Universidad de la República

Facultad de Ingeniería

Fundamentos de Optimización

Segundo obligatorio

Autores:
Federico Bello
Julieta Umpierrez

28 de septiembre de 2025

UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

Índice

1. Parte 1

Se dice que un conjunto $A \subseteq \mathbb{R}^n$ es convexo sii el segmento de recta que une un par de puntos cualquiera esta también incluido en A, es decir:

$$\forall x_1, x_2 \in A \text{ se cumple que: } \lambda x_1 + (1-\lambda)x_2 \in A \text{ con } \lambda \in (0,1)$$

1.1. Parte a

Demostración. Sean A_1 y A_2 dos conjuntos convexos cualquiera, sea $A_3 = A_1 \cap A_2$ y sean $x_1, x_2 \in A_3$. Por lo tanto, $x_1, x_2 \in A_1$ y $x_1, x_2 \in A_2$ Utilizando la definición de conjunto convexo:

$$\lambda x_1 + (1-\lambda)x_2 \in A_1, \lambda \in (0,1)$$

$$\lambda x_1 + (1-\lambda)x_2 \in A_2, \lambda \in (0,1)$$

Por lo que: $\lambda x_1 + (1-\lambda)x_2 \in A_1 \cap A_2$ con $\lambda \in (0,1)$ □

1.2. Parte b

Intuitivamente, se puede ver que si la intersección de dos conjuntos convexos es convexa, la intersección de N conjuntos también lo es, ya que se pueden tomar intersecciones de pares de conjuntos, definir nuevos conjuntos y repetir iterativamente. Por ejemplo, si se tiene A_1, A_2, A_3 convexos, se vio en la parte anterior que $S = A_1 \cap A_2$ es convexo, por lo que por el mismo razonamiento $S \cap A_3$ es convexo.

Demostración. Siguiendo un razonamiento similar al usado en la demostración anterior, para todo par de puntos $x_1, x_2 \in \bigcap_{i \in \mathbb{N}}^{N} A_i$ se cumple que:

$$\lambda x_1 + (1-\lambda)x_2 \in A_i, \lambda \in (0,1)$$

para todo $i = 1, 2, ...N$, por lo tanto:

$$\lambda x_1 + (1-\lambda)x_2 \in \bigcap_{i \in \mathbb{N}}^{N} A_i, \lambda \in (0,1)$$

□

2. Parte 2

2.1. Parte a

Para esta parte se implemento el método de descenso por gradiente con dirección de máximo descenso y paso fijo siguiendo el algoritmo 1. Esto es tomar la dirección de descenso como $-\nabla f(x_k)$. En cuanto al paso óptimo se recuerda que dados m y M, mínimo y máximo valor propio de la matriz Hessiana respectivamente, se tiene que $\alpha_{opt} = \frac{2}{m+M}$, mientras que la convergencia esta asegurada si $\alpha < \frac{2}{M}$. El código del mismo se puede encontrar en el notebook adjunto en la entrega.

Algorithm 1 Descenso por Gradiente

Require: α el valor del paso y tol la tolerancia
  Sea $\mathbf{x_0}$ el valor inicial elegido con algún criterio y $g(\mathbf{x})$ el gradiente de la función a optimizar
  $\mathbf{x} \leftarrow \mathbf{x_0}$
  while $g(\mathbf{x}) > tol$ **do**
    $\mathbf{x} \leftarrow \mathbf{x} - \alpha g(\mathbf{x})$
  end while
  return x

2.2. Parte b

Para esta parte se implemento el método de descenso por gradiente acelerado utilizando el algoritmo 2. A su vez se tomo el parámetro $\beta = \frac{k-1}{k+2}$ como se vio en el teórico. El código del mismo se puede encontrar en el notebook adjunto en la entrega. La inicialización de y_0 se realizo con x_0 para que en el primer paso sea igual a la sección 2.1 y luego se tome la corrección utilizando la memoria.

Algorithm 2 Descenso por Gradiente Acelerado (Nesterov)

Require: α el valor del paso y tol la tolerancia
  Sea $\mathbf{x_0}$ el valor inicial elegido con algún criterio y $g(\mathbf{x})$ el gradiente de la función a optimizar
  $\mathbf{x} \leftarrow \mathbf{x_0}$
  $\mathbf{y} \leftarrow \mathbf{x_0}$
  while $g(\mathbf{x}) > tol$ **do**
    $\mathbf{x_{ant}} \leftarrow \mathbf{x}$
    $\mathbf{x} \leftarrow \mathbf{y} - \alpha g(\mathbf{y})$
    $y \leftarrow x + \beta(\mathbf{x} - \mathbf{x_{ant}})$
  end while
  return x

2.3. Parte c

Aclaración previa: La función de Rosenbrock no es una función convexa, por lo que la teoría dada en el curso no aplica para la misma. Sin embargo, si es localmente convexa en un entorno al mínimo, por lo que tomando un punto inicial cercano a este es posible ignorar el problema.

2.3.1. Experimentos con tamaños de paso

Dado que se realizaron todos los experimentos con paso fijo se tiene que en el caso de $f(x) = \frac{1}{2}||Ax-b||^2$, la hessiana es $\nabla^2 f(x) = A^T A$ por lo que el alpha óptimo es fijo. Sin embargo, en el caso de la función Rosenbrock la hessiana depende de x por lo que el alpha óptimo no es fijo. Es por esto que para el caso de $f(x) = \frac{1}{2}||Ax-b||^2$ se compara el alpha óptimo con otros valores entre $(0, 2/M)$ ya que $2/M$ es el máximo valor para el cual la convergencia esta asegurada y en el caso Rosenbrock se eligieron pasos que generaran convergencia, probando distintos valores.

En la figura 1 se adjuntan las curvas de error para [1] $f(x) = \frac{1}{2}||Ax-b||^2$ en función de las iteraciones para varios valores de alpha en el caso de descenso por gradiente con dirección de máximo descenso (GDSD). Se puede evidenciar como para el alpha óptimo se genera la convergencia mas rápida en termino de iteraciones. Se observa también la tendencia de que, cuanto menor el alfa, mas lenta es la convergencia, ya que en cada paso se avanza menos. Sin embargo, como toda regla, tiene excepciones, ya que si se toma un alfa demasiado grande (0.00021) la cantidad de iteraciones aumenta drásticamente.

[1] Con eje vertical logarítmico, este comentario aplica para todas las gráficas de error en el presente informe

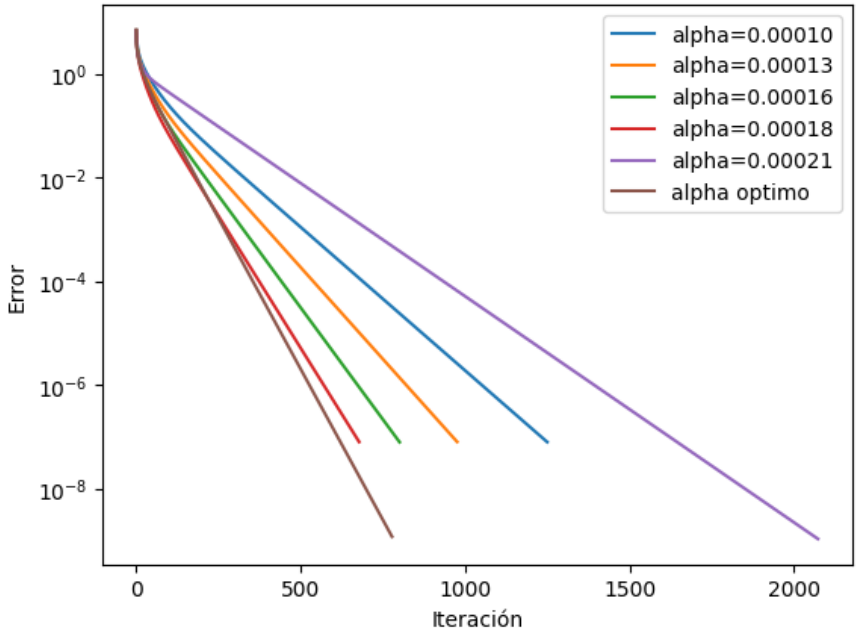


Figura 1: Gráfica de error en función de iteraciones para varios valores de alpha para $f(x) = \frac{1}{2}||Ax - b||^2$ (GDSD)

Luego en la figura 2 se adjunta la experimentación con distintos valores de alpha, utilizando el método de Nesterov para $f(x) = \frac{1}{2}||Ax - b||^2$. En esta se puede apreciar el efecto de la elección de alpha en la cantidad de iteraciones. Sobre todo viendo como al aumentar alpha se llega a menor error en menos iteraciones, es importante aclarar que, al igual al caso anterior, ese alpha no se puede aumentar infinitamente y que cerca del valor para la curva violeta comienza a diverger. Otra cosa a destacar es la diferencia en la forma entre estas curvas en comparación con las vistas para gradiente descendiente. En Nesterov estamos frente a la presencia de lóbulos, lo cual es completamente esperable dado que no se baja en la dirección de máximo descenso, o, aun mas, no tiene por que ser una dirección de descenso. Por lo tanto, el error no tiene por que ser monótono decreciente.

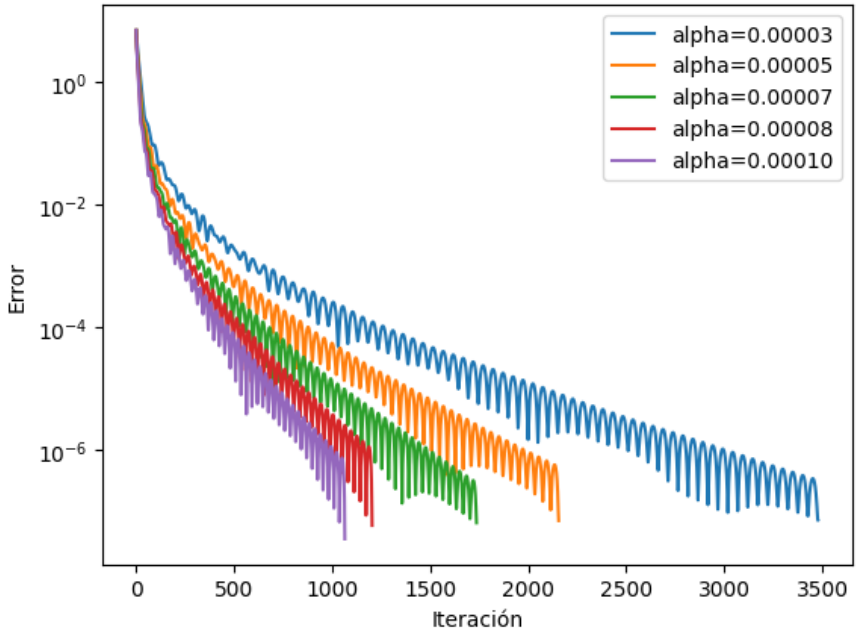


Figura 2: Gráfica de error en función de iteraciones para varios valores de alpha para $f(x) = \frac{1}{2}||Ax - b||^2$ (Nesterov)

Pasando a la función de Rosenbrock se tiene en las figuras 3 y 4 las curvas de error en función de la cantidad de iteraciones para dirección de máximo descenso y para Nesterov respectivamente. Para esta función en particular no es posible utilizar paso fijo con un alpha óptimo, dado que la Hessiana depende del punto. Por lo tanto, lo que aparece en la figura como alpha óptimo es el algoritmo correspondiente modificado para utilizar un tamaño de paso variable, evaluando la Hessiana en cada iteración. Para Nesterov, l β se eligió como fue aclarado en la sección 2.2. Las observaciones en este caso son las mismas que para la función anterior, remarcando como en general el aumento de alpha genera una disminución en la cantidad de iteraciones pero teniendo en cuenta que el valor mas grande de alpha presentado es cercano al limite superior de convergencia.

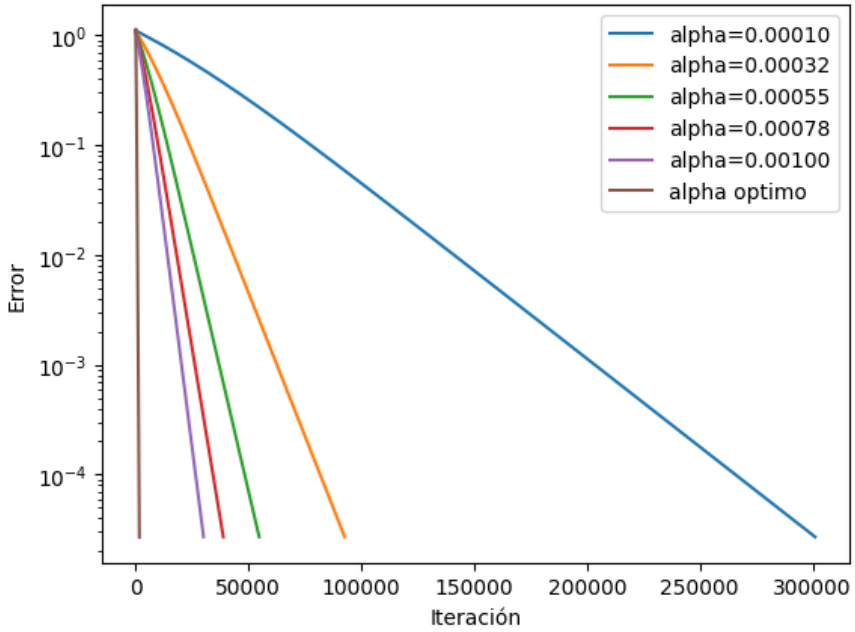


Figura 3: Gráfica de error en función de iteraciones para varios valores de alpha para Rosenbrock (GDSD)

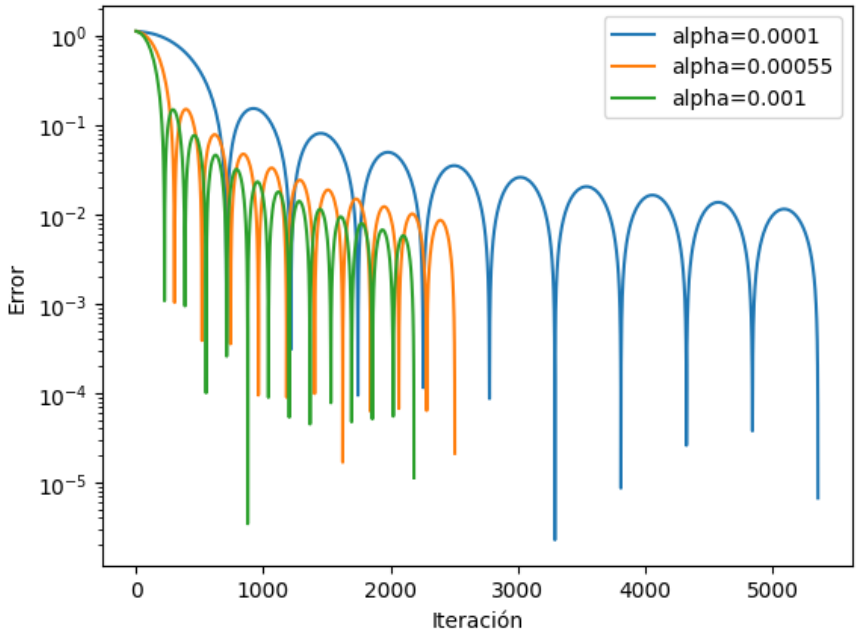


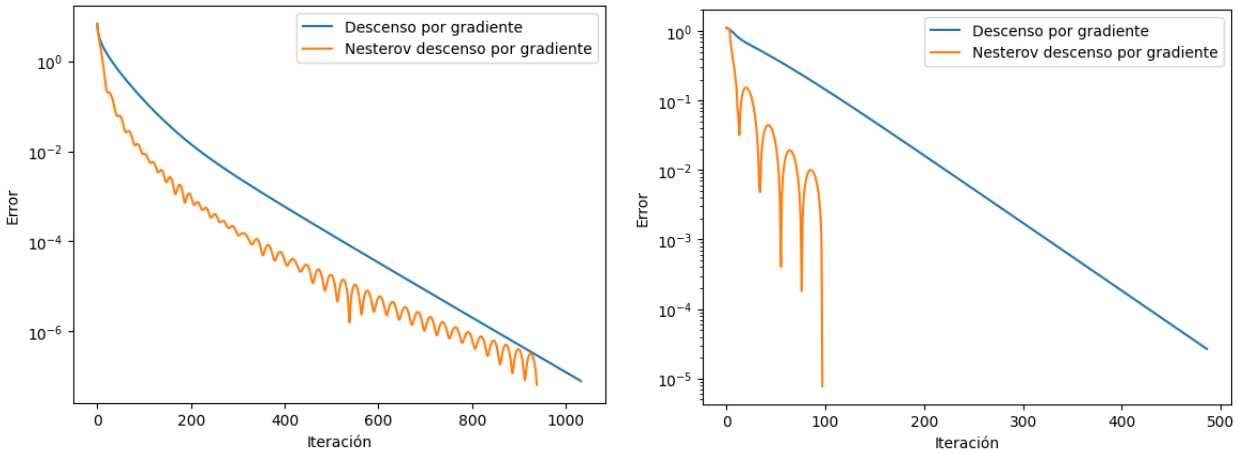
Figura 4: Gráfica de error en función de iteraciones para varios valores de alpha para Rosenbrock(Nesterov)

2.3.2. Comparación de métodos

Si bien a priori parece fácil realizar la comparación, la cantidad de iteraciones y el tiempo tienen una fuerte dependencia de la realización, ya que dependen del punto inicial. Es por esto que se realizaron varias realizaciones, analizando la frecuencia en la que cada uno 'gana' en las respectivas métricas. Para el caso de $f(x) = \frac{1}{2}||Ax - b||^2$ al realizar el experimento con 1000 inicializaciones aleatorias con valores entre 0 y 1, se obtiene que el descenso por gradiente Nesterov llega al óptimo con menos iteraciones en la amplia mayoría de casos, sin embargo, en tiempo, tienen resultados casi idénticos, incluso siendo levemente mejor descenso por

gradiente con dirección de máximo descenso. Esto se debe a que, si bien el descenso por gradiente acelerado llega al óptimo en menos iteraciones, es necesario realizar mas cuentas en cada paso. Esto es verdad siempre que la inicialización sea cercana al punto óptimo. Cuanto mas se aleje el punto inicial mejor es el descenso por gradiente acelerado en ambas métricas, ya que mayor es la diferencia en cantidad de iteraciones.

En la figura 5 se presenta la función de error en función de las iteraciones para $f(x) = \frac{1}{2}||Ax - b||^2$ y para Rosenbrock con ambos métodos. En el caso de $f(x)$ se inicializo con valores entre 0 y 1 y se utilizo un valor de alpha ligeramente menor al óptimo para que Nesterov también sea convergente ($\alpha_{opt} - 1 \times 10^{-4}$). En el caso de Rosenbrock se utilizo un valor que garantizaba convergencia en ambos (6×10^{-2}).



(a) Función $f(x) = \frac{1}{2}||Ax - b||^2$ (b) Función de Rosenbrock

Figura 5: Comparación de GDSD con Nesterov para distintas funciones

Lo primero a destacar es como para ambas gráficas se nota el efecto de agregar el momentum de Nesterov en la cantidad de iteraciones. Sin embargo es claro que el mayor efecto se da en la función de Rosenbrock. Recordar que para GDSD la tasa de convergencia es lineal, es decir, $1 - \frac{2}{\mathcal{K}+1}$ [2], mientras que para Nesterov es $1 - \frac{2}{\sqrt{\mathcal{K}}+1}$. Estas ecuaciones implican que cuanto peor condicionada es la matriz, mayor es el beneficio de utilizar Nesterov. Esto implica que la Hessiana en Rosenbrock queda mucho peor condicionada que la de la otra función, esto además se podíra intuir por las curvas de nivel de Rosenbrock. Es por esto que es mas evidente el cambio en la tasa de convergencia al utilizar Nesterov en la función de Rosenbrock que usando $f(x) = \frac{1}{2}||Ax - b||^2$.

A modo de resumen, en el caso general, si bien utilizar Nesterov implica la necesidad de hacer mas cálculos en cada iteración esta 'inversión' suele ser positiva, ya que se llega al óptimo en menos iteraciones y usualmente en menos tiempo. A priori, no parece necesario cambiar el nombre acelerado.

2.3.3. ¿Los métodos son de descenso?

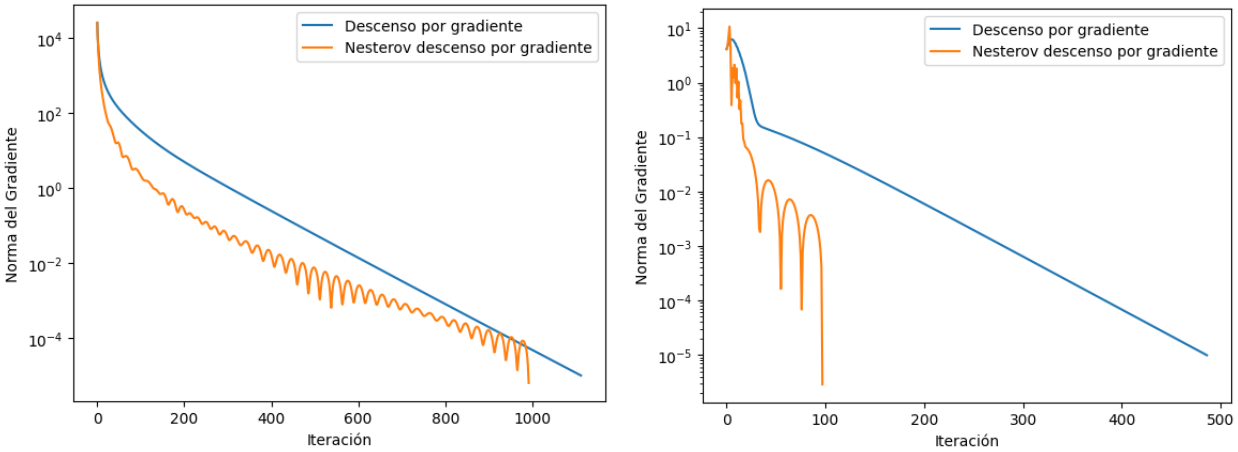
Se dice que un método es de descenso si para cada paso de la iteración se cumple [1]boyd2004convex:

$$f(x^{(k+1)}) < f(x^{(k)})$$

excepto cuando $x^{(k)}$ es óptimo, donde f es la función a minimizar y $x^{(k)}$ el punto en la k-estima iteración. No es difícil ver que en el caso de funciones convexas, es equivalente esta condición a que el error sea estrictamente decreciente.

[2] $\mathcal{K}$ es el numero de condición de la Hessiana

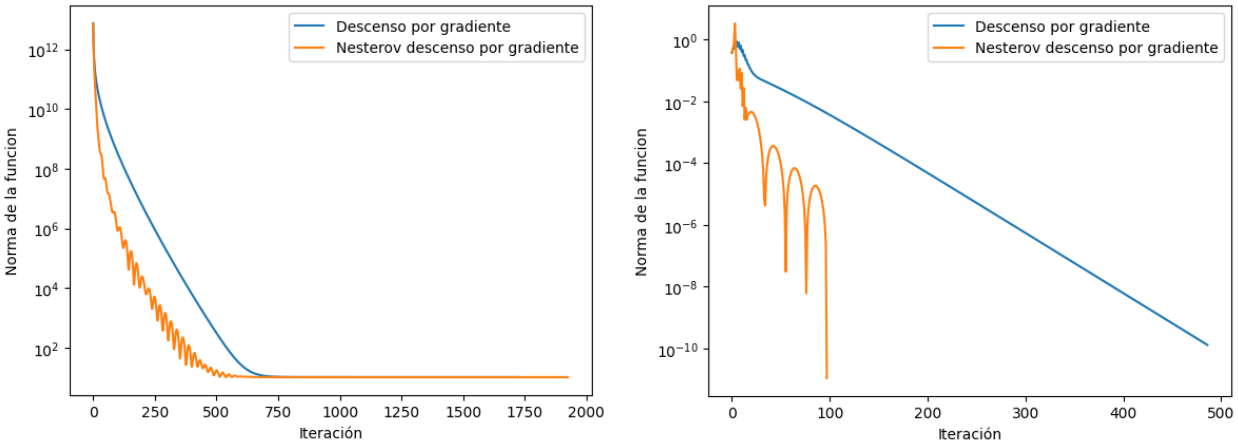
En la figura 6 se observa como varia la norma del gradiente en función de las iteraciones para cada a una de las funciones. Se ve claramente como para cualquiera de los dos métodos el valor del gradiente tiende a cero, llegando al mínimo valor de la función. Sin embargo, viendo la figura 5 se ve que mientras que el descenso por gradiente en dirección de máximo descenso obtiene una función estrictamente decreciente, Nesterov no lo hace, ya que al tener memoria, no necesariamente en cada paso se decrementa el valor de la función.



(a) Función $f(x) = \frac{1}{2}||Ax - b||^2$ (b) Función de Rosenbrock

Figura 6: Norma del gradiente en los distintos métodos para distintas funciones

Esto puede verse directamente en la figura 7, donde se observa claramente que para ambos casos el GDSD resulta en una sucesión estrictamente decreciente, mientras que al utilizar el método acelerado, si bien tiende al mínimo llega a este mismo oscilando. Notar que, debido al comentario realizado sobre la convexidad de Rosenbrock, al comienzo no se tiene un método de descenso. Luego, al irse acercando al mínimo la función se comporta como una función convexa por lo que se puede aplicar el concepto de método de descenso.



(a) Función $f(x) = \frac{1}{2}||Ax - b||^2$ (b) Función de Rosenbrock

Figura 7: Norma de la función en los distintos métodos para distintas funciones

2.3.4. Curvas de nivel en Rosenbrock

En la figura 8 se observan las curvas de nivel de la función de Rosenbrock y, superpuestas, las respectivas trayectorias de cada uno de los métodos de descenso. Para aplicar los métodos se modifico el alpha a uno lejano del óptimo, con el objetivo de poder apreciar mejor la diferencia entre ambos. En las primeras iteraciones (arriba a la derecha en la imagen) se observa claramente la diferencia entre ambos métodos. Mientras el descenso por gradiente en la dirección de máximo descenso toma pasos demasiado grandes, avanzando en zig-zag, Nesterov soluciona esto, recordando el gradiente del paso anterior, consiguiendo así evitar el zig-zagueo.

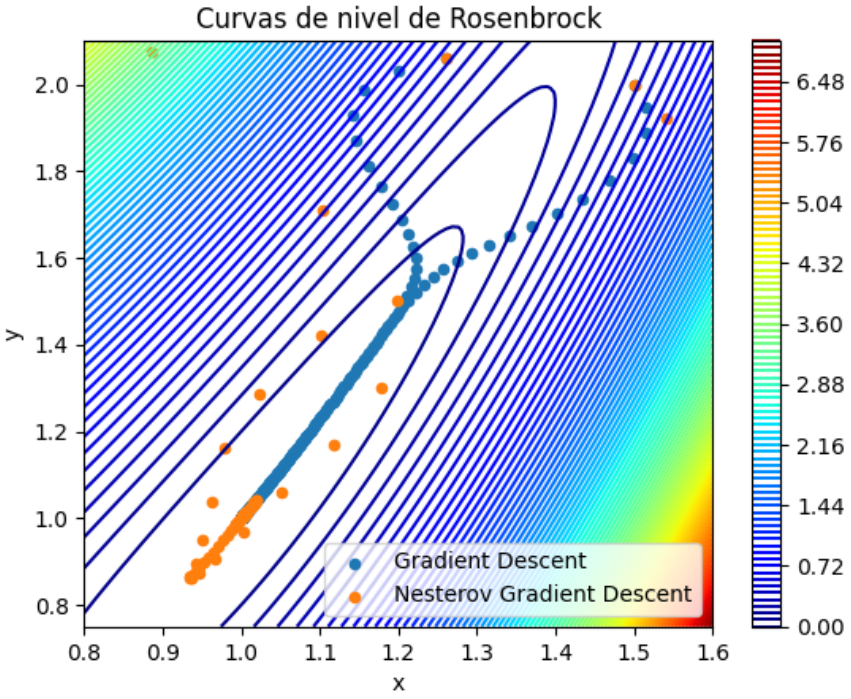


Figura 8: Curvas de nivel de la función de Rosenbrock con las trayectorias de cada método

2.4. Parte d

Para esta parte se implemento el método de descenso por gradiente con dirección de máximo descenso y paso decreciente siguiendo el algoritmo 3. Esto es tomar la dirección de descenso como $-\nabla f(x_k)$, mientras que para la eleccion del tamaño del paso es necesario elegir una sucesión decreciente a_k, cuya serie $\sum_{k=0}^{\infty} a_k$ sea convergente. El código del mismo se puede encontrar en el notebook adjunto en la entrega.

Algorithm 3 Descenso por Gradiente con paso Decreciente

Require: a_k una sucesión decreciente cuya serie $\sum_{k=0}^{\infty} a_k$ sea convergente y tol la tolerancia.
  Sea $\mathbf{x_0}$ el valor inicial elegido con algún criterio y $g(\mathbf{x})$ el gradiente de la función a optimizar
  $k \leftarrow 0$
  $\mathbf{x} \leftarrow \mathbf{x_0}$
  while $g(\mathbf{x}) > tol$ **do**
    $\mathbf{x} \leftarrow \mathbf{x} - a_k g(\mathbf{x})$
    $k \leftarrow k + 1$
  end while
  return x

La figura 9 muestra la curva de error en función de la cantidad de iteraciones para este método en comparación con los métodos anteriormente descriptos. Para la misma se utilizo la sucesión $a_k = \frac{0{,}001}{\sqrt{k+1}}$, donde k es iteración actual. La razón de esta elección tan rebuscada a primera vista es que si bien matemáticamente alcanza con las condiciones anteriormente mencionadas, computacionalmente si los primeros pasos son demasiado grandes se llega rápidamente a un *overflow*. En la figura se observa claramente la desventaja de este método respecto a los anteriores, y es que si no se elige correctamente la sucesión, los primeros pasos pueden llevar a alejarse del mínimo en lugar de acercarse, siendo iteraciones desperdiciadas. Además, si la sucesión tiende a 0 muy rápidamente, el método demorara en converger, teniendo así un compromiso en la sucesión a elegir. En la figura también puede verse como el método no es de descenso.

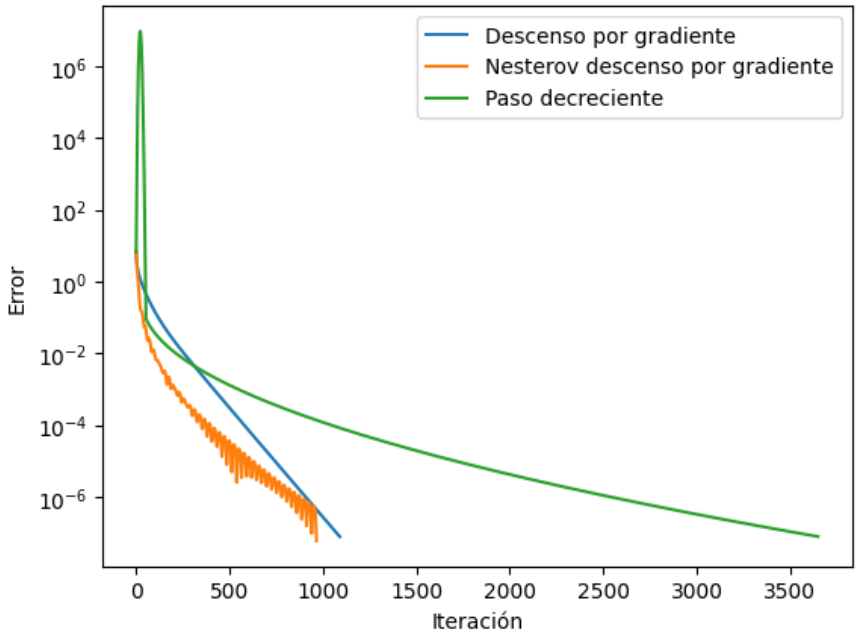


Figura 9: Gráfica de error en función de iteraciones para método con paso decreciente para $f(x)$

Sin embargo, no todo es negativo, una de las principales ventajas del descenso por gradiente con paso decreciente es que permite el uso de un tamaño de paso inicial mayor, lo que puede resultar en una convergencia mas rápida. A medida que el algoritmo se acerca a la solución, el tamaño del paso se va reduciendo gradualmente, lo que permite un ajuste más fino cerca del mínimo. Esto puede ser particularmente útil en funciones en donde el gradiente es chico lejos del mínimo.

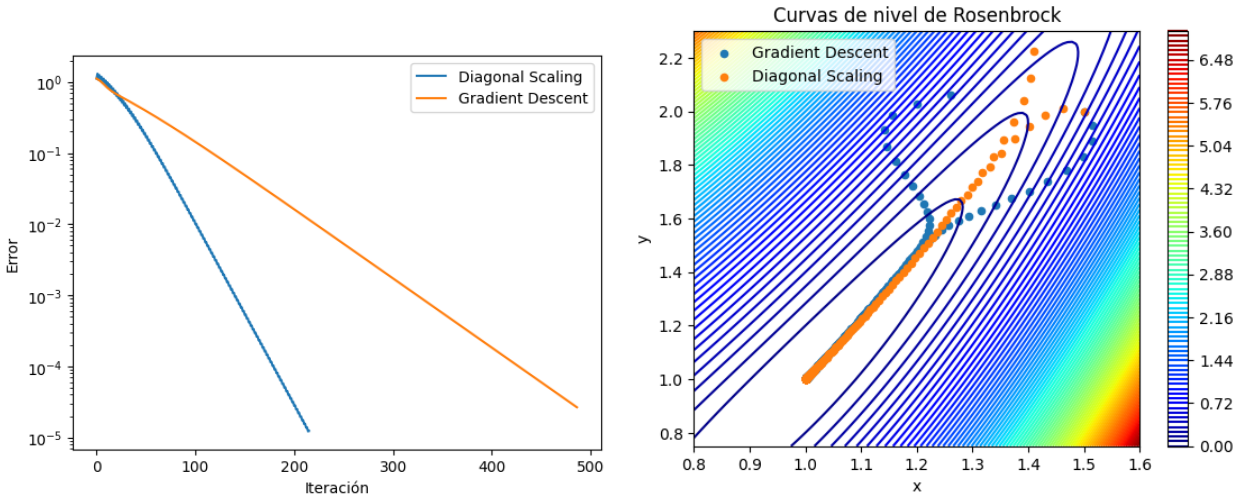
2.5. Parte e

Para esta ultima parte se implemento el método de descenso por gradiente con *diagonal scaling* y paso fijo siguiendo el algoritmo 4. Esto es tomar la dirección de descenso como $-D_H^{-1}(\mathbf{x})\nabla f(x_k)$, donde $D_H(\mathbf{x})$ es la diagonal de la matriz Hessiana evaluada en el punto $\mathbf{x}$. El código del mismo se puede encontrar en el notebook adjunto en la entrega.

Este método tiene la clara desventaja de precisar información de segundo orden, lo cual no siempre es posible debido a la naturaleza de cada problema. Sin embargo, tiene la ventaja de lograr adaptarse a la geometría de cada función, ponderando cada coordenada del descenso y obteniendo una trayectoria mas directa. La figura 10 compara este ultimo método y el descenso por gradiente en dirección de máximo descenso. En particular, la figura 10b muestra, el efecto de agregar la información de segundo orden, resultando en una convergencia mucho mas directa.

Algorithm 4 Descenso por Gradiente con diagonal scaling y paso fijo

Require: α el valor del paso , tol la tolerancia y $D_H(\mathbf{x})$ la diagonal de la Hessiana en el punto $\mathbf{x}$.
Sea $\mathbf{x_0}$ el valor inicial elegido con algún criterio y $g(\mathbf{x})$ el gradiente de la función a optimizar
$\mathbf{x} \leftarrow \mathbf{x_0}$
while $g(\mathbf{x}) > tol$ **do**
$\mathbf{x} \leftarrow \mathbf{x} - \alpha D_H^{-1}(\mathbf{x}) g(\mathbf{x})$
end while
return $\mathbf{x}$



(a) Error en función de iteraciones

(b) Trayectoria de cada método

Figura 10: Descenso por gradiente utilizando *diagonal scaling* en comparación a máximo descenso

La ventaja de este método respecto a otros métodos que también agregan información de segundo orden, como el método de Newton, es que al utilizar solamente la diagonal, se obtiene un algoritmo mucho menos costoso computacionalmente, ya que para conseguir la matriz inversa alcanza con invertir cada entrada de la diagonal.

Referencias

[1] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.