UC | Chile

UC | Chile

Algebra Lineal Aplicada para Ciencia de Datos

Clase 12. Penalización ℓ^1

1. Motivación
2. Formulación del problema
3. Resolución usando métodos proximales

Motivación

Motivación

La **inestabilidad** de la solución de mínima norma Euclideana al sistema

$$Ax = y$$

con $y \in \mathbb{R}^m$ y $A \in \mathbb{R}^{m \times n}$ ancha se puede mitigar usando **penalización cuadrática**

Esta solución se determina resolviendo el problema

$$\underset{x \in \mathbb{R}^n}{\text{minimizar}} \quad \|r(x)\|_2^2 + \lambda \|x\|_2^2$$

donde $r(x) = y - Ax$ es el **residuo** y $\lambda > 0$ es el **parámetro de penalización**

Motivación

La solución se puede representar en términos de la SVD de A como

$$x^\star = \sum_{i=1}^{r} \frac{\sigma_i}{\sigma_i^2 + \lambda}(u_i \cdot y)v_i.$$

Uno de las desventajas de esta solución es que el **número de coeficientes distintos de cero es grande**, esto es, es **densa**

Esto quiere decir que el vector y es aproximadamente una superposición de un gran número de columnas de A

$$y \approx x_1 a_1 + \ldots + x_n a_n$$

Motivación

Sin embargo, en varias aplicaciones es razonable pensar que los datos fueron generados como una **superposición de sólo algunas columnas**

En este caso, nos interesa estabilizar la solución promoviendo una gran cantidad de coeficientes $x_1, \dots, x_n$ **iguales a cero**

Para formalizar esta noción, definimos el **soporte** de $x \in \mathbb{R}^n$ como

$$\mathbf{sop}(x) = \{i : x_i \neq 0\}$$

y el **tamaño del soporte** como el número de elementos en $\mathbf{sop}(x)$

Motivación. Ejemplo

El vector

$$x = \begin{bmatrix} 1 \\ -1 \\ 0 \\ 4 \\ 0 \\ 5 \end{bmatrix}$$

tiene soporte

$$\mathbf{sop}(x) = \{1, 2, 4, 6\}$$

y el tamaño de su soporte es 4

Motivación

Por lo tanto, buscamos una solución aproximada al sistema

$$Ax = y$$

tal que su soporte $\mathbf{sop}(x)$ sea de tamaño pequeño y que sea **estable** a perturbaciones en los datos

En este caso, decimos que la solución es **rala** o ***sparse***

UC | Chile

Formulación del problema

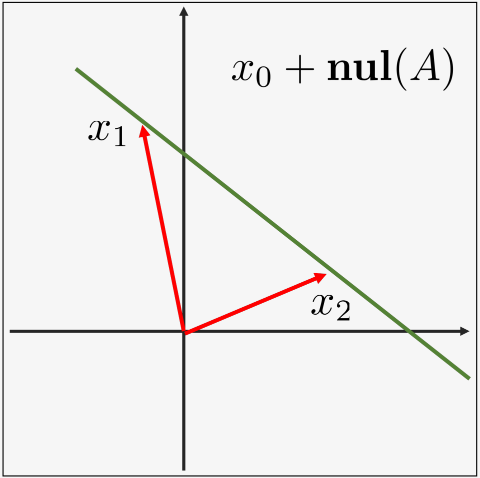
Formulación del problema

Para promover la raleza en la solución, penalizamos por la **norma** ℓ^1

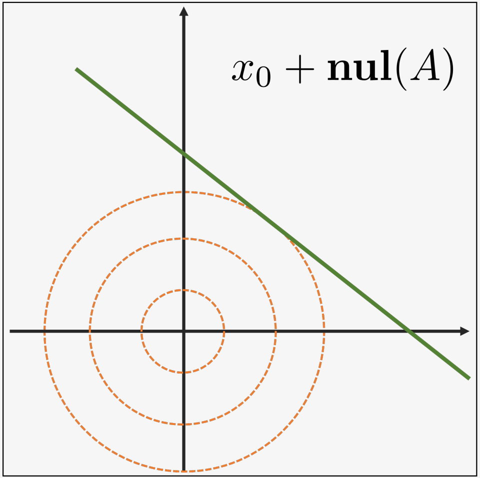
$$\|x\|_1 := \sum_{i=1}^{n} |x_i|$$

Esta es una estrategia popular para promover soluciones que son ralas

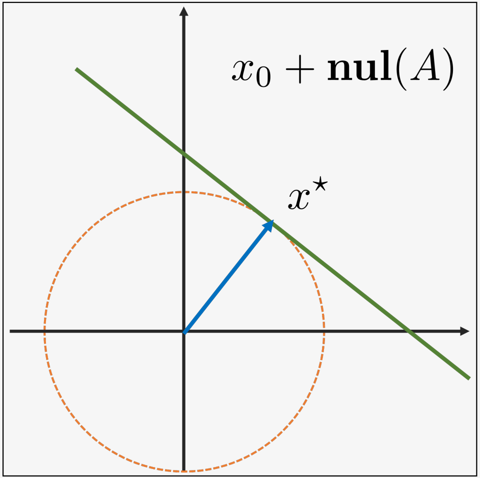
Formulación del problema



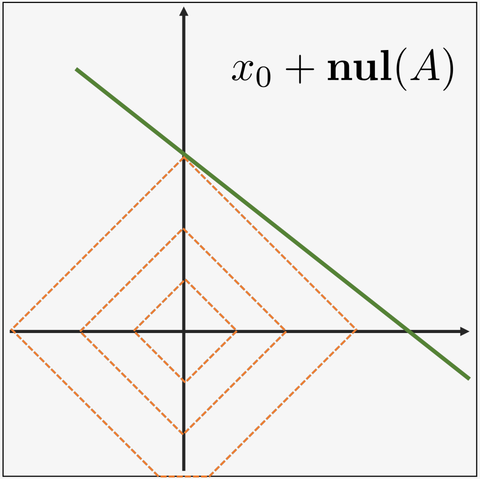
Formulación del problema



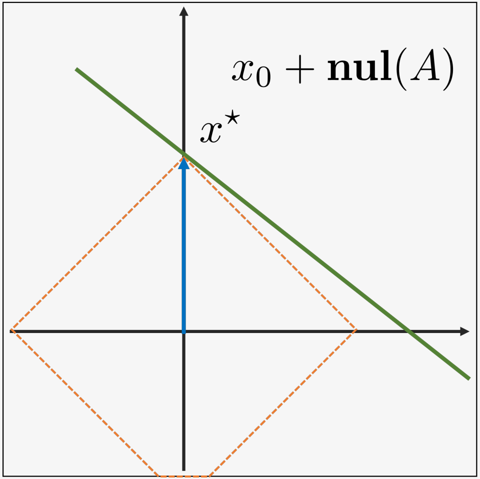
Formulación del problema



Formulación del problema



Formulación del problema



Formulación del problema

Buscamos un vector que encuentre un balance entre la **magnitud del residuo** y **su norma** ℓ^1

Para ello, resolvemos

$$\underset{x \in \mathbb{R}^n}{\text{minimizar}} \quad \|r(x)\|_2^2 + \lambda \|x\|_1$$

donde $\lambda > 0$ es un **parámetro de penalización** que controla este balance

Formulación del problema

$$\underset{x \in \mathbb{R}^n}{\text{minimizar}} \quad \|r(x)\|_2^2 + \lambda \|x\|_1$$

Este problema se conoce con distintos nombres

- En **estadística** se conoce como **regresión LASSO** por ***Least Absolute Shrinkage and Selection Operator***
- En **análisis de señales** se conoce como ***Basis Pursuit Denoising*** **(BPDN)**
- En **métodos Bayesianos** este problema permite determinar el **máximo *a posteriori*** **(MAP)** cuando se usa un *a priori* Laplace y la perturbación es Gaussiana

Resolución usando métodos proximales

Resolución usando métodos proximales

La penalización ℓ^1 no permite encontrar una solución de manera explícita como en el caso de la penalización cuadrática

Por lo tanto, debemos usar un **método iterativo** que define una sucesión de estimadores $x^{(0)}, x^{(1)}, \ldots$ de la solución óptima

Desde ahora escribimos

$$f(x) = \|y - Ax\|_2^2 + \lambda\|x\|_1, \quad r^{(k)} = r(x^{(k)}) \quad \text{y} \quad f^{(k)} = f(x^{(k)}).$$

En este contexto, el índice k es la **iteración**

Resolución usando métodos proximales

En la iteración k buscamos un vector x que reduzca el valor $f^{(k)}$

¿Cómo reducimos la magnitud del residuo?

Vemos que

$$\begin{aligned}
\|r(x)\|_2^2 &= \|y - Ax\|_2^2 \\
&= \|y - A(x^{(k)} + (x - x^{(k)}))\|_2^2 \\
&= \|(y - Ax^{(k)}) - A(x - x^{(k)})\|_2^2 \\
&= \|r^{(k)}\|_2^2 - 2(y - Ax^{(k)}) \cdot A(x - x^{(k)}) + \|A(x - x^{(k)})\|_2^2 \\
&= \|r^{(k)}\|_2^2 - 2A^\top(y - Ax^{(k)}) \cdot (x - x^{(k)}) + \|A(x - x^{(k)})\|_2^2.
\end{aligned}$$

Resolución usando métodos proximales

Si $A = U\Sigma V^\top$ es la SVD completa de A entonces

$$
\begin{aligned}
\|A(x - x^{(k)})\|_2^2 &= (U\Sigma V^\top)(x - x^{(k)}) \cdot (U\Sigma V^\top)(x - x^{(k)}) \\
&= (\Sigma V^\top)(x - x^{(k)}) \cdot (\Sigma V^\top)(x - x^{(k)}) \\
&= \sum_{i=1}^{r} \sigma_i^2 (v_i \cdot (x - x^{(k)}))^2 \\
&\leq \sigma_1^2 \textstyle\sum_{i=1}^{n} (v_i \cdot (x - x^{(k)}))^2 \\
&= \sigma_1^2 V^\top (x - x^{(k)}) \cdot V^\top (x - x^{(k)}) \\
&= \sigma_1^2 \|x - x^{(k)}\|_2^2
\end{aligned}
$$

Resolución usando métodos proximales

Por lo tanto

$$\begin{aligned}\|r(x)\|_2^2 &= \|r^{(k)}\|_2^2 - 2A^\top(y - Ax^{(k)}) \cdot (x - x^{(k)}) + \|A(x - x^{(k)})\|_2^2 \\ &\leq \|r^{(k)}\|_2^2 + g^{(k)} \cdot (x - x^{(k)}) + \sigma_1^2 \|x - x^{(k)}\|_2^2.\end{aligned}$$

donde definimos

$$g^{(k)} := -2A^\top(y - Ax^{(k)}).$$

Resolución usando métodos proximales

En tal caso,

$$\begin{aligned} \|r(x)\|_2^2 &\leq \|r^{(k)}\|_2^2 + g^{(k)} \cdot (x - x^{(k)}) + \sigma_1^2 \|x - x^{(k)}\|_2^2 \\ &= \|r^{(k)}\|_2^2 + \sigma_1^2 \left\| x - \left(x^{(k)} - \frac{1}{2\sigma_1^2} g^{(k)} \right) \right\|_2^2 - \frac{1}{4\sigma_1^2} \|g^{(k)}\|_2^2. \end{aligned}$$

Por lo tanto,

$$f(x) \leq f^{(k)} + \lambda \|x\|_1 + \sigma_1^2 \left\| x - \left(x^{(k)} - \frac{1}{2\sigma_1^2} g^{(k)} \right) \right\|_2^2 - \frac{1}{4\sigma_1^2} \|g^{(k)}\|_2^2 - \lambda \|x^{(k)}\|_1.$$

Resolución usando métodos proximales

Nuesta mejor alternativa para encontrar un candidato que reduzca el valor de $f^{(k)}$ es resolver

$$\underset{x \in \mathbb{R}^n}{\text{minimizar}} \quad \lambda \|x\|_1 + \sigma_1^2 \left\| x - \left(x^{(k)} - \frac{1}{2\sigma_1^2} g^{(k)} \right) \right\|_2^2.$$

Sorprendentemente podemos resolver este problema de manera explícita

Resolución usando métodos proximales

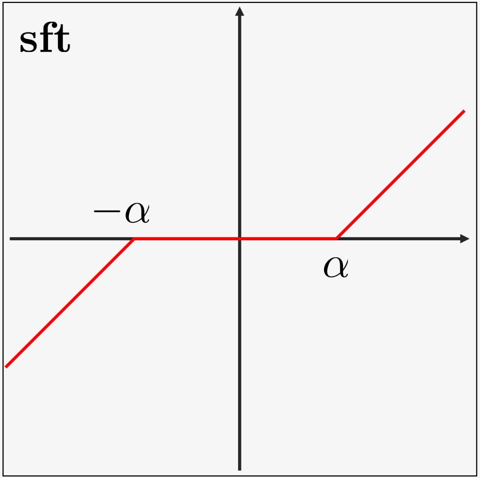
Definimos para $t \in \mathbb{R}$ y $\alpha \geq 0$ la función ***soft-thresholding*** o de **umbral suave** como

$$\mathbf{sft}(t, \alpha) = \begin{cases} t - \alpha & t > \alpha \\ 0 & |t| \leq \alpha \\ t + \alpha & t < -\alpha \end{cases}$$

Abusamos levemente la notación y escribimos para $x \in \mathbb{R}^n$

$$\mathbf{sft}(x, \alpha) = \begin{bmatrix} \mathbf{sft}(x_1, \alpha) \\ \vdots \\ \mathbf{sft}(x_n, \alpha) \end{bmatrix}$$

Resolución usando métodos proximales



Resolución usando métodos proximales

La solución a

$$\underset{x\in\mathbb{R}^n}{\text{minimizar}} \quad \lambda\|x\|_1 + \sigma_1^2 \left\| x - \left(x^{(k)} - \frac{1}{2\sigma_1^2} g^{(k)} \right) \right\|_2^2 .$$

es el vector

$$x^{(k+1)} = \mathbf{sft}\left(x^{(k)} - \frac{1}{2\sigma_1^2} g^{(k)}, \frac{\lambda}{2\sigma_1^2} \right)$$

La sucesión que se genera usando este método **converge a la solución óptima**

Resolución usando métodos proximales

Para resolver

$$\underset{x \in \mathbb{R}^n}{\text{minimizar}} \quad \|r(x)\|_2^2 + \lambda \|x\|_1$$

generamos la sucesión dada por

$$x^{(k+1)} = \mathbf{sft}\left(x^{(k)} - \frac{1}{2\sigma_1^2} g^{(k)}, \frac{\lambda}{2\sigma_1^2}\right)$$

para $x^{(0)}$ dado

Este es el **método de gradiente proximal**

Resolución usando métodos proximales

En la práctica resulta óptimo usar

$$x^{(k+1)} = \mathbf{sft}\left(x^{(k)} - \frac{1}{\sigma_1^2} g^{(k)}, \frac{\lambda}{\sigma_1^2}\right)$$

para $x^{(0)}$ dado

Observe que este paso es 2 veces más grande que el anterior

UC | Chile