Optimización

Análisis Convexo: Subdiferenciales

Estudiante:

Roberto Vásquez Martínez

Profesor:

Dr. Joaquín Peña Acevedo

Universidad de Guanajuato, Guanajuato

10 de Junio de 2022

Introducción

A lo largo del curso de Optimización se han desarrollado métodos en los cuales se utiliza la derivada y segundas derivadas, que con ayuda del Teorema de Taylor, permiten resolver el problema de optimización de forma local, usando aproximaciones lineales y cuadráticas.
Sin embargo, si $f : \mathbb{R}^n \to \mathbb{R}$ es la función objetivo y no es diferenciable no podemos utilizar ninguno de estos métodos vistos en clase para resolver el problema de optimización.
Un caso importante es considerar f una función convexa, donde vimos condiciones necesarias y suficientes para el punto óptimo de f cuando esta función es diferenciable. Relajando esta condición, analizaremos el caso cuando f es convexa pero no diferenciable.
Así, quisieramos generalizar la aproximación lineal para el caso de funciones convexas no diferenciables.
El tema de este escrito esta divido en dos partes: En primer lugar, se desarrollaran y estudiará el tipo de objetos necesarios para generalizar la idea de aproximación lineal y cuales sería en consecuencia las condiciones de optimalidad análogas al caso diferenciable, pero en el contexto de funciones convexas. Así pues, se desarrollará la teoría esencial para estos objetivos desde la formalidad del Análisis Funcional.
Por otro lado, pondremos todo este arsenal de herramientas en práctica en un problema de interés general: Regresión LASSO y hablando de forma más general, en el mundo de los modelos sparse.
Discutiremos la importancia de estos modelos en el mundo científico para justificar que el estudio del Análisis Convexo juega un papel fundamental en la Optimización y en consecuencia en la Modelación Matemática.

1. Sublinealidad

El objetivo de esta sección será estudiar que tipo de funciones debemos considerar para generalizar la idea, tan importante, de aproximación lineal.
Este tipo de aproximación se llevará acabo a través de *funciones sublineales*. Una definición natural que debilita la propiedad de linealidad es la siguiente.

Definición 1.1 (Sublinealidad)**.**
Una función $\sigma : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ se dice sublineal si se cumple

$$\sigma(t_1x_1 + t_2x_2) \leqslant t_1\sigma(x_1) + t_2\sigma(x_2) \quad \forall (x_1, x_2) \in \mathbb{R}^n \times \mathbb{R}^n \quad \text{y} \quad (t_1, t_2) \in \mathbb{R}^+ \times \mathbb{R}^+.$$

El sufijo *sub* se deriva del sentido de la desigualdad ya que en el caso lineal se alcanza la igualdad.

Observación 1.1.
Claramente, a partir de la Definición 1.1 se ve que toda función sublineal es convexa (basta considerar $t_1 + t_2 = 1$).

Ejemplo 1.1.
Consideremos $\|\cdot\| : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ cualquier norma en el espacio euclideano $\mathbb{R}^n$. Por la desigualdad del triángulo claramente $\|\cdot\|$ es una función sublineal.

En Análisis Funcional, se tiene un resultado que caracteriza a los funcionales lineales sobre espacios de Hilbert. Este resultado es el *Teorema de Representación de Riesz*, nos dice que $l : \mathbb{R}^n \to \mathbb{R}$ funcional lineal, existe un único $s \in \mathbb{R}^n$ tal que

$$l(x) = \langle s, x \rangle,$$

donde $\langle \cdot, \cdot \rangle$ representa el producto interior en $\mathbb{R}^n$.
Existe un teorema de representación equivalente para funciones sublineales, pero nos debemos reestringir a una clase especial de estas funciones. Este será tema de la siguiente sección.

1.1. Funciones sublineales cerradas

En primera instancia, debemos considerar las siguientes definiciones

Definición 1.2 (Epígrafe)**.**
Sea $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ no idénticamente $+\infty$. La epígrafe de f la definimos como el conjunto

$$\operatorname{epi} f = \{(x, r) \in \mathbb{R}^n \times \mathbb{R} \ : \ r \geqslant f(x)\}.$$

La clase de funciones que consderamos es la siguiente

Definición 1.3 (Función Cerrada)**.**
Sea $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ no idénticamente $+\infty$. Decimos que f es cerrada si $\operatorname{epi} f$ es un conjunto cerrado en la topología producto.

Por otro lado, decimos que f es continua por debajo en x si

$$\liminf_{y \to x} f(y) \geqslant f(x).$$

Una propiedad equivalente para que $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ sea cerrada es que f sea *continua por debajo* en $\mathbb{R}^n$.
La idea de buscar un representación de las funciones sublineales es para simplificar el problema de optimización, pues al igual que usamos la derivada como aproximación lineal a f en el caso diferenciable, en este caso utilizaremos funciones sublineales para aproximar a f función convexa.
La forma de las funciones sublineales que utilizaremos como aproximación local queda presentada en la siguiente definición.

Definición 1.4 (Función Soporte)**.**
Sea $S \subset \mathbb{R}^n$ no vacío y la función $\sigma_S : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ definida como

$$\sigma_S(x) = \sup\{\langle s, x \rangle \ : \ s \in S\}.$$

Decimos que σ_S es la función soporte del conjunto S.

Observación 1.2.
Podemos ver que el espacio donde actua σ_S y S son duales entre sí.

La siguiente proposición nos dice que en efecto las funciones soporte pertenecen a la clase de funciones cerradas que estamos considerando

Proposición 1.1.
Una función soporte es cerrada y sublineal.

Demostración. En primer lugar, veamos que σ_S es sublineal. Para $(t_1, t_2) \in \mathbb{R}^+ \times \mathbb{R}^+$ notamos que para todo $x_1, x_2 \in \mathbb{R}^n$

$$t_1\sigma_S(x_1) \geqslant t_1\langle s, x_1\rangle,$$

y

$$t_2\sigma_S(x_2) \geqslant t_2\langle s, x_2\rangle.$$

Además, por linealidad del producto interior

$$\sigma_S(t_1x_1 + t_2x_2) \geqslant t_1\langle s, x_1\rangle + t_2\langle s, x_2\rangle.$$

Por la condición de supremos tenemos

$$\sigma_S(t_1x_1 + t_2x_2) \leqslant t_1\sigma_S(x_1) + t_2\sigma_S(x_2),$$

lo que demuestra que σ_S es sublineal.
Para ver que σ_S es cerrada, por las observaciones previas, basta con ver que

$$\liminf_{y\to x}\sigma_S(y) \geqslant \sigma_S(x) \quad \forall x \in \mathbb{R}^n,$$

ya que $\sigma_S < +\infty$ pues $\sigma_S(0) = 0$.
Sea

$$z_n := \inf\{\sigma_S(\xi) \,:\, \xi \in B(x, 1/n)\backslash\{x\}\},$$

se puede ver que z_n es una sucesión monótona creciente tal que

$$\lim_{n\to\infty} z_n = \liminf_{y\to x} f(y).$$

Observamos que para todo $\varepsilon > 0$ existe $\xi_n \in B(x, 1/n)\backslash\{x\}$ tal que

$$z_n \leqslant \sigma_S(\xi_n) < z_n + \varepsilon.$$

Por lo tanto

$$z_n + \varepsilon > \langle s, \xi_n\rangle \quad \forall s \in S.$$

Notamos que

$$z_n + \varepsilon > \langle s, \xi_n\rangle = \langle s, \xi_n - x\rangle + \langle s, x\rangle.$$

Por la continuidad del producto interior y del hecho $\xi_n \to x$ tenemos

$$\liminf_{y \to x} f(y) + \varepsilon = \lim_{n \to \infty} z_n + \varepsilon \geqslant \langle s, x \rangle.$$

Como $\varepsilon > 0$ fue arbitrario entonces

$$\liminf_{y \to x} f(y) \geqslant \langle s, x \rangle \quad \forall s \in S,$$

luego $\liminf_{y \to x} f(y) \geqslant \sigma_S(x)$, que es lo que queríamos probar. ■

Una propiedad importante de las funciones soporte que resalta el tipo de conjuntos de $\mathbb{R}^n$ a los que nos podemos restringir para generalizar la condiciones de optimalidad que se tienen en el caso diferenciable es la siguiente.

Proposición 1.2.
Para $S \subset \mathbb{R}^n$ no vacío se tiene

$$\sigma_S = \sigma_{\overline{S}} = \sigma_{\operatorname{co} S},$$

en particular

$$\sigma_S = \sigma_{\overline{\operatorname{co} S}},$$

donde $\operatorname{co} S$ *es la envolvente convexa de S.*

El resultado anterior se sigue de la continuidad y linealidad del producto interior. A partir de este resultado podemos restringir el concepto de función soporte a conjuntos cerrados y convexos, que al final son las propiedades necesarias para generalizar la noción de aproximación lineal.
Finalmente, presentamos el teorema de representación para funciones sublineales cerradas, esto nos permitirá entender en que sentido se generaliza la optimización revisada a profundidad en el caso diferenciable.
Además, tenemos el siguiente resultado

Proposición 1.3.
Para $S \subset \mathbb{R}^n$ no vacío y σ_S su función soporte. Se cumple que

$$s \in \overline{\operatorname{co} S} \Leftrightarrow \{s \in \mathbb{R}^n \,:\, \langle s, d \rangle \leqslant \sigma_S(d) \forall d \in \mathbb{R}^n\}$$

Este resultado nos permite entender el siguiente teorema de representación y su relación con los conjuntos cerrados y convexos de $\mathbb{R}^n$.

Teorema 1.1 (Representación de Funciones Sublineales Cerradas)**.**
Sea $\sigma : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ función sublineal cerrada y

$$S_\sigma = \{s \in \mathbb{R}^n \,:\, \langle s, d \rangle \leqslant \sigma(d) \quad \forall d \in \mathbb{R}^n\}.$$

Entonces σ es la función soporte de S_σ.

Por la Proposición 1.3 que S_σ es un cojunto cerrado y convexo, por lo que este teorema de representación para funcionales cerrados genera una biyección entre los conjuntos cerrados y convexos en $\mathbb{R}^n$.

La prueba de Teorema 1.1 utiliza el hecho de que toda función cerrada y convexa es el supremo de las funciones afines la minorizan, cuando nos restringimos al caso particular de una función sublineal cerrada σ consideramos solo a los funcionales lineales que minorizan a σ, por lo que en efecto σ es la función soporte de S_σ.

2. Subdiferencial

Con la teoría desarrollada definiremos formalmente la generalización del gradiente para funciones convexas no necesariamente diferenciables.
En la sección anterior permitimos funciones en con valores en los reales extendidos, sin embargo, en lo que sigue, al menos que se especifique lo contrario, supondremos $f : \mathbb{R}^n \to \mathbb{R}$ convexa.
Al igual que en el caso diferenciable, tenemos la noción de *derivada direccional*, que de manera similar nos permitirá generalizar la aproximación que tenemos en el sentido del Teorema de Taylor y así construir métodos de descenso para este caso.
En primer lugar, dado $x, d \in \mathbb{R}^n$

$$q(t) = \frac{f(x+td) - f(x)}{t} \quad \text{para } t > 0. \tag{1}$$

Para f convexa, sabemos que q es creciente y localmente Lipschitz continua en 0, luego la siguiente definición tiene sentido.

Definición 2.1 (Derivada direccional)**.**
Sean $x, d \in \mathbb{R}^n$ y $f : \mathbb{R}^n \to \mathbb{R}$ función convexa. Entonces la derivada direccional de f en x en la dirección d es

$$f'(x,d) := \lim_{t \downarrow 0} q(t) = \inf\{q(t) \ : \ t > 0\}.$$

Tenemos el siguiente resultado que nos mostrará la importancia del teorema de representación que fue el tema central de la sección anterior.

Proposición 2.1.
Para $x \in \mathbb{R}^n$ fija. La función $f'(x,\cdot)$ es finita y sublineal.

Demostración. Una definición equivalente a sublinealidad es que la función sea convexa y homogénea positivamente (saca escalares positivos). Probaremos ambas propiedades para verficar que $f'(x,\cdot)$ es sublineal.
En primer lugar, consideremos $d_1, d_2 \in \mathbb{R}^n$ y $\alpha_1, \alpha_2 \in \mathbb{R}$ tal que $\alpha_1 + \alpha_2 = 1$. De la convexidad de f tenemos

$$\begin{aligned} f(x + t(\alpha_1 d_1 + \alpha_2 d_2)) - f(x) &= f(\alpha_1(x + td_1) + \alpha_2(x + td_2)) - \alpha_1 f(x) - \alpha_2 f(x) \\ &\leqslant \alpha_1[f(x + td_1) - f(x)] + \alpha_2[f(x + td_2) - f(x)] \quad \forall t, \end{aligned}$$

Dividiendo por $t > 0$ y haciendo $t \downarrow 0$ obtenemos

$$f'(x, \alpha_1 d_1 + \alpha_2 d_2) \leqslant \alpha_1 f'(x, d_1) + \alpha_2 f'(x, d_2),$$

de lo que se sigue la convexidad de $f'(x,\cdot)$.
Ahora probamos la homogeneidad positiva. Aquí basta observar que para $\lambda > 0$

$$f'(x,\lambda d) = \lim_{t\downarrow 0} \lambda \cdot \frac{f(x+\lambda t d) - f(x)}{\lambda t} = \lambda \lim_{\tau \downarrow 0} \frac{f(x+\tau d) - f(x)}{\tau} = \lambda f'(x,d),$$

lo que completa la prueba de que $f'(x,\cdot)$ es sublineal.
Finalmente, para $\|d\| = 1$, como f es convexa y finita entonces es Lipschitz continua en x, en particular, existe $\varepsilon > 0$ y $L > 0$ tal que

$$|f(x+td) - f(x)| \leqslant Lt \quad \forall 0 \leqslant t < \varepsilon,$$

luego $|f'(x,d)| \leqslant L$, ahora para $d \neq 0$ se tiene que

$$\left| f'\left(x, \frac{d}{\|d\|}\right)\right| \leqslant L,$$

por homogeneidad

$$|f'(x,d)| \leqslant L\|d\|, \tag{2}$$

y así tenemos que $f'(x,d)$ finita para toda $d \in \mathbb{R}^n$. ■

Una hecho importante y clave para definir el subdiferencial viene enunciado en la siguiente proposición

Proposición 2.2.
Para todo x se cumple $f'(x,\cdot)$ es una función sublineal cerrada.

Demostración. Al ser f convexa existe $\delta > 0$ tal que f es Lipschitz con constante L en $B_\delta(x)$ que es la bola de abierta de radio δ y centro x.
Sean $d_1, d_2 \in \mathbb{R}^n$ con $d_1, d_2 \neq 0$, notamos que existe $t' > 0$ de forma que

$$y + t'd_1, y + t'd_2 \in B_\delta(x) \quad \text{con } y \in B_\delta(x),$$

y de hecho esto se cumple para todo $0 < t < t'$
Notamos que

$$|f(y+td_1) - f(y+td_2)| = \left|[f(y+td_1) - f(y)] - [f(y+td_2) - f(y)]\right|,$$

Usando que f es Lipschitz en $B_\delta(x)$ con constante L tenemos que

$$|f(y+td_1) - f(y+td_2)| \leqslant Lt\|d_1 - d_2\|,$$

dividiendo entre t y haciendo $t \downarrow 0$ tenemos que

$$|f'(y,d_1) - f'(y,d_2)| \leqslant L\|d_1 - d_2\|,$$

y este argumento se puede aplicar para todo x en el dominio de f, que al estar considerando f finita este dominio es $\mathbb{R}^n$.
Por lo tanto $f'(x,\cdot)$ es continua para toda $x \in \mathbb{R}^n$, en particular es continua por debajo que equivale a ser cerrada. ■

Finalmente, podemos definir el concepto que generaliza al gradiente en funciones convexas finitas en general. Utilizando el Teorema 1.1 y la Proposición 2.2 nos permiten introducir el siguiente concepto.

Definición 2.2 (Subdiferencial I)**.**
El subdiferencial $\partial f(x)$ de f en x es el conjunto no vacío del cual $f'(x,\cdot)$ es función soporte, i.e.

$$\partial f(x) = \{s \in \mathbb{R}^n \ : \ \langle s, d\rangle \leqslant f'(x,d) \quad \forall d \in \mathbb{R}^n\}.$$

Un vector $s \in \partial f(x)$ decimos que es un subgradiente de f en x.

Observación 2.1.
Por la Proposición 1.3 el conjunto $\partial f(x)$ es cerrado y convexo. Más aún, de la condición tipo Lipschitz (2) tenemos que para $s \in \partial f(x)$ con $s \neq 0$

$$\|s\|^2 = \langle s, s\rangle \leqslant L\|s\|,$$

luego

$$\|s\| \leqslant L,$$

entonces $\partial f(x)$ es un conjunto cerrado y acotado de $\mathbb{R}^n$, por el Teorema de Heine-Borel $\partial f(x)$ es compacto, que es una propiedad ideal en el contexto de optimización.

Del detalle de esta definición es que necesitaremos calcularlo para generalizar las condiciones de optimalidad, deberíamos entonces poder calcular la derivada direccional y posteriormente determinar el conjunto de la cual esta derivada direccional es conjunto soporte. Para facilitar estos cálculos, presentamos una definición más directa.

Definición 2.3 (Subdiferencial II)**.**
El subdiferencial $\partial f(x)$ es el conjunto de vectores s que satisfacen

$$f(y) \geqslant f(x) + \langle s, y - x\rangle \quad \forall y \in \mathbb{R}^n.$$

A continuación probamos la equivalencia entre las Definiciones 2.2 y 2.3

Proposición 2.3.
Las Definiciones 2.2 y 2.3 son equivalentes.

Demostración. Sea $s \in \partial f(x)$ dado por la Definición 2.2, luego

$$\langle s, d\rangle \leqslant f'(x,d) \quad \forall d \in \mathbb{R}^n.$$

Al ser $f'(x,d) = \text{ínf}_{t>0}\, q(t)$ con q como (1) entonces

$$\langle s, d\rangle \leqslant \frac{f(x+td) - f(x)}{t} \quad \forall d \in \mathbb{R}^n \text{ y } t > 0.$$

Como $\varphi : \mathbb{R}^n \times \mathbb{R}^+ \to \mathbb{R}^n$ definida como

$$\varphi((d,t)) = x + td,$$

es un mapeo sobre entonces se tiene que s es subgradiente de f en x en el sentido de la Definición 2.2 si y sólo si

$$f(y) \geqslant f(x) + \langle s, y - x \rangle,$$

que es justo la condición que define el subdiferencial en la Definición 2.3. ■

Estas equivalencia nos dice que podemos considerar aproximaciones afines o lineales para definir el subdiferencial que coincide con la intuición de querer generalizar la idea de linealidad que proporciona la derivada en este contexto.
A partir de la Proposición 2.3 podemos presentar el siguiente resultado que enuncia condiciones de optimalidad en el lenguaje de subdiferenciales.

Teorema 2.1 (Condiciones de Minimalidad)**.**
Sea $f : \mathbb{R}^n \to \mathbb{R}$ convexa. Las siguientes condiciones son equivalentes

(i) *f es minimizada en x_* sobre $\mathbb{R}^n$.*
(ii) $0 \in \partial f(x_*)$.
(iii) *$f'(x_*, d) \geqslant 0$ para toda $d \in \mathbb{R}^n$.*

Demostración. (i)⇔(ii): Esta equivalencia es clara a partir de la Definición 2.3 de subdiferencial pues

$$f(y) \geqslant f(x_*) \quad \forall y \in \mathbb{R}^n,$$

si y solo si

$$f(y) \geqslant f(x_*) + \langle y - x_*, 0 \rangle \quad \forall y \in \mathbb{R}^n,$$

lo que es equivalente a $0 \in \partial f(x_*)$.
(ii)⇔(iii): Esta proposición lógica se sigue de la Definición 2.2 ya que

$$f'(x_*, d) \geqslant 0 \quad \forall d \in \mathbb{R}^n,$$

si y sólo si

$$f'(x_*, d) \geqslant \langle 0, d \rangle \quad \forall d \in \mathbb{R}^n,$$

que es equivalente a $0 \in \partial f(x_*)$ según la Definición 2.2. ■

Observación 2.2.
Si x_* es mínimo local de f se sigue cumpliendo (iii) y por el Teorema 2.1 tenemos que la convexidad implica que x_* es mínimo global.

Observación 2.3.
La propiedad $0 \in \partial f(x)$ es una generalización de $\nabla f(x) = 0$ cuando f es diferenciable en x.

Finalmente, mostramos un ejemplo de una función convexa importante en el caso de optimización pues aparece en el contexto de regularización y selección de variables, tema que discutiremos en la última sección. Este ejemplo es sobre la función valor absoluto.

Ejemplo 2.1.
Consideremos $f : \mathbb{R} \to \mathbb{R}$ como $f(x) = |x|$ la función valor absoluto. Sabemos que f es diferenciable en $\mathbb{R}\setminus\{0\}$ con $\partial f(x)$ igual a 1 o -1 según $x > 0$ o $x < 0$, respectivamente. Todo se reduce a hallar $\partial f(0)$.
Buscamos los $s \in \mathbb{R}$ tal que

$$|y| \geqslant s \cdot y \quad \forall y \in \mathbb{R}.$$

Si $y > 0$ entonces $s \leqslant 1$, mientras que si $y < 0$ entonces $s \geqslant -1$, por lo que

$$\partial f(0) = [-1, 1].$$

Por lo tanto

$$\partial f(x) = \begin{cases} 1 & \text{si } x > 0 \\ -1 & \text{si } x < 0 \\ [-1,1] & \text{si } x = 0. \end{cases}$$

El subgradiente en 0 se ve como en la siguiente

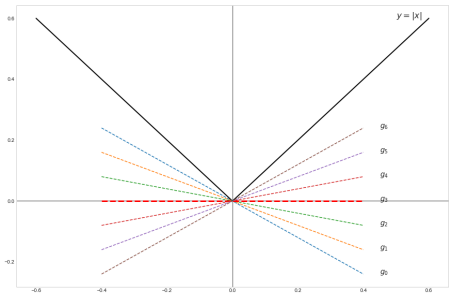

Figura 1: Subdiferencial de $|x|$ en $x = 0$

Aquí $\partial f(0)$ representa el conjunto de las pendientes de las aproximaciones lineales del valor absoluto en $x = 0$.

3. Método de Descenso

En los métodos de búsqueda en línea buscamos convergencia al un óptimo x_* a partir de una sucesión definida de forma recursiva

$$x_{k+1} = x_k + \alpha_k d_k,$$

con d_k un *dirección de descenso* en x_k, de forma que $f(x_{k+1}) < f(x_k)$. Para obtener esta dirección de descenso, en el caso de descenso máximo utilizamos $d_k = -\nabla f(x_k)$, esto si f es diferenciable. Sin embargo, estas direcciones de descenso deben estar definidas de otra forma cuando f es no diferenciable y supondremos convexidad para poder utilizar la idea de sublinealidad.
Recordamos la idea de *dirección de descenso*.

Definición 3.1 (Dirección de Descenso)**.**
Una dirección de descenso para la función convexa f en x es un $d \in \mathbb{R}^n$ tal que existe $t > 0$ de forma que

$$f(x + td) < f(x).$$

Tenemos así que estas direcciones son aquellas por las cuales debemos movernos respecto a x para decrementar el valor de la función objetivo. Como conceptualizamos el proceso de optimización como un problema de minimización precisamente por eso se llaman direcciones de descenso.
A continuación mostramos un resultado que relaciona las direcciones de descenso con toda la teoría desarrollada en las secciones previas.

Teorema 3.1 (Condiciones de descenso)**.**
Una dirección de descenso d de f en x está definida por cualquiera de las siguientes propiedades

(i) $f'(x, d) < 0$.
(ii) $\langle s, d\rangle < 0 \quad \forall s \in \partial f(x)$.

Demostración. Las condiciones (i) y (ii) son equivalente por la Definición 2.2 del subdiferencial y porque de la Observación 2.1 sabemos que $\partial f(x)$ es un conjunto compacto.
Ahora para ver que estas condiciones equivalen a que d sea dirección de descenso basta ver que d dirección de descenso equivale a $f'(x, d) < 0$.
Si d dirección de descenso existe $t' > 0$ tal que

$$f(x + t'd) - f(x) < 0,$$

de la Definición 2.1 se tiene que

$$f'(x, d) \leqslant \frac{f(x + t'd) - f(x)}{t'} < 0,$$

que es lo buscado.
Ahora si $f'(x, d) < 0$ por la condición de ínfimo tenemos que exite $t' > 0$ de modo que

$$f'(x, d) \leqslant q(t') < 0,$$

con q como en (1), lo que implica

$$f(x + t'd) < f(x),$$

luego d es dirección de descenso. ■

Geométricamente una dirección de descenso de f en x corresponde a un hiperplano que separa los conjuntos cerrados convexos $\partial f(x)$ y $\{0\}$.
Ahora definimos el problema que define la dirección de máximo descenso, por conveniencia normalizamos estas direcciones para que la solución de máximo descenso exista por compacidad.

Definición 3.2 (Dirección de Máximo Descenso)**.**
Sea $\|\cdot\|$ una norma en $\mathbb{R}^n$. Una dirección de máximo descenso normalizada de f en x, con respecto a la norma $\|\cdot\|$, es una solución al problema de optimización

$$\text{mín}\{f'(x,d) \,:\, \|d\| = 1\}, \tag{3}$$

o de forma equivalente, podemos escribir en notación min-máx este problema de optimización ya que $f'(x,\cdot)$ es función soporte de $\partial f(x)$

$$\min_{\|d\|=1} \max_{s\in\partial f(x)} \langle s, d\rangle,$$

y ese máximo se alcanza gracias a que $\partial f(x)$ es un compacto.

Observamos que existe al menos una solución a (3) pues $f'(x,\cdot)$ es continua y por el Teorema de Valores Extremos existe solución al problema (3) en el compacto $\{d \in \mathbb{R}^n \,:\, \|d\| = 1\}$.
Por lo tanto, esto nos permite definir de forma precisa la generalización del método de descenso máximo en este contexto.

Algoritmo 3.1 (Método de Descenso Máximo)**.**
Empezamos con $x_1 \in \mathbb{R}^n$. Ajustamos $k = 1$. Repetimos los siguientes pasos:

(i) **(Criterio de Paro)** $0 \in \partial f(x_k)$
(ii) **(Dirección de Descenso)** Para la norma $\|\cdot\|$ en $\mathbb{R}^n$. Hallar d_k solución de (3)
(iii) **(Búsqueda en línea)** Encontrar $t_k > 0$ de forma que $x_{k+1} := x_k + t_k d_k$ es tal que

$$f(x_{k+1}) < f(x_k)$$

(iv) Hacer $k = k + 1$ y volver a (i).

Todo lo desarrollado anteriormente nos permite afirmar que este método está bien definido: En (ii) la dirección hallada cumplirá $f'(x_k, d_k) < 0$ por el Teorema 2.1 ya que $0 \notin \partial f(x_k)$ si hemos llegado a ese paso. Por el Teorema 3.1, la dirección d_k es de descenso por lo que también el paso (iii) (backtracking) está bien definido.

4. Regresión LASSO

En esta última sección veremos como estos resultados teóricos nos ayudan a entender, generar soluciones e incluso introducir condiciones de optimalidad en un aplicación importante en la cual nos encontramos con un problema de optimización convexo no diferenciable. Este problema es: Regresión LASSO.
La idea de LASSO (Tibshirani, 1996) surge en el contexto de selección de variables y como propuesta de un modelo más ralos en un mundo en el que tenemos fenónemos con una gran cantidad de variables, pero quizás un subconjunto más pequeñas de ellas conforman la esencia del fenómeno. Por ejemplo, en Medicina, no podríamos esperar que 30,000 o más genes en el cuerpo humano están directamente involucrados en el

proceso del desarrollo del cáncer. El enfoque es entonces buscar modelos más simples, evitando el sobreajuste y consiguiendo una mejor entendimiento del objeto de estudio.
En el modelo de regresión lineal clásico tenemos un vector $Y = [y_1, \dots, y_N]^T$ de variables dependientes que se quieren explicar a través de un conjunto de variables explicativas concentradas en la matriz,

$$\mathrm{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \dots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{Np} \end{bmatrix},$$

luego el problema de regresión lineal supone una relación lineal entre Y y las variables representadas por las columnas de X de forma que

$$Y = \beta_0 \mathrm{e} + X\beta + \varepsilon, \text{ con } \varepsilon \text{ término de error,}$$

donde, $\mathrm{e} = [1, 1, \dots, 1]^T \in \mathbb{R}^N$, β es el vector de coeficientes asociados a cada variable mientras que $\beta_0 \in \mathbb{R}$ es el intercepto. Generalmente se supone $\varepsilon \sim \mathrm{N}(0, \sigma \mathrm{I})$ con I la matriz identidad de $N \times N$.
Si $\tilde{X} = [e, X]$. La solución por mínimos cuadrados es

$$\hat{\beta}_{OLS} = \mathrm{argmin}_{\beta \in \mathbb{R}^{p+1}} \left\{ \frac{1}{2N} \|Y - X\beta\|_2^2 \right\},$$

cuando $p + 1 < N$ y $\tilde{X}$ es de rango completo entonces la solución se obtiene de forma cerrada como

$$\hat{\beta}_{OLS} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \begin{bmatrix} \beta_0 \\ \beta \end{bmatrix}.$$

Sin embargo, cuando $p \gg N$ no tenemos esta propiedad y probablemente caegamos en un modelo sobreajustado.
En ese sentido, para forzar que algunas variables no se vean representadas en el modelo (su correspondiente índice en β es 0) necesitamos obtener un estimador *regularizado*.
La regularización que propone LASSO considera la norma $\|\cdot\|_1$ en ℓ_1, entonces para inducir una solución rala se propone resolver el problema

$$\hat{\beta}_{LASSO} = \mathrm{argmin}_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p} \left\{ \frac{1}{2N} \|Y - \beta_0 e - X\beta\|_2^2 \right\} \quad \text{sujeto a } \|\beta\|_1 \leqslant t, \tag{4}$$

para algún $t > 0$ que representa una cota superior de la influencia de cada variable, lo que permite tener resultados menos dispares en magnitud de los coeficientes y posiblemente los coeficientes mayor influencia obliguen a otros a ser 0.
Centrando la variable Y y normalizando las columnas en X podemos omitir el intercepto β_0 (este sería 0 después de la estimación), luego el problema de regresión LASSO en su forma langrangiana es

$$\hat{\beta}_{LASSO} = \mathrm{argmin}_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2N} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} \quad \text{para algún } \lambda > 0. \tag{5}$$

Por la dualidad del método de multiplicadores de Lagrange existe una correspondencia biunívoca entre t y λ.

Observación 4.1.
El término $\frac{1}{2N}$ que aparece en las funciones objetivos no hace ninguna diferencia, sin embargo, es conveniente pues induce una reparametrización de los valores λ de forma que estos sean comparables para diferentes tamaños de muestra N.

4.1. Solución de LASSO

Ahora derivaremos una solución para obtener el estimador $\hat{\beta}_{LASSO}$ en (5) a partir de la teoría desarrollada en las secciones previas.
En lo que sigue haremos uso de una función conocida como umbral suave (en inglés: *Soft thresholding*), que definimos a continuación.

Definición 4.1 (Soft Thresholding)**.**
Para $\lambda > 0$, definimos la función soft thresholding con respecto a λ, denotada $S_\lambda : \mathbb{R} \to \mathbb{R}$, como

$$S_\lambda(x) = \operatorname{sgn}(x)(|x| - \lambda)^+,$$

donde sgn es la función signo y $(\cdot)^+$ representa a la función parte positiva.

La gráfica de esta función se muestra en la siguiente figura

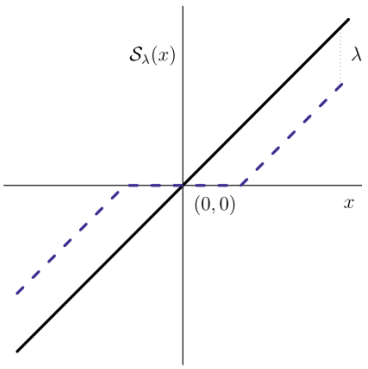

Figura 2: Función soft thresholding (línea punteada) y la función identidad (línea sólida)

Con esta definición construiremos una solución a LASSO, en donde suponemos previamente que el vector Y está centrado así como las columnas de X están centradas y tienen norma euclideana 1.

4.1.1. Un solo predictor

Con un solo predictor ($p = 1$) el problema de regresión LASSO es el siguiente

$$\min_{\beta \in \mathbb{R}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - z_i\beta)^2 + \lambda|\beta| \right\} \quad \text{con } \lambda > 0,$$

con $z_i := x_{i1}$ y $\beta := \beta_1$.

Sea $g : \mathbb{R} \to \mathbb{R}$ definida como

$$g(\beta) = \frac{1}{2N} \sum_{i=1}^{N} (y_i - z_i \beta)^2,$$

y $f : \mathbb{R} \to \mathbb{R}$ como

$$f(\beta) = g(\beta) + \lambda |\beta|.$$

Por lo tanto, el problema de regresión LASSO con un sólo predictor es equivalente a

$$\min_{\beta \in \mathbb{R}} \{f(\beta)\}. \tag{6}$$

A continuación, demostramos cual es la solución a este problema en el siguiente teorema.

Teorema 4.1 (LASSO un solo predictor)**.**
La solución a (6) *es*

$$\hat{\beta}_{LASSO} = S_\lambda\left(\frac{1}{N}\langle Z, Y\rangle\right),$$

donde $Z = [z_1, \dots, z_N]$ *y* $Y = [y_1, \dots, y_N]$.

Demostración. Notamos que g y la función valor absoluto son funciones convexas. Ahora, observamos que el operador subdiferencial ∂ es aditivo y homogéneo positivo, por lo que

$$\partial f(\beta) = \partial g(\beta) + \lambda \partial |\beta|.$$

Como g es diferenciable entonces $\partial g(\beta) = \{g'(\beta)\}$, luego

$$\partial f(\beta) = -\frac{1}{N}\langle Z, Y\rangle + \beta + \lambda \partial |\beta|.$$

Por otro lado, $\lambda \partial |\beta|$ ya lo hemos calculado en el Ejemplo 2.1.
Por el Teorema 2.1 basta hallar $\hat{\beta} \in \mathbb{R}$ tal que

$$0 \in -\frac{1}{N}\langle Z, Y\rangle + \hat{\beta} + \lambda \partial |\hat{\beta}|.$$

Si $\hat{\beta} = 0$ entonces se debe cumplir que

$$0 \in \left[-\lambda - \frac{1}{N}\langle Z, Y\rangle, \lambda - \frac{1}{N}\langle Z, Y\rangle\right],$$

por lo que $0 \in \partial f(0)$ si y sólo si

$$\left|\frac{1}{N}\langle Z, Y\rangle\right| \leqslant \lambda.$$

Ahora, si $\hat{\beta} \neq 0$ entonces $0 \in \partial f(\hat{\beta})$ equivale a que se cumpla la ecuación

$$\hat{\beta} = \frac{1}{N}\langle Z, Y\rangle - \lambda \operatorname{sgn}(\hat{\beta}),$$

luego, si $\hat{\beta}$ satisface la ecuación anterior entonces

$$\begin{aligned} \operatorname{sgn}(\hat{\beta}) = 1 \quad &\Leftrightarrow \quad \frac{1}{N}\langle Z, Y\rangle > \lambda, \\ \operatorname{sgn}(\hat{\beta}) = -1 \quad &\Leftrightarrow \quad \frac{1}{N}\langle Z, Y\rangle < -\lambda. \end{aligned}$$

Por lo tanto,

$$\hat{\beta} = \begin{cases} \frac{1}{N}\langle Z, Y\rangle - \lambda & \text{si } \frac{1}{N}\langle Z, Y\rangle > \lambda, \\ 0 & \text{si } \left|\frac{1}{N}\langle Z, Y\rangle\right| \leqslant \lambda, \\ \frac{1}{N}\langle Z, Y\rangle + \lambda & \text{si } \frac{1}{N}\langle Z, Y\rangle < -\lambda, \end{cases}$$

y es fácil ver que lo anterior equivale a

$$\hat{\beta} = S_\lambda\left(\frac{1}{N}\langle Z, Y\rangle\right).$$

Concluimos así que $0 \in \partial f(\hat{\beta})$ si y sólo si $\hat{\beta} = S_\lambda\left(\frac{1}{N}\langle Z, Y\rangle\right)$. ■

El Teorema 4.1 nos da una solución cerrada, que es un óptimo globar, para el problema de regresión LASSO con $p = 1$, esto nos permitirá construir una solución para LASSO en el caso general.

4.1.2. Múltiples predictores

Como obtenemos de forma cerrada la solución para un sólo predictor y la función objetivo de LASSO es convexa podemos utilizar el *método de descenso por coordenadas* para obtener el óptimo global del problema LASSO general.
En la iteración $k+1$ debemos hallar para $j = 1, 2, \ldots, N$

$$\beta_j^{(k+1)} = \operatorname{argmin}_{\beta_j \in \mathbb{R}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \sum_{r \neq j} x_{ir}\hat{\beta}_r - x_{ij}\beta_j)^2 + \lambda \sum_{r \neq j} |\hat{\beta}_r| + \lambda |\beta_j| \right\} \tag{7}$$

donde denotamos por $\hat{\beta}_r$ con $r = 1, 2, \ldots, p$ a los estimadores de β_r hasta ese momento. En el esquema de descenso por coordenadas en la iteración $k+1$ cada estimador de β_j se irá actualizando cíclicamente en algún orden previamente definido.
Sea

$$r_i^{(k,j)} := y_i - \sum_{r \neq j} x_{ir}\hat{\beta}_r \quad \text{para } i = 1, 2, \ldots, N,$$

el problema de optimización anterior en la iteración $k+1$ correspondiente a la coordenada j equivale a

$$\beta_j^{(k+1)} = \operatorname{argmin}_{\beta_j \in \mathbb{R}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (r_i^{(k,j)} - x_{ij}\beta_j)^2 + \lambda |\beta_j| \right\},$$

que es el problema LASSO de un solo predictor.

Si X_j es la j-ésima columna de X y $R_k^{(j)} := [r_1^{(k,j)}, \dots, r_N^{(k,j)}]$ entonces la solución del problema de optimización anterior según el Teorema 4.1 es

$$\beta_j^{(k+1)} = S_\lambda\left(\frac{1}{N}\left\langle X_j, R_k^{(j)}\right\rangle\right),$$

ahora si $R_k = [r_1, \dots, r_N]$ con $r_i^{(k)} = y_i - \sum_{r=1}^{p} x_{ip}\hat{\beta}_r$ entonces con un poco de cálculo se puede ver que esta solución es equivalente a

$$\beta_j^{(k+1)} = S_\lambda\left(\hat{\beta}_j + \frac{1}{N}\left\langle X_j, R_k\right\rangle\right). \tag{8}$$

Podemos así describir en detalle el algoritmo que resuelve LASSO vía descenso por coordenadas.

Algoritmo 4.1 (Solución LASSO descenso por coordenadas)**.**

Sea $\lambda > 0, \tau > 0$ una tolerancia, M un número máximo de iteraciones y $\beta^{(0)} = [\beta_j^{(0)}, \dots, \beta_p^{(0)}]$ un punto inicial. Ajustamos $k = 1$.

(i) **(Criterio de Paro)** $\|\beta^{(k)} - \beta^{(k-1)}\| < \tau$ o $k > M$.

(ii) Para $j = 1, \dots, p$

  (a) $\hat{\beta} := [\beta_1^{(k)}, \dots, \beta_{j-1}^{(k)}, \beta_j^{(k-1)}, \beta_{j+1}^{(k-1)}, \dots, \beta_p^{(k-1)}]$ con coordenadas denotadas por $\hat{\beta}_r$ para $r = 1, 2, \dots, p$.

  (b) $R_k := Y - X\hat{\beta}$

  (c) Resolvemos el problema (7) a través de (8), luego

$$\beta_j^{(k)} := S_\lambda\left(\hat{\beta}_j + \frac{1}{N}\left\langle X_j, R_k\right\rangle\right)$$

(iii) Hacer $k = k + 1$ y volver a (i).

Podemos inicializar el Algoritmo 4.1 considerando $\beta^{(0)}$ la solución por mínimos cuadrados de la regresión lineal clásica, ya que, finalmente, es la solución que queremos regularizar.

4.1.3. Ejemplo: Crime Data

Implementamos el Algoritmo 4.1 en Python 3.8.10 y lo probamos en el conjunto de datos *Crime data* (Ver Hastie y col., 2015, pp.10). El archivo que contiene esta base de datos se puede obtener en el sitio Hastie.su.domain. La variable Y en este caso será la columna `crime rate`, que consiste en la tasa de criminalidad por millón de habitantes.

Se tienen 5 predictores, que corresponderán a las las columnas de X, la descripción de cada variable es

(i) `funding`: presupuesto municipal para seguridad, en $ pér cápita.

(ii) `hs`: porcentaje de ciudadanos mayores a 25 años con preparatoria terminada.

(iii) `not-hs`: porcentaje de ciudadanos entre 16 y 19 años que no están en preparatoria y no se graduaron de preparatoria.

city	funding	hs	not-hs	college	college4	crime rate
1	40	74	11	31	20	478
2	32	72	11	43	18	494
3	57	70	18	16	16	643
4	31	71	11	25	19	341
5	67	72	9	29	24	773
⋮	⋮	⋮	⋮	⋮		
50	66	67	26	18	16	940

Figura 3: Tasa de criminalidad para $N = 50$ ciudades de EUA.

(iv) `college`: porcentaje de ciudadanos entre 18 y 24 años cursando estudios universitarios.

(v) `college4`: procentaje de ciudadanos mayores de 25 años con al menos 4 años de universidad.

En el Algoritmo 4.1 suponemos λ conocido, sin embargo, lo podemos calcular vía *Validación Cruzada*. Esto también lo implementamos en el notebook Solucion_LASSO.ipynb.

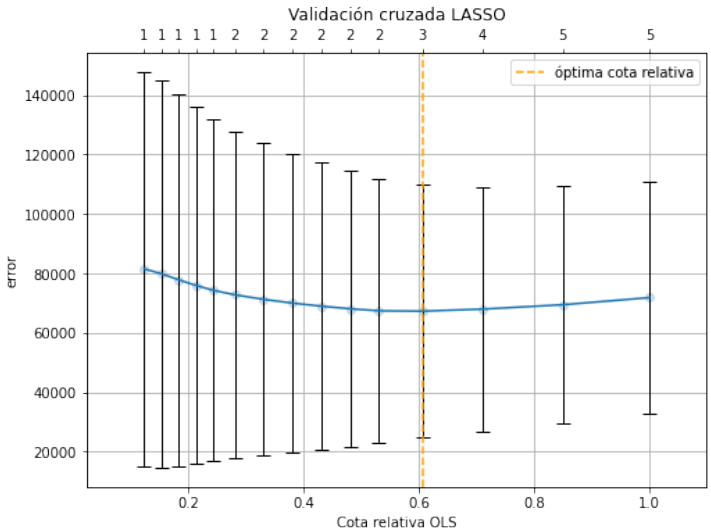

Figura 4: Marcamos con línea segmentada donde se obtiene el mejor error. Esto sugiere una selección de 3 variables

En la parte superior de la Figura 4 ponemos el número de coeficientes no 0 con respecto al a cada valor de λ. En el eje x no ponemos los valores de λ sino de los valores de $\frac{\|\hat{\beta}_{LASSO}(\lambda)\|}{\|\hat{\beta}_{OLS}\|}$ para cada λ, que lo podemos entender como la cota relativa respecto a mínimos cuadrados o como el encogimiento que hacemos al vector solución LASSO respecto a $\hat{\beta}_O LS$. Si este valor es cercano a 0 la restricción es fuerte y pocas variables o ninguna tendrá la posibilidad de ser representada, mientras que si es cercano a 1 la solución LASSO es parecida a la solución por mínimos cuadrados entonces todas las

variables quedarían con alguna señal.
Según la media error obtenido para cada λ probado en validación cruzada, el mejor desempeño se alcanza cuando

$$\frac{\|\hat{\beta}_{LASSO}(\lambda)\|}{\|\hat{\beta}_{OLS}\|} = 0.6,$$

que es parecido al valor obtenido en Hastie y col., 2015 página 13 y en este caso

$$\lambda = 25.71.$$

Ahora, mostramos la traza LASSO, que no es más que la grafica de como varían los coeficientes respecto a la cota relativa, más aún no dice el grado de importancia de cada una de estas variables según el peso de los coeficientes (que son comparables por la estandarización).

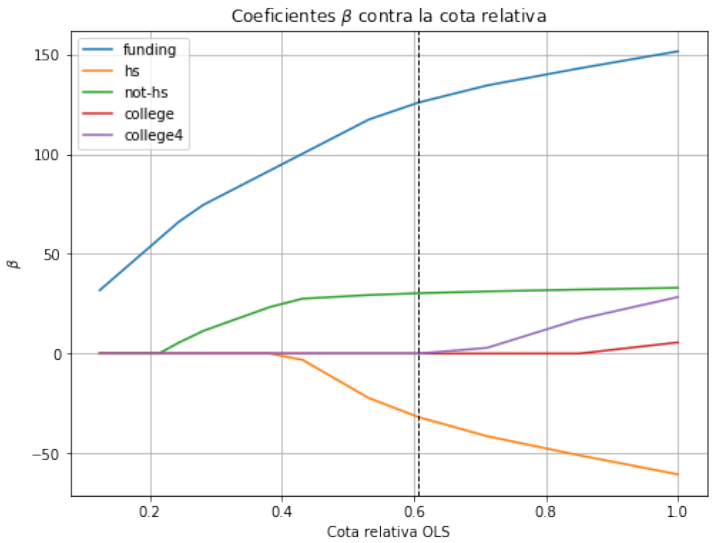

Figura 5: Traza LASSO respecto a la cota relativa

Al igual que en la Figura 4, marcamos con línea segmentada la cota relativa con mejor desempeño, y como vimos en la Figura 4, esta recta vertical corta a 3 coeficientes no 0, lo que rectifica que LASSO sugiere seleccionar solo las variables `funding,hs,not-hs`, y precisamente la que parece tener más importancia de todas ellas es `funding`.

Conclusiones

A lo largo de este escrito, vimos una teoría que permite generalizar algunas de las ideas de optimización en el caso diferenciable, centrándonos en el caso convexo.

Desarrollamos la teoría con la mayor precisión posible en el alcance de este proyecto para poder definir formalmente el concepto de subdiferencial (que generaliza el concepto de derivada para funciones convexas) y construir condiciones de optimalidad global para el caso de funciones convexas finitas, así generalizar las ideas de aproximación lineal que nos propociona el cálculo diferencial con la idea de sublinealidad. Asimismo, con las ideas desarrolladas sobre la sublinealidad como aproximación (estilo Taylor) a funciones convexas pudimos construir con precisión una generalización del Método de Descenso Máximo, tema que fue de gran interés a lo largo del curso de Optimización, aunque solo se estudió para el caso de funciones diferenciables.
Con estas ideas, pudimos proponer una solución al problema de regresión LASSO, que forma parte de una gran cantidad de modelos llamados *sparse*, a través de la teoría del Cálculo Subdiferencial. Precisamente por la importancia de estos modelos sparse en los campos de la medicina y problemas de alta dimensionalidad en general ($p \gg N$) este tipo de teoría es de sumo interés y reafirma aún más lo deseable que es la convexidad en el mundo de la optimización y el papel del Análisis Funcional para entender y proponer mejores métodos.

Referencias

Hastie, T., Tibshirani, R. & Wainwright, M. (2015). *Statistical Learning with Sparsity: The LASSO and Generalizations* (F. Bunea & et al, Eds.). Taylor & Francis Group.

Hiriart-Urruty, J. & Lemaréchal, C. (1993). *Convex Analysis and Minimization Algorithms I: Fundamentals* (M. Artin & et al, Eds.). Springer-Verlag.

Nocedal, J. & Wright, S. (2006). *Numerical Optimization* (T. Mikosch & et al, Eds.). Springer.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the LASSO. *Journal of the Royal Statistical Society*, *58*(1).