

Aceleración del método de Barzilai-Borwein
Edgar Osvaldo López Zúñiga
Giovanni Gamaliel López Padilla

Introducción

El estudio de la solución de la optimización de una función cuadrática de la forma

$$f(x) = \frac{1}{2}x^T Ax - b^T x \quad (1)$$

donde $b \in \mathbb{R}^n$ y A es una matriz positiva definida en $\mathbb{R}^{n \times n}$, es equivalente a la solución del sistema de ecuaciones

$$Ax = b$$

El problema empieza a tener dificultades cuando el tamaño de la matriz aumenta y esta contiene ceros en sus elementos (caso sparse). Esto provoca que la factorización por Cholesky sea impráctica por el tiempo de ejecución o el espacio en memoria. En 1988 Barzilai y Borwein¹ propusieron dos tamaños de paso para mejorar el desempeño de métodos de descenso de gradiente. La elección de los tamaños de paso propuestos están basados en métodos Cuasi-Newton, donde el tamaño de paso (α_k) es remplazada por la matriz D_k tal que

$$D_k = \alpha_k I \quad (2)$$

La matriz D_k es una aproximación de la inversa del hessiano. El tamaño de paso es calculado a partir de la optimización de D_k^{-1} (BB1) y D_k (BB2) tal que satisfagan la ecuación de la secante desde un punto de vista de mínimos cuadrados (ecuación 3).

$$\min_{D=\alpha I} \|D^{-1}s_{k-1} - y_{k-1}\| \quad \min_{D=\alpha I} \|s_{k-1} - Dy_{k-1}\| \quad (3)$$

donde $s_{k-1} = x_k - x_{k-1}$ y $y_{k-1} = g_k - g_{k-1}$.

Las soluciones del problema son las descritas en la ecuación 4.

$$\alpha_k^{BB1} = \frac{s_{k-1}^T s_{k-1}}{s_{k-1}^T y_{k-1}} \quad \alpha_k^{BB2} = \frac{s_{k-1}^T y_{k-1}}{y_{k-1}^T y_{k-1}} \quad (4)$$

Considerando la desigualdad de Cauchy-Schwarz observa que cuando $s_{k-1}^T y_{k-1}$ es mayor a cero, entonces se cumple que $\alpha_k^{BB1} \geq \alpha_k^{BB2}$. Esto por ello, que se suele llamar paso largo a Barzilai-Borwein al paso α_k^{BB1} y paso corto a α_k^{BB2} . Para una función cuadrática, el paso α_k^{BB1} es el tamaño de paso de máximo descenso con retardo de un paso y α_k^{BB2} será el paso del método de mínimo gradiente. En 1993, Raydan demuestra la convergencia del método para el caso cuadrático,² y en 1997³ introdujo una estrategia global basada en una búsqueda lineal no monótona, la cual establece la convergencia global para el método de Barzilai-Borwein (BB) para los casos no cuadráticos. El método BB no asegura la convergencia cuando la función objetivo es fuertemente convexa. Para ello existen varios métodos para estabilizar la convergencia del problema. Uno de los métodos para estabilizar a estos métodos es la elección del tamaño de paso en cada iteración de la siguiente forma:

$$\alpha_k = \min \alpha_k^{BB}, \Delta \quad (5)$$

donde Δ es un valor fijo. En el artículo de Oleg Burdakov⁴ realiza varios experimentos con esta estrategia y obtuvo que para la función de rosenbrock, el valor de Δ con mejores resultados fue 0.1. Por esta misma razón se siguen explorando estrategias que complementen al método BB para la convergencia global en funciones fuertemente convexas.

Desarrollo

El método (BB) tiene convergencia R-superlineal para minimizar funciones cuadráticas bidimensionales fuertemente convexas y R-lineal para el caso general n-dimensional.⁵ El método de BB cuenta con la propiedad de reducir los valores de la función objetivo de manera no monótona. Esta propiedad es una característica intrínseca la cual es la razón de su eficiencia. Sin embargo, es importante para los métodos de gradiente mantener la monotonicidad. Debido a la eficiencia del algoritmo BB y la complejidad de obtener un tamaño de paso α_k^{SD} para el caso general, el trabajo se motiva en la búsqueda de una forma de aceleración para el método de Barzilai-Borwein incorporando pasos monótonos. Considérese el problema de acelerar los métodos de descenso de gradiente que generan secuencias de iterados de la forma

$$x_{k+1} = x_k - \alpha_k g_k$$

para resolver el problema

$$\min_{x \in \mathbb{R}^n} f(x)$$

donde $f : \mathbb{R}^n \rightarrow \mathbb{R}$ es continuamente diferenciable, $g_k = \nabla f(x_k)$ y $\alpha_k > 0$ es el tamaño de paso sobre la dirección del gradiente. En particular se usará una función cuadrática (ecuación 1) utilizando el tamaño de paso mostrado en la ecuación 6.

$$\alpha_k(\Psi(A)) = \frac{g_{k-1}^T g_{k-1}}{g_{k-1}^T \Psi(A) A g_{k-1}} \quad (6)$$

Donde $\Psi(\cdot)$ es una función real analítica en $[\lambda_1, \lambda_n]$ que se puede expresar como una serie de potencias (ecuación 7).

$$\Psi(z) = \sum_{k=-\infty}^{\infty} c_k z^k, \quad c_k \in \mathbb{R} \quad \text{tal que} \quad 0 < \sum_{k=-\infty}^{\infty} c_k z^k < +\infty \quad \forall z \in [\lambda_1, \lambda_n] \quad (7)$$

Los valores de λ_1 y λ_n es el eigenvalor mínimo y máximo respectivamente. Se puede observar que los dos tamaños de paso de Barzilai-Borwein α_k^{BB1} y α_k^{BB2} se pueden calcular a partir de un tamaño de paso más general tomando a $\Psi(A) = I$ y $\Psi(A) = A$ en la ecuación 6.

Tamaño de paso

Se tiene que, dos gradientes consecutivos generados por el método de máximo descenso no se mantienen para métodos que utilizan el tamaño de paso general antes definido. Se observa que el método BB es invariante ante traslaciones y rotaciones cuando se minimizan funciones cuadráticas,⁶ por lo que se puede asumir para simplicidad que la matriz A tiene la forma mostrada en la ecuación 8.

$$A = \text{diag}\{\lambda_1, \dots, \lambda_n\}, \quad \text{donde} \quad 0 < \lambda_1 \leq \dots \leq \lambda_n \quad (8)$$

Se ha mostrado que una familia de métodos de gradiente entre los que se incluye máximo descenso y mínimo gradiente reducirán asintóticamente sus búsquedas en un subespacio bidimensional. Se explotaron algunas propiedades ortogonales en este subespacio bidimensional para acelerar los métodos.⁷ Es por ello que si se quiere acelerar la convergencia de los métodos con tamaño de paso α_k (ecuación 6) en un subespacio de menor dimensionalidad se deben mantener propiedades ortogonales. Suponiendo que para un $k > 0$, existe un q_k que satisface

$$(I - \alpha_{k-1} A) q_k = g_{k-1} \quad (9)$$

donde q_k es invariante ante rotaciones y traslaciones, por lo que se puede asumir que la matriz A corresponde a una matriz diagonal de una función cuadrática. Con esto, se tiene el lema 1 que presenta una propiedad para la derivación del tamaño de paso.

Lema 1. *Suponiendo que la secuencia $\{g_k\}$ se obtiene aplicando un método de gradiente con pasos como el paso general (α_k) para minimizar una función cuadrática y q_k satisface que $(I - \alpha_{k-1}A)q_k = g_{k-1}$. Entonces*

$$q_k^T \Psi(A)g_{k+1} = 0.$$

Para probar el lema 1 se tiene que:

$$\begin{aligned} q_k^T \Psi(A)g_{k+1} &= q_k^T \Psi(A)(I - \alpha_k A)(I - \alpha_{k-1} A)g_{k-1} \\ &= q_k^T (I - \alpha_k A) \Psi(A)(I - \alpha_{k-1} A)g_{k-1} \\ &= g_{k-1}^T \Psi(A)(I - \alpha_k A)g_{k-1} \\ &= g_{k-1}^T \Psi(A)g_{k-1} - \alpha_k g_{k-1}^T \Psi(A)A g_{k-1} \\ &= g_{k-1}^T \Psi(A)g_{k-1} \left[\frac{g_{k-1}^T \Psi(A)g_{k-1}}{g_{k-1}^T \Psi(A)A g_{k-1}} - \alpha_k \right] \\ &= 0 \end{aligned}$$

con lo que se demuestra que el vector q_k^T y g_{k+1} son ortogonales bajo $\Psi(A)$.

Ahora suponemos que los vectores $\Psi^r(A)q_{k-1}$ y $\Psi^{1-r}(A)g_k$ son vectores no nulo, con $r \in (R)$. Considerando el problema de minimizar una función f en un subespacio bidimensional generado por

$$u = \frac{\Psi^r(A)q_{k-1}}{|\Psi^r(A)q_{k-1}|} \quad v = \frac{\Psi^r(A)g_k}{|\Psi^r(A)g_k|} \quad (10)$$

donde u y v forman una base ortogonal para \mathbb{R}^2 . Definiendo una función ϕ como

$$\phi(t, l) := f \left(x_k + t \frac{\Psi^r(A)q_{k-1}}{\|\Psi^r(A)q_{k-1}\|} + l \frac{\Psi^r(A)g_k}{\|\Psi^r(A)g_k\|} \right) \quad (11)$$

al expandir en una serie de Taylor obtenemos que

$$\phi(t, l) = f(x_k) + \nabla^T f(x_k) [tu + lv] + \frac{1}{2} [tu + lv]^T \nabla^2 f(x_k) [tu + lv] \quad (12)$$

Tomando a una matriz B_k como en la ecuación 13.

$$B_k = (u, v)^T \quad (13)$$

Con las ecuaciones 10 y 13 podemos escribir la ecuación 12 como en la ecuación 14.

$$\phi(t, l) = f(x_k) + g_k^T B_k^T \begin{pmatrix} t \\ l \end{pmatrix} + \frac{1}{2} \begin{pmatrix} t \\ l \end{pmatrix}^T B_k A B_k^T \begin{pmatrix} t \\ l \end{pmatrix} \quad (14)$$

Definiendo el vector ϑ como

$$\vartheta_k = B_k g_k = \begin{pmatrix} \frac{g_k^T \Psi^r(A)q_{k-1}}{|\Psi^r(A)q_{k-1}|} \\ \frac{g_k^T \Psi^{1-r}(A)g_k}{|\Psi^{1-r}(A)g_k|} \end{pmatrix} \quad (15)$$

y al hessiano de f en el paso k como

$$H_k = B_k A B_k^T = \begin{pmatrix} \frac{q_{k-1}^T \Psi^{2r}(A) A q_{k-1}}{\|\Psi^r(A) q_{k-1}\|^2} & \frac{q_{k-1}^T \Psi^r(A) A g_k}{\|\Psi^r(A) q_{k-1}\| \|\Psi^{1-r}(A) g_k\|} \\ \frac{q_{k-1}^T \Psi^r(A) A g_k}{\|\Psi^r(A) q_{k-1}\| \|\Psi^{1-r}(A) g_k\|} & \frac{g_k^T \Psi^{2(1-r)}(A) A g_k}{\|\Psi^r(A) g_k\|^2} \end{pmatrix} \quad (16)$$

Entonces la función $f\phi(t, l)$ se puede escribir como en la ecuación 17.

$$\phi(t, l) = f(x_k) + \vartheta_k^T \begin{pmatrix} t \\ l \end{pmatrix} + \frac{1}{2} \begin{pmatrix} t \\ l \end{pmatrix}^T H_k \begin{pmatrix} t \\ l \end{pmatrix} \quad (17)$$

Denotando las componentes de H_k por $H_k^{(ij)}$, $i, j = 1, 2$ y notando que $B_k B_k^T = I$ se tiene el teorema 1.

Teorema 1 (Terminación finita). *Suponga que un método de gradiente se aplica para minimizar una función cuadrática bidimensional con α_k dado por el paso general mencionado anteriormente para todas las $k \neq k_0$ y usa el tamaño de paso*

$$\tilde{\alpha}_{k_0} = \frac{2}{\left(H_{k_0}^{(11)} + H_{k_0}^{(22)}\right) + \sqrt{\left(H_{k_0}^{(11)} - H_{k_0}^{(22)}\right)^2 + 4\left(H_{k_0}^{(12)}\right)^2}}$$

en la iteración k_0 -ésima donde $k_0 \geq 2$. Entonces, el método encontrará el minimizador en máximo $k_0 + 3$ iteraciones.

Para demostrar el teorema 1 suponemos que x_k no es un minimizador para toda $k = 1, \dots, k_0 + 1$. Para simplificar durante la prueba, se utilizará como notación k para referirse a k_0 .

Es necesario observar que $\tilde{\alpha}_k$ satisface la ecuación 18.

$$\tilde{\alpha}_k^2 \Delta - \tilde{\alpha}_k \left(H_k^{(11)} + H_k^{(22)}\right) + 1 = 0 \quad (18)$$

en donde $\Delta = \det(H_k) = \det(A) > 0$. Sea

$$\Theta = \left(H_k^{(12)} \vartheta_k^{(1)} + H_k^{(22)} \vartheta_k^{(2)}\right) \vartheta_k^{(1)} - \left(H_k^{(11)} \vartheta_k^{(1)} + H_k^{(12)} \vartheta_k^{(2)}\right) \vartheta_k^{(2)}$$

en donde $\vartheta_k^{(i)}$ son las componentes de ϑ_k . Multiplicando Θ a la ecuación 18 se tiene lo siguiente

$$\tilde{\alpha}_k^2 \Delta \Theta - \tilde{\alpha}_k \left(H_k^{(11)} + H_k^{(12)}\right) \Theta + \Theta = 0,$$

que exactamente es

$$\begin{aligned} ab &= cd \\ a &= H_k^{(22)} v_k^1 - H_k^{(12)} v_k^{(2)} - \tilde{\alpha}_k \Delta \vartheta_k^{(1)} \\ b &= \vartheta_k^{(2)} - \tilde{\alpha}_k (H_k^{(12)} \vartheta_k^{(1)} + H_k^{(22)} \vartheta_k^{(2)}) \\ c &= H_k^{(11)} v_k^2 - H_k^{(12)} v_k^{(1)} - \tilde{\alpha}_k \Delta \vartheta_k^{(1)} \\ d &= \vartheta_k^{(1)} - \tilde{\alpha}_k (H_k^{(11)} \vartheta_k^{(1)} + H_k^{(12)} \vartheta_k^{(2)}). \end{aligned}$$

Consideremos dos vectores $\mathbf{A} = (a_1, a_2)^T$ y $\mathbf{B} = (b_1, b_2)^T$, tales que se cumple $a_1 b_2 = a_2 b_1$. Sustituyendo b_2 en \mathbf{B} , tenemos lo siguiente

$$\mathbf{B} = \begin{pmatrix} b_1 \\ \frac{b_1 a_2}{a_1} \end{pmatrix} = \frac{b_1}{a_1} \mathbf{A}$$

por los que podemos decir que \mathbf{B} es paralelo a \mathbf{A} . Haciendo uso del resultado anterior podemos decir que el vector

$$\begin{pmatrix} H_k^{(22)} \vartheta_k^{(1)} - H_k^{(12)} \vartheta_k^{(2)} - \tilde{\alpha}_k \Delta \vartheta_k^{(1)} \\ H_k^{(11)} \vartheta_k^{(2)} - H_k^{(12)} \vartheta_k^{(1)} - \tilde{\alpha}_k \Delta \vartheta_k^{(2)} \end{pmatrix}$$

es paralelo a

$$\begin{pmatrix} v_k^{(1)} - \tilde{\alpha}_k (H_k^{(11)} \vartheta_k^{(1)} + H_k^{(12)} \vartheta_k^{(2)}) \\ v_k^{(2)} - \tilde{\alpha}_k (H_k^{(12)} \vartheta_k^{(1)} + H_k^{(22)} \vartheta_k^{(2)}) \end{pmatrix}$$

Por ende el vector $\vartheta_k + H_k(-\tilde{\alpha}_k \vartheta_k)$ es paralelo a $H_k^{-1} \vartheta_j - \tilde{\alpha}_k \vartheta_k$. Es decir

$$\vartheta_k + H_k(-\tilde{\alpha}_k \vartheta_k) = \gamma (H_k^{-1} \vartheta_j - \tilde{\alpha}_k \vartheta_k) \quad (19)$$

donde $\gamma \neq 0 \in \mathbb{R}$. Si multiplicamos por la derecha a la ecuación 19 por B_k^T se tiene que

$$B_k^T [\vartheta_k + H_k(-\tilde{\alpha}_k \vartheta_k)] = \gamma B_k^T (H_k^{-1} \vartheta_j - \tilde{\alpha}_k \vartheta_k).$$

Se sabe que $B_k^T B_k = I$, $\vartheta_k = B_k g_k$ y que $H_k = B_k A B_k^T$. Además $g_{k+1} = g_k + \tilde{\alpha}_k A g_k$. Haciendo uso de este conocimiento, se obtiene que

$$\begin{aligned} g_k + A(-\tilde{\alpha}_k B_k^T \vartheta_k) &= \gamma (B_k^T H_k^{-1} \vartheta_k + \tilde{\alpha}_k g_k) \\ g_k + A(-\tilde{\alpha}_k g_k) &= \gamma (B_k^T H_k^{-1} \vartheta_k + \tilde{\alpha}_k g_k) \\ g_{k+1} &= \gamma (B_k^T H_k^{-1} \vartheta_k + \tilde{\alpha}_k g_k) \\ g_{k+1} &= \gamma (B_k^T B_k A^{-1} B_k^T \vartheta_k + \tilde{\alpha}_k g_k) \\ g_{k+1} &= \gamma (A^{-1} g_k + \tilde{\alpha}_k g_k) \end{aligned}$$

Factorizando A^{-1} se tiene

$$g_{k+1} = \gamma A^{-1} (g_k + \tilde{\alpha}_k A g_k)$$

Considerando que $g_{k+1} = g_k + \tilde{\alpha}_k A g_k$ llegamos a

$$g_{k+1} = \gamma A^{-1} g_{k+1}$$

Es decir g_{k+1} es un eigenvector de la matriz A . Por hipótesis, sabemos que x_{k+2} no es un minimizador, así que $g_{k+2} \neq 0$ y el algoritmo no se detendrá en la $k+2$ -ésima iteración. Entonces, calculando α_{k+2} como

$$\alpha_{k+2} = \frac{g_{k+1}^T \Psi(A) g_{k+1}}{g_{k+1}^T \Psi(A) A g_{k+1}} = 1/\lambda$$

tenemos que

$$\begin{aligned} g_{k+3} &= (I - \alpha_{k+2}A)g_{k+2} \\ &= (1 - \alpha_{k+2}\lambda)g_{k+2} = 0 \end{aligned}$$

lo que implica que x_{k+3} debe ser el minimizador. Si tomamos $k_0 = 2$ en el teorema, el tamaño de paso encontrará al minimizador exacto en máximo 5 iteraciones cuando se tiene una función cuadrática bidimensional fuertemente convexa.

Corolario 1. *Suponga que un método de gradiente se aplica a una función cuadrática bidimensional con α_{k_0+m} para $k_0 \geq 2$, algún entero positivo m y con $\alpha_{k_0+m} = \tilde{\alpha}_{k_0}$ para toda $k \neq k_0 + m$. Entonces, el método se detendrá en máximo $k + m + 3$ iteraciones.*

Si hacemos $\Psi(A) = I$, $\Psi(A) = A$ y $r = 1/2$ en la ecuación 6 y tomando $k_0 = k$ se puede derivar los siguientes tamaños de paso:

$$\tilde{\alpha}_k^{BB1} = \frac{2}{\frac{q_{k-1}^T A q_{k-1}}{\|q_{k-1}\|^2} + \frac{1}{\alpha_k^{SD}} + \sqrt{\left(\frac{q_{k-1}^T A q_{k-1}}{\|q_{k-1}\|^2} - \frac{1}{\alpha_k^{SD}}\right)^2 + \frac{4(q_k^T A g_k)^2}{\|q_{k-1}\|^2 \|g_k\|^2}}} \quad (20)$$

y

$$\tilde{\alpha}_k^{BB2} = \frac{2}{\frac{1}{\hat{\alpha}_{k-1}} + \frac{1}{\alpha_k^{MG}} + \sqrt{\left(\frac{1}{\hat{\alpha}_{k-1}} - \frac{1}{\alpha_k^{MG}}\right)^2 + \Gamma_k}} \quad (21)$$

respectivamente, en donde

$$\hat{\alpha}_k = \frac{q_k^T A q_k}{q_k^T A^2 q_k} \quad \text{y} \quad \Gamma_k = \frac{4(q_{k-1}^T A^2 g_k)^2}{q_{k-1}^T A q_{k-1} g_k^T A g_k} \quad (22)$$

Ahora, de 20 y 21 se tiene

$$\tilde{\alpha}_k^{BB1} \leq \min \left\{ \alpha_k^{SD}, \frac{\|q_{k-1}\|^2}{q_{k-1}^T A q_{k-1}} \right\} \quad \text{y} \quad \tilde{\alpha}_k^{BB2} \leq \min \{ \alpha_k^{MG}, \hat{\alpha}_{k-1} \} \quad (23)$$

En consecuencia, tanto $\tilde{\alpha}_k^{BB1}$ como $\tilde{\alpha}_k^{BB2}$ son pasos cortos y monótonos que reducen tanto el valor de la función como el gradiente respectivamente, y del teorema 1 se puede concluir que si insertamos estos pasos en los métodos BB1 y BB2, se contará con terminación finita para minimizar funciones cuadráticas bidimensionales fuertemente convexas. Estos tamaños de paso conforman el método de gradiente no monótono adaptable (ANGM).

Los tamaños de paso generados por el método BB1 pueden estar muy alejados de los recíprocos de los eigenvalores más grandes de la matriz Hessiana A para la función cuadrática, por lo que los tamaños de paso α_k^{BB1} pueden ser muy grandes para lograr reducir de forma efectiva las componentes del gradiente g_k que corresponden a los primeros eigenvalores más grandes. Por lo que haciendo uso de $g_{k+1}^{(j)} = (1 - \alpha_k \lambda_j)g_k$ se puede observar que estas componentes del gradiente pueden ser reducidas cuando se toman tamaños de paso pequeños. También vale la pena mencionar que cuando los métodos de gradiente utilizan tamaños de paso largos y cortos de forma adaptativa, normalmente tienen un mejor desempeño que si se implementaran estos pasos alternando entre ellos. Entonces, para lograr desarrollar métodos de gradiente que combinen los dos pasos monótonos de BB con el tamaño de paso corto dado por 18 el artículo extiende la propiedad ortogonal desarrollada en el lema 1 y la propiedad de terminación finita del teorema 1.

Lema 2 (Propiedad de ortogonalidad generalizada). *Suponga que un método de gradiente con tamaños de paso de la forma 6 se aplica a minimizar una función cuadrática 1. En particular, al $k-1$ -ésimo paso y al k -ésimo se utilizan dos tamaños de paso $\alpha_{k-1}(\Psi(A))$ y $\alpha_k(\Psi_1(A))$, respectivamente, donde Ψ y Ψ_1 pueden ser dos funciones analíticas diferentes usadas en 6. Si $q_k \in \mathbb{R}^n$ satisface*

$$(I - \alpha_{k-1}(\Psi(A)A))q_k = g_{k-1}, \quad (24)$$

entonces se tiene

$$q_k^T \Psi_1(A)g_{k+1} = 0. \quad (25)$$

Teorema 2 (Terminación finita generalizada). *Suponga que se aplica un método de gradiente para minimizar una función cuadrática bidimensional 1 con α_k dado por 6 para todo $k \neq k_0$ y $k \neq k_0 - 1$, y usa los tamaños de paso $\alpha_{k-1}(\Psi_1(A))$ y $\alpha_k(\Psi_1(A))$ las iteraciones $k-1$ -ésima y la k -ésima, respectivamente, donde $k_0 \geq 2$. Entonces el método encontrará el minimizador en máximo $k_0 + 3$.*

El método de gradiente no-monótono adaptativo (ANGM) toma el paso largo de BB α_k^{BB1} cuando $\alpha_k^{BB2}/\alpha_k^{BB1} \geq \tau_1$ para algún $\tau_1 \in (0, 1)$. De otra manera, toma un tamaño de paso α_k^{BB2} o $\tilde{\alpha}_k^{BB2}$ dependiendo del cociente $\|g_{k-1}\|/\|g_k\|$. El paso α_k^{BB2} minimiza al gradiente tal que

$$\alpha_k^{BB2} = \alpha_{k-1}^{MG} = \arg \min_{\alpha \in \mathbb{R}} \|g_{k-1} - \alpha A g_{k-1}\|.$$

Así, cuando $\|g_{k-1}\|/\|g_k\| > \tau_2$ para algún $\tau_2 > 1$ (la norma del gradiente decrece), puede ser razonable utilizar el tamaño de paso anterior α_{k-1} como aproximación de α_k^{BB2} .

El ANGM utiliza el tamaño de paso monótono $\tilde{\alpha}_k^{BB2}$ cuando $\|g_{k-1}\| > \tau_2 \|g_k\|$, de otra manera se deberían tomar ciertos pasos BB2. El ANGM aplica las siguientes estrategias adaptativas para elegir el tamaño de paso:

$$\alpha_k = \begin{cases} \min\{\alpha_k^{BB2}, \alpha_{k-1}^{BB2}\}, & \text{si } \alpha_k^{BB2} < \tau_1 \alpha_k^{BB1} \text{ y } \|g_{k-1}\| < \tau_2 \|g_k\| \\ \tilde{\alpha}_k^{BB2}, & \text{si } \alpha_k^{BB2} < \tau_1 \alpha_k^{BB1} \text{ y } \|g_{k-1}\| \geq \|g_k\| \\ \alpha_k^{BB1}, & \text{de otro modo} \end{cases} \quad (26)$$

Para el cálculo de $\tilde{\alpha}_k^{BB2}$ es necesario calcular α_k^{MG} , que no es fácil de obtener cuando la función objetivo no es cuadrática. En su lugar, el cálculo de $\tilde{\alpha}_{k-1}^{BB2}$ requiere solo a α_k^{BB2} , de la que es más sencillo disponer, incluso para una función objetivo general. Además, se ha encontrado que los métodos de gradiente que utilizan pasos retardados pueden llevar a mejores desempeños, así reemplazando $\tilde{\alpha}_k^{BB2}$ en 26 por $\tilde{\alpha}_{k-1}^{BB2}$ se obtiene una variante llamada ANGR1:

$$\alpha_k = \begin{cases} \min\{\alpha_k^{BB2}, \alpha_{k-1}^{BB2}\}, & \text{si } \alpha_k^{BB2} < \tau_1 \alpha_k^{BB1} \text{ y } \|g_{k-1}\| < \tau_2 \|g_k\| \\ \tilde{\alpha}_{k-1}^{BB2}, & \text{si } \alpha_k^{BB2} < \tau_1 \alpha_k^{BB1} \text{ y } \|g_{k-1}\| \geq \|g_k\| \\ \alpha_k^{BB1}, & \text{de otro modo} \end{cases} \quad (27)$$

Ahora, para simplificar el ANGR1 se puede reemplazar $\tilde{\alpha}_{k-1}^{BB2}$ por su cota superior, es decir, se puede hacer uso de:

$$\alpha_{k-1}^{BB2} \leq \min\{\alpha_k^{BB2}, \hat{\alpha}_{k-2}\}$$

Como resultado de esto, se obtiene otra variante de ANGM a la que se llama ANGR2:

$$\alpha_k \begin{cases} \min\{\alpha_k^{BB2}, \alpha_{k-1}^{BB2}\}, & \text{si } \alpha_k^{BB2} < \tau_1 \alpha_k^{BB1} \text{ y } \|g_{k-1}\| < \tau_2 \|g_k\| \\ \min\{\alpha_k^{BB2}, \hat{\alpha}_{k-2}\}, & \text{si } \alpha_k^{BB2} < \tau_1 \alpha_k^{BB1} \text{ y } \|g_{k-1}\| \geq \|g_k\| \\ \alpha_k^{BB1}, & \text{de otro modo} \end{cases} \quad (28)$$

Es necesario notar que para la implementación de los métodos ANGM, ANGR1 y ANGR2 no es necesario calcular productos de matriz-vector.

$$\hat{\alpha}_k = \frac{q_k^T A q_k}{q_k^T A^2 q_k} = \frac{\alpha_{k-1} q_k^T (q_k - g_{k-1})}{(q_k - g_{k-1})^T (q_k - g_{k-1})}$$

por lo ende, no se necesita calcular ningún producto de matriz-vector para calcular $\hat{\alpha}_{k-1}$ en $\tilde{\alpha}_k^{BB2}$, $\hat{\alpha}_{k-1}$ en $\tilde{\alpha}_{k-1}^{BB2}$ y los tamaños de paso utilizados en ANGR2. Debido a que el cálculo de $A g_k$ se necesita para el término g_{k+1} y Γ_k en $\tilde{\alpha}_k^{BB2}$. Para $\tilde{\alpha}_{k-1}^{BB2}$, se necesita el calculo del término

$$g_{k-1}^T A^2 q_{k-2} = \frac{1}{\alpha_{k-3}} (q_{k-2} - g_{k-3})^T A g_{k-1} \quad A g_{k-1} = \frac{1}{\alpha_{k-1}} (g_{k-1} - g_k)$$

por ende

$$\begin{aligned} \Gamma_{k-1} &= \frac{4((q_{k-2} - g_{k-3})^T A g_{k-1})^2}{\alpha_{k-3}((q_{k-2} - g_{k-3})^T q_{k-2}) g_{k-1}^T A g_{k-1}} \\ &= \frac{4((q_{k-2} - g_{k-3})^T (g_{k-1} - g_k))^2}{\alpha_{k-3} \alpha_{k-1} ((q_{k-2} - g_{k-3})^T q_{k-2}) g_{k-1}^T (g_{k-1} - g_k)} \end{aligned}$$

por lo que no se requiere calcular productir de matriz-vector adicionales para Γ_{k-1} en $\tilde{\alpha}_{k-1}^{BB2}$. Un problema a tener en cuenta es que el cálculo de q_k de forma exacta como se define en 9 puede llegar a ser tan difícil como minimizar la función cuadrática. Es por esto que a la hora de implementar, se puede utilizar una aproximación de q_k notando que satisface la ecuación

$$q_k^T g_k = \|g_{k-1}\|^2 \quad (29)$$

De esta manera, se puede encontrar esa aproximación requiriendo que esta condición se mantenga. Una forma eficiente de encontrar la q_k que satisface la ecuación de secante 29 es tratando a la matriz Hessiana A como una matriz diagonal, y así derivar q_k de $g_{k+1}^{(j)} = (1 - \alpha_k \lambda_j) g_k^{(j)}$ cuando $g_k^{(j)} \neq 0$,

$$q_k^{(i)} = \frac{g_{k-1}^{(i)}}{1 - \alpha_{k-1} \lambda_i} = \frac{(g_{k-1}^{(i)})^2}{g_k^{(i)}}, \quad i = 1, \dots, n \quad (30)$$

y en el caso de que $g_k^{(i)} = 0$ se hace $q_k^{(i)} = 0$.

Resultados

Los métodos de búsqueda antes mencionados fueron probados con las siguientes funciones:

Rosembrock

La función de Rosembrock se define en la ecuación 31.

$$f(x) = \sum_{i=1}^{n-1} 100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2 \quad (31)$$

Su gradiente es calculado de la siguiente manera:

$$\nabla f(x) = \begin{cases} -400x_i(x_{i+1} - x_i^2) & \text{para } i = 1 \\ 200(x_i - x_{i-1}^2 - 400x_i(x_{i+1} - x_i^2) - 2(1 - x_i)) & \text{para } 1 < i < n \\ 200(x_i - x_{i-1}^2) & \text{para } i = n \end{cases} \quad (32)$$

y el hessiano se puede obtener realizando la siguiente operación:

$$\nabla^2 f(x) = \begin{cases} \frac{\partial^2 f}{\partial x_i^2} = -200 & \text{para } x = 1 \\ \frac{\partial^2 f}{\partial x_i^2} = 1200x_i^2 - 400x_{i+1} + 202 & \text{para } 1 < i < n \\ \frac{\partial^2 f}{\partial x_i \partial x_{i+1}} = -400x_i & \text{para } 0 \leq i < n \\ \frac{\partial^2 f}{\partial x_i^2} = 200 & \text{para } i = n \end{cases} \quad (33)$$

Wood

La función de Wood se define en la ecuación 34.

$$f(x) = 100(x_1^2 - x_2)^2 + (x_1 - 1)^2 + (x_3 - 1)^2 + 90(x_3 - x_4)^2 + 10.1((x_2 - 1)^2 + (x_4 - 1)^2) + 19.8(x_2 - 1)(x_4 - 1) \quad (34)$$

donde $x \in \mathbb{R}^4$. Con la función de Wood definida, podemos obtener le gradiente de la función de Wood. El resultado del gradiente de la función de Wood se encuentra en la siguiente ecuación.

$$\nabla f(x) = \begin{cases} \frac{\partial f}{\partial x_1} = 400(x_1^2 - x_2)x_1 + 2(x_1 - 1) \\ \frac{\partial f}{\partial x_2} = -200(x_1^2 - x_2) + 20.2(x_2 - 1) + 19.8(x_4 - 1) \\ \frac{\partial f}{\partial x_3} = 2(x_3 - 1) + 360(x_3^2 - x_4)x_3 \\ \frac{\partial f}{\partial x_4} = -180(x_3^2 - x_4) + 20.2(x_4 - 1) + 19.8(x_2 - 1) \end{cases} \quad (35)$$

De igual forma, se puede obtener el hessiano de la función de Wood. El resultado del Hessiano se calculado en la siguiente ecuación.

$$\nabla^2 f(x) = \begin{cases} \frac{\partial^2 f}{\partial x_1^2} = 400(x_1^2 - x_2) + 800x_1^2 + 2 \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} = \frac{\partial^2 f}{\partial x_2 \partial x_1} = -400x_1 \\ \frac{\partial^2 f}{\partial x_2^2} = 220.2 \\ \frac{\partial^2 f}{\partial x_4 \partial x_2} = \frac{\partial^2 f}{\partial x_2 \partial x_4} = 19.8 \\ \frac{\partial^2 f}{\partial x_3^2} = 720x_3^2 + 360(x_3^2 - x_4) + 2 \\ \frac{\partial^2 f}{\partial x_4 \partial x_3} = \frac{\partial^2 f}{\partial x_3 \partial x_4} = -360x_3 \\ \frac{\partial^2 f}{\partial x_4^2} = 200.2 \end{cases} \quad (36)$$

Lambda

En el artículo de Yakui Huang⁸ se utiliza una función definida de la siguiente manera:

$$f(x) = \frac{1}{2}x^T A x \quad A = \begin{cases} 0 & \text{para } i \neq 0 \\ 10^{\frac{ncond}{n-1}(n-j)} & \text{para } i = j \end{cases}$$

donde $ncond = \log \kappa$ con $\kappa = 10^3$ y $n = 10$. Calculando su gradiente se obtiene que tiene la siguiente forma

$$\nabla f(x) = Ax$$

por consecuente, el hessiano de la función es unicamente la matriz A .

Cuadrática de la forma $\text{diag}\{1, \lambda\}$

En el mismo artículo⁸ se propone una función cuadrática donde la matriz esta definida de la siguiente manera:

$$A = \text{diag}\{1, \lambda\}$$

En nuestro caso tomamos $\lambda = 10$. El gradiente y hessiano de esta función tienen la misma forma que la función lambda.

Porcentaje de contribución

Se definió una función γ para medir el porcentaje de las dos componentes más grandes del gradiente en cada iteración de la optimización. La función γ tiene la siguiente forma

$$\gamma(g_k) = \frac{|g_k^{(1)}| + |g_k^{(2)}|}{\sum_i |g_k^{(i)}|}$$

Realizando el calculo de la función γ con la función lambda en el punto inicial $x = (10, 10, \dots, 10)^T$ se obtuvieron los resultados mostrados en la figura 1 para los diferentes métodos.

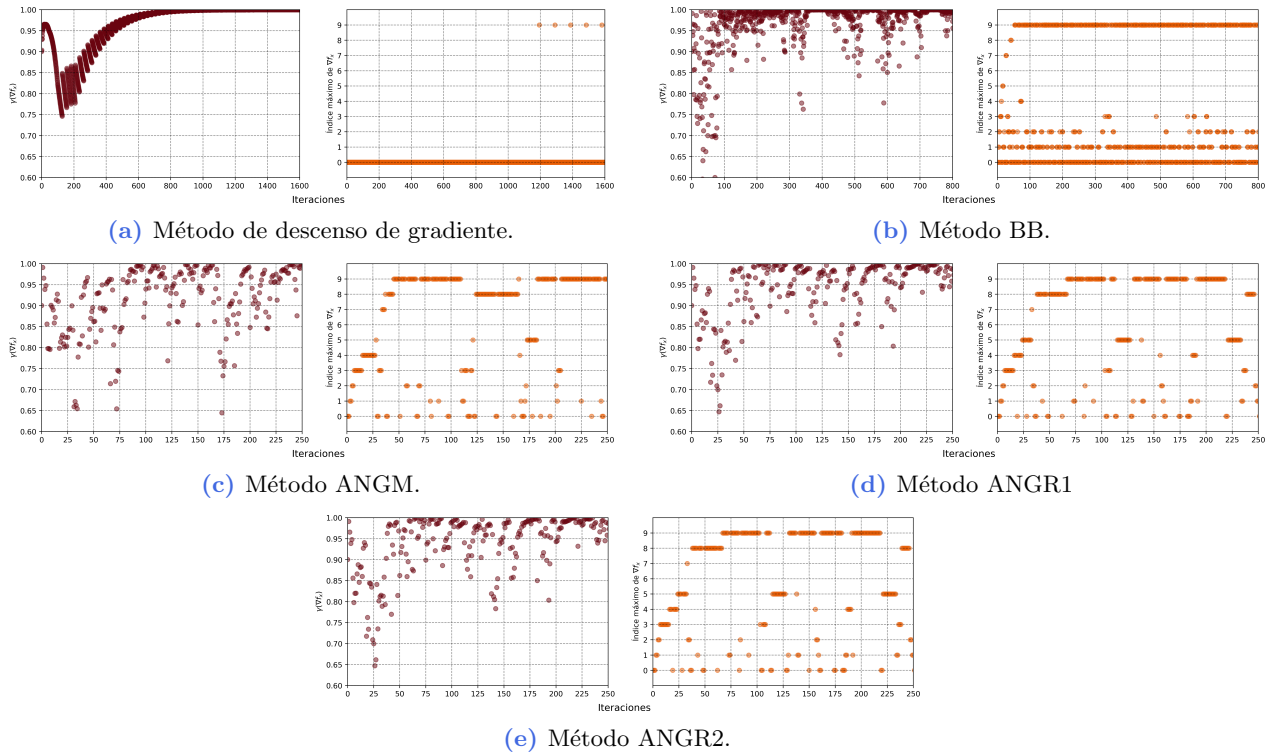


Figura 1: Función γ para diferentes métodos (izquierda) y la posición en el vector de la componente más grande (derecha).

En la tabla 1 se muestra el número de iteraciones en las que se obtuvo un valor mayor a 0.8 para la función γ en cada método de optimización.

Método	$\gamma(g_k) > 0.8$	Total
SD	8057	8104
BB	815	851
ANGM	293	316
ANGR1	240	253
ANGR2	200	245

Tabla 1: Número de iteraciones donde el valor de la función γ tuvo un valor mayor a 0.8 para cada método de optimización.

Estabilidad de resultados

Se realizaron 100 ejecuciones para cada par de función y método de optimización partiendo de un punto aleatorio dado por la distribución normal con media 0 y desviación 10 (ecuación 37).

$$x_0 = \mathcal{N}(\mu = 0; \sigma = 10) \quad (37)$$

En la figura 2 se muestra el valor de la función y la norma del gradiente para la mejor iteración para cada método con la función cuadrática con matriz diagonal.

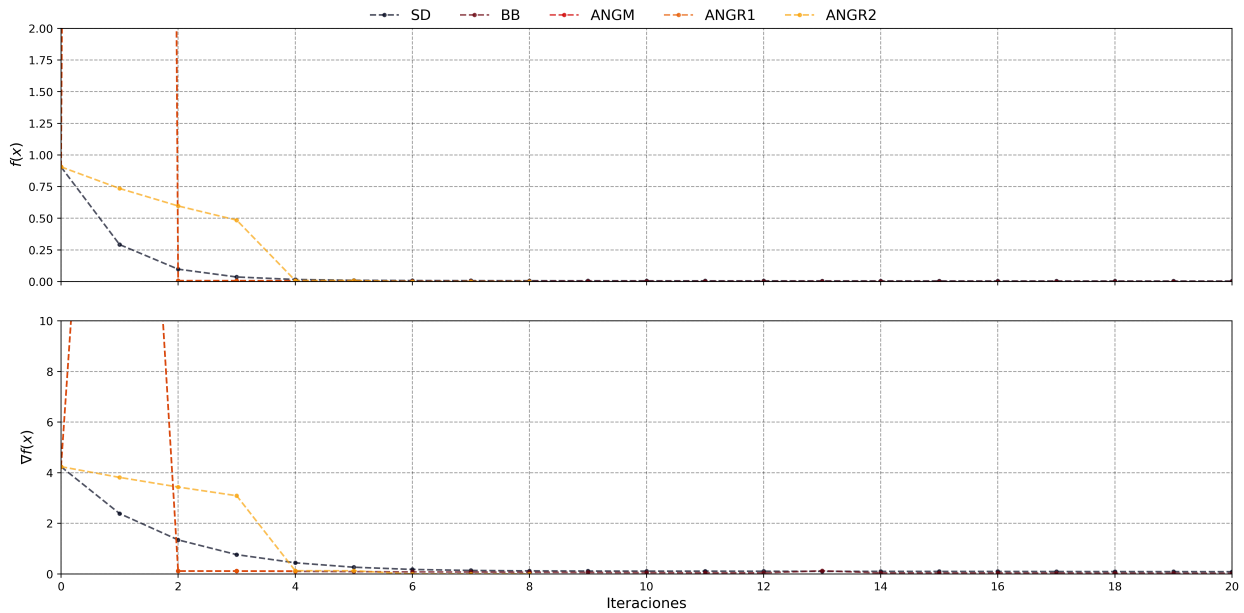


Figura 2: Valor de la función y norma del gradiente en cada iteración de la mejor ejecución de cada método de optimización de la función cuadrática con matriz diagonal.

Se realizó el mismo procedimiento con la función de Rosenbrock. En la figura 3 se muestra el resultado del valor de la función y norma del gradiente para la mejor iteración de cada método.

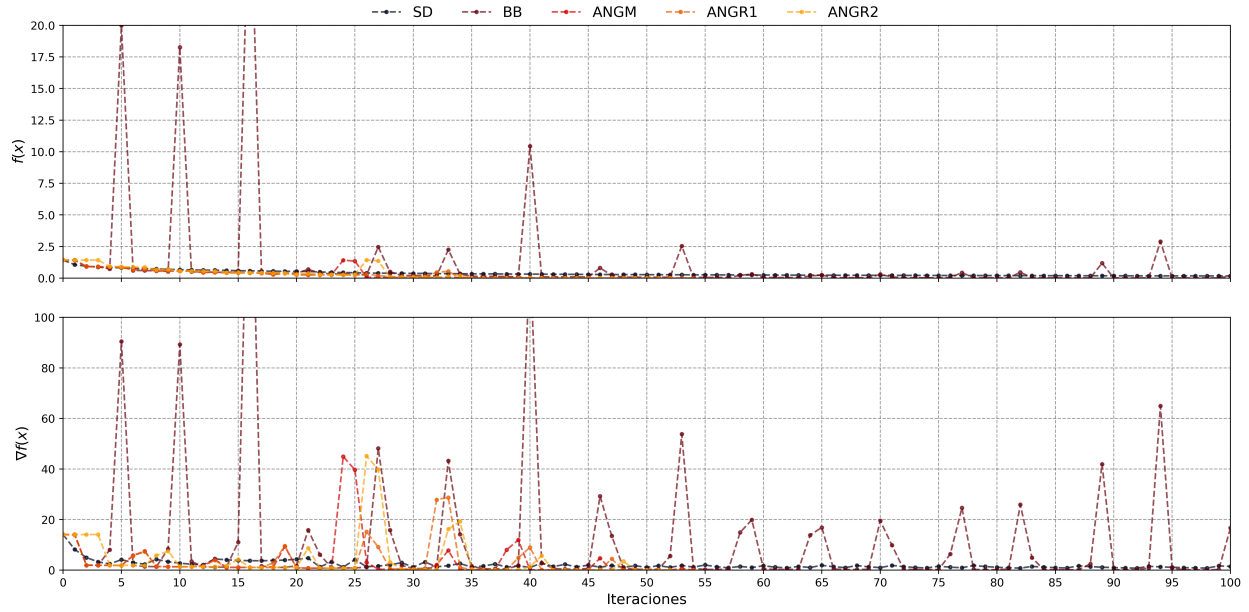


Figura 3: Valor de la función y norma del gradiente en cada iteración de la mejor ejecución de cada método de optimización de la función de Rosenbrock con matriz diagonal.

En las tablas 2 y 3 se muestran las medias y desviaciones estandar para el valor de la función y norma de gradiente en el punto óptimo y el número de iteraciones realizadas por cada método y función.

Función	Valor	SD	BB	ANGM	ANGR1	ANGR2
Lambda	Función	0.000000	0.000000	0.000000	0.000000	0.000000
	Gradiente	0.000001	0.000001	0.000001	0.000001	0.000001
	Iteraciones	7764.810000	2770.340000	252.430000	289.490000	255.830000
Cuadrática	Función	0.000000	0.000000	0.000000	0.000000	0.000000
	Gradiente	0.000001	0.000001	0.000000	0.000000	0.000000
	Iteraciones	713.000000	1531.230000	5.240000	7.360000	9.700000
Rosenbrock	Función	4.161332	42.245667	0.000000	0.000000	0.000000
	Gradiente	0.566027	45.446672	0.000000	0.000000	0.000000
	Iteraciones	10000.000000	1603.210000	68.840000	70.170000	67.150000
Wood	Función	0.000000	3.158103	5.070144	4.561704	8.428735
	Gradiente	0.000004	0.000001	0.000000	0.000000	0.000000
	Iteraciones	9428.690000	1806.430000	360.360000	292.430000	199.210000

Tabla 2: Media de las 100 ejecuciones partiendo de puntos aleatorios.

Función	Valor	SD	BB	ANGM	ANGR1	ANGR2
Lambda	Función	0.000000	0.000000	0.000000	0.000000	0.000000
	Gradiente	0.000000	0.000000	0.000000	0.000000	0.000000
	Iteraciones	612.385412	209.340437	32.553066	28.338322	25.043016
Cuadrática	Función	0.000000	0.000000	0.000000	0.000000	0.000000
	Gradiente	0.000000	0.000000	0.000000	0.000000	0.000000
	Iteraciones	43.810658	96.111137	1.198652	1.856030	2.886751
Rosembrock	Función	14.244453	393.585673	0.000000	0.000000	0.000000
	Gradiente	1.357743	396.058464	0.000000	0.000000	0.000000
	Iteraciones	0.000000	1867.607211	17.317429	20.596192	14.927011
Wood	Función	0.000000	10.092718	12.223427	11.860350	14.404998
	Gradiente	0.000019	0.000000	0.000000	0.000000	0.000000
	Iteraciones	465.597882	633.528762	190.558643	204.762156	119.763512

Tabla 3: Desviaciones estandar de las 100 ejecuciones partiendo de puntos aleatorios.

Conclusiones

Se logró implementar los métodos descritos en el artículo y comprobar para el caso cuadrático considerado que se logra una terminación finita a la hora de minimizar estas funciones, en la gran mayoría de los casos siendo esta terminación en tan solo 5 iteraciones, cuando el método de Barzilai-Borwein original llega a tomar hasta 1500 pasos y máximo descenso 713. También se logró observar con esta función cómo los métodos propuestos, logran reducir las componentes del gradiente correspondientes a los eigenvalores más grandes de la matriz, cuando los otros métodos tienen dificultad para esto, tardando más en llegar al óptimo.

Para la función de Wood, el método de Barzilai y Borwein y los métodos propuestos en el artículo se quedan atrapados en mínimos locales, cuando el método de máximo descenso no lo hace.

En general, los métodos propuestos muestran tener un mejor desempeño a la hora de minimizar las funciones de prueba utilizada, con excepción de la función de Wood y destacando notoriamente en el caso cuadrático bidimensional.

Referencias

- ¹ Barzilai J, Borwein JM. Two-Point Step Size Gradient Methods. IMA Journal of Numerical Analysis. 1988 01;8(1):141–148. Available from: <https://doi.org/10.1093/imanum/8.1.141>.
- ² Raydan M. On the Barzilai and Borwein choice of steplength for the gradient method. IMA Journal of Numerical Analysis. 1993;13(3):321–326. Available from: <https://doi.org/10.1093/imanum/13.3.321>.
- ³ Raydan M. The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem. SIAM Journal on Optimization. 1997;7(1):26–33. Available from: <https://doi.org/10.1137/S1052623494266365>.
- ⁴ Oleg B. Stabilized Barzilai-Borwein Method. JCM. 2019 jun;37(6):916–936. Available from: <https://doi.org/10.4208%2Fjcm.1911-m2019-0171>.
- ⁵ Fletcher R. On the barzilai-borwein method. In: Optimization and control with applications. Springer; 2005. p. 235–256. Available from: https://doi.org/10.1007/0-387-24255-4_10.
- ⁶ Dai YH, Fletcher R. Projected Barzilai-Borwein methods for large-scale box-constrained quadratic programming. Numerische Mathematik. 2005;100(1):21–47. Available from: <https://doi.org/10.1007/s00211-004-0569-y>.
- ⁷ Huang Y, Dai YH, Liu XW, Zhang H. On the asymptotic convergence and acceleration of gradient methods. Journal of Scientific Computing. 2022;90(1):1–29. Available from: <https://doi.org/10.48550/arXiv.1908.07111>.
- ⁸ Huang Y, Dai YH, Liu XW, Zhang H. On the acceleration of the Barzilai-Borwein method. Computational Optimization and Applications. 2022;p. 1–24.