Métodos Clásicos de Regularización

Aprendizaje Automático INF-393 II-2018

Ricardo Ñanculef

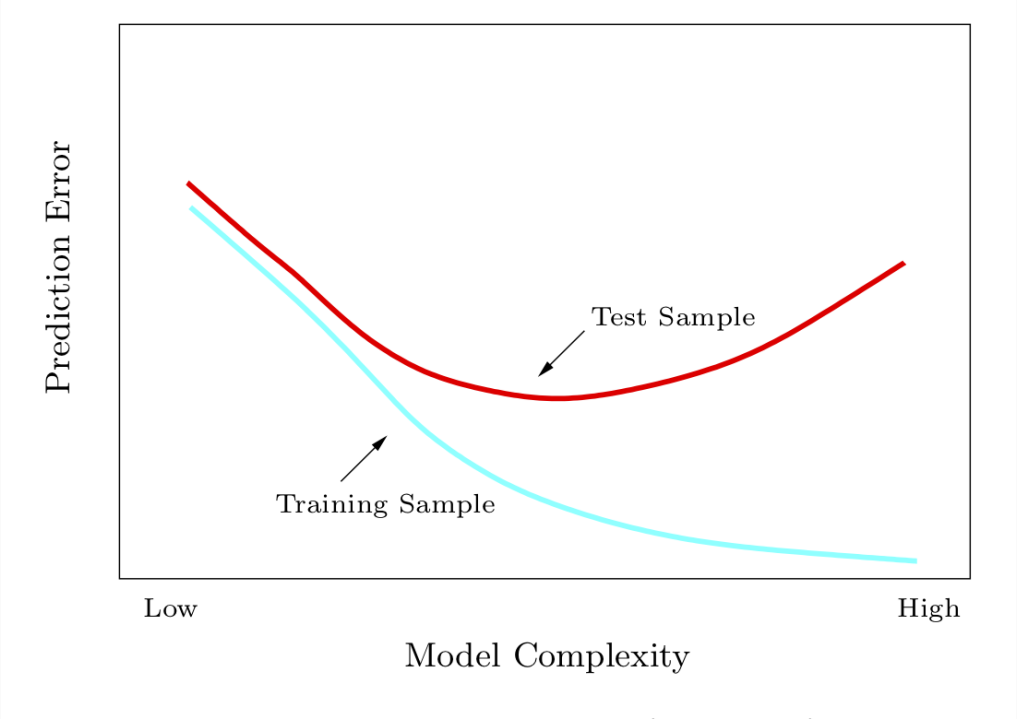
UTFSM Campus San Joaquín

Table of contents

Introducción

Propósito General

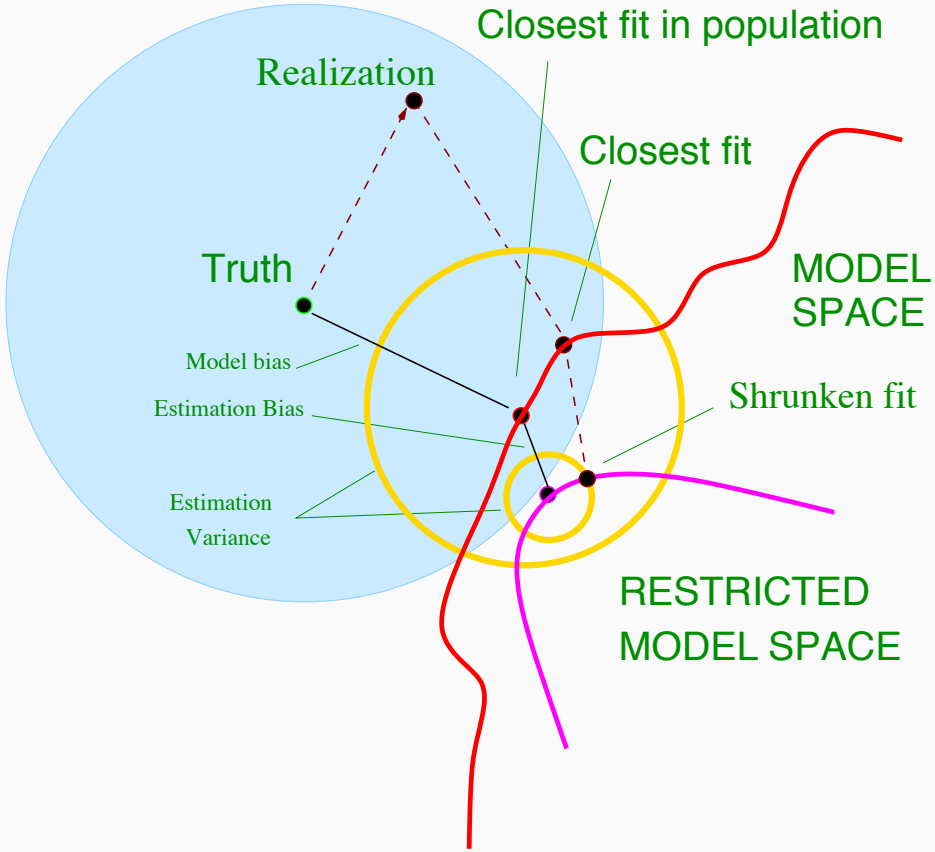
Regularización: Cualquier método que, modificando la forma en que se entrena el modelo, reduzca el riesgo de overfitting, mejorando la capacidad de generalización[1] del modelo obtenido.



[1] capacidad predictiva sobre datos de prueba (no vistos en fase de entrenamiento)

Regularización como Restricción del Espacio de Hipótesis

La inmensa mayoría de los métodos se basan en restringir la complejidad efectiva del espacio de hipótesis explorado por el modelo.



Punto de Vista Probabilistico/Bayesiano

Casi todos los criterios de entrenamiento que hemos revisado hasta el momento, se pueden formular como una métodos de máxima verosimilitud, que buscan optimizar

$$\ell(\theta) = \log P(S|\theta)\,, \tag{1}$$

donde θ denota los parámetros "libres" del modelo. Desde el punto de vista Bayesiano, un método de regularización aparece cuando decidimos optimizar,

$$P(\theta|\text{datos}) = \frac{P(\text{datos}|\theta)\,P(\theta)}{P(\text{datos})} \propto P(\text{datos}|\theta)\,P(\theta) \tag{2}$$

$$\text{a-posteriori} = \frac{\text{verosimilitud} \times \text{a-priori}}{\text{evidencia}}\,.$$

imponiendo un determinado a-priori $P(\theta)$ sobre el espacio de parámetros.

Punto de Vista Probabilistico/Bayesiano

Un a-priori $P(\theta)$ sobre el espacio de parámetros codifica una preferencia sobre las soluciones que debiésemos obtener del entrenamiento (e.g. soluciones suaves, soluciones dispersas, etc).

La gran mayoría de los métodos implementan a-prioris que utilizan una determinada norma sobre θ. Ejemplos clásicos:

- Regularización con la norma ℓ_2: $P(\theta) \propto \exp(-\frac{\lambda}{2}\|\theta\|_{\ell_2}^2)$.
- Regularización con la norma ℓ_1: $P(\theta) \propto \exp(-\frac{\lambda}{2}\|\theta\|_{\ell_1})$.

Alternativamente, es posible imponer un a-priori sobre el espacio de hipótesis directamente. Por ejemplo, $P(f) \propto \exp(-\frac{\lambda}{2}\|\frac{\partial f}{\partial \theta}\|_{\ell_2}^2)$

Regularización con la Normal ℓ_2

Idea Básica: A-priori Gausiano

- Consideremos un problema de aprendizaje con ejemplos $S = \{(x^{(\ell)}, y^{(\ell)})\}_{\ell=1}^{n}$ e hipótesis de la forma $f(x;\theta)$, donde $\theta \in \mathbb{R}^p$ denota el vector de parámetros del modelo.

- La regularización con la norma ℓ_2, también denominada regularización de Tikhonov, aparece cuando consideramos un a-priori $\mathcal{N}(0, \lambda^{-1})$ sobre θ, es decir, una distribución de la forma $P(\theta) \propto \exp(-\frac{\lambda}{2}\|\theta\|_2^2)$.

$\star$ **Intuitivamente, este a-priori codifica la preferencia de que, a no ser que los datos nos demuestren lo contrario, esperamos que muchos de los parámetros del modelo sean 0, es decir, no sean efectivamente utilizados en el modelo.**

Caso Clásico: Modelo Lineal

- Por ejemplo, en un modelo lineal de la forma $y = f(x) + \epsilon$, con $f(x) = w^T x = \sum_i w_i x_i + b$, el vector θ representa los d coeficientes que acompañan a cada uno de los atributos que hemos seleccionado para representar x y aprender y.

- Como $\theta_i = w_i = 0$ implica que el atributo i-ésimo puede ser ignorado completamente en el modelo, el a-priori $\mathcal{N}(0, \lambda^{-1})$ sobre θ, representa la preferencia por usar, en promedio, la menor cantidad de atributos posibles, es decir, por hacer una cuidadosa selección de características.

- El a-priori $\mathcal{N}(0, \lambda^{-1})$ sobre θ, representa también la preferencia de que ninguno de los atributos obtenga un coeficiente significativamente más grande que los demás[2].

[2] Este objetivo interacciona negativamente con el anterior, situación que motivará otro método de regularización.

Penalización de $\|\theta\|_2^2$

Si usamos los datos para optimizar el a-posteriori de θ con un a-priori de la forma $P(\theta) \propto \exp(-\frac{\lambda}{2}\|\theta\|_2^2)$, obtenemos un nuevo criterio de entrenamiento

$$\arg\max \log P(\theta|S) = \arg\max \log P(S|\theta) + \lambda \log P(\theta) \quad (3)$$

$$= \arg\max \ell(\theta) - \lambda\|\theta\|_2^2 \,. \quad (4)$$

- El criterio de entrenamiento tradicional ($\log P(S|\theta)$) tiene ahora un término adicional $\|\theta\|_2^2 = \sum_i \theta_i^2$ que penaliza magnitudes muy alejadas de 0 de alguno de los parámetros.
- El parámetro de regularización $\lambda > 0$ determina la relevancia del a-priori con respecto a la verosimilitud (típicamente $\propto$ error de entrenamiento)

Ridge Regression

- Por ejemplo, bajo los supuestos estándares, la función de log-verosimilitud correspondiente al modelo lineal $y = f(x; w) + \epsilon$ con $f(x; w) = w^T x + b$, toma la forma

$$\ell(w) = -\sum_{\ell} \left(y^{(\ell)} - f(x^{(\ell)}; w) \right)^2 . \tag{5}$$

- El criterio de entrenamiento regularizado toma la forma

$$\min J_\lambda(w) = \sum_{\ell} \left(y^{(\ell)} - f(x^{(\ell)}; w) \right)^2 + \frac{\lambda}{2} \|w\|^2 . \tag{6}$$

- El método obtenido se denomina ridge regression.
- Notemos que no se impone un a-priori sobre b.

Regresión Logística Regularizada

- Por ejemplo, la función de log-verosimilitud correspondiente al clasificador logístico $y = \sigma(f(x; w))$ con $f(x; w) = w^T x + b$, toma la forma

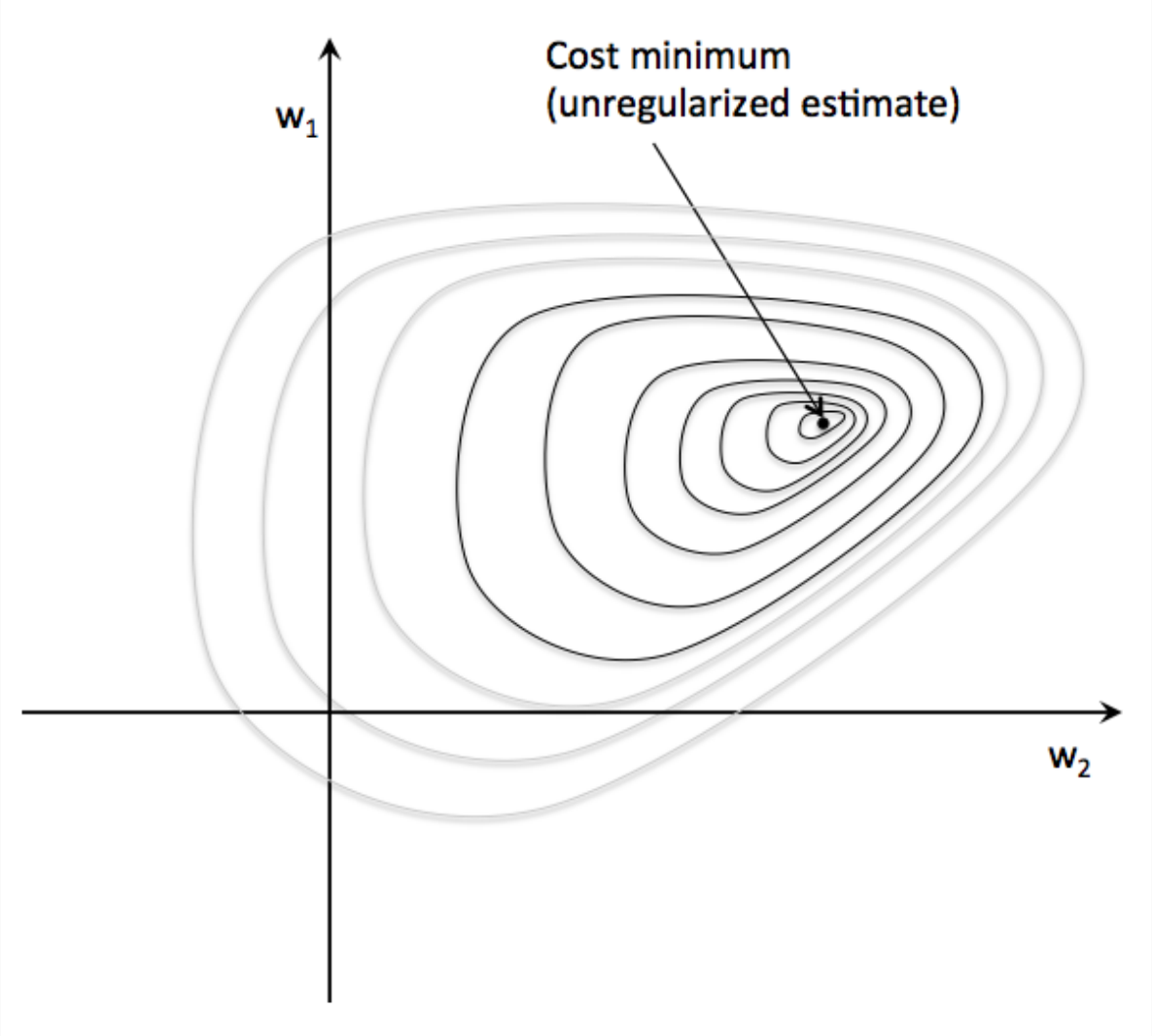
$$\begin{aligned} \ell(w) &= -\sum_{\ell} \left(y^{(\ell)} \sigma(f(x^{(\ell)}; w)) + (1 - y^{(\ell)})(1 - \sigma(f(x^{(\ell)}; w))) \right) \\ &= -\mathrm{KL}\left(y^{(\ell)} \middle|\middle| \sigma(f(x^{(\ell)}; w)) \right) . \end{aligned} \tag{7}$$

- El criterio de entrenamiento regularizado toma la forma

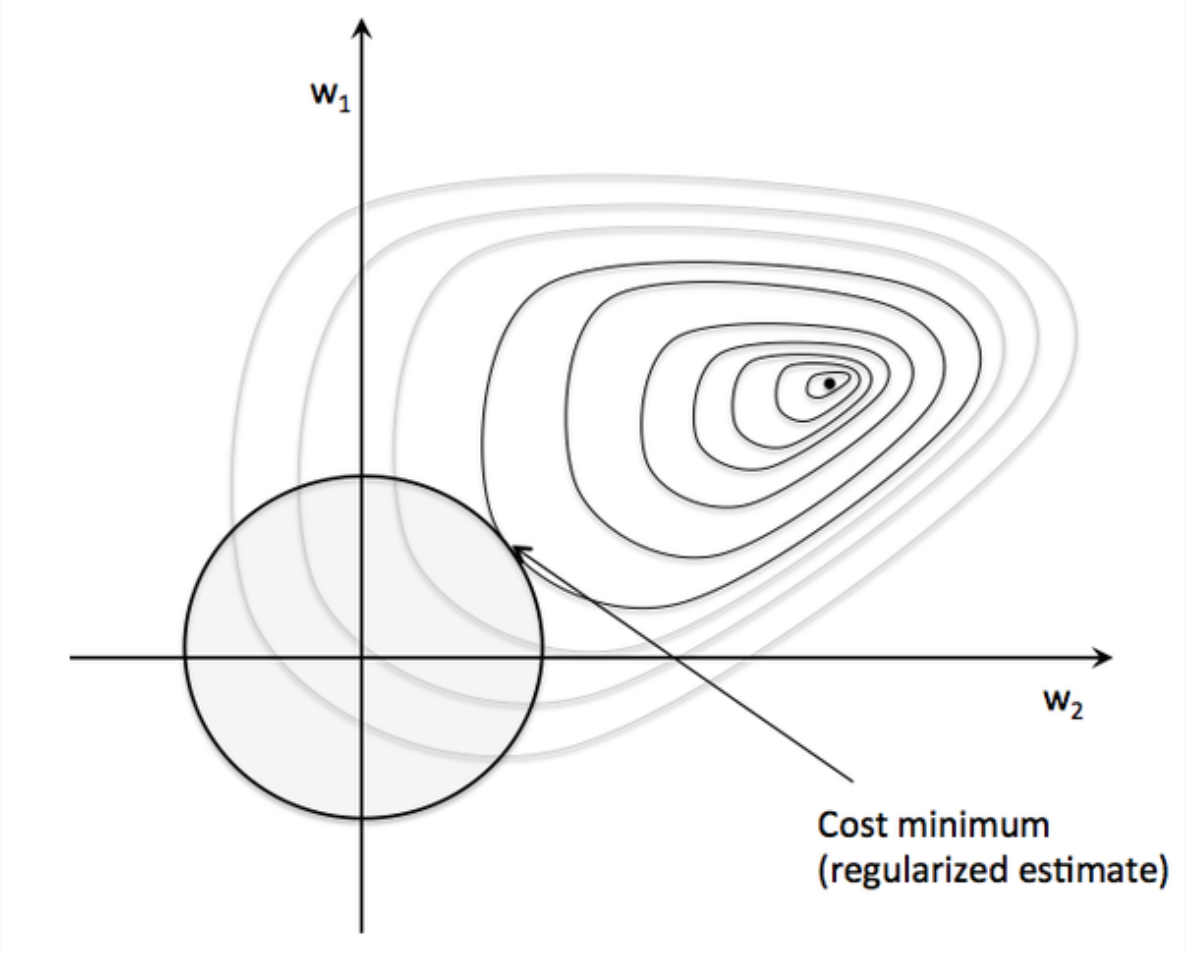
$$\min J_\lambda(w) = \mathrm{KL}\left(y^{(\ell)} \middle|\middle| \sigma(f(x^{(\ell)}; w)) \right) + \frac{\lambda}{2} \|w\|^2 . \tag{8}$$

- El método obtenido es en realidad la versión por defecto de regresión logística en la mayoría de las librerías modernas.

Efecto sobre el Aprendizaje



Efecto sobre el Aprendizaje



Efecto sobre el Aprendizaje

- Consideremos la siguiente versión genérica de la función de entrenamiento regularizada

$$\min J_\lambda(\theta) = J(\theta) + \frac{\lambda}{2}\|\theta\|^2 \,, \tag{9}$$

donde $J(\theta)$ es la función de entrenamiento no-regularizada, que asumiremos diferenciable (y preferentemente convexa).

- Sea θ^* el mínimo de $J(\theta)$ y consideremos una aproximación a segundo orden de $J(\theta)$ en torno a θ^*,

$$\begin{aligned} J(\theta) &\approx J(\theta^*) + (\theta - \theta^*)^T \nabla J(\theta^*) + (\theta - \theta^*)^T H(\theta^*)(\theta - \theta^*) \\ &= J(\theta^*) + (\theta - \theta^*)^T H(\theta^*)(\theta - \theta^*) \,, \end{aligned} \tag{10}$$

donde $\nabla J_i(\theta) = \partial J/\partial\theta_i$ y $H_{ij}(\theta) = \partial^2 J/\partial\theta_i\partial\theta_j$ (notar que ambos están evaluados en $\theta = \theta^*$).

Efecto sobre el Aprendizaje

- El mínimo θ_λ de la función de entrenamiento regularizada debe satisfacer $\nabla J_\lambda(\theta_\lambda) = 0$. Por lo tanto,

$$\nabla J_\lambda(\theta_\lambda) = \nabla J(\theta_\lambda) + \lambda\theta = 0\,, \tag{11}$$

es decir, $(H^* + \lambda I)\,\theta_\lambda = H^*\theta^*$, con $H^* = H(\theta^*)$. Por lo tanto,

$$\theta_\lambda = (H^* + \lambda I)^{-1}\, H^*\theta^*\,, \tag{12}$$

- Consideremos primero el caso en que $H^* = D = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2)$,

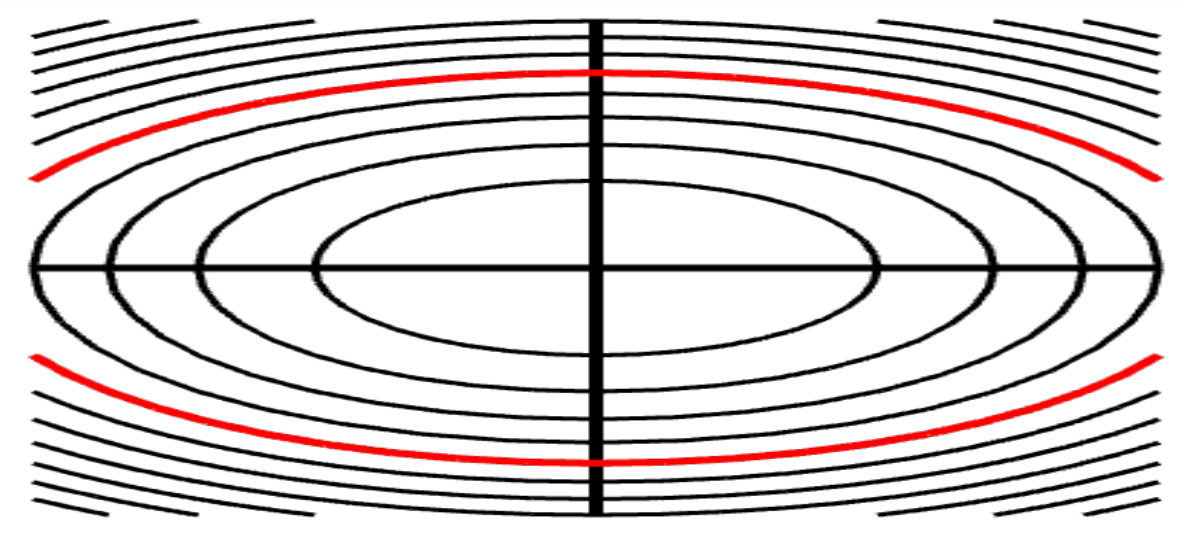
$$\theta_\lambda = (D + \lambda I)^{-1}\, D\theta^* \tag{13}$$

$$\Leftrightarrow \theta_{\lambda,i} = \left(\frac{\sigma_i^2}{\sigma_i^2 + \lambda}\right)\theta_i^*\,.$$

Efecto sobre el Aprendizaje

Si $\lambda > k\sigma_i^2$, tenemos

$$\theta_{\lambda,i} < \left(\frac{1}{k+1}\right)\theta_i^* \,.$$



• Las direcciones más "podadas" por el regularizador son aquellas que corresponden a un valor pequeño de σ_i^2: direcciones del espacio de parámetros donde el objetivo de entrenamiento no varia mucho.

Efecto sobre el Aprendizaje

- Tomando ahora una EVD de H^*, $H^* = VDV^T$, obtenemos

$$\theta_\lambda = V\left(D + \lambda I\right)^{-1} DV^T\theta^* \tag{14}$$

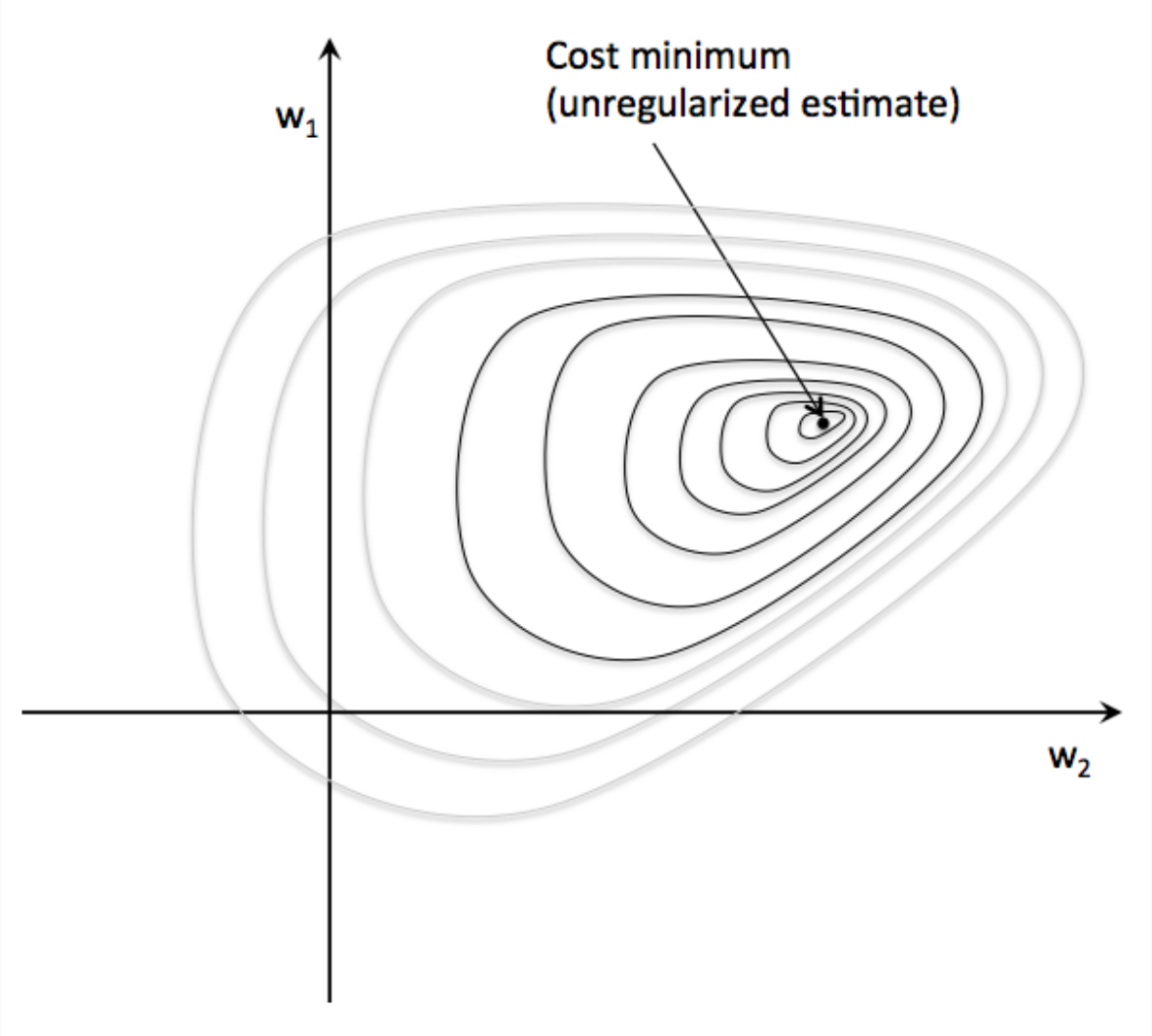
$$\theta_\lambda = \sum_j v_j \left(\frac{\sigma_j^2}{\sigma_j^2 + \lambda}\right) v_j^T\theta^*\,,$$

- Comparando con la solución no regularizada (θ_0)

$$\theta_0 = \sum_j v_j v_j^T\theta^*\,, \tag{15}$$

observamos que el regularizador "poda" (shrink) más significativamente las componentes de θ^* en las direcciones v_j correspondientes a un bajo valor de σ_j^2 (direcciones de la función objetivo que varían menos significativamente).

Efecto sobre el Aprendizaje



Ridge Regression

- En el caso ridge regression, la f.o. no regularizada es de segundo orden

$$J(w) = \sum_{\ell} \left(y^{(\ell)} - f(x^{(\ell)}; w) \right)^2 , \tag{16}$$

de modo que la aproximación (hecha para el análisis) es exacta.

- Para visualizar lo que sucede en este caso, recordemos que $f(x^{(\ell)}; w) = w^T x + b$. Podemos eliminar b del modelo, asumiendo que $\bar{x} = \frac{1}{n} \sum_{\ell} x^{(\ell)} = 0$ y $\bar{y} = 0$. En este caso, el estimador MV de b es $\hat{b} = 0$.

- Con el supuesto anterior, $J(w)$ toma la forma matricial

$$J(w) = (Xw - Y)^2 , \tag{17}$$

donde $X_{(\ell)} = x^{(\ell)}$ y $Y_{(\ell)} = y^{(\ell)}$.

Ridge Regression

- La matriz se segundas derivadas es entonces

$$H = X^T X \,. \tag{18}$$

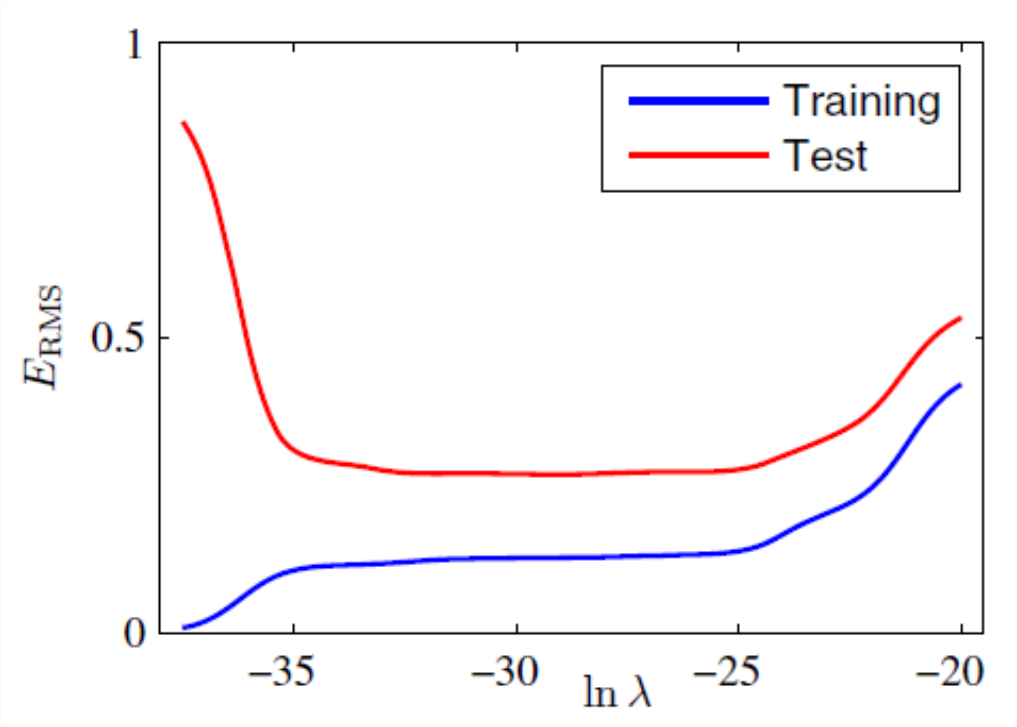
- En el caso de predictores ortogonales, $X^T X = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, con $\sigma_i^2 \propto \text{var}(X^{(i)})$. Por lo tanto, la solución regularizada

$$\theta_{\lambda,i} = \left(\frac{\sigma_i^2}{\sigma_i^2 + \lambda} \right) \theta_i^* \,.$$

preserva mejor los predictores que varian más significativamente entre el conjunto de ejemplos.

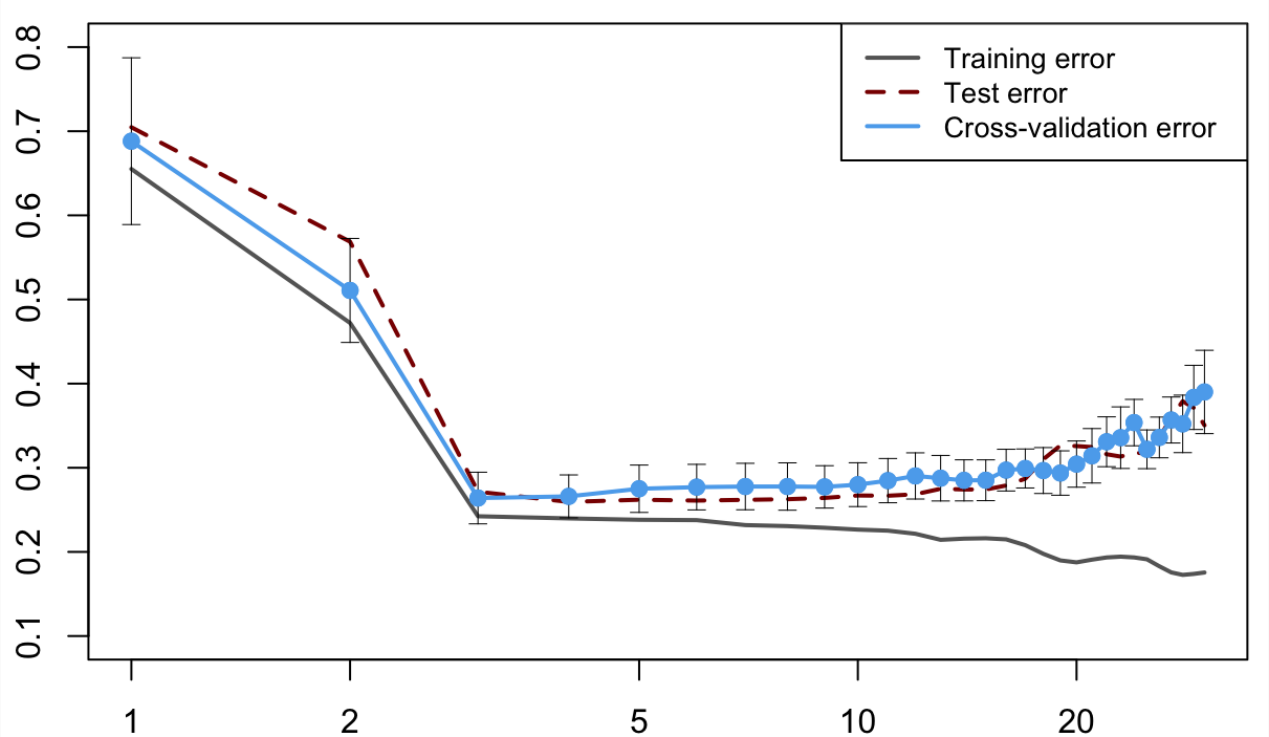
Elección del Parámetro de Regularización

- El parámetro de regularización $\lambda > 0$ determina el grado de "poda" (shrinkage) que sufren los coeficientes de la solución no regularizada.
- A un mayor de λ corresponde siempre un mayor (o igual) error de entrenamiento (valor mayor o igual de $J(\theta)$).



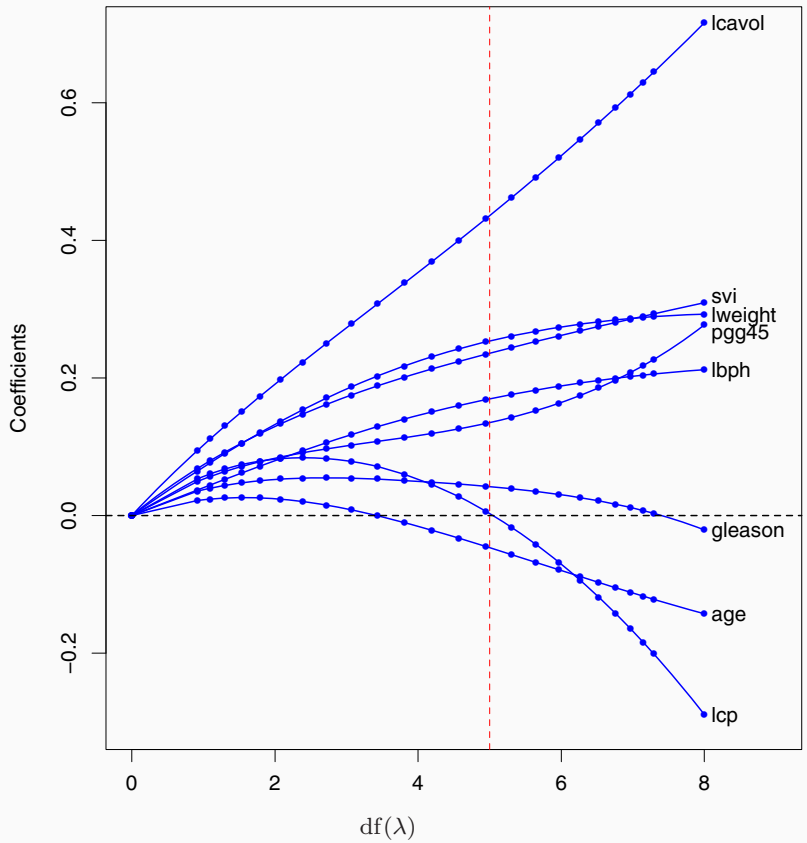
Elección del Parámetro de Regularización

- Por lo tanto la elección debe estar guiada por el error de predicción del modelo regularizado. En la práctica se utiliza un predictor del error de test correspondiente a θ_λ y se elige el valor λ que minimiza esa estimación.



Elección del Parámetro de Regularización

- Existen algoritmos incrementales capaces de encontrar eficientemente las soluciones θ_λ correspondientes a todo un rango de posibles valores de λ. Estos algoritmos permiten visualizar el efecto sobre los coeficientes del modelo, información que puede ser usada para seleccionar atributos.



Aprendizaje vía Gradiente

- Recordemos que una buena parte de los algoritmos de entrenamiento usados en aprendizaje automático están basados en gradiente descendente/ascendente.

Algorithm 1: Gradiente descendente para minimizar $J(\theta)$

$\theta \leftarrow \theta_{\text{init}}$
do
  $\theta \leftarrow \theta - \eta_t \nabla J(\theta)$
while *not convergence*;

- Este algoritmo se puede adaptar fácilmente para introducir el regularizador.

Aprendizaje vía Gradiente

- El algoritmo modificado toma la forma

Algorithm 2: Gradiente Descendente para Minimizar $J(\theta)$

$\theta \leftarrow \theta_{\text{init}}$
do
  $\theta \leftarrow \theta - \eta_t \left(\nabla J(\theta) + \lambda\theta\right)$
while *not convergence*;

- La actualización de θ también se puede escribir
$\theta \leftarrow (1 - \eta_t\lambda)\theta - \eta_t \nabla J(\theta)$.

Regularización con la Normal ℓ_1

Problema de la Regularización ℓ_2

- En la práctica, la regularización con la norma ℓ_2 no genera soluciones verdaderamente dispersas.

- Para entender porqué, notemos primero que , bajo ciertas condiciones de regularidad[3] el criterio de entrenamiento regularizado,

$$\min J_\lambda(\theta) = J(\theta) + \frac{\lambda}{2}\|\theta\|^2 \,, \tag{19}$$

también se puede escribir

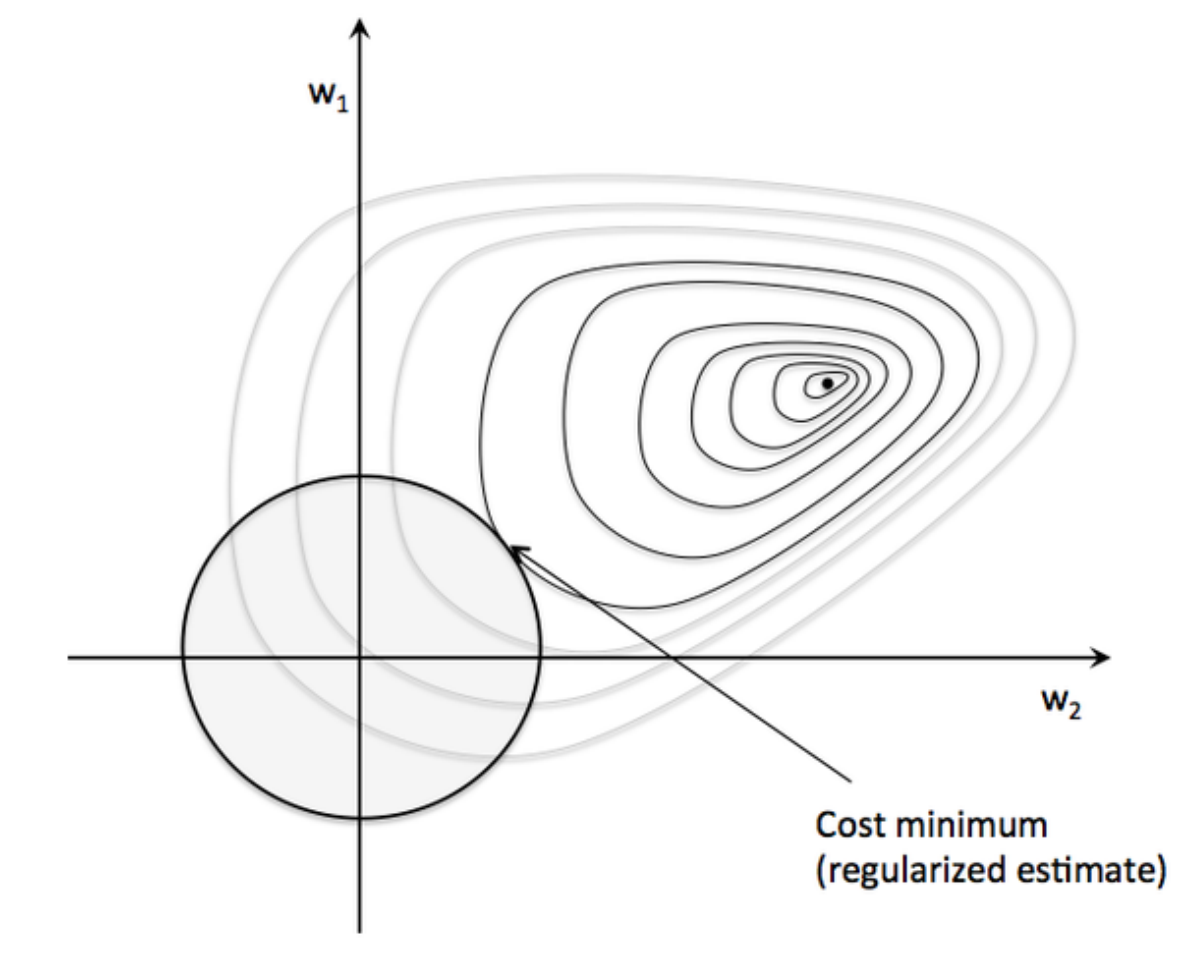
$$\min J(\theta) \text{ s.t. } \|\theta\|^2 < t_\lambda \,, \tag{20}$$

en el sentido de que $\forall\lambda$, $\exists t_\lambda$ tal que la solución θ_{t_λ} del segundo problema es la solución óptima del primero.

[3] Para obtener la equivalencia se requieren las mismas condiciones que hacen suficientes y necesarias las condiciones de Karush-Kuhn-Tucker (KKT).

Problema de la Regularización ℓ_2

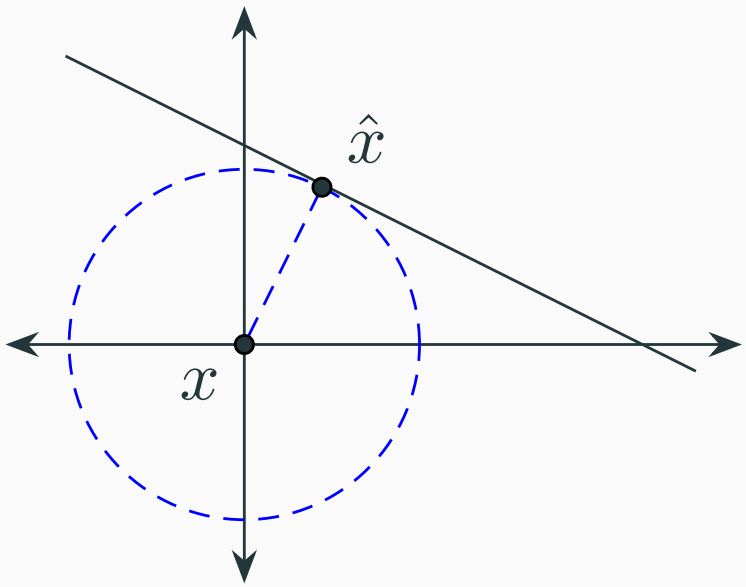
- La restricción $\|\theta\|^2 < \gamma_\lambda$ es satisfecha por una infinidad de soluciones, muchas no dispersas. En efecto, de entre las soluciones factibles, aquella que maximiza la f.o. no regularizada será muy frecuentemente no dispersa.



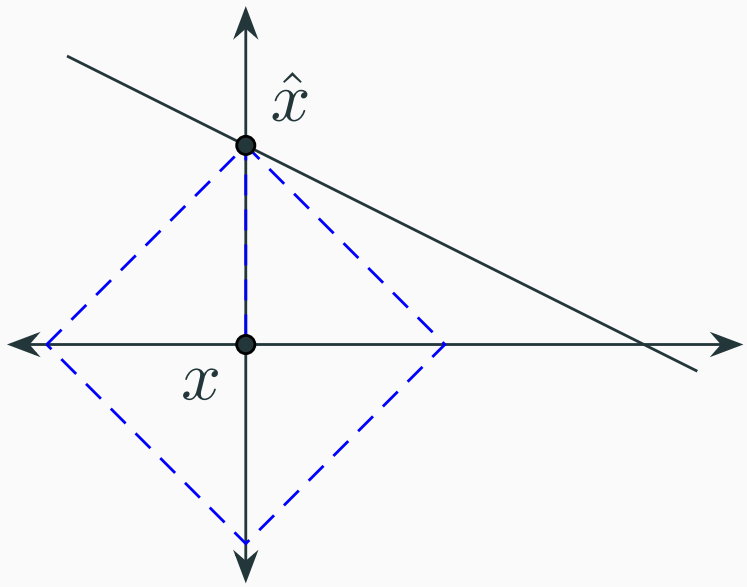
Normas

Idea: Considerar otra norma para θ!

ℓ_p **norms:** $\|\theta\|_{\ell_p} = \sqrt[p]{\sum_i \theta_i^p}$ (cuando $p \in [0, 1)$, pseudo-normas).



$p = 2$

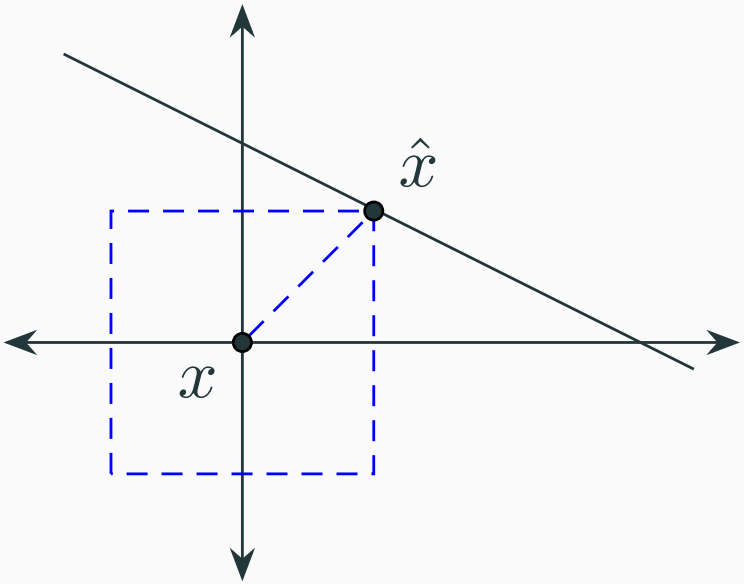


$p = 1$

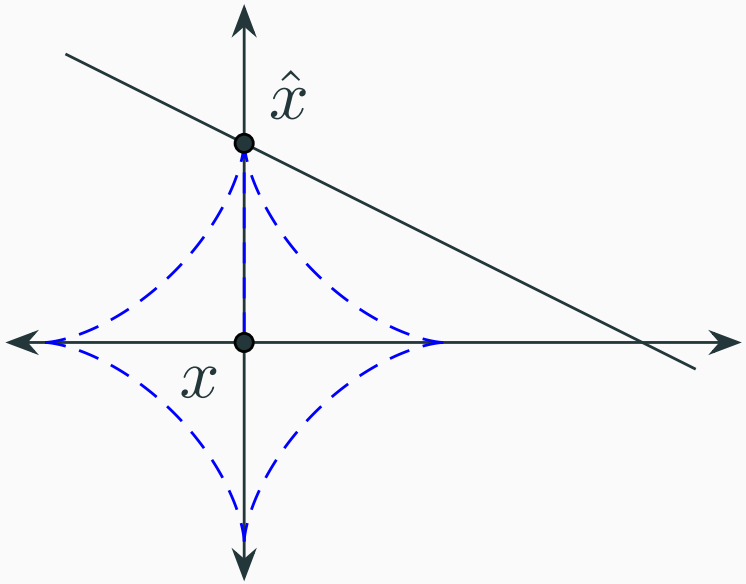
Normas

Idea: Considerar otra norma para θ!

ℓ_∞ **norm:** $\|\theta\|_\infty = \max_i \theta_i$.



$p = \infty$

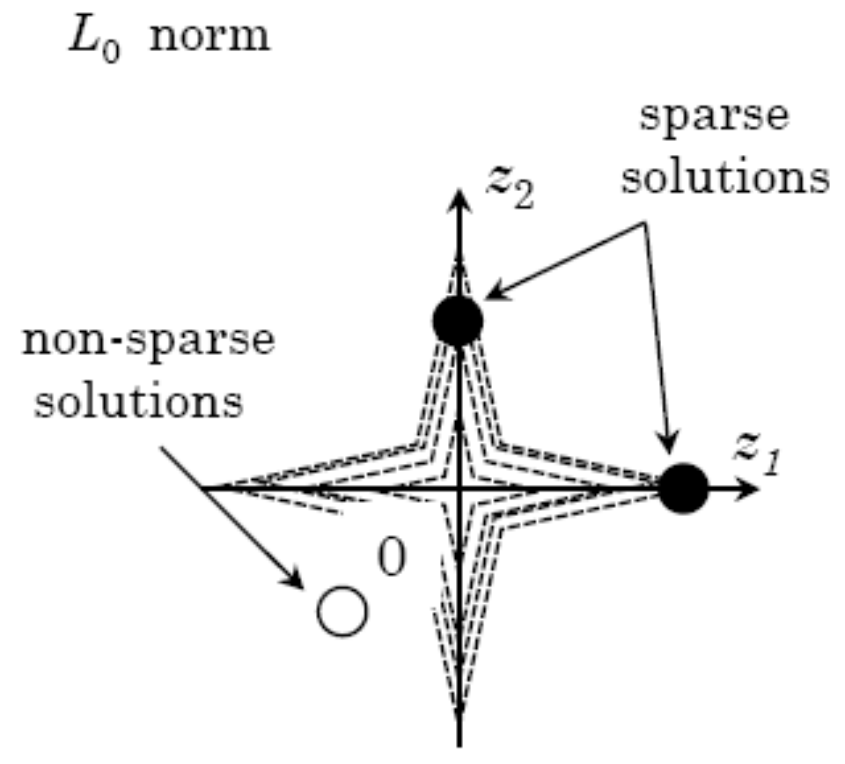


$p = \frac{1}{2}$

Norma Ideal

Idea: La norma que nos conviene considerar si queremos que θ tenga muchas componentes nulas (sea dispersa) es la norma ℓ_0.

ℓ_0 **norm:** $\|\theta\|_0 = \sum_i I(\theta_i > 0)$.



Regularizador Ideal

- Con esta elección, el criterio de entrenamiento regularizado quedaría

$$\min J_\lambda(\theta) = J(\theta) + \lambda \|\theta\|_0 \,, \tag{21}$$

o bien (si preferimos la versión restringida a la penalizada)

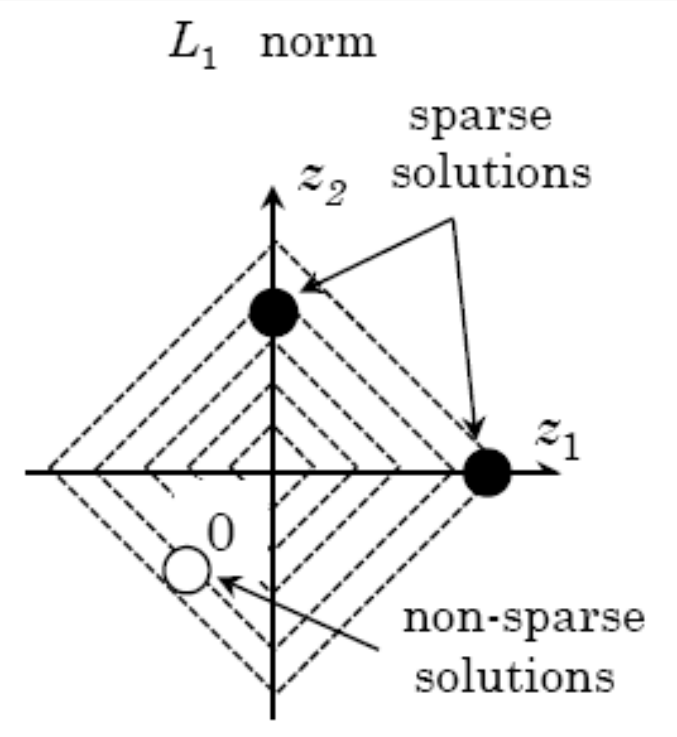
$$\min J(\theta) \text{ s.t. } \|\theta\|_0 \leq \gamma_\lambda \,, \tag{22}$$

- Lamentablemente, el problema de optimización que aparece es combinatorial y NP-duro.

Idea: Considerar una norma más cercana a la ℓ_0, pero que origine un problema más fácil de resolver.

Regularización con la norma ℓ_1

• De entre las normas ℓ_p, la norma ℓ_1 es la más "cercana" a la norma ℓ_0 que cumple con ser convexa (gran ventaja desde el punto de vista de la optimización).



Regularización con la norma ℓ_1

- Con esta elección, el criterio de entrenamiento regularizado quedaría

$$\min J_\lambda(\theta) = J(\theta) + \lambda\|\theta\|_1 \,, \tag{23}$$

Bajo ciertas condiciones de regularidad[4] el problema resultante es equivalente a

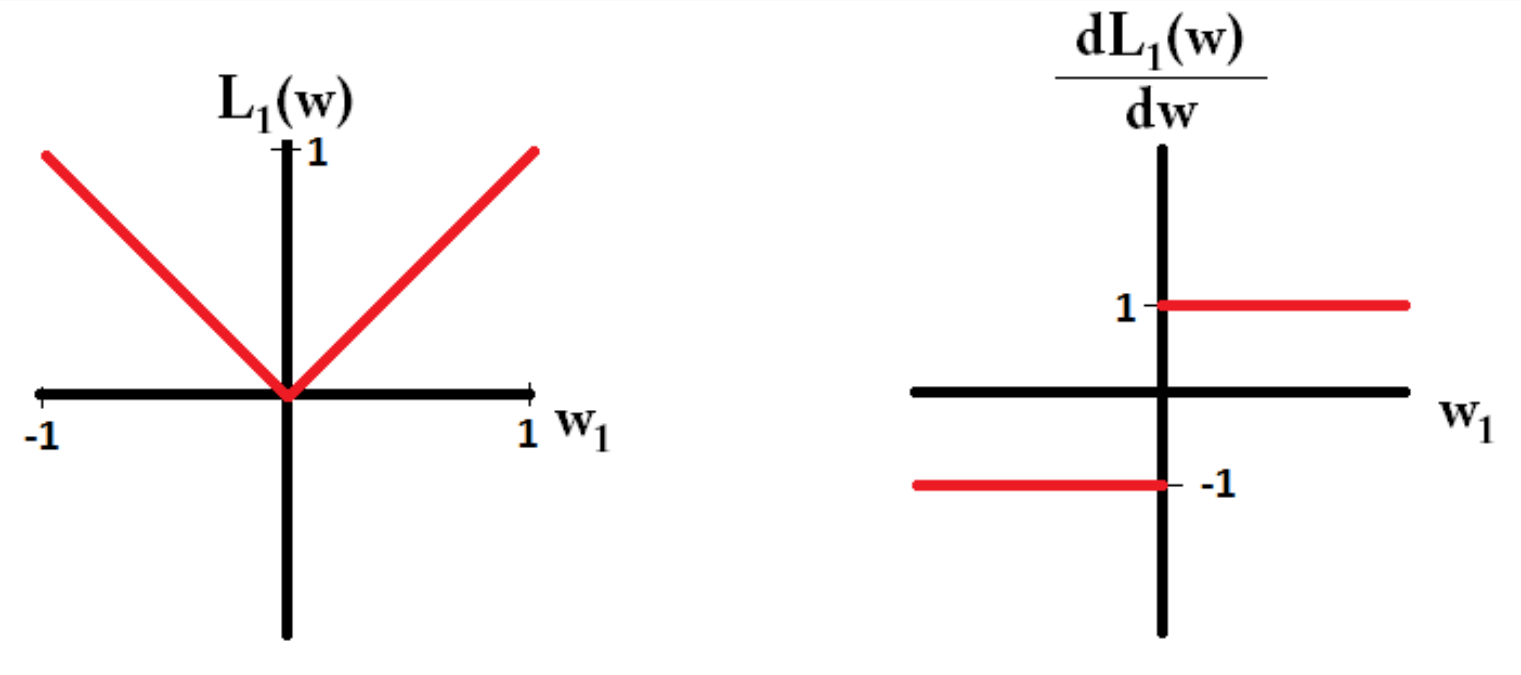
$$\min J(\theta) \text{ s.t. } \|\theta\|_1 < t_\lambda \,, \tag{24}$$

en el sentido de que $\forall\lambda$, $\exists t_\lambda$ tal que la solución θ_{t_λ} del segundo problema es la solución óptima del primero.

[4] Para obtener la equivalencia se requieren las mismas condiciones que hacen suficientes y necesarias las condiciones de Karush-Kuhn-Tucker (KKT). Notemos que como la norma ℓ_1 es convexa, la convexidad de $J(\theta)$, unida a algunas condiciones técnicas fáciles de garantizar, asegura la aplicabilidad de las KKT.

Complicación asociada a la norma ℓ_1

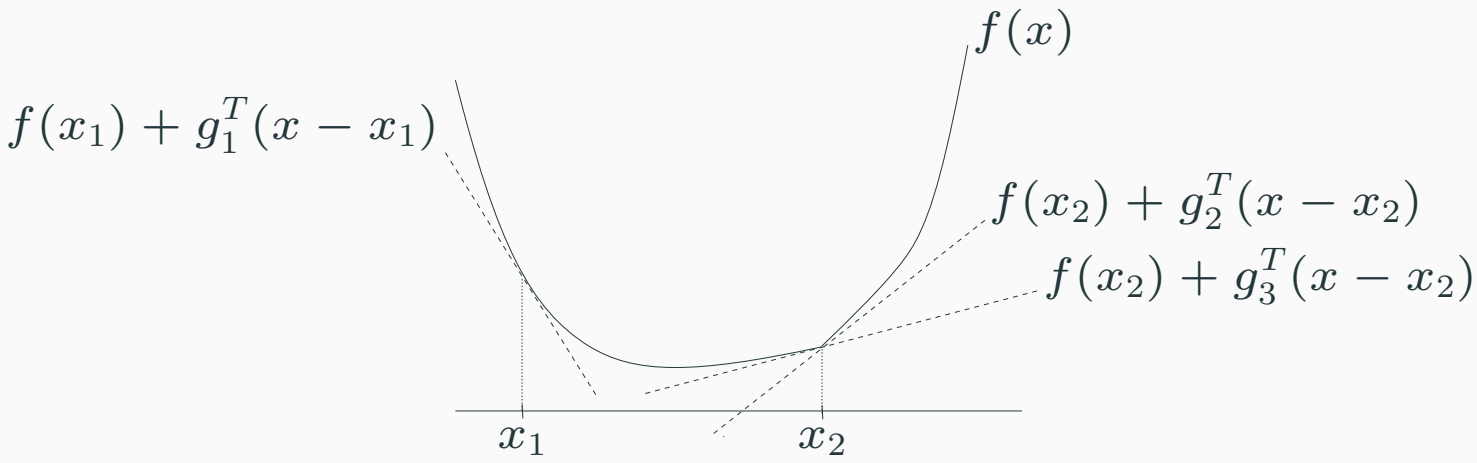
• Una pequeña complicación asociada a la norma ℓ_1 es que no resulta diferenciable en cualquier punto.



• Esto hace que sea ligeramente más complicado expresar las condiciones de optimalidad y que haya que adaptar los algoritmos de aprendizaje basados en gradiente.

Sub-gradiente

- Afortunadamente, si una función no es diferenciable, podría ser todavía sub-diferenciable.



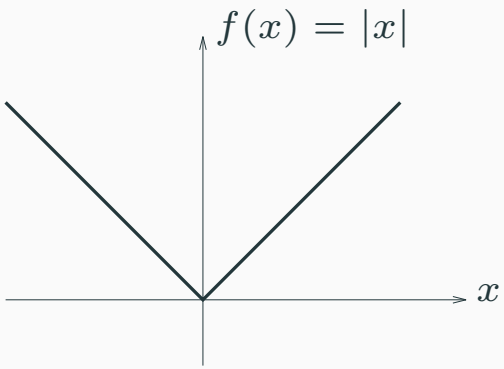
- Una función convexa $h(w)$ se dice sub-diferenciable en w_0 si existe g tal

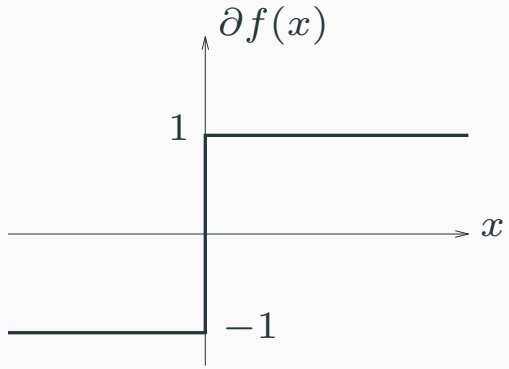
$$h(w) \geq h(w_0) + g^T(w - w_0) \, \forall w \, . \tag{25}$$

En este caso g se denomina un sub-gradiente de h en w_0. El conjunto de todos los sub-gradientes de h en w se anota $\partial h(w)$ y se denomina el sub-diferencial.

Sub-gradiente

- Por ejemplo, la función $h(w) = |w|$ no es diferenciable, pero es sub-diferenciable.





- En el caso de la norma ℓ_1, $\partial \|w\|_1 = \sum_i \partial |w_i|$, con

$$\partial |w_i| = \begin{cases} +1 & w_i > 0 \\ -1 & w_i < 0 \\ [-1, 1] & w_i = 0 \end{cases} \tag{26}$$

Sub-gradiente

- Muchos algoritmos basados en gradiente se pueden adaptar substituyendo los gradientes por sub-gradientes (al costo de una menor tasa/velocidad de convergencia y la obtención de un algoritmo altamente no monótono que obliga a mantener la mejor solución en cache).
- Las condiciones de optimalidad clásicas se pueden adaptar al caso sub-diferenciable. En efecto, para una función convexa, $0 \in \partial h(w)$ si y sólo si w es el mínimo de h.

Efecto sobre el Aprendizaje

- Estamos ahora en condiciones de repetir el análisis que hicimos para el caso de la norma ℓ_2.
- Consideremos entonces la función de entrenamiento regularizada

$$\min J_\lambda(\theta) = J(\theta) + \lambda \|\theta\|_1 \,, \tag{27}$$

donde $J(\theta)$ es la función de entrenamiento no-regularizada, que asumiremos diferenciable y convexa.

- Sea θ^* el mínimo de $J(\theta)$ y consideremos una aproximación a segundo orden de $J(\theta)$ en torno a θ^*,

$$J(\theta) \approx J(\theta^*) + (\theta - \theta^*)^T \, H(\theta^*) \, (\theta - \theta^*) \,.$$

Efecto sobre el Aprendizaje

- El mínimo θ_λ de la función de entrenamiento regularizada debe satisfacer $0 \in \partial J_\lambda(\theta_\lambda)$. Por lo tanto,

$$\partial J_\lambda(\theta_\lambda) = \nabla J(\theta_\lambda) + \lambda \partial \Omega(\theta_\lambda) = 0\,, \tag{28}$$

para algún sub-diferencial $\partial\Omega(\theta)$ de $\|\theta\|_1$.

- Obtenemos $H^*\theta_\lambda + \lambda\partial\Omega(\theta_\lambda) = H^*\theta^*$, con $H^* = H(\theta^*)$.

- Consideremos el caso en que $H^* = D = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2)$, que permite hacer un análisis componente a componente (notemos que $\sigma_i^2 > 0$ porque asumimos $J(\theta)$ convexa)

$$\sigma_i^2 \theta_{\lambda,i} + \lambda \partial\Omega_i(\theta_{\lambda,i}) = \sigma_i^2 \theta_i^*\,. \tag{29}$$

con $\partial\Omega_i$ la i-ésima componente de $\partial\Omega$, que es simplemente el sub-diferencial de $|\theta_i|$.

Efecto sobre el Aprendizaje

- Notemos que, con $H^* = D$,

$$J_\lambda(\theta) \approx J(\theta^*) + \sum_i \sigma_i^2(\theta_i^* - \theta_i)^2 + \sum_i |\theta_i| \,, \tag{30}$$

que nos permite ver claramente que los signos de θ_i^* y $\theta_{\lambda,i}$ deben coincidir o $\theta_{\lambda,i}$ debe ser cero [5].

- Consideremos entonces el caso en que $\theta_i^* > 0$ y $\theta_{\lambda,i} > 0$. En este caso, la condición de optimalidad se transforma en

$$\begin{aligned} \sigma_i^2\theta_{\lambda,i} + \lambda\partial\Omega_i(\theta_{\lambda,i}) &= \sigma_i^2\theta_i^* \qquad (31)\\ \sigma_i^2\theta_{\lambda,i} + \lambda &= \sigma_i^2\theta_i^* \\ \Rightarrow \theta_{\lambda,i} &= \theta_i^* - \frac{\lambda}{\sigma_i^2} \,. \end{aligned}$$

[5] Si sucede por ejemplo que $\theta_i^* > 0$ y $\theta_{\lambda,i} < 0$, podríamos cambiar el signo de $\theta_{\lambda,i} < 0$ obteniendo el mismo valor de $|\theta_i|$ y un valor más pequeño de $\sigma_i^2(\theta_i^* - \theta_i)^2$ ya que θ_i^* y θ_i estarían del mismo lado del 0. Es decir, si sucede que $\theta_i^* > 0$ y $\theta_{\lambda,i} < 0$, podríamos cambiar el signo de $\theta_{\lambda,i} < 0$ obteniendo un menor valor de $J_\lambda(\theta)$. Esto contradice que θ_λ sea el mínimo de $J_\lambda(\theta)$.

Efecto sobre el Aprendizaje

- Como podríamos tener $\theta_i^* > 0$ y $\theta_{\lambda,i} = 0$, el caso $\theta_i^* > 0$ se resume en

$$\theta_{\lambda,i} = \max\left(\theta_i^* - \frac{\lambda}{\sigma_i^2}, 0\right) = \max\left(|\theta_i^*| - \frac{\lambda}{\sigma_i^2}, 0\right) \tag{32}$$

$$= \text{sign}(\theta_i^*)\ \max\left(|\theta_i^*| - \frac{\lambda}{\sigma_i^2}, 0\right)\ . \tag{33}$$

- El caso en que $\theta_i^* < 0$ y $\theta_{\lambda,i} < 0$ es similar,

$$\sigma_i^2\theta_{\lambda,i} - \lambda\partial\Omega_i(\theta_{\lambda,i}) = \sigma_i^2\theta_i^* \tag{34}$$

$$\sigma_i^2\theta_{\lambda,i} - \lambda = \sigma_i^2\theta_i^*$$

$$\Rightarrow \theta_{\lambda,i} = \theta_i^* + \frac{\lambda}{\sigma_i^2}\ .$$

Efecto sobre el Aprendizaje

- Como podríamos tener $\theta_i^* < 0$ y $\theta_{\lambda,i} = 0$, el caso $\theta_i^* < 0$ se resume en

$$\theta_{\lambda,i} = \min\left(\theta_i^* + \frac{\lambda}{\sigma_i^2}, 0\right) = -\max\left(-\theta_i^* - \frac{\lambda}{\sigma_i^2}, 0\right) \tag{35}$$

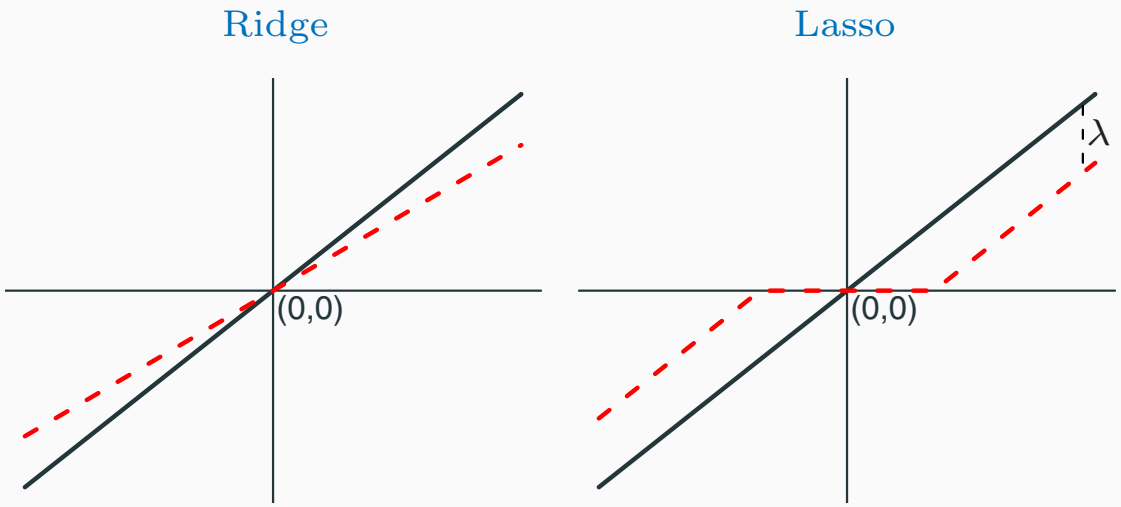
$$= \text{sign}(\theta_i^*)\ \max\left(|\theta_i^*| - \frac{\lambda}{\sigma_i^2}, 0\right)\ . \tag{36}$$

- Si $\theta_i^* = 0$, claramente $\theta_{\lambda,i} = 0$, por lo que todos los casos se resumen en

$$\theta_{\lambda,i} = \text{sign}(\theta_i^*)\ \max\left(|\theta_i^*| - \frac{\lambda}{\sigma_i^2}, 0\right)\ . \tag{37}$$

Regularización ℓ_1 vs ℓ_2

En el caso ortogonal vemos claramente la diferencia entre regularizar con la norma ℓ_2 (izquierda) vs la norma ℓ_2 (derecha).



El operador

$$T_\gamma(\theta) = \text{sign}(\theta)\ \max\left(|\theta| - \gamma, 0\right)\ , \tag{38}$$

se denomina **soft thresholding operator**.

Lasso

- En el caso regression lineal, la f.o. no regularizada toma la forma

$$J(w) = \sum_{\ell} \left(y^{(\ell)} - f(x^{(\ell)}; w) \right)^2 , \tag{39}$$

de modo que la regresión regularizada con la norma ℓ_1 correspondería a minimizar la f.o.

$$J_\lambda(w) = \sum_{\ell} \left(y^{(\ell)} - f(x^{(\ell)}; w) \right)^2 + \lambda \|w\|_{\ell_1} \tag{40}$$

$$= \sum_{\ell} \left(y^{(\ell)} - f(x^{(\ell)}; w) \right)^2 + \lambda \sum_i |w_i| . \tag{41}$$

- Esta forma de regresión lineal se denomina **lasso**.

Regularización vs Feature Selection

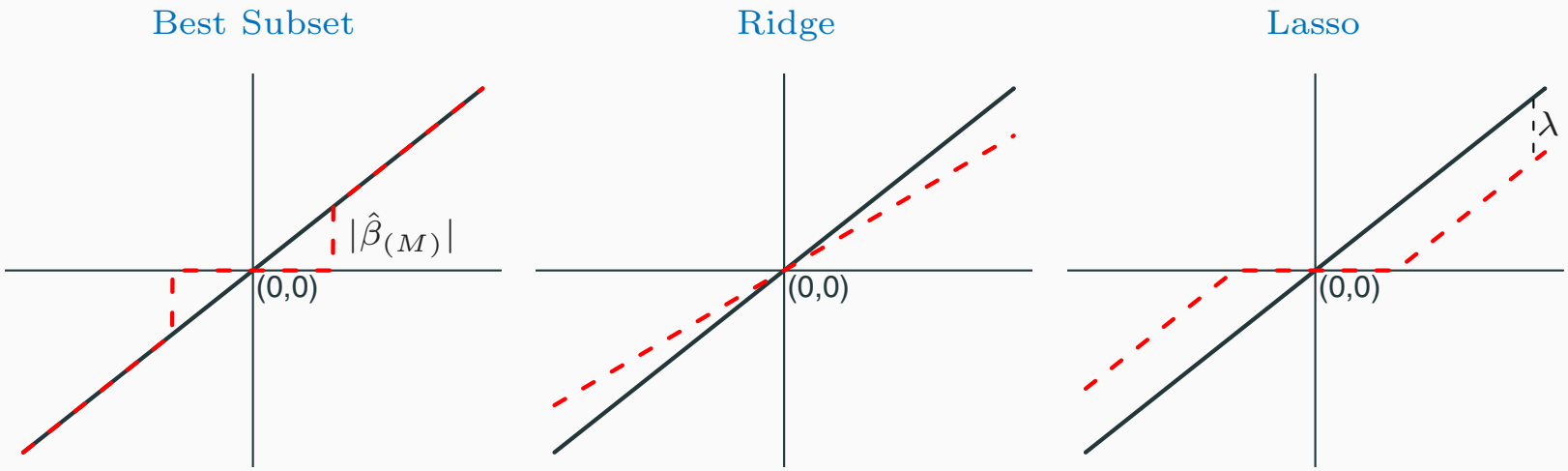
Notemos que en el caso ortogonal, la solución de la regresión lineal no regularizada toma la forma

$$w^* = (X^T X)^{-1} X^T Y \Rightarrow w_i^* = \frac{X_{(i)}^T Y}{\sigma_i^2} \tag{42}$$

Remover el coeficiente (en un proceso de selección de atributos) es equivalente a aplicar el soft thresholding operator con $\gamma = w_i^*$. Es decir, lasso se puede considerar la versión continua de un proceso discreto de selección de atributos que tiene la capacidad de "podar" de manera suave el coeficiente, hasta eventualmente eliminarlo.

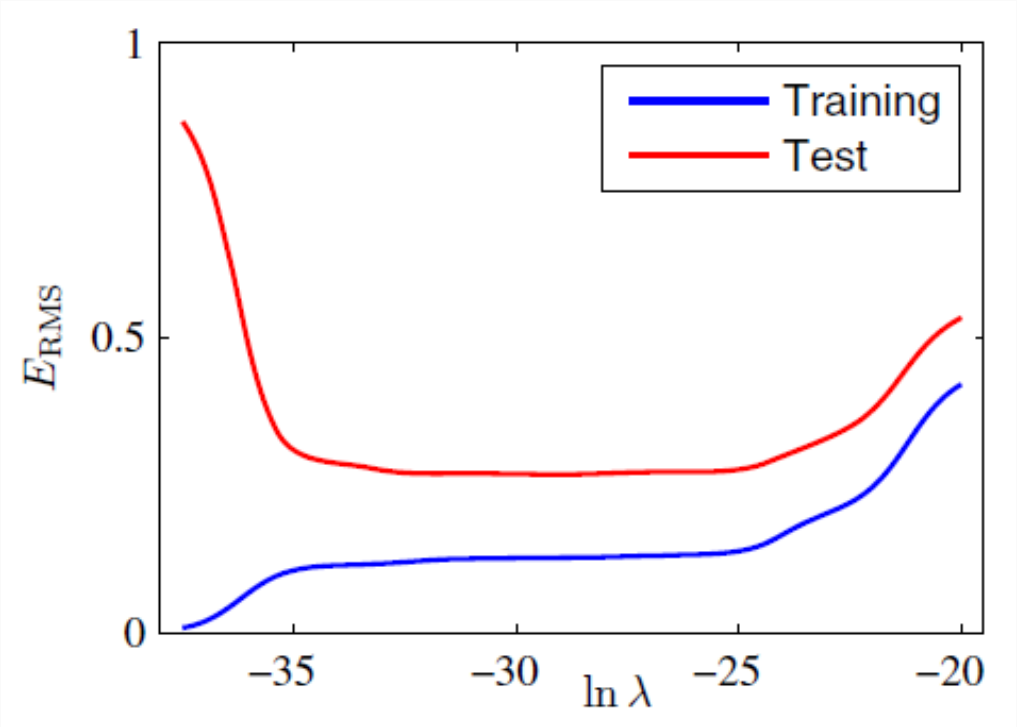
Regularización vs Feature Selection

Gráficamente



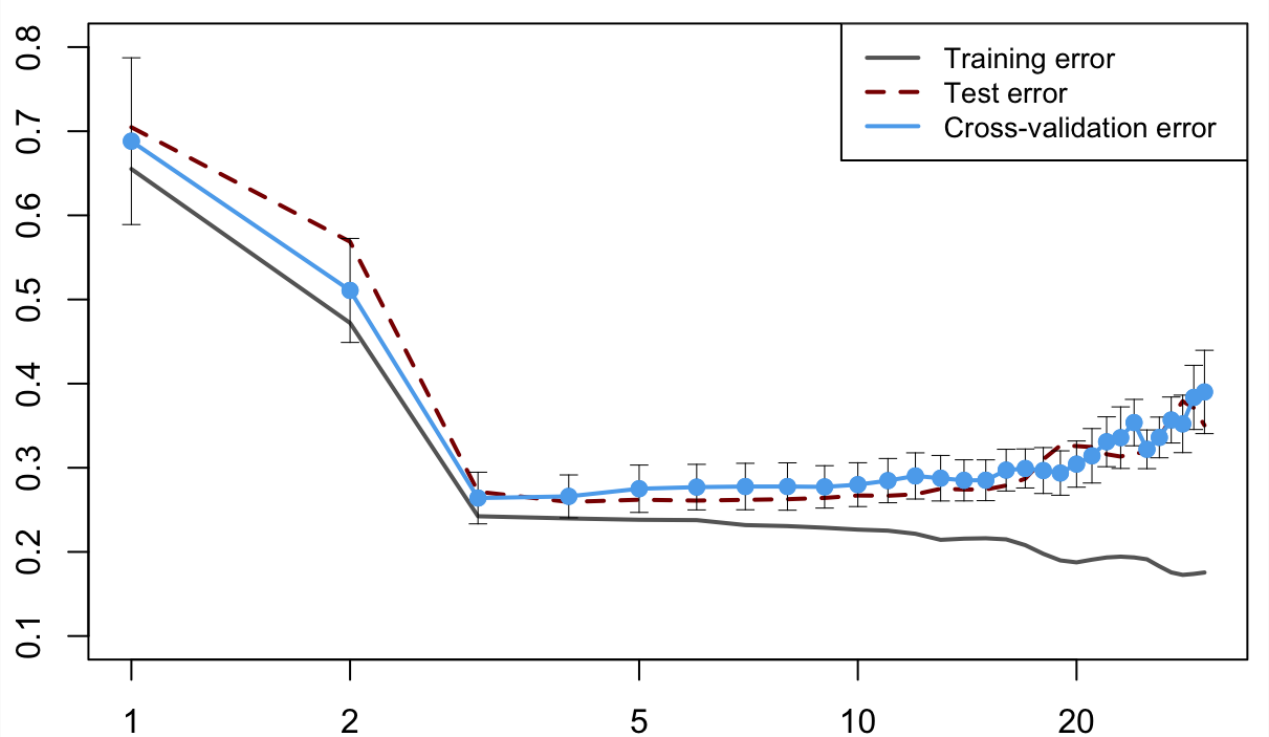
Elección del Parámetro de Regularización

- El parámetro de regularización $\lambda > 0$ determina el grado de "poda" (shrinkage) que sufren los coeficientes de la solución no regularizada.
- A un mayor de λ corresponde siempre un mayor (o igual) error de entrenamiento (valor mayor o igual de $J(\theta)$).



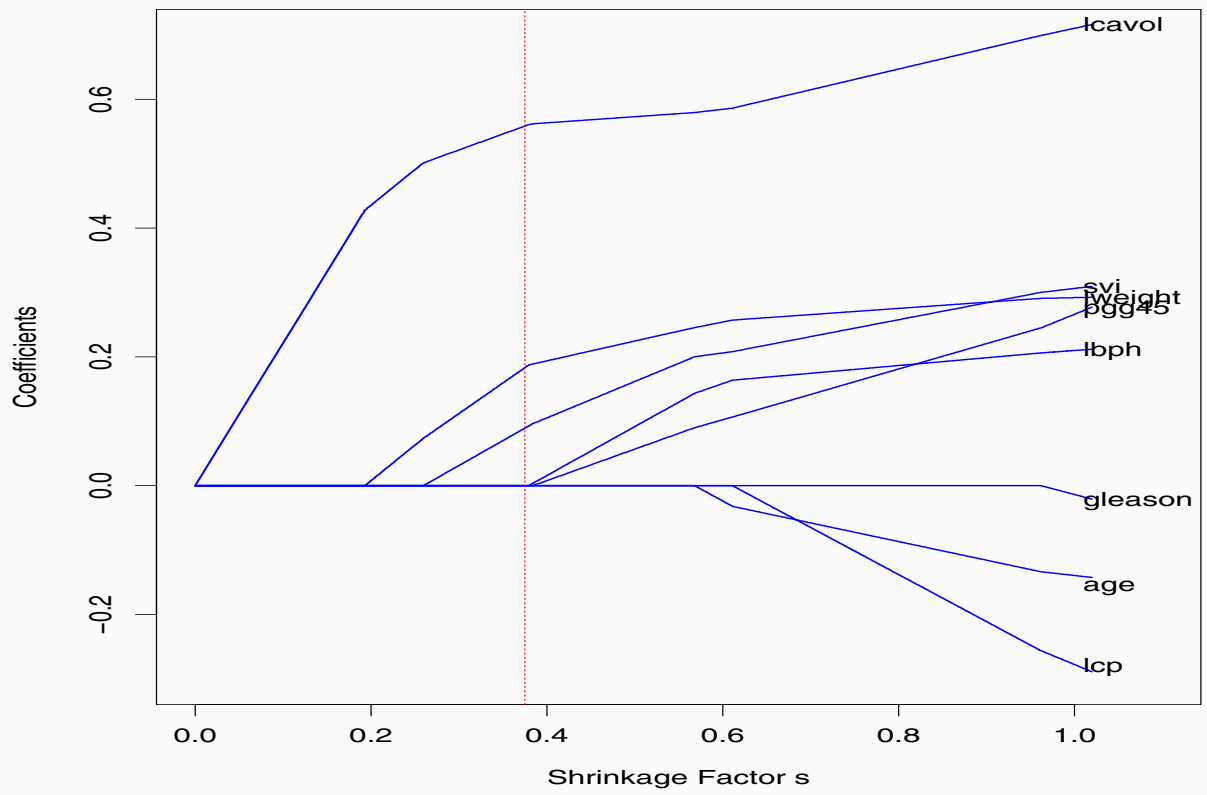
Elección del Parámetro de Regularización

- Por lo tanto la elección debe estar guiada por el error de predicción del modelo regularizado. En la práctica se utiliza un predictor del error de test correspondiente a θ_λ y se elige el valor λ que minimiza esa estimación.



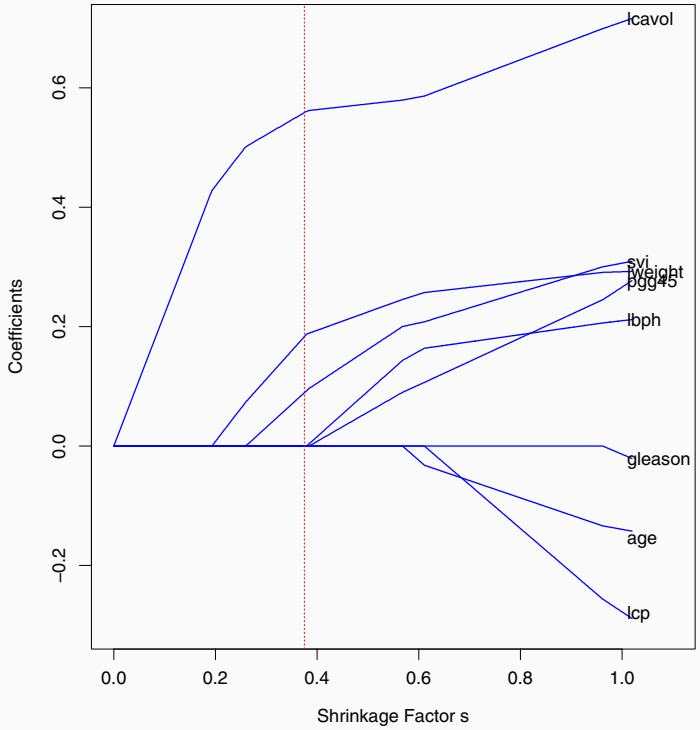
Regularization Path

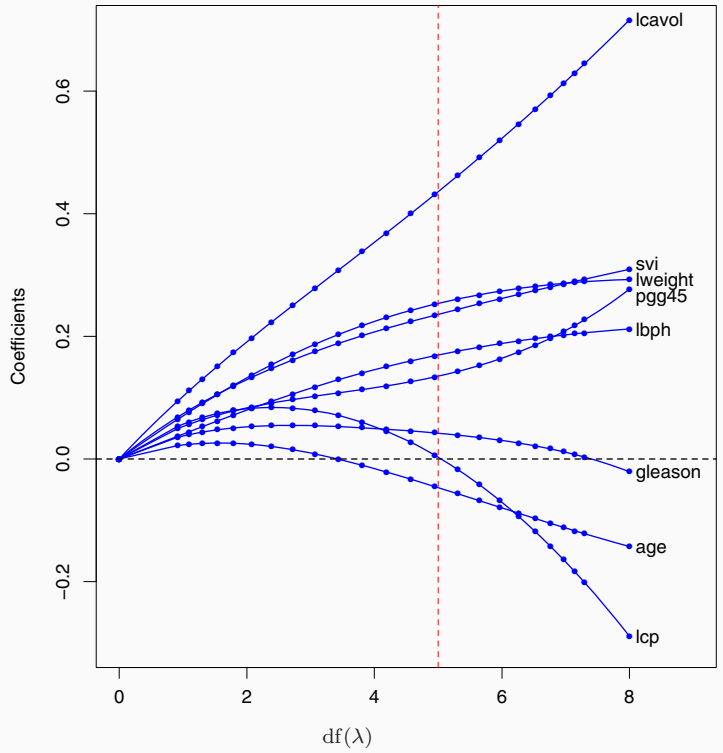
• Existen algoritmos incrementales capaces de encontrar eficientemente las soluciones correspondientes a todo un rango de posibles valores del parámetro de regularización. Esta información que puede ser usada para seleccionar atributos.



$s = t/\|w\|_{\ell_1}$ for $t \in [0, \|w\|_{\ell_1}]$.

Regularization Path ℓ_1 vs ℓ_2





$s = t/\|w\|_{\ell_1}$ for $t \in [0, \|w\|_{\ell_1}]$

$\text{df}(\lambda) = \sum_i (\sigma_i^2/\sigma_i^2 + \lambda)$.