Estudio Experimental del Gradiente Descendiente y sus Variantes en Problemas de Optimización

Wasinger, Pablo Tiago[1], *Sidiropulos, Felipe*[2]

Departamento de Ingeniería, Universidad de San Andrés, Victoria, Buenos Aires, Argentina.
E-mails: [1]*pwasinger@udesa.edu.ar,* [2]*fisidiropuloslamouret@udesa.edu.ar*

Abstract:
Este trabajo presenta un análisis exhaustivo del método de gradiente descendiente y sus variantes en dos contextos diferentes de optimización. Primero, se estudia el comportamiento de estos métodos en la minimización de la función de Rosenbrock bidimensional, un problema clásico que presenta desafíos particulares debido a su geometría no convexa. Se implementan y comparan cuatro variantes: gradiente descendiente básico, método de Newton, backtracking line search y momentum, evaluando su rendimiento con diferentes puntos iniciales y tasas de aprendizaje. Los resultados muestran que el método de Newton exhibe la convergencia más rápida y robusta, mientras que el gradiente descendiente básico muestra una alta sensibilidad a la elección de la tasa de aprendizaje. En segundo lugar, se aplica el gradiente descendiente al problema de regresión lineal utilizando el conjunto de datos California Housing, comparando su desempeño con la solución analítica de la pseudoinversa. Se demuestra que, si bien la tasa de aprendizaje teórica $\eta = 1/\sigma_1^2$ resulta demasiado conservadora, un ajuste empírico a $\eta = 10000/\sigma_1^2$ permite alcanzar resultados equivalentes a la solución analítica. Este trabajo destaca la importancia del balance entre las garantías teóricas y las consideraciones prácticas en la implementación de métodos de optimización.

Keywords: Optimización Numérica, Gradiente Descendiente, Función de Rosenbrock, Regresión Lineal, Método de Newton, Backtracking Line Search, Momentum, Cuadrados Mínimos, Análisis de Convergencia, Tasa de Aprendizaje.

1 Introducción

La optimización numérica es un pilar fundamental en el campo del análisis numérico y el aprendizaje automático, desempeñando un papel crucial en la resolución de problemas complejos en diversas áreas de la ciencia y la ingeniería. Entre los diversos métodos de optimización, el gradiente descendiente destaca por su simplicidad conceptual y su versatilidad, siendo ampliamente utilizado en aplicaciones que van desde el entrenamiento de redes neuronales hasta la resolución de problemas de regresión.

En este trabajo, nos enfocamos en el estudio y análisis del método de gradiente descendiente y sus variantes, explorando su comportamiento en dos contextos diferentes pero complementarios. En primer lugar, abordamos la optimización de la función de Rosenbrock en dos dimensiones, un problema clásico que, debido a su particular geometría no convexa y su valle estrecho, representa un desafío significativo para los métodos de optimización. Este problema nos permite evaluar la eficacia y robustez de diferentes variantes del método, incluyendo el uso del Hessiano, momentum y backtracking line search.

En segundo lugar, aplicamos estos métodos al problema práctico de regresión lineal utilizando el conjunto de datos California Housing, donde exploramos la minimización del error cuadrático medio. Este caso de estudio nos permite contrastar la implementación práctica del gradiente descendiente con la solución analítica proporcionada por el método de la pseudoinversa, analizando aspectos cruciales como la elección de la tasa de aprendizaje y su impacto en la convergencia del método.

El informe está estructurado de la siguiente manera: comenzamos con un marco teórico que presenta los fundamentos matemáticos del gradiente descendiente y sus variantes, incluyendo las condiciones necesarias para su convergencia. Posteriormente, en la sección de desarrollo experimental, describimos detalladamente la metodología empleada para cada problema, incluyendo la implementación de los diferentes métodos y las estrategias de evaluación. En la sección de resultados, presentamos y analizamos los hallazgos obtenidos, comparando el rendimiento de los diferentes métodos y discutiendo las implicaciones prácticas de nuestras observaciones. Finalmente, concluimos con una síntesis de los resultados más relevantes y sus implicaciones para la implementación práctica de métodos de optimización.

2 Marco Teórico

2.1 Método de Gradiente Descendente

El método de Gradiente Descendente es un algoritmo iterativo fundamental en optimización y aprendizaje automático, diseñado para minimizar una función objetivo $f(\mathbf{x})$. Su principal enfoque es ajustar los parámetros de manera iterativa siguiendo la dirección del gradiente negativo de $f(\mathbf{x})$, que indica el camino de descenso más pronunciado.

2.2 Definición Formal

Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función objetivo diferenciable. El Gradiente Descendente se define mediante la siguiente regla de actualización:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla f(\mathbf{x}_k), \tag{1}$$

donde:

- $\mathbf{x}_k \in \mathbb{R}^n$ representa la estimación de los parámetros en la iteración k,
- $\eta > 0$ es la tasa de aprendizaje o tamaño del paso, un parámetro que controla la magnitud de cada ajuste,
- $\nabla f(\mathbf{x}_k)$ es el gradiente de f evaluado en $\mathbf{x}_k$, que señala la dirección del mayor incremento local de f.

El gradiente $\nabla f(\mathbf{x})$ es un vector compuesto por las derivadas parciales de f con respecto a cada componente de $\mathbf{x}$. Matemáticamente, se expresa como:

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}. \tag{2}$$

2.3 Convergencia del Método

La convergencia del Gradiente Descendente depende de varios factores, entre ellos:

- **Tasa de aprendizaje η:** Si η es demasiado grande, el algoritmo puede divergir, oscilando alrededor del mínimo. Si η es muy pequeño, la convergencia puede ser excesivamente lenta.
- **Propiedades de $f(\mathbf{x})$:** Si $f(\mathbf{x})$ es convexa y tiene un gradiente Lipschitz-continuo (es decir, que el gradiente no cambia bruscamente), se puede garantizar la convergencia hacia un mínimo global. Para funciones no convexas, el método puede converger a un mínimo local o a un punto de silla.
- **Condiciones iniciales:** La elección del punto inicial $\mathbf{x}_0$ puede influir en la rapidez y el éxito del algoritmo.

2.4 Variantes del Gradiente Descendente

El método de Gradiente Descendente tiene varias variantes diseñadas para mejorar su desempeño, ya sea en términos de velocidad de convergencia, estabilidad o adaptabilidad a diferentes tipos de problemas. En esta sección, se describen en detalle tres variantes destacadas: el uso de Backtracking para ajustar dinámicamente la tasa de aprendizaje, el Gradiente Descendente con Momento y el uso del Hessiano para determinar el tamaño del paso.

2.4.1 *Método con Backtracking*

El método de Backtracking es una estrategia adaptativa para seleccionar la tasa de aprendizaje η en cada iteración. Su principal objetivo es ajustar η de forma que garantice una disminución suficiente en el valor de la función objetivo $f(\mathbf{x})$, evitando pasos demasiado grandes que puedan causar divergencia o demasiado pequeños que ralenticen la convergencia.

La regla general de Backtracking se basa en la condición de Armijo, que establece que η debe satisfacer:

$$f(\mathbf{x}_k - \eta \nabla f(\mathbf{x}_k)) \leq f(\mathbf{x}_k) - c\eta \|\nabla f(\mathbf{x}_k)\|^2, \tag{3}$$

donde $c \in (0, 1)$ es un parámetro fijo, típicamente pequeño (por ejemplo, $c = 10^{-4}$). El proceso se desarrolla de la siguiente manera:

1. Iniciar con un valor inicial $\eta_0 > 0$ grande.
2. Reducir η multiplicándolo por un factor $\beta \in (0, 1)$ (por ejemplo, $\beta = 0.5$) hasta que la condición de Armijo se cumpla.

Este enfoque asegura que el tamaño del paso se adapte dinámicamente a las características locales de f, equilibrando estabilidad y rapidez de convergencia.

2.4.2 *Gradiente Descendente con Momento*

El Gradiente Descendente con Momento busca acelerar la convergencia, especialmente en regiones donde el gradiente tiene una variación pequeña, como en los valles angostos de funciones no convexas. La idea principal es acumular una "inercia" basada en los gradientes previos para suavizar las oscilaciones en la dirección de descenso.

El método introduce una variable de acumulación $\mathbf{v}_k$ que combina el gradiente actual y el acumulado de iteraciones previas. La regla de actualización se define como:

$$\mathbf{v}_{k+1} = \gamma \mathbf{v}_k + \eta \nabla f(\mathbf{x}_k), \tag{4}$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{v}_{k+1}, \tag{5}$$

donde:

- $\gamma \in [0, 1)$ es el coeficiente de momento, que controla cuánto peso se da a los gradientes acumulados (típicamente $\gamma = 0.9$).
- η es la tasa de aprendizaje.

El término $\mathbf{v}_k$ actúa como una fuerza impulsora que acelera la convergencia en direcciones consistentes y reduce las oscilaciones transversales, mejorando el rendimiento en problemas complejos.

2.4.3 *Método con el Hessiano para el Tamaño del Paso*

En esta variante, el tamaño del paso η se calcula utilizando información de segunda derivada de la función objetivo, representada por la matriz Hessiana $H(\mathbf{x})$. Este enfoque utiliza la actualización de Newton:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - H(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k), \tag{6}$$

donde:

- $H(\mathbf{x}_k)$ es la matriz Hessiana de $f(\mathbf{x})$, que contiene las segundas derivadas parciales:

$$H(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}.$$

- $H(\mathbf{x}_k)^{-1}$ actúa como un factor de escala direccional, ajustando dinámicamente el tamaño del paso en cada dirección.

Este método, conocido también como Gradiente Descendente de Newton, tiene una tasa de convergencia cuadrática en vecindades del mínimo, superando a la convergencia lineal del Gradiente Descendente estándar. Sin embargo, calcular y almacenar $H(\mathbf{x})$ es computacionalmente costoso para problemas de alta dimensionalidad, lo que limita su aplicabilidad a casos con pocos parámetros o a problemas en los que se puede aproximar H eficientemente.

2.5 Aplicaciones

El Gradiente Descendente es fundamental en múltiples áreas, incluyendo:

- Entrenamiento de modelos de aprendizaje automático, como redes neuronales profundas.
- Optimización de funciones objetivo en problemas de gran escala.
- Resolución de problemas de regresión y clasificación.

El método destaca por su flexibilidad y capacidad de adaptarse a diferentes escenarios de optimización, convirtiéndose en una herramienta indispensable en el ámbito científico y tecnológico.

2.6 Importancia de la Convexidad en el Método de Gradiente Descendente

La convexidad de la función objetivo $f(\mathbf{x})$ juega un papel crucial en la eficacia del Método de Gradiente Descendente, ya que afecta directamente la garantía de convergencia al mínimo global.

2.6.1 Definición de Función Convexa

Una función $f : \mathbb{R}^n \to \mathbb{R}$ es convexa si, para todo par de puntos $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ y $\lambda \in [0, 1]$, se cumple:

$$f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y}). \tag{7}$$

Geométricamente, esto implica que la línea recta entre dos puntos cualesquiera del gráfico de f se encuentra por encima o en el gráfico de la función.

2.6.2 Ventajas de la Convexidad

Para el Método de Gradiente Descendente, la convexidad aporta las siguientes ventajas:

- **Existencia y unicidad del mínimo global:** En funciones estrictamente convexas, el mínimo global es único. Esto significa que el algoritmo, independientemente de su punto de inicio, converge hacia el mismo resultado si la tasa de aprendizaje está correctamente ajustada.
- **Dirección garantizada de descenso:** En una función convexa, el gradiente siempre apunta hacia una región donde los valores de f son menores. Esto asegura que cada iteración del método de gradiente lleve hacia el mínimo global o a una región cercana a este.

2.6.3 Riesgos en Funciones No Convexas

Cuando $f(\mathbf{x})$ no es convexa, surgen varias complicaciones:

- **Mínimos locales y puntos de silla:** El algoritmo puede detenerse en un mínimo local o en un punto de silla, lo que impide alcanzar el mínimo global.
- **Oscilaciones:** Sin convexidad, el gradiente puede no guiar de manera consistente hacia una dirección de mejora, especialmente si la función tiene regiones planas o abruptas.
- **Sensibilidad al punto inicial:** La trayectoria de convergencia depende significativamente del punto inicial, aumentando la probabilidad de resultados subóptimos.

En resumen, la convexidad de $f(\mathbf{x})$ no solo facilita la convergencia, sino que también proporciona garantías teóricas sobre la calidad de la solución alcanzada por el Método de Gradiente Descendente. Esto explica por qué las funciones convexas son preferidas en problemas de optimización y aprendizaje automático.

2.7 Condición de Corte para el Gradiente Descendente

Una parte fundamental en la implementación del Método de Gradiente Descendente es la definición de una condición de corte que determine cuándo detener el proceso iterativo. Una de las condiciones más utilizadas es:

$$\|\nabla f(x, y)\| < \epsilon, \tag{8}$$

donde:

- $\|\nabla f(x, y)\|$ es la norma del gradiente de $f(x, y)$ en el punto actual (x, y),
- $\epsilon > 0$ es un umbral predefinido que controla la precisión deseada.

2.7.1 Fundamento Teórico

El gradiente $\nabla f(x, y)$ de una función $f(x, y)$ indica la dirección y magnitud del mayor incremento local de f. En el contexto de minimización de una función convexa, el algoritmo busca puntos donde $\nabla f(x, y) \approx 0$, ya que estos puntos corresponden a **Mínimos locales** si $\nabla f(x, y) = 0$ ya que la matriz Hessiana asociada es definida positiva en ese punto.

Por lo tanto, una norma del gradiente pequeña ($\|\nabla f(x, y)\| < \epsilon$) indica que las iteraciones han alcanzado una región donde el valor de f no disminuye significativamente, lo que es característico de estar cerca de un punto crítico.

2.7.2 Razonamiento Práctico

La condición $\|\nabla f(x, y)\| < \epsilon$ tiene sentido práctico por las siguientes razones:

- **Criterio de estabilidad:** Cuando el gradiente es pequeño, los ajustes en (x, y) durante cada iteración son mínimos, lo que sugiere que el algoritmo ha alcanzado una región de baja pendiente y, por ende, está cerca de un punto óptimo.
- **Balance entre precisión y costo computacional:** Elegir un ϵ adecuado permite detener el algoritmo antes de realizar iteraciones adicionales que no aportan mejoras significativas, optimizando el tiempo de cómputo.
- **Flexibilidad en la precisión deseada:** El valor de ϵ puede ajustarse según la necesidad del problema. Valores pequeños de ϵ son ideales cuando se busca alta precisión, mientras que valores mayores son aceptables en aplicaciones donde una aproximación razonable del óptimo es suficiente.

2.7.3 Limitaciones y Consideraciones

Si bien $\|\nabla f(x, y)\| < \epsilon$ es una condición efectiva, es importante tener en cuenta lo siguiente:

- **Elección de ϵ:** Un ϵ demasiado grande puede detener el algoritmo prematuramente, mientras que un ϵ muy pequeño puede incrementar innecesariamente el costo computacional.

En resumen, la condición $\|\nabla f(x, y)\| < \epsilon$ es un criterio intuitivo y efectivo para determinar la convergencia del Método de Gradiente Descendente, ya que se basa en la cercanía a puntos críticos de $f(x, y)$, balanceando precisión y eficiencia computacional.

2.8 Descomposición en Valores Singulares (SVD)

La Descomposición en Valores Singulares (SVD, por sus siglas en inglés) es una técnica fundamental en álgebra lineal que descompone una matriz en tres componentes esenciales. Dada una matriz $A \in \mathbb{R}^{m \times n}$, su descomposición SVD se expresa como:

$$A = U \Sigma V^T, \tag{9}$$

donde:

- $U \in \mathbb{R}^{m \times m}$ es una matriz ortogonal que contiene los vectores singulares izquierdos de A,
- $\Sigma \in \mathbb{R}^{m \times n}$ es una matriz diagonal rectangular cuyos elementos no negativos en la diagonal son los valores singulares de A,
- $V \in \mathbb{R}^{n \times n}$ es una matriz ortogonal que contiene los vectores singulares derechos de A.

2.8.1 Proceso de Descomposición

El proceso de SVD se desarrolla de la siguiente manera:

1. **Cálculo de $A^T A$ y AA^T:** Se calculan las matrices $A^T A \in \mathbb{R}^{n \times n}$ y $AA^T \in \mathbb{R}^{m \times m}$, ambas simétricas y semidefinidas positivas. Estas matrices comparten los mismos valores propios no nulos, lo que es clave para la construcción de las matrices V y U.
2. **Obtención de vectores y valores propios:** Se calculan los valores propios λ_i y vectores propios v_i de $A^T A$, y los valores propios λ_i y vectores propios u_i de AA^T. Los valores propios satisfacen $\lambda_i = \sigma_i^2$, donde σ_i son los valores singulares de A.
3. **Construcción de las matrices U, Σ y V:** Los vectores propios normalizados v_i forman las columnas de V, y los vectores propios normalizados u_i forman las columnas de U. La matriz Σ es diagonal, con los valores singulares $\sigma_i = \sqrt{\lambda_i}$ ordenados de mayor a menor en la diagonal.

4. **Reducción de la dimensión:** Para comprimir los datos, se puede truncar Σ eliminando los valores singulares más pequeños y sus vectores asociados en U y V. Esto retiene solo los componentes principales de mayor magnitud, minimizando la pérdida de información y proporcionando una aproximación de menor rango de A.

2.9 Regresión por Cuadrados Mínimos

La regresión por cuadrados mínimos es un método estadístico utilizado para ajustar un modelo lineal a un conjunto de datos, minimizando la suma de los cuadrados de los errores. Dado un conjunto de datos $X \in \mathbb{R}^{n \times p}$ con n observaciones y p variables predictoras, y un vector de respuestas $y \in \mathbb{R}^n$, el objetivo es encontrar un vector de coeficientes $\beta \in \mathbb{R}^p$ que minimice el error:

$$\min_{\beta} \|X\beta - y\|_2^2. \tag{10}$$

2.9.1 Formulación Matemática

El problema de minimización se resuelve encontrando $\hat{\beta}$ que satisface:

$$\hat{\beta} = \arg\min_{\beta} \|X\beta - y\|_2^2. \tag{11}$$

La solución óptima $\hat{\beta}$ se obtiene resolviendo las ecuaciones normales:

$$X^T X \hat{\beta} = X^T y. \tag{12}$$

Si $X^T X$ es invertible, la solución explícita es:

$$\hat{\beta} = (X^T X)^{-1} X^T y. \tag{13}$$

2.9.2 Interpretación de la Solución

El vector de coeficientes $\hat{\beta}$ representa la contribución de cada variable predictora al modelo lineal. Cada componente de $\hat{\beta}$ indica el cambio esperado en la variable de respuesta y por unidad de cambio en la correspondiente variable predictora, manteniendo las demás variables constantes.

El ajuste por cuadrados mínimos busca minimizar la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos por el modelo. Este método es eficiente y proporciona la mejor aproximación lineal en términos de error cuadrático medio, siempre que se cumplan las suposiciones de linealidad, independencia, homocedasticidad y normalidad de los errores.

2.10 Elección de la tasa de aprendizaje en cuadrados mínimos

En problemas de mínimos cuadrados lineales, donde la función objetivo se define como:

$$f(\mathbf{w}) = \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2, \tag{14}$$

la elección de la tasa de aprendizaje $\eta = \frac{1}{\sigma^2}$, donde σ es el mayor valor singular de la matriz $\mathbf{X}$, resulta especialmente significativa. A continuación, se detallan las razones teóricas que sustentan esta elección.

2.10.1 Curvatura de la función objetivo

La función objetivo $f(\mathbf{w})$ es convexa y su gradiente está dado por:

$$\nabla f(\mathbf{w}) = \mathbf{X}^\top(\mathbf{X}\mathbf{w} - \mathbf{y}). \tag{15}$$

El término $\mathbf{X}^\top\mathbf{X}$ define la curvatura de $f(\mathbf{w})$, y su mayor valor propio $\lambda_{\max}$ determina la máxima curvatura en cualquier dirección. En el caso de mínimos cuadrados, los valores propios de $\mathbf{X}^\top\mathbf{X}$ están relacionados con los cuadrados de los valores singulares de $\mathbf{X}$. Por lo tanto, el mayor valor propio de $\mathbf{X}^\top\mathbf{X}$ es σ^2, donde σ es el mayor valor singular de $\mathbf{X}$.

2.10.2 Tasa de aprendizaje óptima

Para garantizar la estabilidad y una convergencia eficiente en el método de gradiente descendente, es fundamental que la tasa de aprendizaje η no exceda el inverso del mayor valor propio de la matriz Hessiana de $f(\mathbf{w})$. En este caso, la matriz Hessiana es $\mathbf{X}^\top\mathbf{X}$. Por lo tanto, para minimizar $f(\mathbf{w})$, se establece la condición:

$$\eta \leq \frac{1}{\lambda_{\max}} = \frac{1}{\sigma^2}. \tag{16}$$

Elegir $\eta = \frac{1}{\sigma^2}$ permite realizar actualizaciones que se ajustan de manera óptima a la curvatura de la función, logrando un equilibrio entre la rapidez de convergencia y la estabilidad.

2.10.3 Beneficios en problemas de mínimos cuadrados

La elección de $\eta = \frac{1}{\sigma^2}$ ofrece las siguientes ventajas específicas en problemas de mínimos cuadrados:

- **Convergencia garantizada:** La tasa de aprendizaje está directamente ligada a la máxima curvatura de $f(\mathbf{w})$, evitando oscilaciones o divergencia.
- **Velocidad óptima:** En problemas cuadráticos, esta elección asegura que el gradiente descendente converge en el menor número posible de iteraciones para un esquema fijo.
- **Simplicidad computacional:** Los valores singulares de $\mathbf{X}$ pueden calcularse previamente, lo que simplifica la determinación de η.

2.11 Función Objetivo del Error Cuadrático Medio (ECM)

La regresión lineal busca minimizar la función objetivo conocida como error cuadrático medio (ECM), definida como:

$$\text{ECM}(\mathbf{w}) = \frac{1}{n}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2,$$

donde:

- $\mathbf{X} \in \mathbb{R}^{n \times (d+1)}$ es la matriz de características con n muestras y d atributos, con una columna adicional de 1s para considerar la ordenada al origen.
- $\mathbf{y} \in \mathbb{R}^n$ es el vector de valores objetivo (valor medio de las viviendas).
- $\mathbf{w} \in \mathbb{R}^{d+1}$ es el vector de coeficientes desconocidos (variables de decisión) que buscamos optimizar.

2.11.1 Análisis de la Convexidad de la Función

Para garantizar que los métodos de optimización, como el gradiente descendente, converjan a un mínimo global, es importante verificar la convexidad de la función objetivo.

El ECM puede reescribirse como:

$$\text{ECM}(\mathbf{w}) = \frac{1}{n}\left(\mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top\mathbf{X}\mathbf{w} + \mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w}\right).$$

El término cuadrático $\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w}$ domina la estructura de la función, ya que los demás términos son lineales o constantes respecto a $\mathbf{w}$. Para evaluar la convexidad, consideremos la matriz Hessiana de la función:

$$\mathbf{H} = \nabla^2\text{ECM}(\mathbf{w}) = \frac{2}{n}\mathbf{X}^\top\mathbf{X}.$$

Dado que $\mathbf{X}^\top\mathbf{X}$ es una matriz semidefinida positiva 2.11.2 (su valor propio mínimo es no negativo), se concluye que la

función ECM($\mathbf{w}$) es convexa. Esto asegura que cualquier mínimo local también es un mínimo global, lo que hace que el gradiente descendente sea una técnica adecuada para encontrar la solución óptima.

Además, la convexidad implica que la tasa de aprendizaje puede ser ajustada para garantizar una convergencia eficiente, como se explicó previamente al considerar $\eta = \frac{1}{\sigma^2}$, donde σ^2 es el mayor valor propio de $\mathbf{X}^\top\mathbf{X}$.

2.11.2 Propiedad de semidefinida positiva de $\mathbf{X}^\top\mathbf{X}$

La matriz $\mathbf{X}^\top\mathbf{X}$ es semidefinida positiva porque, por definición, para cualquier vector $\mathbf{z} \in \mathbb{R}^{d+1}$, se cumple que:

$$\mathbf{z}^\top(\mathbf{X}^\top\mathbf{X})\mathbf{z} = (\mathbf{X}\mathbf{z})^\top(\mathbf{X}\mathbf{z}) = \|\mathbf{X}\mathbf{z}\|_2^2 \geq 0.$$

El producto escalar $\|\mathbf{X}\mathbf{z}\|_2^2$ representa la norma al cuadrado de un vector, que siempre es no negativa. Esto demuestra que $\mathbf{X}^\top\mathbf{X}$ no tiene valores propios negativos, ya que sus valores propios están directamente relacionados con el valor cuadrático asociado. Por lo tanto, $\mathbf{X}^\top\mathbf{X}$ es una matriz semidefinida positiva. Esta propiedad es clave en el análisis de la convexidad de la función objetivo, ya que asegura que el término cuadrático domina de manera convexa.

3 Desarrollo Experimental

3.1 Optimización mediante Gradiente Descendiente en Dos Dimensiones

En esta sección, exploramos el comportamiento y eficacia del método de gradiente descendiente y sus variantes en la optimización de funciones no lineales. Para este análisis, utilizamos la función de Rosenbrock bidimensional, también conocida como "función banana" debido a su característica forma curva. Esta función se define como $f(x, y) = (a - x)^2 + b(y - x^2)^2$, donde empleamos los valores $a = 1$ y $b = 100$. La función de Rosenbrock es particularmente interesante para el análisis de métodos de optimización debido a su valle parabólico no convexo, que presenta desafíos significativos para los algoritmos de optimización.

3.1.1 Visualización y Análisis Preliminar

Para comprender mejor el comportamiento de la función objetivo, comenzamos con una visualización tridimensional y un análisis de sus propiedades. Este paso es fundamental para anticipar posibles desafíos en la optimización y para interpretar los resultados posteriores. La visualización se realiza en el dominio $x \in [-2, 2]$ y $y \in [-1, 3]$, región que contiene el mínimo global y exhibe las características más relevantes de la función. Este análisis visual nos permite identificar las regiones donde los métodos de optimización podrían enfrentar mayores dificultades, particularmente en el valle curvo característico de la función.

3.1.2 Análisis de Convexidad

Previo a la implementación de los métodos de optimización, realizamos un análisis de convexidad mediante el estudio del Hessiano de la función. Este análisis es crucial para comprender las garantías teóricas de convergencia y para anticipar el comportamiento de los diferentes métodos de optimización. La no convexidad de la función de Rosenbrock implica que los métodos de optimización podrían enfrentar desafíos significativos en su convergencia hacia el mínimo global.

3.1.3 Implementación del Gradiente Descendiente

El estudio sistemático del método de gradiente descendiente se realiza utilizando un punto inicial estándar en $(-1.5, 2)$, estableciendo como criterio de convergencia una norma del gradiente menor a 10^{-4}. Para analizar la sensibilidad del método a la tasa de aprendizaje, experimentamos con un conjunto de valores que van desde 0.0015 hasta 0.00205. La selección de este rango específico nos permite observar el comportamiento del método tanto en regímenes de convergencia lenta como en situaciones donde podrían surgir inestabilidades.

3.1.4 Variantes del Método de Optimización

La exploración se extiende a diversas variantes del método de gradiente descendiente. Implementamos el método del Hessiano, que aprovecha la información de segundo orden para determinar un tamaño de paso adaptativo. También estudiamos la incorporación de momentum, una técnica que añade un término de inercia para acelerar la convergencia y ayudar a superar oscilaciones locales. Adicionalmente, implementamos el método de backtracking line search, que ajusta dinámicamente el tamaño de paso para optimizar la convergencia en cada iteración.

3.1.5 Análisis Comparativo con Diferentes Puntos Iniciales

Para evaluar la robustez y eficiencia de los métodos implementados, diseñamos un conjunto de experimentos que utiliza múltiples puntos iniciales distribuidos estratégicamente en el espacio de búsqueda. Los puntos seleccionados incluyen posiciones cercanas y lejanas al mínimo global, así como puntos en diferentes regiones del valle característico de la función. Para cada punto inicial, registramos el número de iteraciones necesarias hasta alcanzar la convergencia y analizamos las trayectorias de optimización resultantes. Este diseño experimental nos permite evaluar sistemáticamente la sensibilidad de cada método a las condiciones iniciales y su capacidad para encontrar el mínimo global de la función de Rosenbrock.

3.2 Cuadrados Mínimos mediante Gradiente Descendiente

En esta sección, abordamos el problema de regresión lineal utilizando el conjunto de datos California Housing, que contiene información demográfica y económica de diferentes regiones de California. El objetivo es desarrollar un modelo lineal que permita predecir el valor medio de las viviendas a partir de características demográficas. Este problema se aborda mediante la minimización del error cuadrático medio (ECM), definido como

$$ECM(w) = \frac{1}{n}\|y - Xw\|_2^2 \tag{17}$$

donde $X \in \mathbb{R}^{n\times(d+1)}$ representa la matriz de características, $y \in \mathbb{R}^n$ el vector de valores objetivo, y $w \in \mathbb{R}^{d+1}$ el vector de coeficientes a optimizar.

3.2.1 Preparación y Partición de Datos

Para garantizar una evaluación robusta del modelo, implementamos una estrategia de división de datos donde el 80% de las muestras se destina al conjunto de entrenamiento y el 20% restante al conjunto de prueba. La selección de las muestras se realiza de manera aleatoria para evitar sesgos en la distribución de los datos. Previo al análisis, realizamos una estandarización de las variables predictoras, sustrayendo la media y dividiendo por la desviación estándar calculadas exclusivamente sobre el conjunto de entrenamiento. Este proceso de estandarización es crucial para asegurar que todas las variables contribuyan equitativamente al modelo y para mejorar la estabilidad numérica del proceso de optimización.

3.2.2 Solución Analítica y Análisis de Convexidad

Como punto de referencia, implementamos la solución analítica del problema de cuadrados mínimos mediante el cálculo de la pseudoinversa, expresada como $w = (X^T X)^{-1} X^T y$. Esta solución nos proporciona el óptimo global del problema y sirve como benchmark para evaluar la eficacia del método de gradiente descendiente. Paralelamente, realizamos un análisis de convexidad del ECM, verificando que la función objetivo es estrictamente

convexa cuando la matriz de diseño X tiene rango completo, lo cual garantiza la unicidad de la solución y la convergencia del método de gradiente descendiente.

3.2.3 *Implementación del Gradiente Descendiente*

Para la optimización mediante gradiente descendiente, exploramos diferentes tasas de aprendizaje basadas en el mayor valor singular de la matriz de diseño X. Específicamente, experimentamos con tres escalas: $\eta = 1/\sigma_1$, $\eta = 100/\sigma_1$ y $\eta = 10000/\sigma_1$, donde σ_1 es el mayor valor singular de X. La elección de estas tasas de aprendizaje se fundamenta en el análisis espectral de la matriz de diseño, donde $1/\sigma_1$ representa una cota inferior teórica para garantizar la convergencia del método. Este diseño experimental nos permite evaluar el comportamiento del algoritmo tanto en regímenes de convergencia conservadores como en situaciones más agresivas.

4 Resultados

4.1 Optimización mediante Gradiente Descendiente en Dos Dimensiones

4.1.1 *Visualización de la Función de Rosenbrock*

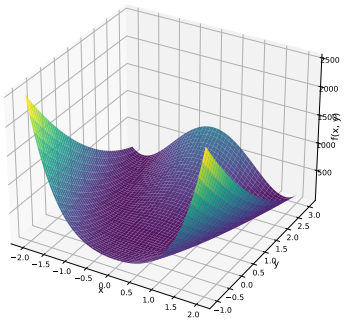

Fig. 1: Visualización 3D de la función de Rosenbrock.

La visualización de la función de Rosenbrock en tres dimensiones (Figura 1) revela su característica forma de "valle", con un mínimo global en $(1, 1)$. El análisis visual confirma la no convexidad de la función, lo cual se verifica posteriormente mediante el estudio de su matriz Hessiana. Esta no convexidad plantea un desafío particular para los métodos de optimización, ya que el valle tiene una curvatura pronunciada pero relativamente plana en la dirección del mínimo.

4.1.2 *Análisis del Gradiente Descendiente Básico*

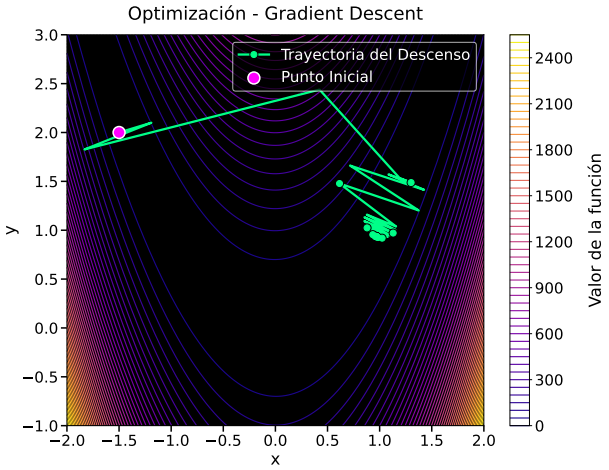

Fig. 2: Trayectoria del método de gradiente descendiente con $\eta = 0.002$.

La trayectoria del método de gradiente descendiente con una tasa de aprendizaje $\eta = 0.002$ (Figura 2) exhibe un comportamiento característico del método: inicialmente, realiza pasos grandes en la dirección del valle, pero a medida que se acerca al mínimo, comienza a oscilar entre las paredes del valle antes de converger finalmente después de 95 iteraciones. Este comportamiento de "zigzag" es típico del gradiente descendiente en funciones que presentan valles estrechos, donde el método debe equilibrar el avance en la dirección del valle con la necesidad de mantenerse dentro de sus límites.

4.1.3 *Sensibilidad a la Tasa de Aprendizaje*

El análisis de diferentes tasas de aprendizaje revela un comportamiento altamente sensible del método (Figura 3). Para $\eta = 0.0015$, el algoritmo avanza tan lentamente que no logra converger dentro de las 3000 iteraciones establecidas como límite. Con $\eta = 0.0019$, se observa convergencia después de aproximadamente 2000 iteraciones. La tasa óptima encontrada fue $\eta = 0.002$, que logra la convergencia en solo 95 iteraciones. Sin embargo, pequeños incrementos en la tasa de aprendizaje degradan rápidamente el rendimiento: con $\eta = 0.00201$, el número de iteraciones aumenta a 250, y con $\eta = 0.00205$, se requieren más de 2500 iteraciones.

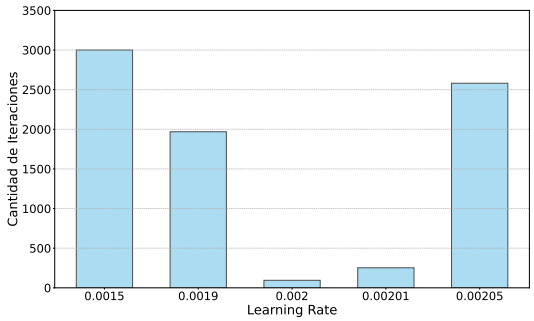

Fig. 3: Convergencia del método de gradiente descendiente con distintos η

Este comportamiento se explica por la geometría de la función: tasas de aprendizaje muy pequeñas resultan en un avance excesivamente cauteloso, mientras que tasas ligeramente mayores al óptimo provocan que el algoritmo "rebote" entre las paredes del valle con mayor intensidad, retrasando la convergencia.

4.1.4 *Análisis de Métodos Alternativos*

El método de Newton (Figura 4) demuestra una eficiencia notable al converger en solo 7 iteraciones. Esto se debe a su capacidad para utilizar la información de curvatura proporcionada por la matriz Hessiana, permitiéndole adaptar la dirección y tamaño de los pasos de manera óptima.

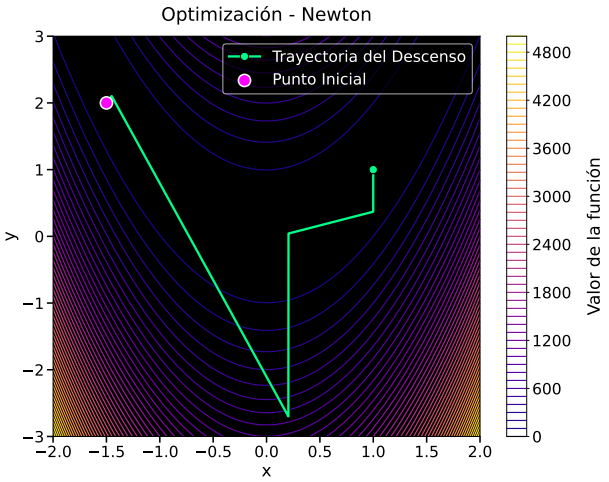

Fig. 4: Trayectoria del método de Newton.

El método de backtracking line search (Figura 5) exhibe un rendimiento intermedio, convergiendo en aproximadamente 20 iteraciones. Su capacidad para ajustar dinámicamente el tamaño del paso le permite evitar las oscilaciones excesivas características del gradiente descendiente básico.

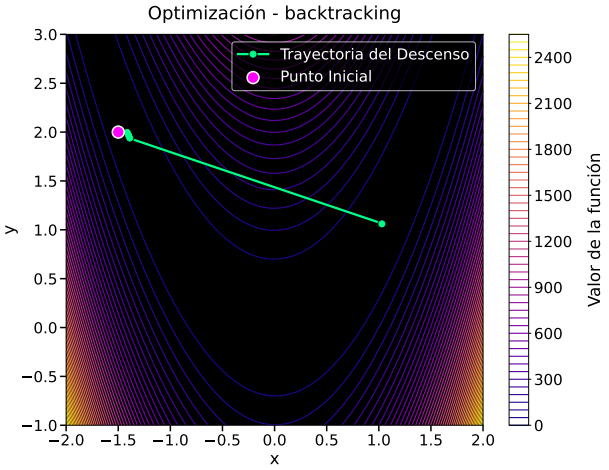

Fig. 5: Trayectoria del método con Backtracking.

La variante con momentum (Figura 6), si bien requiere más iteraciones (369) que el método de Newton o backtracking, muestra una trayectoria más suave. El término de momentum ayuda a mantener una dirección consistente de descenso, reduciendo las oscilaciones pero a costa de una convergencia más lenta.

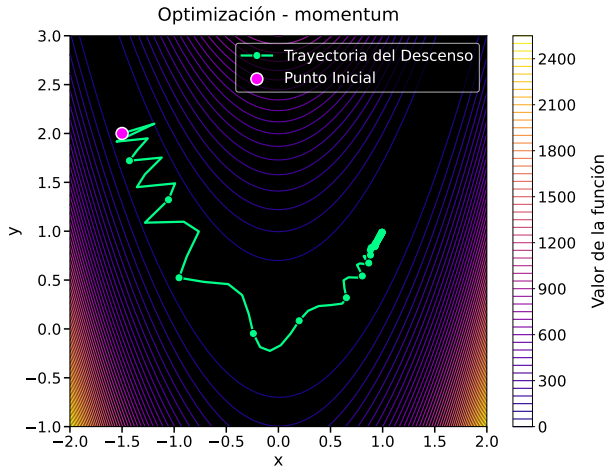

Fig. 6: Trayectoria del método con Momentum.

4.1.5 *Análisis Comparativo con Diferentes Puntos Iniciales*

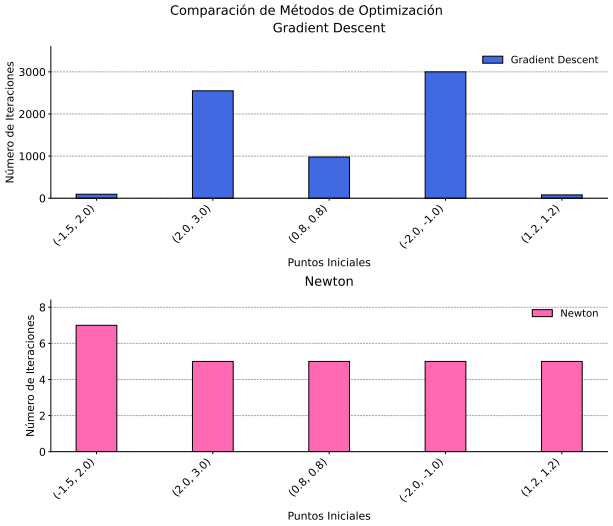

Fig. 7: Comparación del número de iteraciones requeridas por cada método para diferentes puntos iniciales.

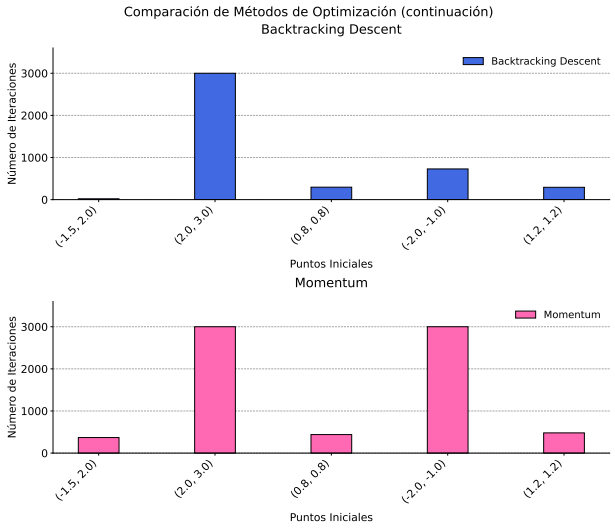

Fig. 8: Comparación del número de iteraciones requeridas por cada método para diferentes puntos iniciales.

La comparación de los cuatro métodos con diferentes puntos iniciales (Figura 7) revela patrones interesantes en su comportamiento. El método de Newton mantiene un rendimiento consistentemente bueno, requiriendo entre 5 y 7 iteraciones independientemente del punto inicial. Esto demuestra su robustez y eficiencia en la navegación del espacio de búsqueda.

El gradiente descendiente básico y el momentum muestran una mayor variabilidad en su rendimiento, con algunos puntos iniciales requiriendo más de 2500 iteraciones para converger. Particularmente, los puntos iniciales $(2.0, 3.0)$ y $(-2.0, -1.0)$ resultan especialmente desafiantes para estos métodos, posiblemente debido a su ubicación en regiones donde el valle de la función es más pronunciado.

El método de backtracking demuestra un comportamiento intermedio, siendo más robusto que el gradiente descendiente básico pero sin alcanzar la consistencia del método de Newton. Su capacidad para ajustar dinámicamente el tamaño del paso le permite manejar mejor los diferentes puntos iniciales, aunque sigue mostrando cierta sensibilidad a la ubicación inicial.

4.2 Gradiente Descendiente para Cuadrados Mínimos

4.2.1 *Comparación inicial entre la pseudoinversa y el gradiente descendente*

En el primer gráfico (Figura 9), se visualizan las predicciones obtenidas utilizando los coeficientes $\mathbf{w}$ de cada método. Se observa claramente cómo las predicciones del gradiente descendente divergen significativamente de los valores esperados, mientras que la pseudoinversa proporciona una aproximación mucho más precisa.

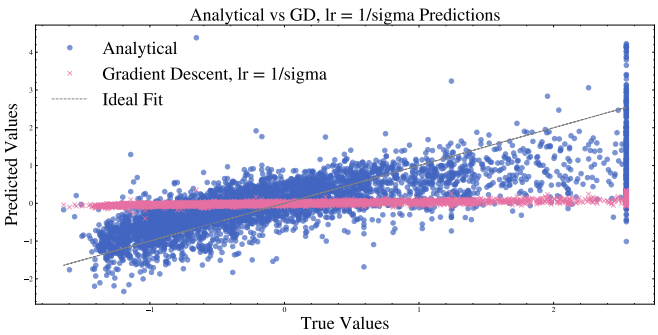

Fig. 9: Predicciones de Pseudoinversa y GD ($\eta = \frac{1}{\sigma^2}$)

Haciendo un análisis del learning rate, vimos que este valor resultó ser extremadamente pequeño ($\eta \approx 2.89 \times 10^{-5}$), lo que condujo a una convergencia lenta y una solución de $\mathbf{w}$ que proporcionaba predicciones notablemente malas en comparación con la solución obtenida mediante la pseudoinversa.

4.2.2 *Análisis del error con distintas tasas de aprendizaje*

Para mejorar la solución obtenida mediante gradiente descendente, evaluamos el error cuadrático medio (ECM) utilizando distintas tasas de aprendizaje. Como se especifica en 3.2.3, probamos valores de $\eta = \frac{100}{\sigma^2}$ y $\eta = \frac{10000}{\sigma^2}$ y obtuvimos el siguiente gráfico:

Fig. 10: Error de predicción con distintos η.

Los resultados de la Figura 10 donde se presenta la evolución del error en función de las distintas tasas de aprendizaje muestran que, al aumentar η, el gradiente descendente convergió más rápidamente hacia una solución adecuada. En particular, con $\eta = \frac{10000}{\sigma^2}$, el error cuadrático medio fue prácticamente idéntico al obtenido con la pseudoinversa.

4.2.3 *Visualización de las predicciones con la tasa de aprendizaje ajustada*

Finalmente, utilizamos $\eta = \frac{10000}{\sigma^2}$ para calcular los coeficientes $\mathbf{w}$ mediante gradiente descendente. Las predicciones generadas con este valor de $\mathbf{w}$ se compararon con las predicciones de la pseudoinversa, mostrando resultados prácticamente idénticos. Esta equivalencia se ilustra claramente en el tercer gráfico (Figura 11), donde mayormente las predicciones son superpuestas, destacando la efectividad de este ajuste en la tasa de aprendizaje.

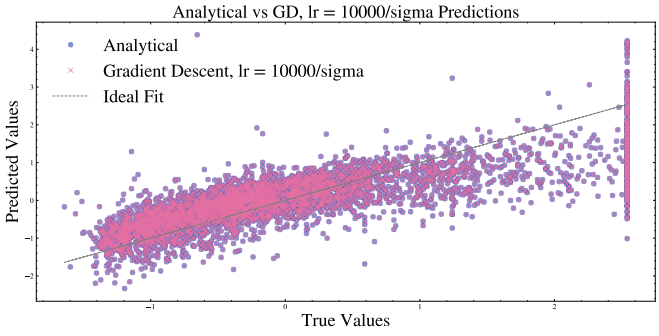

Fig. 11: Predicciones de Pseudoinversa y GD ($\eta = \frac{10000}{\sigma^2}$)

5 Conclusiones

Este trabajo ha explorado en profundidad el comportamiento y la eficacia de diferentes métodos de optimización, centrándose en el gradiente descendiente y sus variantes. A través del análisis de

la función de Rosenbrock y el problema de regresión lineal en el conjunto de datos California Housing, hemos llegado a varias conclusiones significativas.

En el contexto de la optimización bidimensional con la función de Rosenbrock, los resultados demuestran la crucial importancia de la selección adecuada de parámetros y la sensibilidad de los métodos a las condiciones iniciales. El método de Newton demostró ser el más eficiente y robusto, convergiendo consistentemente en 5-7 iteraciones independientemente del punto inicial. Este rendimiento superior se atribuye a su capacidad para utilizar la información de segundo orden de la función objetivo. Por su parte, el método de backtracking line search emergió como una alternativa práctica al gradiente descendiente básico, ofreciendo un buen balance entre eficiencia computacional y velocidad de convergencia, con un promedio de 20 iteraciones hasta la convergencia.

La implementación con momentum, si bien mostró trayectorias más suaves, no superó significativamente al gradiente descendiente básico en términos de eficiencia global, requiriendo aproximadamente 369 iteraciones para converger. Un hallazgo particularmente relevante fue la extrema sensibilidad a la tasa de aprendizaje en el gradiente descendiente básico, donde variaciones de apenas 0.00001 en el valor de η causaron diferencias significativas en el número de iteraciones necesarias para la convergencia.

En cuanto al problema de regresión lineal mediante cuadrados mínimos, el estudio reveló aspectos importantes sobre la implementación práctica del gradiente descendiente. La elección teórica de $\eta = 1/\sigma_1^2$ resultó ser demasiado conservadora, llevando a una convergencia excesivamente lenta y predicciones subóptimas. Sin embargo, el ajuste empírico de la tasa de aprendizaje a $\eta = 10000/\sigma_1^2$ produjo resultados prácticamente idénticos a los obtenidos mediante la solución analítica de la pseudoinversa, demostrando que las garantías teóricas de convergencia pueden ser demasiado restrictivas en la práctica.

La comparación entre los errores cuadráticos medios mostró que, con una tasa de aprendizaje adecuada, el gradiente descendiente puede alcanzar la misma precisión que los métodos directos, aunque requiriendo más iteraciones. Esta observación sugiere que, si bien los métodos de optimización basados en gradiente son herramientas poderosas, su implementación práctica requiere un cuidadoso balance entre la teoría y la experimentación empírica.

Los resultados también destacan la importancia de considerar múltiples enfoques de optimización, ya que diferentes métodos pueden ser más apropiados dependiendo de las características específicas del problema, los requisitos de precisión y los recursos computacionales disponibles. La elección del método y sus parámetros debe considerar no solo las garantías teóricas de convergencia, sino también las características específicas del problema y los requisitos de rendimiento computacional.