Acelerando Barzilai-Borwein

Peralta Rosales Oscar Esaú - Stack Sánchez Pablo Antonio
Optimización I- Maestría en Computación
Centro de Investigación en Matemáticas

***Resumen*—En este proyecto se revisa y presenta un resumen del paper *On the acceleration of the Barzilai-Borwein method* además de la implementación del método de gradiente no monótono adaptativo ANGM ahí presentado junto con sus dos variantes, ANGR1 y ANGR2 usados resolver problemas de optimización sin restriccioes. Estos métodos adaptativos dan algunos pasos no monótonos incluyendo los tradicionales de Barzilai - Borwein y algunos monotonos usando el nuevo tamaño de paso. Los algoritmos propuestos demostraron reducir considerablemente el número de iteraciones necesarias para lograr la convergencia.**

I. Introducción

Los métodos de descenso de gradiente han sido ampliamente utilizados para resolver problemas suaves de optimización sin resricciones

$$\min_{x\in\mathbb{R}^n} f(x) \tag{1}$$

generando una secuencia de iterandos

$$x_{k+1} = x_k - \alpha_k g_k \tag{2}$$

en donde $f : \mathbb{R}^n \to \mathbb{R}$ es continua y diferenciable, $g_k = \nabla f(x_k)$ y $\alpha_k > 0$ es el tamaño de paso en la dirección contrria al gradiente. El método más clásico para calcular α_k se conoce como el paso exacto (SD)

$$\alpha_k^{SD} = \arg\min_{\alpha\in\mathbb{R}} f(x_k - \alpha g_k) \tag{3}$$

Aunque el paso SD tiene localmente la mayor reducción en la dirección negativa del gradiente, en la practica con frecuencia no tiene un buen desempeño. Teoricamente, cuando f es una función cuadrática y extrictamente convexa como

$$f(x) = \frac{1}{2}x^T Ax - b^T x \tag{4}$$

en donde $b \in \mathbb{R}$ y $A \in \mathbb{R}^{n\times n}$ es simétrica y postiva definida, el método SD converge de forma Q-lineal y tendrá un efecto de zigzag entre dos direcciones ortogonales.
Barzilai y Borwein propusieron las siguientes formas de calcular el tamaño de paso, mejorando significativamente el desempeño de los métodos de descenso de gradiente:

$$\alpha_k^{BB1} = \frac{s_{k-1}^T s_{k-1}}{s_{k-1}^T y_{k-1}} \quad y \quad \alpha_k^{BB2} = \frac{s_{k-1}^T y_{k-1}}{y_{k-1}^T y_{k-1}} \tag{5}$$

en donde $s_{k-1} = x_k - x_{k-1}$ y $y_{k-1} = g_k - g_{k-1}$. Cuando la función objetivo es cuadrática (4), el tamaño de paso a_k^{BB1} es exactamente el tamaño de paso SD pero desfasado por una iteración, mientras a_k^{BB2} será exactamente el tamaño de paso de minimo gradiente (MG):

$$\alpha_k^{BB1} = \frac{g_{k-1}^T g_{k-1}}{g_{k-1}^T A g_{k-1}} = \alpha_{k-1}^{SD}$$
$$\alpha_k^{BB2} = \frac{g_{k-1}^T A g_{k-1}}{g_{k-1}^T A^2 g_{k-1}} = \alpha_{k-1}^{MG}$$

Está demostrado que el método de Barzilai-Borwein (BB) converge R-superlineal al minimizar funciones cuadráticas extricamente convexas de dos dimensiones y con R-lineal para el caso general de n dimensiones.
La propiedad intrínseca de reducir la función objetivo no monotónicamente es la que ocasiona la eficiencia del método de BB. Sin embargo, ha sido señalado que mantener la monotonidad es importante para los métodos de descenso de gradiente, de ahí que en el mencionado trabajo se busque mejorar y acelerar el algoritmo BB incorporando algunos pasos monotónicos.

Primeramente, se considera acelerar los métodos de descenso de gradiente (2) para funciones cuadráticas (4) usando el siguiente tamaño de paso

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

en donde $\Psi(.)$ es una función analítica real en $[\lambda_1, \lambda_n]$ que puede ser expresado por una serie de Laurent

$$\Psi(z) = \sum_{k=-\infty}^{\infty} c_k z^k, \qquad c_k \in \mathbb{R}$$

tal que $0 < \sum_{k=-\infty}^{\infty} c_k z^k < +\infty$ paa todo $z \in [\lambda_1, \lambda_n]$. Aqui λ_1 y λ_n son los eigenvalores más pequeño y más grande de A. El método (6) es no monótono y los dos pasos de BB α_k^{BB2} y α_k^{BB2} se pueden obtener al hacer $\Psi(A) = I$ y $\Psi(A) = A$ respectivamente.

II. Metodología

II-A. Derivación del nuevo tamaño de paso

Obsérvese que el método (6) es invariante a translaciones y rotaciones cuando se minimizan funciones cuadráticas, por lo tanto para el análisis se asume sin perdida de generalidad que la matriz A es diagonal.

$$A = diag\{\lambda_1, \ldots, \lambda_n\} \tag{7}$$

En otros articulos se ha demostrado que una familia de métodos de gradientes incluyendo a SD y MG asintoticamente reduciran sus búsquedas a un subespacio de 2 dimensiones y pueden ser acelerados al explotar ciertas características de ortogonalidad en este subespacio. De igual forma, podemos acelerar la familia (6) de métodos de descenso de gradiente en un subespacio menor si se cumplen algunas propiedades de ortogonalidad.
Suponga que, para una $k > 0$ esxiste q_k que satisface

$$(I - \alpha_{k-1}A)q_k = g_{k-1} \tag{8}$$

Ahora, supóngase que la secuencia $\{g_k\}$ se obtiene al aplicar el método de gradiente (2) con tamaño de paso (6) para minimizar una función cuadrática (4) y q_k satisface (8), entonces tenemos

$$q_k^T \Psi(A) g_{k+1} = 0 \tag{9}$$

Lo anterior muestra una propiedad generalizada de ortogonalidad para q_k y g_{k+1}, que es una propiedad clave para derivar el nuevo tamaño de paso.
Supongamos que tanto $\Psi^r(A)q_{k-1}$ y $\Psi^{1-r}(A)g_k$ son vectores diferentes de cero, en donde $r \in \mathbb{R}$. Ahora minimizamos en la función f en un subespacio bidimensional generado por $\frac{\Psi^r(A)q_{k-1}}{||\Psi^r(A)q_{k-1}||}$ y $\frac{\Psi^{1-r}(A)g_k}{||\Psi^{1-r}(A)g_k||}$, y sea

$$\begin{aligned}\rho(t,l) &:= f(x_k + t\frac{\Psi^r(A)q_{k-1}}{||\Psi^r(A)q_{k-1}||} + l\frac{\Psi^{1-r}(A)g_k}{||\Psi^{1-r}(A)g_k||}) \\ &= f(x_k) + \vartheta_k^T \begin{pmatrix} t \\ l \end{pmatrix} + \frac{1}{2}\begin{pmatrix} t \\ l \end{pmatrix} + H_k \begin{pmatrix} t \\ l \end{pmatrix}\end{aligned} \tag{10}$$

en donde

$$\vartheta_k = B_k g_k = \begin{pmatrix} \frac{g_k^T \Psi^r(A)q_{k-1}}{||\Psi^r(A)q_{k-1}||} \\ \frac{g_k^T \Psi^{1-r}(A)g_k}{||\Psi^{1-r}(A)g_{k-1}||} \end{pmatrix} \tag{11}$$

con

$$B_k = \left(\frac{\Psi^r(A)q_{k-1}}{||\Psi^r(A)q_{k-1}||}, \frac{\Psi^{1-r}(A)g_k}{||\Psi^{1-r}(A)g_k||}\right)$$

y

$$H_k = B_k A B_k^T = \begin{pmatrix} \frac{q_{k-1}^T \Psi^{2r}(A) A q_{k-1}}{||\Psi^r(A)q_{k-1}||^2} & \frac{q_{k-1}^T \Psi(A) A g_k}{||\Psi^r(A)q_{k-1}||||\Psi^{1-r}(A)g_k||} \\ \frac{q_{k-1}^T \Psi(A) A g_k}{||\Psi^r(A)q_{k-1}||||\Psi^{1-r}(A)g_k||} & \frac{q_{k-1}^T \Psi^{2(1-r)}(A) A g_k}{||\Psi^{1-r}(A)g_k||^2} \end{pmatrix} \tag{12}$$

Denotamos los componentes de H_k con $H_k^{(ij)}$, $i, j = 1, 2$.
Note que $B_k B_k^T = I$ cuando $g_k^T \Psi(A) q_{k-1} = 0$.
Suponga que un método de gradiente (2) se aplica para minimizar una función cuadrática de dos dimensión (4) con α_k dado por (6) para toda $k \neq K_0$ y usa el tamaño de paso.

$$\tilde{\alpha}_{k_0} = \frac{2}{\left(H_{k_0}^{(11)} + H_{k_0}^{(22)}\right) + \sqrt{\left(H_{k_0}^{(11)} + H_{k_0}^{(22)}\right)^2 + 4\left(H_{k_0}^{(12)}\right)^2}} \tag{13}$$

Note que haciendo $\Psi(A) = I$, $\Psi(A) = A$ y $r = \frac{1}{2}$ en (12), y haciendo $k_0 = k$ en (13), podemos derivar los siguientes dos tamaños de pasos:

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

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

en donde

$$\hat{a}_k = \frac{q_k^T A q_k}{q_k^T A^2 q_k} \quad y \quad \Gamma_k = \frac{4\left(q_{k-1}^T A^2 g_k\right)^2}{q_{k-1}^T A q_{k-1} g_k^T A g_k} \tag{14}$$

Con base en el análisis anterior, se propone un método de gradiente no monótono adaptativo (ANGM) y sus dos variantes, ANGR1 y ANGR2 [1] para resolver problemas de optimización sin restriccioes. Estos métodos adaptativos dan algunos pasos no monótonos incluyendo los pasos tradicionales de BB (5) y algunos pasos monotonos usando el nuevo tamaño de paso.

ANGM aplica la siguiente estrategia para escoger el tamaño de paso:

$$\alpha_k = \begin{cases} \min\{\alpha_k^{BB2}, \alpha_{k-1}^{BB1}\} & \text{si } \alpha_k^{BB2} < \tau_1 \\ & \text{y } ||g_{k-1}|| < \tau_2||g_k|| \\ \tilde{\alpha}_k^{BB2} & \text{si } \alpha_k^{BB2} < \tau_1 \\ & \text{y } ||g_{k-1}|| \geq \tau_2||g_k|| \\ \alpha_k^{BB1} & \text{otro caso} \end{cases} \tag{15}$$

Notemos que para obtener $\tilde{\alpha}_k^{BB2}$ es necesario calcular α_k^{MG} lo que resulta complicado cuando la función objetivo no es cuadrática. En cambio, el cálculo de $\tilde{\alpha}_{k-1}^{BB2}$ solo requiere α_k^{BB2} que es fácil de obtener. Por lo tanto para la primera variante de ANGM, simplemente se reemplaza $\tilde{\alpha}_k^{BB2}$ por $\tilde{\alpha}_{k-1}^{BB2}$. Así, ANGR1 aplica la siguiente estrategia para escoger el tamaño de paso:

$$\alpha_k = \begin{cases} \min\{\alpha_k^{BB2}, \alpha_{k-1}^{BB1}\} & \text{si } \alpha_k^{BB2} < \tau_1 \\ & \text{y } ||g_{k-1}|| < \tau_2||g_k|| \\ \tilde{\alpha}_{k-1}^{BB2} & \text{si } \alpha_k^{BB2} < \tau_1 \\ & \text{y } ||g_{k-1}|| \geq \tau_2||g_k|| \\ \alpha_k^{BB1} & \text{otro caso} \end{cases} \tag{16}$$

Por otro lado, dado que el cálculo de $\tilde{\alpha}_{k-1}^{BB2}$ necesita $\hat{\alpha}_{k-2}$ y τ_{k-1}, y además $\tilde{\alpha}_{k-1}^{BB2} \leq min\{\alpha_k^{BB2}, \hat{\alpha}_{k-2}\}$. Asi, para simplificar ANGR1, se reemplaza $\tilde{\alpha}_{k-1}^{BB2}$ por su cota supe-

rior. ANGR2 aplica la siguiente estrategia para escoger el tamaño de paso:

$$\alpha_k = \begin{cases} \min\{\alpha_k^{BB2},\, \alpha_{k-1}^{BB1}\} & \text{si } \alpha_k^{BB2} < \tau_1 \\ & \text{y } ||g_{k-1}|| < \tau_2 ||g_k|| \\ \min\{\alpha_k^{BB2}, \hat{\alpha}_{k-2}\} & \text{si} \alpha_k^{BB2} < \tau_1 \\ & \text{y } ||g_{k-1}|| \geq \tau_2 ||g_k|| \\ \alpha_k^{BB1} & \text{otro caso} \end{cases} \quad (17)$$

Notemos que para los nuevos 3 métodos, ANGM, ANGR1 y ANGR2 es necesario calcular q_k para obtener los tamaños de pasos. Sin embargo, calcular q_k exactamente de (8) puede ser tan difícil como minimicar la función cuadrática. Nótese que el q_k que satisface (8) también satisface la ecuación de la secante.

$$q_k^T g_k = ||g_{k-1}||^2$$

Por lo tanto se puede encontrar una aproximación de q_k. Una manera eficiente es tratar al Hessiano A como matriz diagonal (7) y derivar q_k de (8), que es cuando $g_k^{(i)} \neq 0$

$$q_k^{(i)} = \frac{g_{k-1}^{(i)}}{1 - \alpha_{k-1}\lambda_i} = \frac{\left(g_{k-1}^{(i)}\right)^2}{g_k^{(i)}}, \;\; i = 1, \ldots, n$$

y simplemente hacemos $q_k^{(i)} = 0$, si $g_k^{(i)} = 0$. En resumen, la aproximación de q_k se puede calcular como:

$$q_k^{(i)} = \begin{cases} \frac{\left(g_{k-1}^{(i)}\right)^2}{g_k^{(i)}} & \text{si } g_k^{(i)} \neq 0 \\ 0 & \text{si } g_k^{(i)} = 0. \end{cases} \quad (18)$$

Se programaron estas 3 variantes y se probaron con las funciones de Wood y Rosenbrock con los puntos iniciales utilizados habitualmente en las tareas, posteriormente se realizaron 100 corridas con valores iniciales aleatorios variando el parámetro τ_1 con la función de Wood. Los resultados se muestran en la siguiente sección.

III. Resultados

En esta sección se presentan las comparaciones numéricas entre los métodos ANGM, ANGR1, ANGR2 y los tradicionales BB1 y BB2. Para las primeras pruebas se utilizó un valor de $\tau_1 = 0{,}4$ y $\tau_2 = 1$.
En primer lugar se probó la función de Wood con el siguiente punto inicial

$$x^0 = [-3, -1, -3, -1]^T$$

El cuadro I muestra la comparación entre el promedio del número de iteraciones, la norma del gradiente y el tiempo con los 3 métodos propuestos y los BB con tamaños de paso BB1 y BB2.

Algorithm 1: Algoritmo ANGRM, ANGR1 y ANGR2

Result: x^*
α_0 <- Proponer
$k = 0$
while $||g_k|| > tol$ **do**
  $g_k = f(x_k)$
  if *k==0* **then**
    Usar $\alpha_k = \alpha_0$
  end
  else if *Min de iteraciones para ANGRM, ANGR1 o ANGR2* **then**
    Usar (15), (16) o (17) para calcular α_k
  end
  else
    Usar (5) para calcular α_k
  end
  $x_{k+1} = x_k + \alpha_k d_k$
  $k = k + 1$
end

Cuadro I
Resultados promedio de 100 ejecuciones de los métodos con la Función Wood

	BB1	BB2	ANGRM	ANGR1	ANGR2
Iteraciones	7234	379	242	398	323
$\|\|\nabla f(x)\|\|$	8.96e-07	2.50e-07	2.67e-09	2.65e-07	1.43e-07
Tiempo (s)	0.3456	0.028	0.026	0.041	0.029

Se observa que el número más grande de iteraciones y tiempo de ejecución se obtuvo con BB1, el método que proporcionó el mejor desempeño fue el ANGRM. En general es posible notar que a pesar de que wood es una función de pocas dimensiones los métodos propuestos se comportan de mejor forma.
Posteriormente se probó con la función de Rosembrock con el siguiente punto inicial

$$x^0 = [-1{,}2, 1, 1, \ldots, 1, -1{,}2, 1]$$

En el cuadro II se presentan los resultados.

Cuadro II
Resultados promedio de 100 ejecuciones de los métodos con la Funcion Rosembrock

	BB1	BB2	ANGRM	ANGR1	ANGR2
Iteraciones	1087	1012	329	334	299
$\|\|\nabla f(x)\|\|$	7.18e-07	9.56e-07	8.22e-07	1.81e-07	3.55e-07
Tiempo (s)	1.41	1.35	0.81	0.72	0.57

Con la función de Rosembrock las diferencias en el desempeño de los algoritmos son más notorias. El cuadro II muestra que los tiempos de ejecución para los 2 tamaños de paso del BB superan por casi el doble el de los 3 métodos propuestos.

El número de iteraciones tanto del BB1 como del BB2 es alrededor de 3 veces mayor.

Cuadro III
Función Wood puntos aleatorios promedio de 100 corridas

	ANGM			**ANGR1**			**ANGRR2**		
t1	*iter*	$\|\|\nabla f(x)\|\|$	*Tiempo (s)*	*iter*	$\|\|\nabla f(x)\|\|$	*Tiempo (s)*	*iter*	$\|\|\nabla f(x)\|\|$	*iter*
0.1	201.73	3.885e-07	1.681e-02	207.24	3.837e-07	1.576e-02	224.08	4.240e-07	1.442e-02
0.2	169.15	3.991e-07	1.402e-02	185.11	3.482e-07	1.400e-02	112.13	3.999e-07	7.267e-03
0.3	73.85	3.761e-07	6.321e-03	83.3	3.843e-07	6.352e-03	82.21	3.400e-07	5.364e-03
0.4	89.39	3.578e-07	7.679e-03	89.17	3.021e-07	6.899e-03	79.93	3.750e-07	5.338e-03
0.5	82.28	3.337e-07	6.963e-03	84.95	3.590e-07	6.546e-03	88.11	3.405e-07	5.743e-03
0.6	78.43	2.942e-07	6.642e-03	88.97	3.638e-07	6.832e-03	84.6	3.327e-07	5.558e-03
0.7	74.86	3.409e-07	6.361e-03	91.47	3.107e-07	7.013e-03	89.14	2.921e-07	5.777e-03
0.8	88.96	2.767e-07	7.517e-03	93.02	3.024e-07	7.107e-03	92.94	2.595e-07	6.112e-03
0.9	100.85	2.820e-07	8.722e-03	101.53	2.914e-07	7.821e-03	93.06	2.672e-07	6.227e-03

Para darnos una mejor idea del verdadero desempeño de los algoritmos propuestos, se realizaron 100 corridas de la función de Wood con punto inicial aleatorio, además se varió el parámetro τ_1 de 0.1 a 0.9 en incrementos de 0.1. El promedio de los resultados obtenidos se muestra en el cuadro III. La función de Resombrock fue omitida puesto que es una función en general no convexa y comúnmente los puntos aleatoreos iniciales seleccionados hacen que los algoritmos no convergan; recuérdese que el algoritmo de Barzilai-Borwein (BB) no garantiza convergencia en problemas no convexos ni suavemente convexos.

De forma general se aprecia que tanto valores muy pequeños de τ_1 como valores muy grande deterioran el desempeño de ANGM, ANGR1, ANGR2.

El menor número de iteraciones en ANGM y ANGR1 se obtuvo con un valor de $\tau_1 = 0{,}3$, mientras que para ANGR1 se consiguió con $\tau_1 = 0{,}4$.

En el cuadro IV se presentan los resultados obtenidos para el BB1 y BB2.

Cuadro IV
Función Wood puntos aleatorios promedios 100 corridas

BB1			**BB2**		
iter	$\|\|\nabla f(x)\|\|$	*Tiempo (s)*	*iter*	$\|\|\nabla f(x)\|\|$	*Tiempo (s)*
509.31	4.437e-07	1.614e-02	136.62	5.345e-07	4.170e-03

Si se comparan los tiempos de ejecución de estos 2 métodos contra los 3 propuestos durante las 100 corridas, se observa que el obtenido para BB1 es un ordén de magnitud superior, mientras que BB2 tiene un desempeño similar a ANGM, ANGR1 y ANGR2 aunque en un mayor número de iteraciones.

IV. Conclusiones

Se mostró un resumen y resultados obtenidos del método desarrallado por Yakui Huang, Yu-Hong Dai, Xin-Wei Liu y Hongchao Zhang basado en una mejora el método ya conocido de Barzilai-Borwein, en pro de aprovechar la particularidad de evitar calcular el tamaño de paso exacto α^{SD} lo cual puede resultar muy costoso e introducir pasos monótonos con tal de retener monotonicidad y ayudar a acelerar este algoritmo.

Se presentaron 3 variantes del algoritmo, la primera ANGM, que bajo este análisis aún hace uso de la matriz A para el cálculo del paso de gradiente, la primer modificación al usar un tamaño de paso retardado permite deshacerse del uso de está matriz llevando al método ANGR1, y posteriormente usando como cota $\tilde{\alpha}_{k-1}^{BB2} \leq min\{\alpha_k^{BB2}, \hat{\alpha}_{k-2}\}$ permite derivar la tercera versión, ANGR2. Los resultados mostrados en la sección anterior demuestran que se logra una mejora significante en comparación del método de Barzilai-Borwein normal.

Referencias

[1] Yakui Huang, Yu-Hong Dai, Xin-Wei Liu, Hongchao Zhang, On the acceleration of the Barzilai-Borwein method, arXiv:2001.02335 .