8.5 Muy bien!

Optimización y Descenso por Gradiente

Joaquín Cadeiras y Lucas Pini

Universidad de San Andrés, Buenos Aires, Argentina

1er Semestre 2024

Resumen

En este trabajo se aplicaron métodos numéricos, específicamente el algoritmo de descenso por gradiente, para resolver un sistema de ecuaciones lineales del tipo $Ax = b$. Se definieron dos funciones de costo, una estándar y otra regularizada. Se resolvió el problema utilizando tanto el descenso por gradiente como la descomposición en valores singulares (SVD). Se analizaron las soluciones obtenidas tras 1000 iteraciones del algoritmo de gradiente descendente, con un paso determinado. Se compararon los resultados y se estudió la influencia de la regularización. Los resultados mostraron que el método de gradiente descendente, con la correcta elección de parámetros, converge a soluciones comparables a las obtenidas mediante SVD.

1. Introducción

En la actualidad, la resolución de sistemas de ecuaciones lineales es una tarea fundamental en numerosas aplicaciones científicas y de ingeniería. Métodos numéricos como el descenso por gradiente han demostrado ser herramientas eficaces para abordar estos problemas, especialmente en contextos donde las soluciones exactas no son factibles debido a la dimensionalidad o la naturaleza de los datos. La optimización mediante métodos iterativos permite manejar matrices y vectores grandes y dispersos de manera eficiente.

La motivación detrás de este trabajo radica en explorar y comparar diferentes técnicas de optimización aplicadas a la resolución de sistemas infradeterminados. Particularmente, se busca investigar cómo la regularización, una técnica comúnmente empleada para evitar el sobreajuste y mejorar la estabilidad de las soluciones, afecta el rendimiento y la precisión de los algoritmos de descenso por gradiente. Además, se pretende analizar la eficiencia del descenso por gradiente en comparación con métodos tradicionales como la descomposición en valores singulares (SVD), proporcionando una visión integral de las ventajas y limitaciones de cada enfoque en escenarios prácticos.

2. Métodos pre-existentes

2.1. Algoritmo de Descenso por Gradiente

El descenso por gradiente [1] es un un método iterativo que, dado una función de costo, actualiza los parámetros de x en la dirección opuesta al gradiente. Esto lo hacemos con el objetivo de minimizar dicha función. Los parámetros del algoritmo se realizan de la siguiente manera:

$$x_{t+1} = x_t - s\nabla F(x_t) \tag{1}$$

Donde s es la tasa de aprendizaje y $\nabla F(x_t)$ es el gradiente de la función de costo en el punto x_t.

2.2. Regularización L2

Debido a que el problema puede presentado tiene más incógnitas que ecuaciones, se aplica una regularización que agrega un termino que dependa de la norma-2 del vector al cuadrado a la función de costo donde δ^2 es un parámetro nuevo a elegir. Dando como resultado:

$$F_2(x) = F(x) + \delta^2|x|_2^2 \tag{2}$$

2.3. Método de SVD[2][3]

El método de SVD (Descomposición en Valores Singulares) es un método que se utiliza para descomponer una matriz en tres matrices, U, S y V. $A \in \mathbb{R}^{n\times m}$, en el caso que $m \geqslant n$ se puede descomponer de la siguiente manera:

$$A = U\Sigma V^T = \begin{pmatrix} u_{11} & \cdots & u_{1n} \\ \vdots & \ddots & \vdots \\ u_{n1} & \cdots & u_{nn} \end{pmatrix} \left(\begin{array}{ccc} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \\ \hline & 0 & \end{array}\right) \begin{pmatrix} v_{11} & \cdots & v_{1m} \\ \vdots & \ddots & \vdots \\ v_{m1} & \cdots & v_{mm} \end{pmatrix} \tag{3}$$

Su expresión económica es:

$$A = U\Sigma V^T = \begin{pmatrix} u_{11} & \cdots & u_{1n} \\ \vdots & \ddots & \vdots \\ u_{n1} & \cdots & u_{nn} \end{pmatrix} \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \end{pmatrix} \begin{pmatrix} v_{11} & \cdots & v_{1n} \\ \vdots & \ddots & \vdots \\ v_{m1} & \cdots & v_{mn} \end{pmatrix}^T \tag{4}$$

3. Cuadrados mínimos mediante descenso por gradiente

3.1. Definición de la Función de Costo

El principal objetivo de este trabajo fue aplicar el algoritmo de gradiente descendiente (ver 2.1) al problema de encontrar la solución al sistema $Ax = b$. Donde se definió una matriz $A \in \mathbb{R}^{5\times100}$, $x \in \mathbb{R}^{100}$ y $b \in \mathbb{R}^5$. Una vez establecido el problema a resolver, se planteó la siguiente función de costo:

$$F(x) = (Ax - b)^T(Ax - b). \tag{5}$$

3.2. Resolución del problema

Se tomó la decisión de comparar tres diferentes soluciones al problema planteado. Por un lado resolver utilizando simplemente el método de Gradient Descent (2.1) con la función de costo (5) y por otro resolverlo aplicando una regularización a la función de costo que permite agregarle un termino que dependa de la norma-2 del vector al cuadrado de la función. Por ultimo se compararon ambas resoluciones con una respuesta analítica exacta obtenida mediante SVD (2.3).

Tanto para la función costo a la que llamaremos $F_1(x)$ como a su regularización que llamaremos $F_2(x)$ se les calculo sus respectivos gradientes y el Hessiano asociado a cada una, dando como resultado las siguientes ecuaciones:

Gradiente para $F_1(x)$:

$$\nabla F_1(x) = 2A^T(Ax - b). \tag{6}$$

Gradiente para $F_2(x)$:

$$\nabla F_2(x) = \nabla F(x) + 2\delta_2 x. \tag{7}$$

Hessiano de F_1:

$$H(F_1) = 2A^TA \tag{8}$$

Hessiano de F_2:

$$H(F_2) = 2A^TA + 2\delta_2\mathbb{I} \tag{9}$$

Posteriormente, para la implementación del descenso por gradiente, tomando un valor de condición inicial aleatorio, se utilizó la formula iterativa definida en la explicación del método de gradiente

descendiete (2.1) con un paso s óptimo [4] definido en función del mayor autovalor del Hessiano de F en las ecuaciones (8) y (9):

$$s = \frac{1}{\lambda_{\max}} \tag{10}$$

La elección de este valor de s se debe a que proporciona una tasa de aprendizaje óptima. Por un lado asegurando que el paso sea suficientemente pequeño para garantizar la convergencia del algoritmo pero a su vez lo suficientemente grande para obtener una convergencia relativamente rápida.

En el caso de $F_2(x)$ fue necesario definir distintos valores de δ^2 y comparar los resultados obtenidos posteriormente. Permitiendo así tomar una decisión sobre cual es el el valor mas apropiado para la resolución de este problema.

Las simulaciones fueron realizadas en un principio en el margen de las 1000 iteraciones, veremos que en algunos de los casos esta cantidad no fue suficiente para la convergencia de los métodos y parámetros utilizados. Por ende, luego se realizaron mas iteraciones con el fin de detallar mas claramente la eficiencia de los métodos y la comparación de los mismos en distintos gráficos.

Finalmente, se utilizó la descomposición en valores singulares (SVD) (2.3) para obtener una solución analítica del problema con la cual comparar las otras dos resoluciones.

3.3. Resultados y Análisis

A continuación se detallaron los resultados obtenidos de las simulaciones realizadas a partir de las condiciones previamente mencionadas. En primer lugar, compararemos los datos obtenidos con respecto a la convergencia de la norma de $F_1(X)$ y $F_2(X)$, a partir de la evolución del descenso por gradiente. Véase la Figura (1).

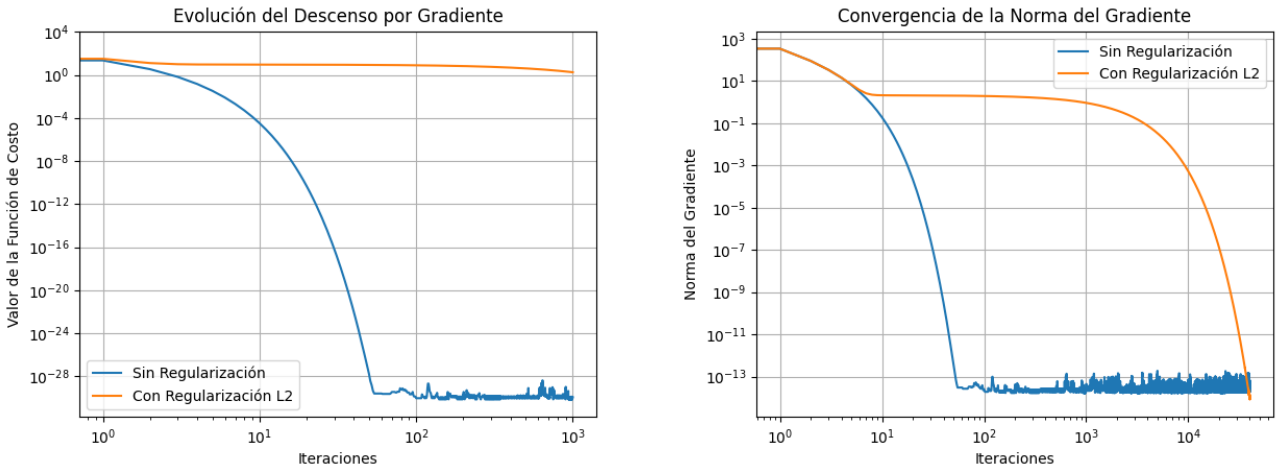

Figura 1: Gráfico de la evolución del Método de Descenso y la Norma del Gradiente

Se observó que tanto el descenso por gradiente sin regularización como el descenso por gradiente con regularización L2 muestran una tendencia decreciente de la función de costo. Esto pareciera indicar que ambos algoritmos convergen hacia una solución. Se pudo ver que F_1 llega a 0 (10^{-30} que podemos ver el error de redondeo de máquina) mientras que F_2 se estabiliza mucho antes en $1,7$, ambos habiendo empezado en 20 y 30 respectivamente. Aunque parezca mejor F_1, llegar a valores de costo extremadamente bajos puede ser señal de estar sobreajustando. En cambio, usando la regularización L2, se logra prevenir esto llegando a un costo bajo, que no es la mejor solución, pero la solución mas *generalizable*.

Otra diferencia que podemos notar entre ambos métodos es la velocidad de convergencia de los mismos. El método sin regularización muestra una disminución mas rápida de la función de costo, lo que podría sugerir que dicho método es mas eficiente en la minimización de la función para este problema si no tenemos en cuenta el problema previamente mencionado del sobreajuste.

En cuanto al gráfico de la convergencia de las normas de los gradientes, nuevamente mostró que ambos métodos se acercan al mínimo, proporcionando información adicional de la convergencia del algoritmo. La disminución constante de la norma de los gradientes indica que el algoritmo esta progresando en dirección a la minimización de la función de costo. Cuando ambas normas del gradiente se aproximan a cero sugiere que el algoritmo ha alcanzado un punto cercano al mínimo local o global de la función de costo. Analizando estas normas se puede dar cuanta de la estabilidad del algoritmo. El método con regularización muestra una disminución mas estable de la norma, lo que indica por consecuencia una convergencia mas estable y sin oscilaciones.

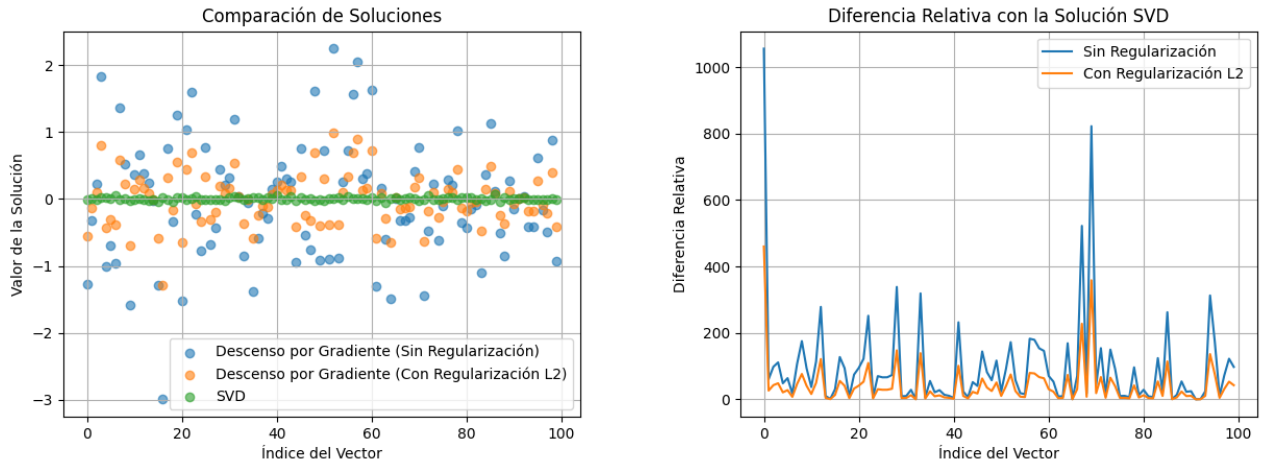

Figura 2: Gráfico de la comparación de las soluciones y la diferencia absoluta

Al momento de comparar ambos métodos con la solución obtenida analíticamente mediante SVD, se graficaron dos distintas comparaciones de los resultados obtenidos. El primer gráfico de la Figura (2) mostró como los valores de solución de el método con regularización se aproximan mas en promedio a la solución exacta. Mientras que si vemos los puntos del gradiente sin regularización que están marcados en azul, se puede ver como varían y se alejan significativamente de los valores esperados para algunos puntos.

Por otra parte, el segundo gráfico donde se puede ver el error relativo con respecto a SVD, que la función de graficada en naranja este siempre por debajo de la graficada en azul, deja en claro que la regularización L2 devuelve menos error para cualquier iteración del algoritmo.

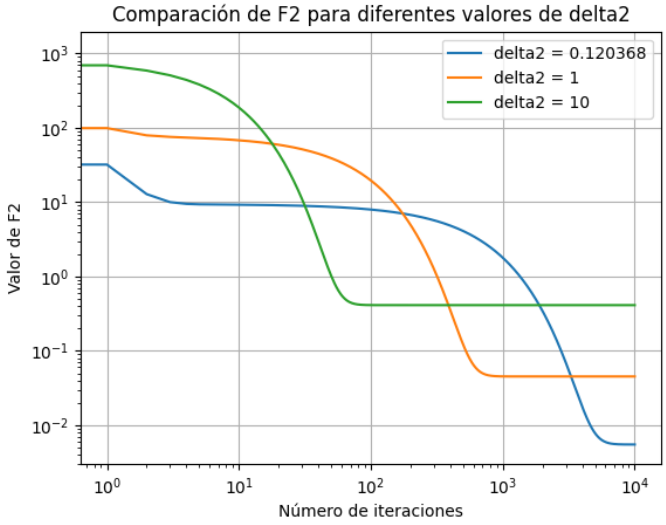

Figura 3: Valores de F_2 para $\delta_2 = 10^{-2}\sigma_{max}, 1, 10$

En la Figura (3) podemos ver que con δ_2 mayores, la convergencia es más rápida, pero converge a un valor más grande, mientras que con δ_2 menores hay una convergencia mas lenta, pero converge a valores más chicos. Esto se debe a que a mayor la restricción, más evita sobreajustes pero al costo de una convergencia menos precisa a los datos.

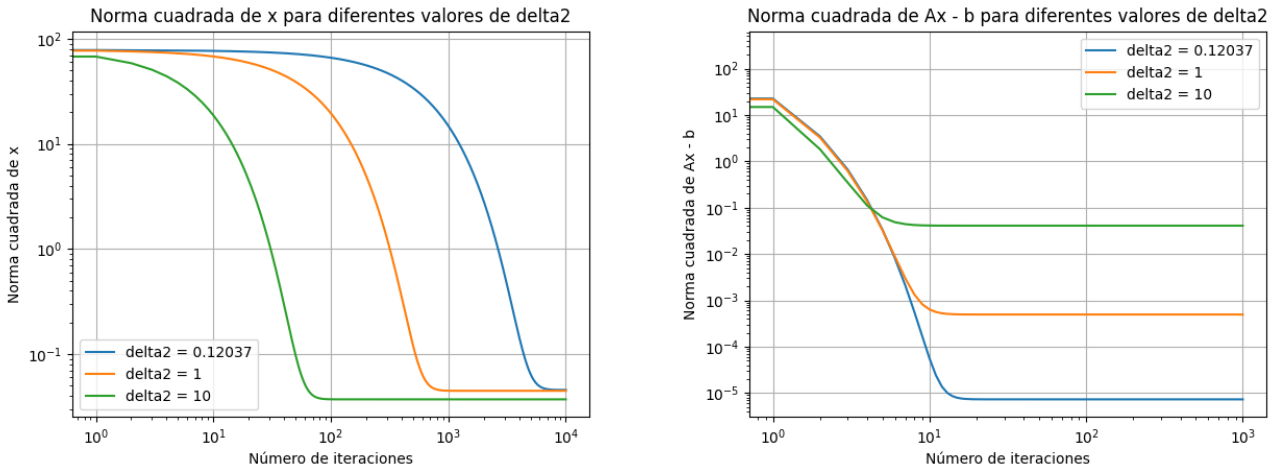

Figura 4: $\|x\|_2^2$ y $\|Ax - b\|_2^2$ para $\delta_2 = 10^{-2}\sigma_{max}, 1, 10$

En la primera gráfica de la Figura (4) podemos ver que para diferentes valores de δ_2 la convergencia de $\|x\|_2^2$ es más rápida, sin embargo todos convergen a valores similares. En la segunda gráfica, que representa la norma cuadrada del residuo, ocurre lo opuesto. Observamos que para diferentes valores de δ_2 la rapidez de convergencia es similar, pero convergen a valores muy distintos. Llegamos a la conclusión que el valor de delta afecta directamente al residuo final, donde valores de δ_2 chicos llevan a un residuo final chico y viceversa.

4. Conclusión

La incorporación de la regularización L2 y la elección de la tasa de aprendizaje s como $s = \frac{1}{\lambda_{\text{máx}}}$ demostraron ser estrategias efectivas para mejorar tanto la convergencia como la estabilidad del algoritmo de descenso por gradiente. En particular, la regularización L2 mostró claros beneficios en términos de evitar el sobreajuste y mejorar la generalización del modelo, siendo δ_2 el parámetro principal para conseguir esto.

Este estudio también subraya la importancia de ajustar correctamente los parámetros del algoritmo y de considerar técnicas de regularización adecuadas para obtener soluciones de buena calidad en problemas de regresión lineal.

Referencias

[1] S.P. Boyd y L. Vandenberghe. *Convex Optimization*. Berichte über verteilte messysteme pt. 1. Cambridge University Press, 2004. ISBN: 9780521833783. URL: https://books.google.com.ar/books?id=mYm0bLd3fcoC.

[2] L.N. Trefethen y D. Bau. *Numerical Linear Algebra*. Other Titles in Applied Mathematics. Society for Industrial y Applied Mathematics, 1997. ISBN: 9780898713619. URL: https://books.google.com.ar/books?id=4Mou5YpRD_kC.

[3] Gene H. Golub y Christian H. Reinsch. "Singular value decomposition and least squares solutions". En: *Numerische Mathematik* 14 (1970), págs. 403-420. URL: https://api.semanticscholar.org/CorpusID:123532178.

[4] Philip Wolfe. "Convergence Conditions for Ascent Methods". En: *SIAM Review* 11.2 (1969), págs. 226-235. DOI: 10.1137/1011036. eprint: https://doi.org/10.1137/1011036. URL: https://doi.org/10.1137/1011036.