. \tag{1}$$

Summing Eq. (1) over all values of i and recalling that $\sum_{i=1}^n \lambda_i = 1$ gives

$$f^*(y) + \sum_i \lambda_i f(x_i) \geq y \sum_i \lambda_i x_i$$

which holds for all values of $y \in \mathbb{R}$ and in particular

$$\begin{aligned} \sum_i \lambda_i f(x_i) &\geq \sup_y\{y \sum_i \lambda_i x_i - f^*(y)\} \\ &= f(\sum_i \lambda_i x_i), \end{aligned}$$

where the equality follows from the definition of the convex conjugate and the fact that $f = f^{**}$. This proves the inequality.

(c) (Solution Provided by Carl Doersch)

i. Observe that $f''(x) = \frac{1}{x} > 0$ for $x > 0$, so f is convex. Next, we compute the convex conjugate

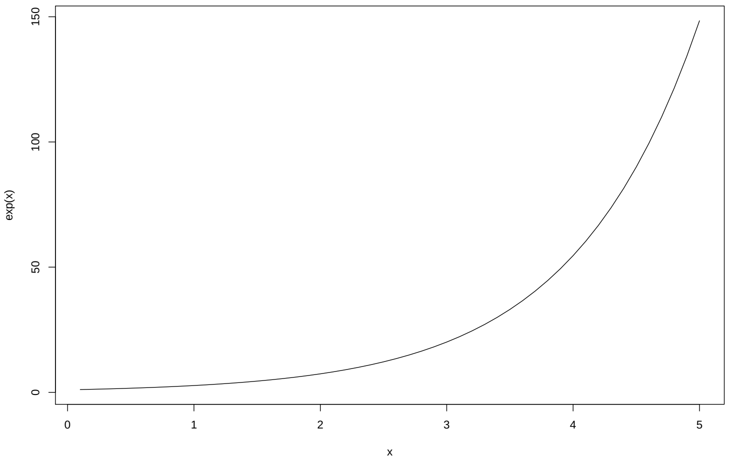
$$f^*(y) = \sup_x \{xy - x\ln x + x\}.$$

$\frac{\partial}{\partial x}(xy - x\ln(x) + x) = y - \ln(x) = 0$. So, $x^* = \exp(y)$.
$f^*(y) = (\exp(y)y - \exp(y)y + \exp(y)) = \exp(y)$.

Bi-Conjugate function $f^{**}(x) = \sup_y(xy - \exp(y))$.
For $x < 0$, $f^{**}(x)$ is unbounded since $y \to -\infty$.
For $x = 0$, $f^{**}(x) = 0$.
For $x > 0$, $\frac{\partial}{\partial y}(xy - \exp(y)) = x - \exp(y) = 0$. So, $y^* = \ln(x)$.
$f^{**}(x) = (x\ln(x) - x)$.

$$f^{**}(x) = \begin{cases} \infty & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ x\ln(x) - x & \text{if } x > 0 \end{cases}$$

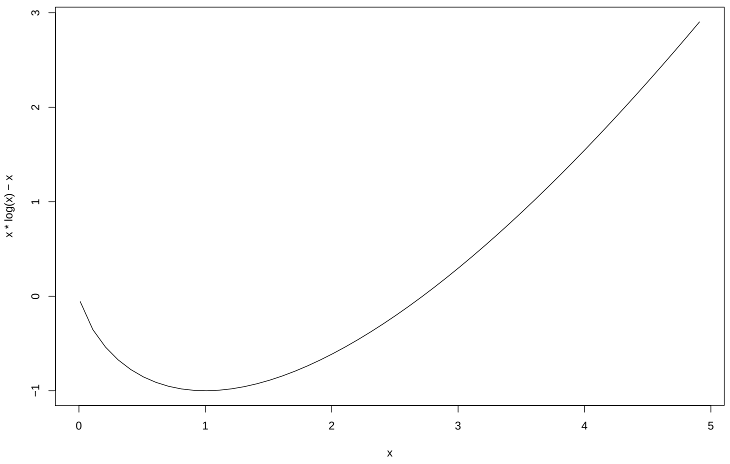
So f^{**} is almost equal to f except at $x = 0$.



ii. This function is the pointwise supremum over a set of linear functions so it is convex. Define the indicator function $g(z) = \mathbb{I}\{z \in C\}$. Observe that $g^*(x) = \sup_z \{x^T z - g(z)\} = \sup_{z\in C} x^T z = f(x)$, so that $g^{**} = f^*$. But since g is a closed and convex function $g = g^{**}$ and $f^*(z) = \mathbb{I}\{z \in C\}$. The dual of f^* is f since f is closed and convex.

iii.

$$f^*(y) = \sup_x \{yx - (\max(1 - x, 0))^2\}.$$



It is easy to see that for $y > 0$, $f^*(y) = \infty$. Otherwise, for $y \leq 0$,

$$\sup_{x \leq 1} \{yx - (\max(1-x,0))^2\} = \sup_{x \leq 1} \{yx - (1-x)^2\} = \frac{y^2}{4} + y$$

and

$$\sup_{x > 1} \{yx - (\max(1-x,0))^2\} = \sup_{x > 1} \{yx\} = y.$$

Therefore

$$f^*(y) = \begin{cases} y + y^2/4 & y \leq 0 \\ \infty & y > 0 \end{cases}$$

$f^{**}(x) = \sup_y \{xy - f^*(y)\}$. Note that the slope of f^* at zero is 1, so for $x < 1$, xy lies entirely underneath $f^*(y)$ except at 0. For $x > 1$, the slopes match somewhere on $y < 0$. Thus, we take the derivative and solve to find:

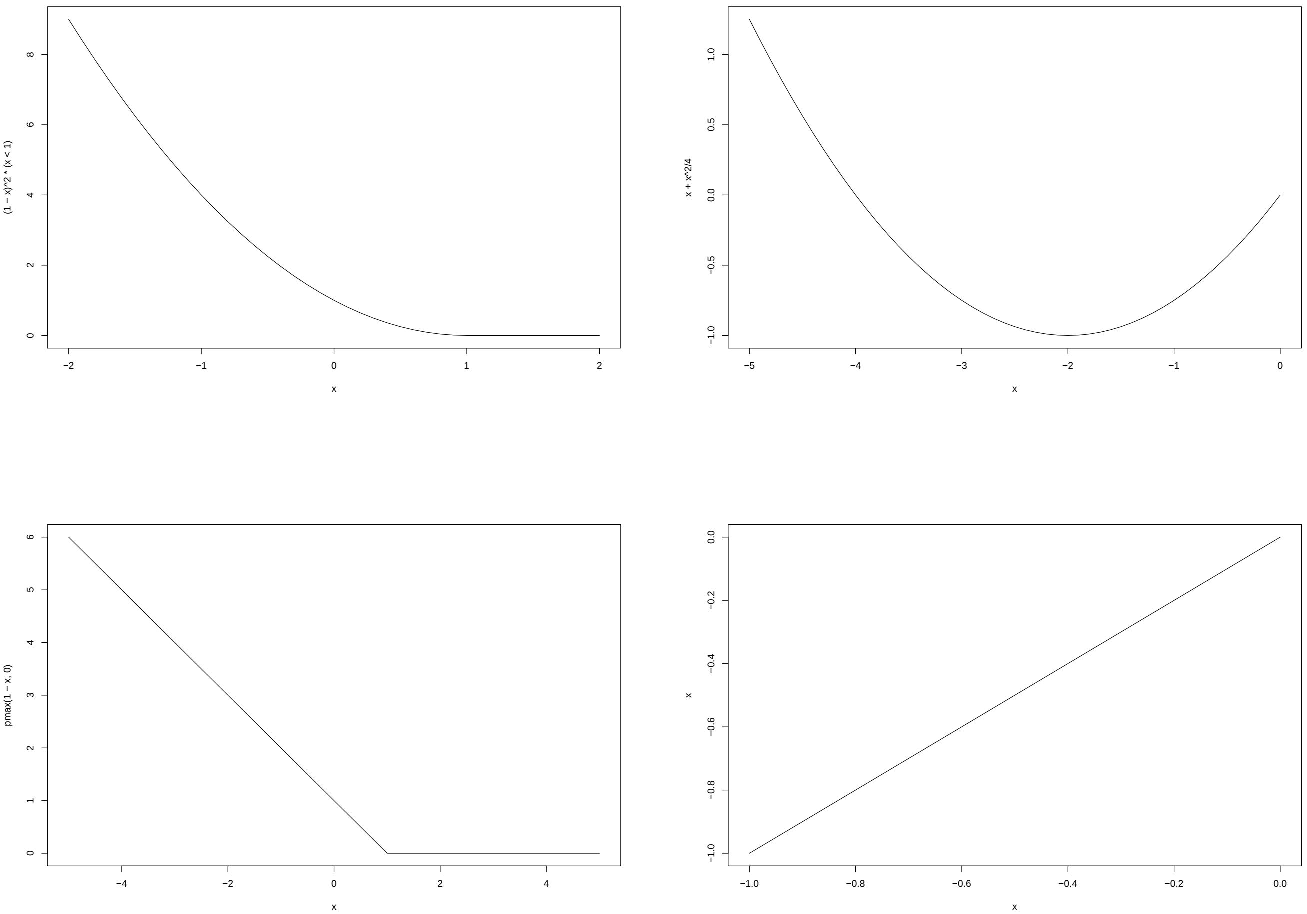
$$f^{**}(x) = \begin{cases} (1-x)^2 & x \geq 1 \\ 0 & x > 1 \end{cases}$$

Since $f^{**} = f$, f is convex.

iv. f is a convex function as a pointwise maximum of two convex functions. The conjugate $f^*(y) = \infty$ when $y < -1$ (for $x \to \infty$) or $y > 0$ (for $x \to -\infty$). For any other values of y, we can split the supremum into $x > 1$ and $x \leq 1$, and conclude that $f^*(y) = y$. Combining the two, the convex conjugate is given as

$$f^*(y) = \begin{cases} y & \text{if } y \in [-1, 0] \\ \infty & \text{otherwise.} \end{cases}$$

$f^{**}(x) = \sup_y \{xy - f^*(y)\} = \sup_{y \in [-1,0]} \{xy - y\} = \max(1-x, 0) = f(x)$



(d) i.

$$\begin{aligned} f^*(B) &= \sup_A \{\text{tr}(AB) - f(A)\} \\ &= \sup_A \{\text{tr}(AB) - \text{tr}(AS) + \log \text{det}A\} \\ &= \sup_A \{\text{tr}(A(B-S)) + \log \text{det}A\} \end{aligned} \tag{2}$$

Differentiating wrt A and setting to 0,

$$(B-S)^T + (A^{-1})^T = 0,$$

which gives $\hat{A} = (S-B)^{-1}$. Note that, since $f(A)$ is defined only for positive definite matrices, we require that $(S-B)^{-1} \succ 0$, else $f^*(B) \to \infty$. Substitute $\hat{A}$ back in $f^*(B)$,

$$\begin{aligned} f^*(B) &= \text{tr}(-(B-S)^{-1}(B-S)) + \log \text{det}(S-B)^{-1} \\ &= \text{tr}(-I_p) - \log \text{det}(S-B), \end{aligned}$$

where I_p is the identity matrix in p dimensions. Thus,

$$f^*(B) = \begin{cases} -p - \log \text{det}(S-B) & \text{if } (S-B)^{-1} \succ 0 \\ \infty & \text{otherwise.} \end{cases}$$

ii. Using the Lagrangian, we can formulate the dual as

$$\begin{aligned}
&\sup_{\lambda} \inf_{A} \{\text{tr}(AS) - \log \text{det}A + \lambda(\text{tr}(A) - L)\} \\
&\quad = \sup_{\lambda} \left\{ - \sup_{A} \{-\text{tr}(AS) + \log \text{det}A - \lambda\text{tr}(A)\} - \lambda L \right\} \\
&\quad = \sup_{\lambda} \left\{ - \sup_{A} \{-\text{tr}(AS) + \log \text{det}A + \text{tr}(A(-\lambda))\} - \lambda L \right\} \\
&\quad = \sup_{\lambda} \left\{ - \sup_{A} \{\text{tr}(A(-\lambda I_p)) - \text{tr}(AS) + \log \text{det}A\} - \lambda L \right\} \\
&\quad = \sup_{\lambda} \{-f^*(-\lambda I_p) - \lambda L\} \\
&\quad = \sup_{\lambda} \{p + \log \text{det}(S + \lambda I_p) - \lambda L\}\,.
\end{aligned}$$

This is the dual of our problem, subject to $\lambda \geq 0$.

3. (Subdifferentials and thresholding)

(a) i. Define

$$\mathcal{L}(x) := (Z - x)^2 + \lambda|x|.$$

A point $\hat{\mu}$ is a global minimum if and only if $0 \in \partial\mathcal{L}(\hat{\mu})$, i.e.,

$$0 \in -2(Z - x) + \lambda\partial|\cdot|(\hat{\mu}). \tag{3}$$

Consider 3 cases ($\hat{\mu} < 0, \hat{\mu} = 0, \hat{\mu} > 0$):

A. $\hat{\mu} > 0$. Then Eq. (3) reduces to $-2(Z - \hat{\mu}) + \lambda = 0$, which implies

$$\hat{\mu} = Z - \frac{\lambda}{2}.$$

Note that $\hat{\mu} > 0$ if $Z > \frac{\lambda}{2}$

B. $\hat{\mu} < 0$. Then Eq. (3) reduces to $-2(Z - \hat{\mu}) - \lambda = 0$, which implies

$$\hat{\mu} = Z + \frac{\lambda}{2}.$$

Note that $\hat{\mu} < 0$ if $Y < -\frac{\lambda}{2}$

C. $\hat{\mu} = 0$. Then, $\partial|\cdot|(\hat{\mu}) \in [-1, 1]$ and Eq. (3) reduces to :

$$-2Z - \lambda c = 0$$

where $c \in [-1, 1]$, which holds for

$$-\frac{\lambda}{2} \leq Z \leq \frac{\lambda}{2}.$$

Thus combining the three cases, we get

$$\hat{\mu} = \text{sign}(Z)\left(|Z| - \frac{\lambda}{2}\right)_+$$

ii. Define

$$\mathcal{L}(x) := (Z - x)^2 + \lambda x^2,$$

which is differentiable and

$$\frac{\partial\mathcal{L}}{\partial x} = -2(Z - x) + 2\lambda x.$$

Setting $\partial\mathcal{L}/\partial x = 0$ and solving for x gives

$$\hat{\mu} = \frac{Z}{1 + \lambda}.$$

iii. This penalty results in a hard-thresholding rule,

$$\hat{\mu} = \begin{cases} Z & \text{if } Z^2 \geq \lambda \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Define

$$\mathcal{L}(x) := (Z - x)^2 + \lambda \, \mathbb{I}\{x \neq 0\}.$$

Observe that $\mathcal{L}(0) = Z^2$ and $\mathcal{L}(Z) = \lambda$, so that $\hat{\mu} = 0$ as long as $\lambda > Z^2$.
Remark: In this question, you cannot use the subdifferential calculus, since $\mathbb{I}\{x \neq 0\}$ is not a convex function.

For $\mathrm{pen}(x) = x^2$, we will not have $\hat{\mu} = 0$ unless $\lambda \to \infty$. This can be seen from the fact that Z has a density and $\hat{\mu} = \frac{Z}{1+\lambda}$ so that $\mathbb{P}[\hat{\mu} = 0] = 0$. Note that this holds irrespective of the true mean μ. The situation is different for $\mathrm{pen}(x) = |x|$, when $\hat{\mu} = 0$ whenever $|Z| \leq \lambda/2$. When $\mu = 0$, then $\mathbb{P}[|Z| \geq \lambda/2] \leq 2\mathbb{P}[\mathcal{N}(0,1) \geq \lambda/(2\sigma)] = 2(1 - \Phi^{-1}(\lambda/(2\sigma)))$, where $\Phi(\cdot)$ is the CDF of the standard normal distribution, so that $\hat{\mu} = 0$ with high probability. When $\mu \neq 0$, then $\mathbb{P}[|Z| \geq \lambda/2] \leq 2\mathbb{P}[\mathcal{N}(0,1) \geq (\lambda - |\mu|)/(2\sigma)] = 2(1 - \Phi^{-1}((\lambda - |\mu|/(2\sigma)))$, so that $\hat{\mu} \neq 0$ as long as $|\mu| > \lambda$. From the discussion, we can conclude that there is no correspondence between the squared penalty and the Lasso penalty, unless $\lambda \to \infty$ or $\lambda = 0$.

(b) Define

$$\mathcal{L}(x) := \frac{1}{2} \sum_{j=1}^{G} \|Z_{\mathcal{G}_j} - x_{\mathcal{G}_j}\|^2 + \lambda_1 \sum_{j=1}^{G} \sum_{i=1}^{p_j} |x_{j,i}| + \lambda_2 \sum_{j=1}^{G} \|x_{\mathcal{G}_j}\|_2.$$

Observe that

$$\begin{aligned} \mathcal{L}(x) &= \sum_{j=1}^{G} \mathcal{L}_{\mathcal{G}_j}(x_{\mathcal{G}_j}) \\ &= \sum_{j=1}^{G} \left\{ \frac{1}{2} \|Z_{\mathcal{G}_j} - x_{\mathcal{G}_j}\|^2 + \lambda_1 \sum_{i=1}^{p_j} |x_{j,i}| + \lambda_2 \|x_{\mathcal{G}_j}\|_2 \right\}, \end{aligned}$$

so that the optimization can be done over each group of variables $\mathcal{G}_j$ $(j = 1, \ldots, G)$ separately. Without loss of generality, we can assume that there is only one group of variables, i.e., $G = 1$, and write:

$$\hat{\mu} = \arg\min_{x \in \mathbb{R}^p} \frac{1}{2} \|Z - x\|^2 + \lambda_1 \sum_{i=1}^{p} |x_i| + \lambda_2 \|x\|_2,$$

where $Z \in \mathbb{R}^p$. The subdifferential of $\mathcal{L}(\cdot)$ is given as

$$\partial \mathcal{L}(\cdot) = -(Z - x) + \lambda_1 t + \lambda_2 s \tag{5}$$

where $t \in \partial|\cdot|$ and $s \in \partial\|\cdot\|_2$. Recall

$$\partial|\cdot|(x) = \begin{cases} [-1,1]^p & \text{if } x = 0 \\ (\text{sign}(x_j))_{j=1}^p & \text{otherwise,} \end{cases} \tag{6}$$

and

$$\partial\|\cdot\|_2(x) = \begin{cases} B^2(1) & \text{if } x = 0 \\ \left(\frac{x_j}{\|x\|_2}\right)_{j=1}^p & \text{otherwise.} \end{cases} \tag{7}$$

We consider the following three cases separately:

i. $\|\hat{\mu}\|_2 = 0$

ii. $\|\hat{\mu}\|_2 \neq 0$, but $\exists i$ such that $\hat{\mu}_i = 0$

iii. $\|\hat{\mu}\|_2 \neq 0$ and $\forall i$ $\hat{\mu}_i \neq 0$.

From Eq. (5) it follows that $\hat{\mu} = 0$ if and only if for each $i = 1, \ldots, p$ the following holds

$$s_i = \frac{1}{\lambda_2}(Z_i - \lambda_1 t_i), \tag{8}$$

with $\sum s_i^2 \leq 1$ and $t_i \in [-1,1]$. The RHS of Eq. (8) is minimized for

$$t_i = \begin{cases} \frac{Z_i}{\lambda_1} & \text{if } |\frac{Z_i}{\lambda_1}| \leq 1 \\ \text{sign}(\frac{Z_i}{\lambda_1}) & \text{otherwise.} \end{cases} \tag{9}$$

Combining Eq. (8) and (9), we have that $\hat{\mu} = 0$ if and only if

$$\begin{aligned} &\sum_{i=1}^p s_i^2 \leq 1 \quad \wedge \quad \{t_i \in [-1,1]\}_{i=1}^p \\ &\Leftrightarrow \sum_{i=1}^p (Z_i - \lambda_1 t_i)^2 \leq \lambda_2^2 \quad \wedge \quad \{t_i \in [-1,1]\}_{i=1}^p \\ &\Leftrightarrow \sum_{i=1}^p \left(S_{\lambda_1}^{(1)}(Z_i)\right)^2 \leq \lambda_2^2 \\ &\Leftrightarrow \|S_{\lambda_1}^{(1)}(Z)\|_2 \leq \lambda_2. \end{aligned} \tag{10}$$

Next, we consider the case when $\hat{\mu} \neq 0$, but $\hat{\mu}_i = 0$ for some i. From Eq. (5) it follows that $\hat{\mu}_i = 0$ if and only if

$$t_i = \frac{Z_i}{\lambda_1} \in [-1,1]. \tag{11}$$

Finally, consider the case when $\hat{\mu} \neq 0$. From Eq. (5), for each $i = 1, \ldots, p$ we can express $\hat{\mu}_i$ as

$$\hat{\mu}_i + \lambda_2 \frac{\hat{\mu}_i}{\|\hat{\mu}\|_2} = Z_i - \lambda_1 t_i \tag{12}$$

$$\hat{\mu}_i = \frac{Z_i - \lambda_1 t_i}{1 + \frac{\lambda_2}{\|\hat{\mu}\|_2}}. \tag{13}$$

Squaring Eq. (13) and summing over all $i = 1, \ldots, p$, gives

$$\sum_i \hat{\mu}_i^2 = \|\hat{\mu}\|^2 = \frac{\sum_i (Z_i - \lambda_1 t_i)^2}{\left(1 + \frac{\lambda_2}{\|\hat{\mu}\|_2}\right)^2}, \tag{14}$$

which can be solved for $\|\hat{\mu}\|_2$ giving

$$\|\hat{\mu}\|_2 = \sqrt{\sum_i (Z_i - \lambda_1 t_i)^2} - \lambda_2. \tag{15}$$

We plug the expression for $\|\hat{\mu}\|_2$ back into Eq. (13) and obtain

$$\begin{aligned} \hat{\mu}_i &= \left(\sqrt{\sum_i (Z_i - \lambda_1 \text{sign}(Z_i))^2} - \lambda_2\right) \frac{Z_i - \lambda_1 \text{sign}(Z_i)}{\sqrt{\sum_i (Z_i - \lambda_1 \text{sign}(Z_i))^2}} \\ &= \left(\|S_{\lambda_1}^{(1)}(Z)\|_2 - \lambda_2\right) \frac{S_{\lambda_1}^{(1)}(Z_i)}{\|S_{\lambda_1}^{(1)}(Z)\|_2}. \end{aligned} \tag{16}$$

Combing Eq. (10), (11) and (16), we obtain

$$\hat{\mu} = \left(\|S_{\lambda_1}^{(1)}(Z)\|_2 - \lambda_2\right)_+ \frac{S_{\lambda_1}^{(1)}(Z)}{\|S_{\lambda_1}^{(1)}(Z)\|_2}. \tag{17}$$

The explicit expression given in Eq. (17) can be seen as a combination of two soft-thresholding operations. First, the elements of the vector Z are soft-thresholded to obtain $\tilde{Z} = S_{\lambda_1}^{(1)}(Z)$ and then

$$\hat{\mu} = S_{\lambda_2}^{(2)}(\tilde{Z}) = S_{\lambda_2}^{(2)}(S_{\lambda_1}^{(1)}(Z)). \tag{18}$$

(c) $(\Rightarrow)$

From the definition of subdifferential, for $w \in \partial\Psi(y)$ implies that for all $z \in \mathbb{R}$,

$$\begin{aligned} \Psi(z) &\geq \Psi(y) + w^T(z - y) \\ w^T y - \Psi(y) &\geq w^T z - \Psi(z) \end{aligned} \tag{19}$$

Since Eq. (19) holds for all $z \in \mathbb{R}$,

$$w^T y - \Psi(y) \geq \Psi^*(w). \tag{20}$$

Now, suppose $y \notin \partial\Psi^*(w)$. Then for some $z \in \mathbb{R}$,

$$\begin{aligned} \Psi^*(z) &< \Psi^*(w) + y^T(z - w) \\ y^T w - \Psi^*(w) &< y^T z - \Psi^*(z) \leq \Psi^{**}(y) = \Psi(y) \end{aligned} \tag{21}$$

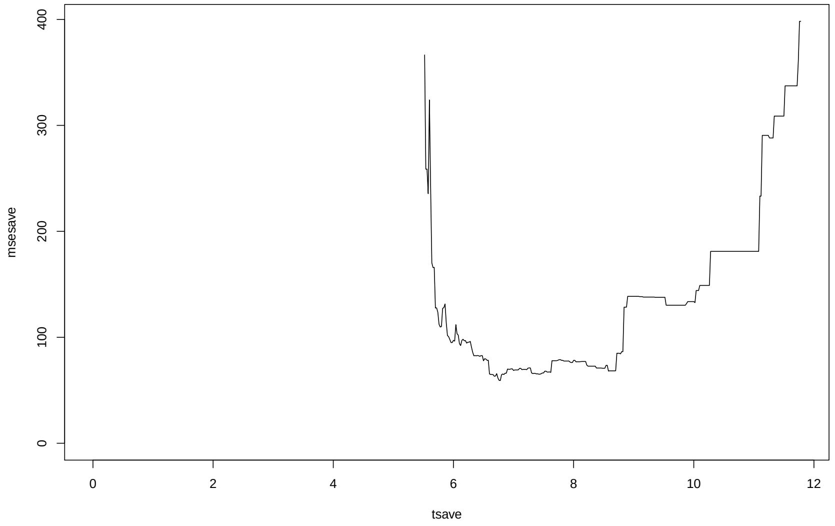
Now, Eq. (21) contradicts Eq. (20) and $y \in \partial\Psi^*(w)$.
$(\Leftarrow)$
For $y \in \partial\Psi^*(w)$ and for all $z \in \mathbb{R}$,

$$\Psi^*(z) \geq \Psi^*(w) + y^T(z - w), \tag{22}$$

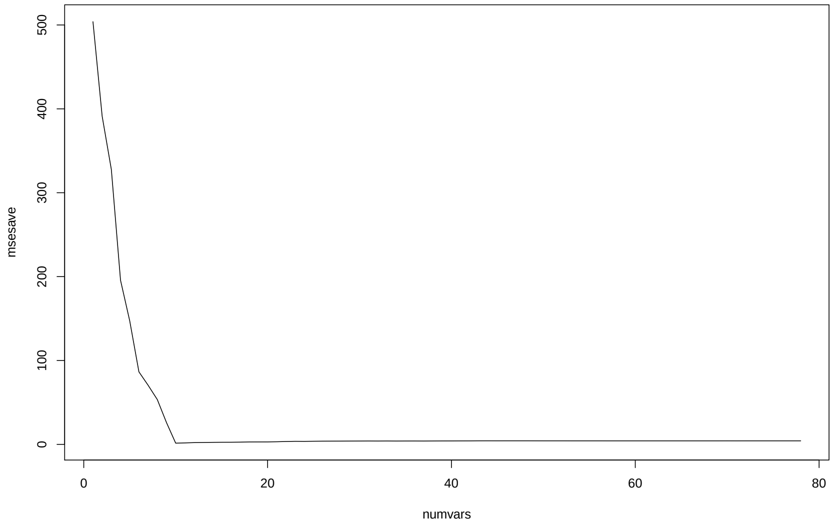
and now the implication follows by repeating the above analysis.

4. (Solution Provided by Carl Doersch)

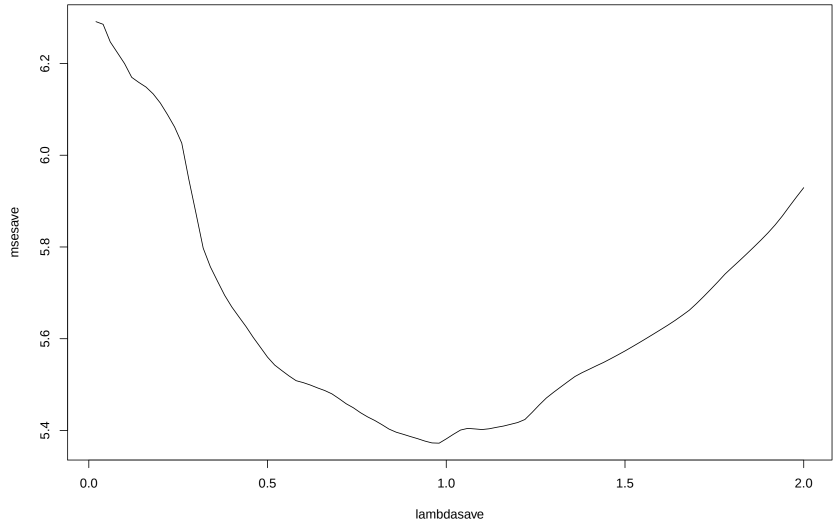
b) Marginal regression



c) Forward stepwise regression



d) Lars



e) First of all, the forward stepwise regression does approximately what we would expect; it begins by adding the first 10 real explanatory variables before steadily adding noise variables with very small coefficients. Thus, after an initial steep reduction in error, the error climbs slowly.

The marginal regression case is a bit more interesting. For large t it behaves similar to the forward stepwise regression case. However, decreasing t does not do the same thing as adding variables in stepwise regression; instead, marginal regression tends to add other variables besides the real explanatory variables before adding all of the real explanatory variables. This happens because some variables may happen to correlate well with one of the 5 explanatory variables. In forward stepwise regression, irrelevant variables wouldn't be added, since the variance would already be explained by the first ten variables variables. The marginal regression adds them without checking if they actually explain the data better than the true explanatory variables. Furthermore, since the variables may actually correlate reasonably well with the data, so that they will enter the final model high coefficients. This could be avoided if we use marginal regression to select variables and then use the ordinary least square regression to fit the variables.

The lasso behaves similarly to the forward stepwise regression, although it is slightly worse, because the lasso tends to shrink the coefficients for the true explanatory variables. However, we can see that it tends to add the true explanatory variables to the model in order of their strength as predictors.The model gets slightly worse as it allows too many parameters to enter the model (lambda gets too small), and it gets arbitrarily bad as lambda gets too large (too much shrinkage).