Aceleración del método de Barzilai-Borwein

Edgar Osvaldo López Zúñiga
Giovanni Gamaliel López Padilla

Introducción

El estudio de la solución de la optmización de una función cuadratica de la forma

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

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 sparce). Esto provoca que la factorización por Cholesky sea impráctica por el tiempo de ejecucción o el espacio en memoria. En 1988 Barzilai y Borwein propusieron dos tamaños de paso para mejorar el desempeño de métodos de descenso de gradiente. La elección de los tamaños de paso propuestos estan basados en métodos Cuasi-Newton. Donde α_k es remplazada por la matriz D_k tal que

$$D_k = \alpha_k I \tag{2}$$

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).

$$\min_{D=\alpha I} ||D^{-1}s_{k-1} - y_{k-1}|| \qquad \min_{D=\alpha I} ||s_{k-1} - Dy_{k-1}|| \tag{3}$$

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.

$$\alpha_k^{BB1} = \frac{s_{k-1}^T s_{k-1}}{s_{k-1}^T y_{k-1}} \qquad \alpha_k^{BB2} = \frac{s_{k-1}^T y_{k-1}}{y_{k-1}^T y_{k-1}} \tag{4}$$

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 e 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 [1], y en 1997 introdujó 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:

$$\alpha_k = \min \alpha_k^{BB}, \Delta \tag{5}$$

donde Δ es un valor fijo. En el artículo de Oleg Burdakov [2] realiza varios experimentos con esta estrategia y obtuvo que para la función de rosembrock, 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 [3]. El método de BB cuenta con la propiedad de reducir los valores de la función objetivo de manera no monóntona. 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

$$\min_{x \in \mathbb{R}^n} f(x) \tag{6}$$

donde $f : \mathbb{R}^n \to \mathbb{R}^n$ 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 usara una función cuadrática (ecuación 1) utilizando el tamaño de paso mostrado en la ecuación 7.

$$\alpha_k(\Psi(A)) = \frac{g_{k-1}^T g_{k-1}}{g_{k-1}^T \Psi(A) A g_{k-1}} \tag{7}$$

Donde $\Psi(\cdot)$ es una función real analítica en $[\lambda_1, \lambda_n]$ que se puede expresar como una serie de potencias (ecuación 8).

$$\Psi(z) = \sum_{k=-\infty}^{\infty} c_k z^k,\ c_k \in \mathbb{R}^n \text{ tal que } 0 < \sum_{k=-\infty}^{\infty} c_k z^k < +\infty \forall z z \in [\lambda_1, \lambda_n] \tag{8}$$

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 7.

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. También se observa que el método BB es invariante ante tralaciones y rotaciones cuando se minimizan funciones cuadráticas [4], por lo que se puede asumir para simplicidad que la matriz A tiene la forma mostrada en la ecuación 9.

$$A = \text{diag}\{\lambda_1, ..., \lambda_n\}, \qquad \text{donde} \qquad 0 < \lambda_1 \leq ... \leq \lambda_n \tag{9}$$

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 [5]. Es por ello que si se quiere acelerar la convergencia de los métodos con tamaño de paso α_k (ecuación 7) en un subespacio de menor dimensionalidad se deben mantener propiedades ortogonales. Suponiendo que para un $k > 0$, existe un q_k que satisface

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

donde q_k es invariante ante rotaciones y traslaciones, por lo que se puede asumiendo 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 $\Psi(A)$.

Por el lema 1 se tiene que g_k^T y q_{k-1} son ortogonales bajo $\Psi(A)$ para cualquier $k > 0$. 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

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

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

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

al expandir en una serie de Taylor obtenemos que

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

Tomando a una matriz B_k como en la ecuación 13.

$$B_k = (u,v)^T \tag{13}$$

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

$$\phi(t,l) = f(x_k) + g_k^T B_k^T \binom{t}{l} + \frac{1}{2}\binom{t}{l}^T B_k A B_k^T \binom{t}{l} \tag{14}$$

Definiendo el vector ϑ como

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

y al hessiano de f en el paso k como

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

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

$$\phi(t,l) = f(x_k) + \vartheta_k^T \binom{t}{l} + \frac{1}{2}\binom{t}{l}^T H_k \binom{t}{l} \tag{17}$$

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 $k = 1, ..., k_0+1$. Para simplificar durante la prueba, se utilizará comom notación k para referirse a k_0.

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

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

en donde $\Delta = \det(H_k) = \det(A) > 0$. Ahora, 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 CHECAR

$$\begin{aligned}&(H_k^{(22)}v_k^1 - H_k^{(12)}v_k^{(2)} - \tilde{\alpha}_k\Delta\vartheta_k^{(1)})[\vartheta_k^{(2)} - \tilde{\alpha}_k(H_k^{(12)}\vartheta_k^{(1)} + H_k^{(22)}\vartheta_k^{(2)})]\\ &= (H_k^{(11)}v_k^2 - H_k^{(12)}v_k^{(1)} - \tilde{\alpha}_k\Delta\vartheta_k^{(1)})[\vartheta_k^{(1)} - \tilde{\alpha}_k(H_k^{(11)}\vartheta_k^{(1)} + H_k^{(12)}\vartheta_k^{(2)})].\end{aligned}$$

Consideremos dos vectores $\boldsymbol{A} = (a_1, a_2)^T$ y $\boldsymbol{B} = (b_1, b_2)^T$, tales que se cumple $a_1b_2 = a_2b_1$. Sustituyendo b_2 en $\boldsymbol{B}$, tenemos lo siguiente

$$\boldsymbol{B} = \begin{pmatrix} b_1 \\ \frac{b_1a_2}{a_1}\end{pmatrix} = \frac{b_1}{a_1}\boldsymbol{A}$$

por los que podemos decir que $\boldsymbol{B}$ es paralelo a $\boldsymbol{A}$. 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

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

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.

Resultados

Conclusiones

Referencias

[1] Raydan M. On the Barzilai and Borwein choice of steplength for the gradient method. IMA Journal of Numerical Analysis. 1993;13(3):321–326.

[2] Oleg B. Stabilized Barzilai-Borwein Method. JCM. 2019 jun;37(6):916–936. Available from: https://doi.org/10.4208%2Fjcm.1911-m2019-0171.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] Sun W, Yuan YX. Optimization theory and methods: nonlinear programming. vol. 1. Springer Science & Business Media; 2006.

[8] 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.